energies

Article

Clustering Analysis for Active and Reactive Energy Consumption Data Based on AMI Measurements

Oscar A. Bustos-Brinez [1,2] and Javier Rosero Garcia [1,*]

[1] EM&D Research Group, Electrical and Electronics Engineering Department, Faculty of Engineering, Universidad Nacional de Colombia, Bogota 111321, Colombia; oabustosb@unal.edu.co

[2] MindLab Research Group, Systems and Industrial Engineering Department, Faculty of Engineering, Universidad Nacional de Colombia, Bogota 111321, Colombia

* Correspondence: jaroserog@unal.edu.co

Abstract: Electrical data analysis based on smart grids has become a fundamental tool used by electrical grid stakeholders to understand the energy consumption patterns of users, although many proposals in this area do not consider reactive energy as another source of useful information regarding distribution costs and threats to the grid. In this regard, the analysis of reactive energy patterns can become an extremely useful addition to existing electrical data analysis frameworks. This work shows the application of a series of clustering techniques over measurements of both active and reactive energy consumption measured for end users from the Colombian electrical network, including an analysis of the efficiency of the network measured by calculating the ratio of active energy to total consumption (power factor) per user. This allows a detailed characterization of users to be compiled, based on the identification of different active and reactive energy consumption behaviors, which could help grid operators to improve overall grid management and to increase the efficiency of their reactive energy compensation strategies.

Keywords: energy analytics; data analysis; electrical grid management; reactive energy; power factor

Academic Editors: Rodolfo Araneo, Massimo Panella and Antonello Rosato

Received: 18 November 2024
Revised: 24 December 2024
Accepted: 30 December 2024
Published: 6 January 2025

Citation: Bustos-Brinez, O.A.; Rosero Garcia, J. Clustering Analysis for Active and Reactive Energy Consumption Data Based on AMI Measurements. *Energies* **2025**, *18*, 221. https://doi.org/10.3390/en18010221

1. Introduction

The increasing use of AMI (advanced metering infrastructure) in conjunction with existing electrical grids has served as the basis for a new concept within electrical grid management known as smart grids [1,2]. In many applications, the efficient automated analysis of large volumes of data generated by AMI devices is crucial, including demand response [3,4], estimation of consumption peaks [5], load capacity analysis [6], and construction of new tariff schemes [7,8]. The analysis of AMI measurements has been an instrumental tool utilized by grid operators and government agencies to enhance policies pertaining to energy generation, distribution, and consumption across various levels of the grid, including final users, intermediate distribution nodes, and generation hubs [9]. In particular, developing countries have prioritized the analysis of energy demand in their networks [10], as more frequent climate variability and economic volatility scenarios could adversely impact power generation and distribution, particularly when considering renewable energy sources.

The majority of AMI-based data analysis focuses exclusively on active energy, as this is what can be converted into useful work in both residential and industrial settings. However, a more comprehensive framework for data analysis should take into account power line and voltage stability losses, such as those caused by electrical devices and transformers whose currents show a phase inequality compared to voltage, i.e., the generation of reactive energy [11,12]. The losses caused by reactive energy generation are frequently overlooked despite their relationship with multiple overloading and instability issues, and the handling of these risks often entails significant costs for distributors and users, particularly in networks with distributed energy resources [13,14]. Reactive energy has typically been addressed through a variety of approaches, including the use of physical devices situated in proximity to its sources [15] and the implementation of optimization algorithms for energy distribution [16,17]. In contrast, there is a relative lack of emphasis on studying reactive energy from the perspective of end users.

Clustering methodologies have been extensively utilized in data analysis on electric networks for multiple purposes, including the detection of frequent consumption patterns among a vast number of users, the facilitation of demand aggregation across various levels of energy distribution, and the identification of known behaviors such as the classification of different types of consumption through the grouping of profiles (in particular, the distinction between residential, commercial, and industrial profiles) [18]. Furthermore, the application of clustering techniques can assist in the identification of infrequent or rare consumption patterns, which may be confined to small, isolated groups or even may not align with any group. These isolated and infrequent data points that deviate from the typical behavior of the users under analysis can be identified as "anomalies" in the context of the original data set [19]. The detection of anomalies using this mechanism, although only taking into account the AMI measurements, does not require additional information about users, such as their geographic location or customer types, allowing for more direct detection that can be easily coupled with automated data generation.

In this study, we adopt a user-centric approach to examine the consumption of active and reactive energy. To this end, we utilize a data set comprising hourly measurements of both active and reactive energy values obtained from AMI meters situated in central Colombia, encompassing a four-month period in 2023. The analyses conducted in this study entail the application of multiple clustering techniques to the active and reactive energy data, independently, with the objective of identifying groups of users exhibiting similar consumption profiles and also instances of infrequent and/or isolated data. The analyses are performed in two distinct ways: firstly, by consumption magnitude and secondly, by profile geometry. An additional analysis considers the power factor (calculated here as the proportion of active energy over total energy consumption) to assess the distribution of users into different groups according to the amount of power lost by each. With these different analyses, it is possible to conduct an in-depth examination of active and reactive energy consumption patterns in a relatively uncommon way. This approach allows for a more informed characterization of users and the detection of unexpected consumption patterns that may indicate problems in the electrical network. These two outcomes can be useful to help in the design of new strategies to compensate for reactive energy generation (also known as power factor correction [20,21]) and thus lead to improvements in the efficiency of distribution and energy consumption.

In summary, this work presents the following main contributions:

1. A systematic evaluation of five different clustering methods applied over the data set of AMI measurements discriminated by month and type (active and reactive), including a description of the clusters obtained in each case, identifying the most common user behaviors in terms of both their magnitude and the geometry of the profiles.
2. An analysis of the clusters of the power factors of users, establishing the relative importance of each cluster and how the active and reactive consumptions of the users behave in each of them.

3. The presentation of a user characterization based on the clusters obtained in the previous steps, built by categorizing users according to the magnitude of their active and reactive energy consumption and the normality (or abnormality) of their consumption profiles.

The article is structured in the following sections. Section 2 presents a review of applications of clustering methods for electrical analysis, particularly when considering data generated through AMI meters. Section 3 establishes the methodology followed to process the data, defining the main features of the data set under analysis, and stating the processing stages carried out on the data from its original presentation to the obtention of relevant information such as consumption profiles. Section 4 details the results obtained when applying the clustering methods to the different types of measurement (active energy, reactive energy, power factor), making a comparison of the results for each month and establishing different groups of users based on the clusters they belong to in each analysis. Finally, the conclusions of the work are presented in Section 5.

2. Background

In essence, the primary objective of clustering methods can be defined as the partitioning of a data set into a number of groups, such that data points within the same group are highly similar and data points in different groups are highly dissimilar. Although clustering methods are widely used in different applications in the electrical sector, this work focuses on the use of these methods to perform classifications of users based on their consumption behavior (commonly represented as numerical structures called "electrical consumption patterns" or ECPs) [22]. Other alternatives for clustering methods rely on an interpretation of consumption that differs from patterns, by analyzing consumption as a time series with strong linear correlation [23]. A search was therefore conducted in multiple articles related to the electrical sector that include the use of clustering methods (especially when focused on the analysis of consumption patterns) and how they have been used in different use cases to analyze and group similar behaviors. It is our intention to include a variety of methods that rely on diverse mathematical formulations, particularly in regard to how they define the cost function to be minimized in order to achieve an optimal separation of groups. This cost function can be defined in terms of multiple mathematical definitions, including predefined similitude measurements (distance-based methods), a definition of density in a particular feature space (density-based methods) or probabilistic modeling (generative methods), among others [24].

K-means is one of the most frequently cited methods in the literature, largely due to its efficiency and straightforward implementation. This method employs a distance metric (commonly, the Euclidean distance) to construct a predefined number of clusters, wherein each cluster is represented by its "centroid", defined as the mean of all its data points [25]. K-means has been employed extensively in the analysis of consumption patterns across a range of scenarios, with a particular focus on identifying recurrent behaviors among end users. In [26], the method is employed to identify five distinct patterns associated with seasonal fluctuations in energy consumption among a small group of users in Portugal. In [27], the authors present a comprehensive methodology based on k-means and apply it to an open-access data set of 5500 London household consumption profiles, identifying three distinct clusters of users. The approach presented in [28] integrates the application of k-means with a statistical analysis of the correlation between hourly measurements, thereby accounting for the temporal component. The joint methods are applied to a considerably larger data set than that used in previous cases, derived from a smart grid situated in a city in southern Denmark. While k-means is a widely utilized method in these applications, there are numerous variations that may demonstrate superior performance in specific scenarios [29]. In particular, bisecting k-means, an algorithm that integrates k-means with the concept of hierarchical cluster structures, has been demonstrated to outperform the original method in data sets with high dimensionality or a large number of clusters [30].

A second clustering method that is frequently compared to k-means is DBSCAN (density-based spatial clustering of applications with noise). In this method, clusters are constructed by defining areas of high and low density in the feature space using two parameters: a radius that defines the vicinity of a point, and a minimum number of points in such a vicinity that determines whether the point is located on a high-density area or not [31,32]. This approach allows the formation of clusters with irregular shapes, in contrast with k-means, and also identifies anomalous points that appear in low-density regions. Due to its additional anomaly detection capability, DBSCAN is often employed as a preliminary step to eliminate noisy data before the application of other methodologies or models. For instance, in [33], DBSCAN is utilized to remove outlier points in a time series of electrical consumption as a previous step to a knowledge discovery classification model. Similarly, in [34], it is used to identify seasonal patterns in consumption before the implementation of an association rules algorithm. Furthermore, DBSCAN has numerous variations, including HDBSCAN (hierarchical DBSCAN), which is frequently utilized in fault detection and to spot attacks on control systems [35,36]. Another variation of the method, known as OPTICS (ordering points to identify clustering structure), enables the formation of clusters with varying densities and is regarded as being more precise in identifying anomalies [37].

3. Methodology

The methodology outlined in this paper is designed to achieve two primary objectives: the characterization of the average consumption of end users, as represented by average consumption profiles, and the identification of anomalous consumption patterns from said profiles. The accomplishment of these objectives is carried out by the implementation of several clustering algorithms, which not only enable the identification of common behaviors that allow the characterization of demand without the necessity for additional geographic or socioeconomic data, but under certain conditions, they can also facilitate the identification of atypical behaviors that can be classified as anomalies. A visual summary of the proposed methodology is presented in Figure 1, and it can be summarized in the following three main stages.

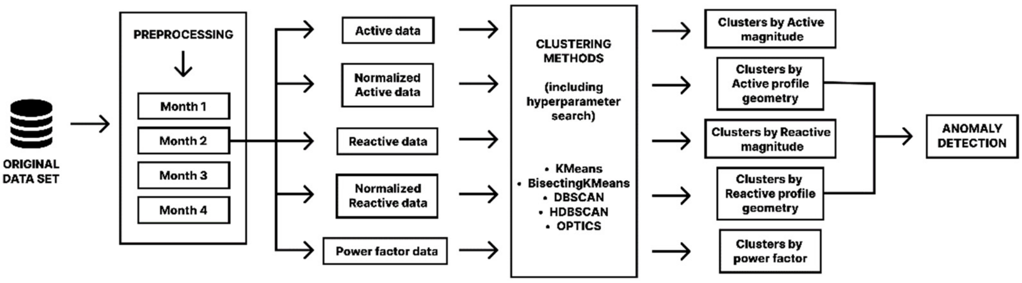

Figure 1. Visual summary of the methodology, presented as a series of consecutive steps. From left to right, it starts with the preprocessing of the original data to discard users with missing values and isolate registers from different months. In each month, information is separated into five measurements (two for active energy, two for reactive energy, and power factor) and each one is passed to the clustering methods, on which a parameter search is performed. Finally, some of the results of the best clusters are used to perform the detection of anomalous consumption patterns.

Preprocessing. The original user consumption data is organized in such a way that a single timeline is constructed for each user, and it is analyzed for missing or unknown values according to the time period included in the data set. If a user's time series contains missing data, no data imputation is performed (to avoid making assumptions that could contaminate the data); rather, the time series is discarded in its entirety. The analysis focuses on monthly time periods in order to understand and characterize medium-term user consumption trends, aiming to reduce the possibility of introducing non-periodic variations on specific days. In this regard, user time series are segmented into monthly intervals, and within each month a 24 h consumption profile is built to encapsulate the consumption behavior for that period. This preprocessing stage is executed independently for the active and reactive energy records contained within the data set.

Search for the best clustering method. In accordance with the active and reactive energy consumption profiles generated in the previous step, five representations are constructed for each user: two active energy representations, one considering magnitude and the other considering only the profile geometry, two reactive energy representations also considering magnitude and geometry, and one representation through the calculation of the power factor, in which both energy profiles are used to calculate the user's energy consumption efficiency at each moment of the day.

The representations for active and reactive energy can be based on raw measurements to consider magnitude (whose units are watt-hours in the case of active energy, and VArh in the case of reactive energy), or a normalization of the consumption profiles can be carried out. In this case, information of consumption magnitude is discarded in favor of a greater focus on the geometry of the curve, its rises and falls. The normalization of a profile is carried out by multiplying the vector that represents it by a scalar value that corresponds to the inverse of its Euclidean norm. The power factor representation is calculated using the following process: the value of a given hour is calculated by taking the active energy value and dividing it by the sum of active and reactive values. The result is a measure of the overall percentage that active energy takes in total consumption, similar to the power factor (PF). In this way, ratios closer to zero indicate a low consumption of active energy compared to the reactive energy, and ratios closer to one indicate a high consumption of active energy compared to the reactive energy. Users where the sum of both measurements is zero at any time are discarded, to avoid indeterminations.

Representations of the same type from all users are then gathered and used as input for five different clustering methods. These methods were selected as part of the state of the art in data analysis of the electricity sector, and are tested several times with different combinations of their parameters. This process allows for the creation of a robust model that best adapts to the particularities of each case. The results are then compared using a performance metric, and the result of the best model is considered the most accurate segmentation for that representation of the data.

Anomaly detection. The generation of groups by the optimal models in each scenario facilitates the identification of anomalous consumption patterns, which are manly identified by the distribution of consumption throughout the day represented by the profiles, rather than by the presence of atypical consumption values (e.g., exceedingly high or low values). The results of the models are utilized solely where the geometry of the profiles is under consideration. Based on these segmentations, it is determined whether there are small and isolated clusters, or if there are data that, due to their rarity, cannot be incorporated into any cluster. These data, excluded from the most prevalent groups, can be interpreted as rare consumption patterns that are distant from the general characteristics of the data set, and thus can be classified as anomalies.

Experimental Setup

The experimental setup developed to carry out the steps of this methodology enables the classification of the algorithms applied to find the best clustering models into two large categories: two distance-based methods and three density-based methods. Within the first category, k-means and bisecting k-means are distinguished. These methods involve a parameter search centered on the k value: the number of clusters. This number is varied between 2 and 15 clusters. The second group of methods comprises DBSCAN, HDBSCAN, and OPTICS, whose parameter searches are characterized by two significant values. The first is the size of the neighborhood of a point, which is assessed to determine its proximity to an area of high point density; it was searched between 0.1 and 1.0. The second is the minimum number of points in that neighborhood from which a point can be included in a cluster; it was searched between 2 and 5.

The performance of these five methods is compared through the use of the Davies–Bouldin metric, a measure that takes into account both the internal consistency of the obtained clusters and the separation between different clusters [38], and that is commonly expressed using the following formula:

$$DB = \frac{1}{N}\sum_{i=1}^{N} \max\left(W_{ij}\right) \tag{1}$$

where N is the number of clusters and $W_{ij} = \frac{R_i+R_j}{d(c_i,c_j)}$ measures the quality of the separation between clusters i-th and j-th, being R_i the average distance of the points of cluster i-th to their mass center, and $d\left(c_i,\ c_j\right)$ the distance between mass centers of clusters i-th and j-th. In this case, smaller values of the metric indicate better segmentations.

Data access and code implementation of the methodology were carried out using the Python programming language. The application of the clustering models, the search for the best parameters and the calculation of the performance metric were carried out using the implementations included in the Scikit-Learn library [39].

4. Results and Discussion

4.1. Data Set Description and Exploratory Analysis

The data set of active and reactive energy measurements utilized in this study is composed of a series of measurements directly obtained from AMI devices associated with the Colombian electrical grid. No additional sources of information regarding the users in the data set are present; only their measurements are included. All data are found as records in a CSV table, and contain the user identification; the type of measurement, active or reactive; the value of the measurement; and the timestamp of the sample. The records span a period of four complete months, and as shown in Figure 1, each month is treated separately. In all cases, there are more active energy measurements than reactive energy.

To ensure data quality and remove missing data, the preprocessing steps outlined in the methodology are applied, resulting in a different number of users in each month. No data imputation is performed, since the elimination of these incomplete data represents a relatively minor impact, with less than 1% of the original users being removed. The final number of valid users for each month is presented in Table 1, where it is discriminated by the type of measurement: active or reactive. The first month shows a slight difference with respect to the other months, showing more active records and fewer reactive records.

Table 1. Number of users with complete measurements, discriminated by month and type of energy.

	Active	**Reactive**
Month 1	66,067	23,667
Month 2	60,605	27,760
Month 3	62,183	27,603
Month 4	62,377	27,879

With the preprocessed data set, a statistical analysis is performed to understand the general behavior of the data as a whole. For each hour of the day, the distribution of measurements is analyzed by using averages, medians and interquartile ranges, and the results for all days can be seen in Figure 2, separated by month and by energy type (active profiles on the left, reactive profiles on the right). In general, the averages (arithmetic means) have higher values than the medians, showing that there is a significant proportion of profiles with notably high values. This is particularly notorious for the reactive energy profiles, where the averages, shown as a green line, are outside the whiskers of the box plots. Although the consumption appears to be fairly similar for all hours of the day, the subtle differences between hours suggest that in general the lowest consumption is found in the early morning (hours 2 to 5) and the highest at night (hours 20 to 22).

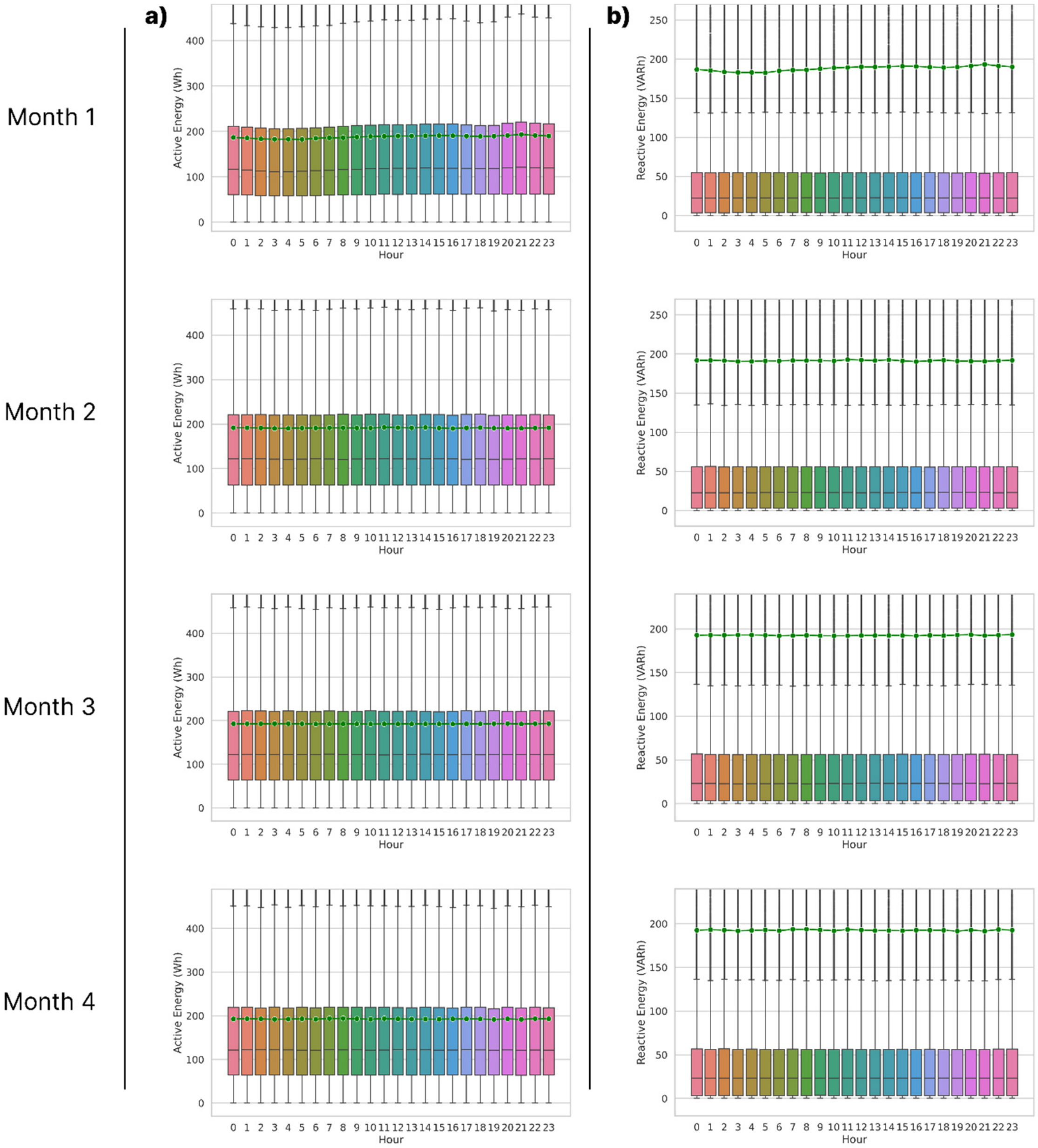

Figure 2. Statistical analysis over consumption profiles (hour by hour), showing interquartile ranges as colored box plots, and average (mean) values as a green line connecting through all hours. Section (**a**) shows the results of this analysis for active profiles, and section (**b**) the results for reactive profiles.

Another statistical analysis that we consider useful beyond the consumption profiles is to discard the separation between hours of the day in favor of considering days of the week. In this way, consumption is summarized not by hour but by day, and consumption that occurs on the same day is grouped together (i.e., all Mondays, all Tuesdays, etc.). The distribution of the measurements (carried out in a similar way to the previous case) is presented in Figure 3. For Sundays, there is a clear decrease in magnitude for all months and both types of measurements, probably related to the holiday status of Sundays in Colombia, when lower consumption would be expected. However, there are more subtle variations between days of the week for both active energy (lower Mondays in month 1, lower Fridays in months 3 and 4) and reactive energy (lower Wednesdays in month 2, higher Saturdays in month 3, higher Mondays in month 4).

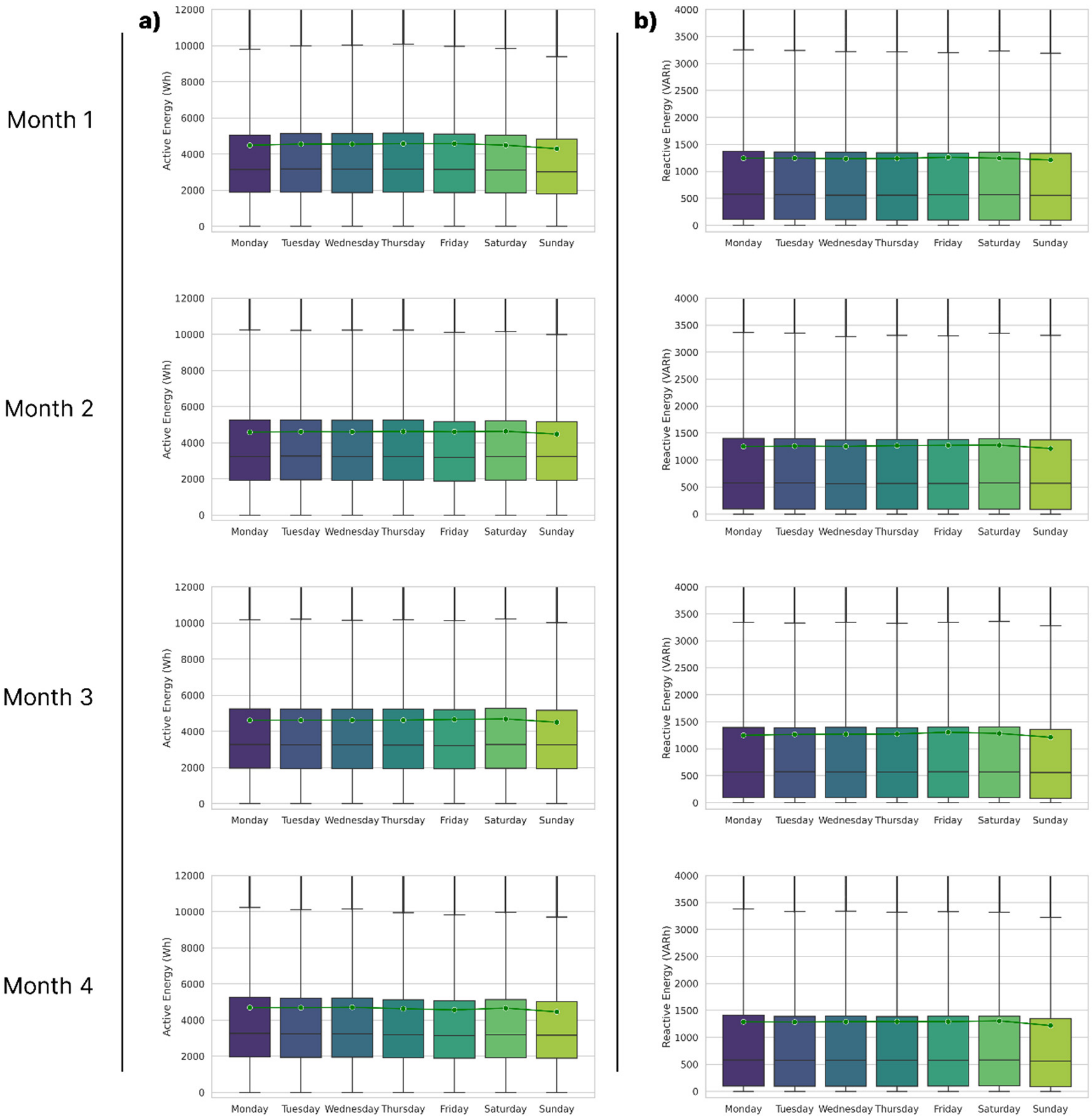

Figure 3. Statistical analysis over consumption data grouped by day of the week, showing interquartile ranges as colored box plots, and average (mean) values as a green line connecting through all week days. Section (**a**) shows the results of this analysis for active profiles, and section (**b**) the results for reactive profiles.

From this exploration of the data, we can determine that despite some seasonal variations, an average consumption profile centered on the 24 h of the day can adequately represent the users in the data set without much loss of information. For each user, average consumption profiles are created independently for active and reactive energy. These profiles are composed of 24 values, where each value corresponds to the user's average consumption in each hour (the first value would be the average of all records within hour 0, the second value would be the average of all records within hour 1, and so on up to hour 23). These lists of 24 values are the consumption profiles (active and reactive) associated with the user. The profiles have been created separately for each month, so each user can have up to eight profiles, two per month.

4.2. Comparison of Clustering Methods

The results obtained after applying the clustering methods are presented below. In total, each method was applied five times in different subsets of the data for each month: raw active data, normalized active data, raw reactive data, normalized reactive data, and calculated proportion (power factor). In each case, a hyperparameter search was performed for each method to improve its performance over the different types of data. The Davies–Bouldin scores of the methods with better parameter values are shown in Table 2. K-means and bisecting k-means perform notably better in the data considering magnitude (with both active and reactive profiles), while OPTICS and HDBSCAN are clearly dominant in normalized profiles. DBSCAN seems to be the worst method overall, showing a good performance in only a few cases.

Table 2. Davies–Bouldin scores of all algorithms over the data set, discriminated by months and energy types, considering absolute values, normalized values and power factor (proportion of active energy over total). The best values of the metric in each case are highlighted in bold.

Data Set	K-means	Bisecting K-means	DBSCAN	OPTICS	HDBSCAN
Month 1—Active	**0.7280**	0.7295	0.8462	1.4112	1.4316
Month 1—Normalized act.	1.1113	1.0842	1.1429	1.1628	**0.4283**
Month 1—Reactive	**0.6467**	**0.6467**	2.0449	2.1159	2.1629
Month 1—Normalized react.	1.7130	1.6714	0.3823	**0.3381**	0.3823
Month 1—Power factor	**0.7449**	0.7493	1.0270	1.2864	1.0424
Month 2—Active	0.6796	**0.6709**	0.8370	0.8418	1.4167
Month 2—Normalized act.	1.1736	1.1736	1.0828	0.6719	**0.4538**
Month 2—Reactive	0.6662	**0.5191**	2.0318	2.0534	2.0609
Month 2—Normalized react.	1.7119	1.7365	0.4015	**0.3696**	0.4015
Month 2—Power factor	**0.7363**	0.7613	0.7861	0.8810	1.0747
Month 3—Active	0.7378	**0.7140**	0.8487	0.8538	1.3511
Month 3—Normalized act.	1.1862	1.1862	1.0759	0.9411	**0.4478**
Month 3—Reactive	**0.4333**	0.7433	1.1437	1.0833	2.0768
Month 3—Normalized react.	1.7491	1.7779	**0.3881**	**0.3881**	**0.3881**
Month 3—Power factor	**0.7363**	0.7416	1.0034	1.3083	1.0795
Month 4—Active	**0.7099**	**0.7099**	0.8293	0.8348	1.5927
Month 4—Normalized act.	1.0199	1.0219	0.3896	**0.3810**	0.4253
Month 4—Reactive	**0.4842**	0.6265	1.9510	0.7387	2.1959
Month 4—Normalized react.	1.7788	1.7550	0.3881	**0.3271**	0.3881
Month 4—Power factor	**0.6932**	0.7098	1.0586	1.3340	1.0648

4.3. Active Energy Analysis

The results of the best methods on the active profiles can be seen in Figure 4. The graphs associated with absolute (raw) energy values can be seen on the left (Figure 4a), and those associated with normalized profiles on the right (Figure 4b). Each cluster is represented either using its centroid (in the case of k-means and bisecting k-means) or the average of all the profiles of each group (for OPTICS and HDBSCAN).

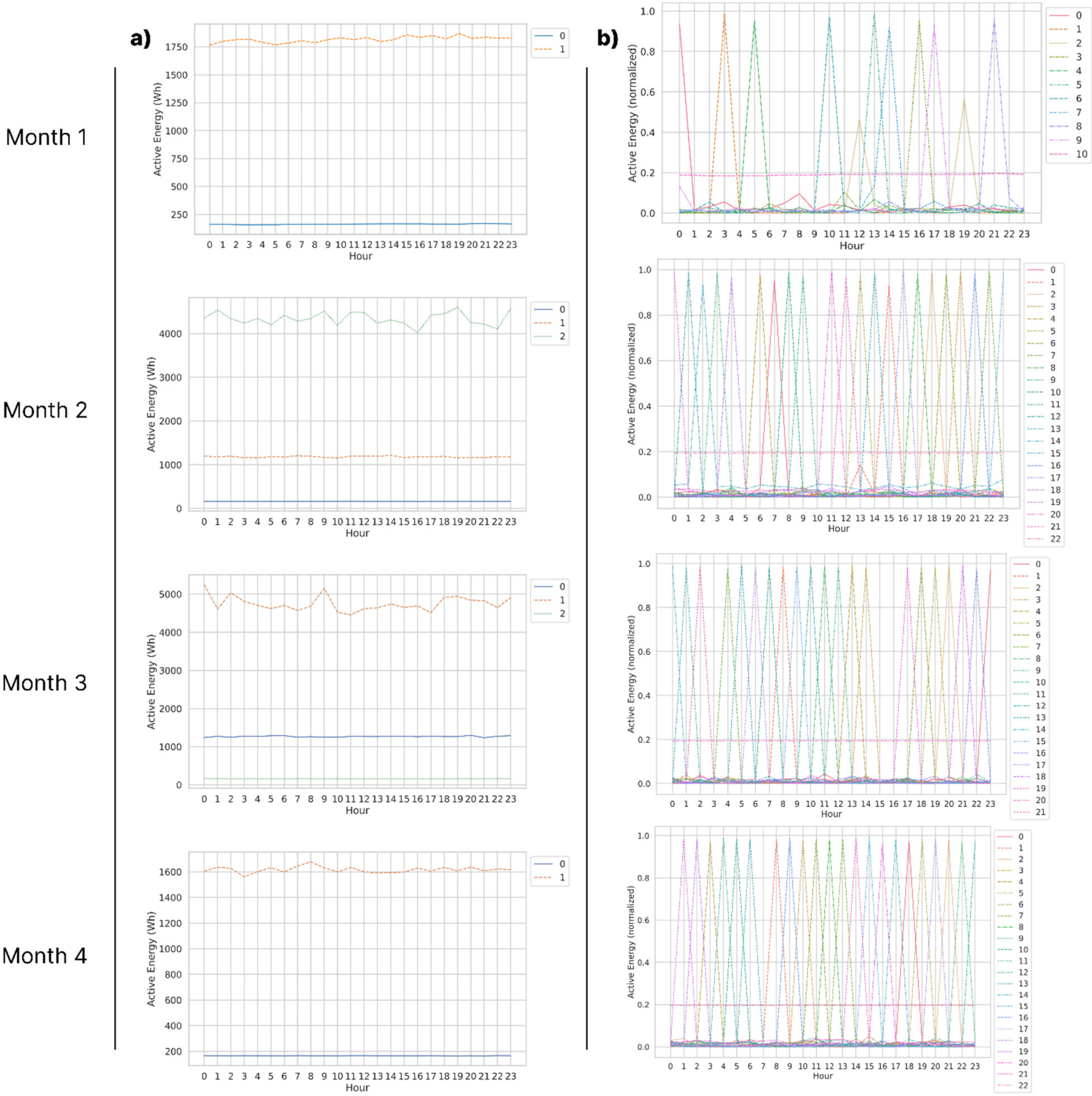

Figure 4. Cluster averages (centroids) obtained by applying the best clustering method on active energy user profiles. Section (**a**) shows the results for active profiles considering magnitude, and section (**b**) shows the results for normalized active profiles.

Absolute energy profiles (Figure 4a) are divided into two or three clusters, depending on the month. In all cases, there is a cluster of low values, whose centroid lies around 200 watt-hours throughout the day. When there are two clusters (the cases of months 1 and 4), the other cluster has a centroid with a slightly more irregular curve, around 1.6 or 1.7 kWh. In the other two months (months 2 and 3), there are three clusters, with the second cluster centered around 1.2 kWh and the third cluster (with a centroid curve that varies notably more than other clusters) between 4 and 5 kWh. This indicates that, in general, users can be separated across some well-defined regions: users with low energy consumption around 150 kWh per month, users with medium energy consumption between 850 and 1300 kWh per month, and users with high energy consumption around 3000 kWh per month or more.

Normalized energy profiles (Figure 4b) show consistent behavior across all months, where in all of them, there is a cluster with a very flat curve around 0.2 (the majority behavior in all cases), and a series of small clusters with a single sharp peak in a specific hour of the day and small values in all other hours. Month 1 is kind of outlier in this scenario since the number of clusters is notably lower than in other months, and one of the peaked clusters shows two peaks in different hours instead of only one in a single hour. These small-peaked clusters can be seen as anomalous, given the majority behavior is notably flatter, and as such, can point to users with problems of unexpected voltage peaks, possible measurement errors, or other vulnerabilities like power theft.

4.4. Reactive Energy Analysis

The centroids of the best methods for reactive energy (both raw/absolute and normalized) can be seen in Figure 5. In a similar way to active energy, absolute reactive profiles (in Figure 5a) are separated into two or three clusters in almost all months, with only month 2 showing four clusters. However, the behaviors in each case are not as comparable as before. Clusters in months 1 and 4 (the ones with only two clusters) differ notably in magnitude, with centroids around 50 and 900 VArh in month 1 to centroids around 25 and 145 VArh in month 4. Months 2 and 3, although both are showing a low cluster around 30 VArh and a very high one around 3000 VArh, differ in the middle region, showing either two clusters located around 250 and 1000 VArh (month 2) or a single one around 600 VArh (month 3). A general pattern in this case should separate users in three regions: a low reactive consumption area from 0 to 36 kVArh per month (the lower clusters in all months), a broad middle reactive consumption area between 90 and 720 kVArh per month (the upper clusters in months 1 and 4, and the middle ones in months 2 and 3), and a high reactive consumption area above 1800 kVArh per month.

Normalized reactive clusters (in Figure 5b) show similar patterns to the active ones, with the majority behavior shown in all cases as a flat line around 0.2 and a series of small clusters with a single peak during a specific hour of the day. The main difference with the active ones is the number of clusters: for active profiles, there are between 21 and 22 clusters (with the notable exception of month 1), but for reactive profiles, there are between 15 and 20 clusters. The behavior of the reactive clusters between the four months does not vary too much, showing differences only in the hours in which the peaks are located. Month 1 is the only one where a cluster with two peaks appears (located in hours 11 and 18), and hour 11 is the only one where there is no peak for any cluster in the other three months. As before, these peaked clusters can be interpreted as anomalous behaviors that could be related to inappropriate compensation strategies, events of overcharge in the grid, or measurement errors.

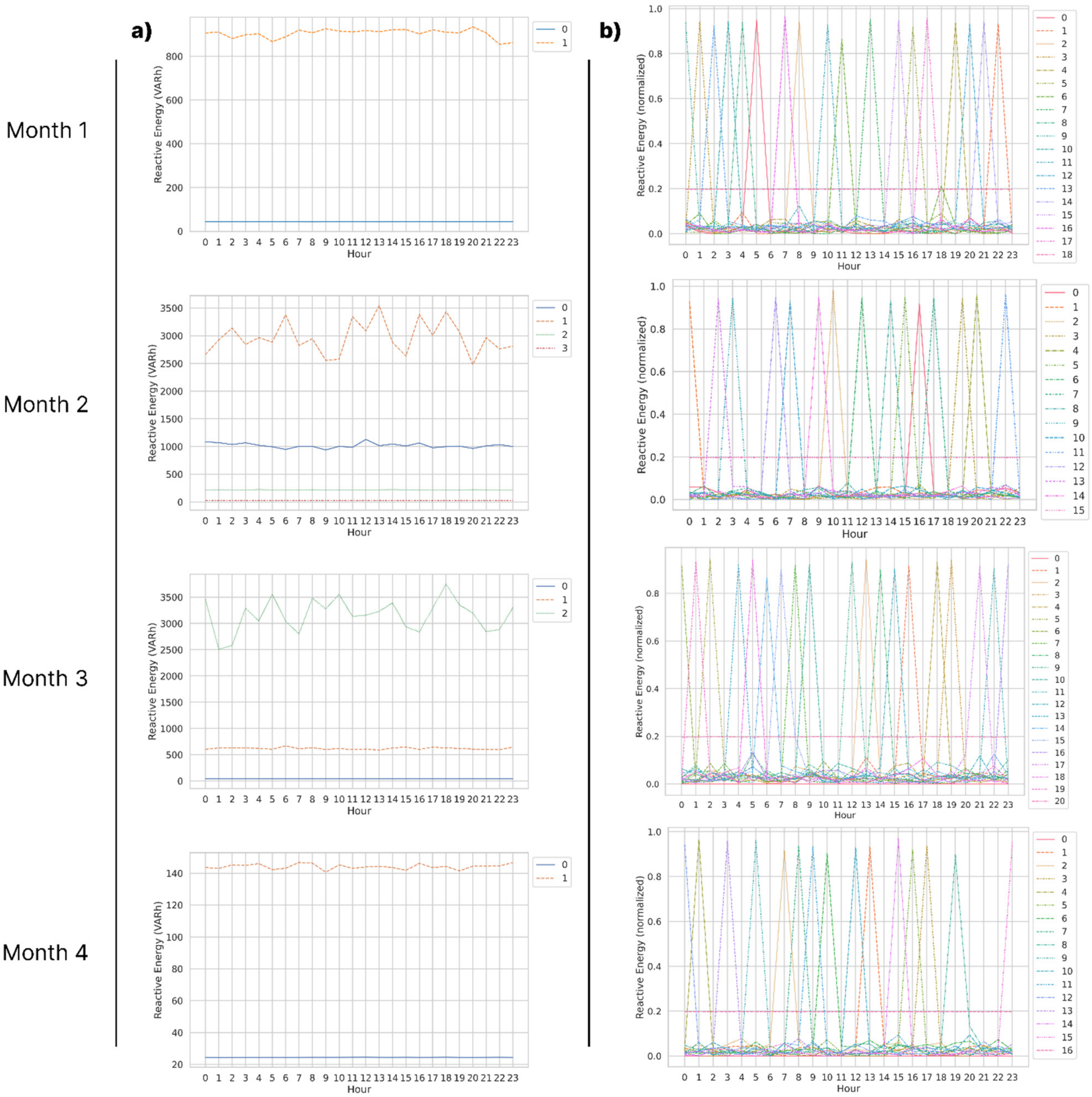

Figure 5. Cluster averages (centroids) obtained by applying the best clustering method on reactive energy user profiles. Section (**a**) shows the results for reactive profiles considering magnitude, and section (**b**) shows the results for normalized reactive profiles.

4.5. Power Factor Analysis

The power factor profiles of a user are obtained by finding the ratio of active to total energy (the sum of active and reactive consumptions) for each hour of the day. The power factor always has values between zero and one, and the values along the day can change, although they tend to stay around certain values. For all of the four months, the best clustering method for power factor profiles was k-means, as can be seen in Table 2. In all of these cases, the search for the optimal number of clusters gave the same result of three clusters with roughly the same behavior. The results of the analysis are presented in Figure 6. The curves of the centroids of the three clusters (an example of which is shown in Figure 6a) are generally flat throughout the day. The upper cluster (in blue) always lies around 0.90, showing a clear advantage of active energy. The middle cluster (in orange) always lies around 0.64, indicating that active and reactive energies are much more similar, although with a slight advantage of active energy. The lower cluster (in green) always lies around 0.11, showing that reactive energy clearly dominates, in contrast to the other clusters. Considering the importance of these three groups for all months (shown in the bar plot in Figure 6b), the relative size of each group is roughly similar in all months, varying only a few decimals in the blue and orange clusters. In general, these two clusters where active energy is bigger than reactive (blue and orange) concentrate 96% of all the data in any given month, while the dominance of reactive energy is only clear in the remaining 4% of the total.

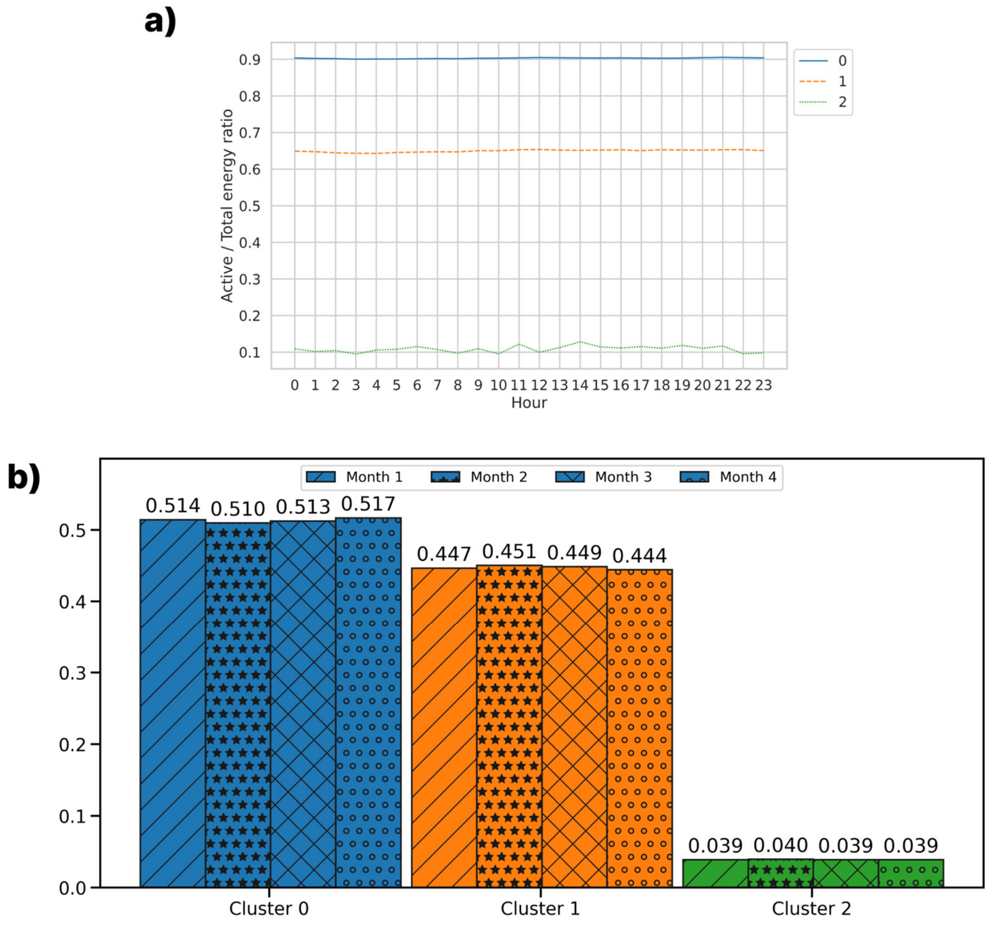

Figure 6. Results of the segmentation of users based on power factor values. Since the results are pretty similar for the four months, section (**a**) shows the centroids of the three clusters obtained in a particular month to showcase the levels around which each cluster is centered. Section (**b**) is a bar plot that presents the relative weight of the three clusters for each month, showing that the total of users is roughly divided 51%/45%/4% among the three clusters.

The distribution of the active and reactive measurements within each of the three groups can be seen in Figure 7. For this analysis, the power factor profiles of each user were summarized in a single average value and discriminated according to the cluster to which they belong. The generated graphs, called violin plots, combine box plots (here, as black rectangles) with a representation of the density of the data similar to a histogram, in order to illustrate the regions that concentrate most of the data. When grouping active and reactive values from the same cluster, a clear difference in shape can be observed between the clusters, given that in the reactive dominant cluster (green), both of the consumption values are much closer to zero, while for the other two groups, the active energy values tend to be distributed in a similar way while the reactive energy values appear in different intervals. With these plots, we can characterize the three clusters: the blue cluster is composed of users with relevant active measurements but relatively small reactive measurements; the orange cluster is composed of users with both relevant active and reactive measurements (with a visible advantage of active energy); and the green cluster is composed of users with both small active and reactive measurements (with a clearer advantage of reactive energy).

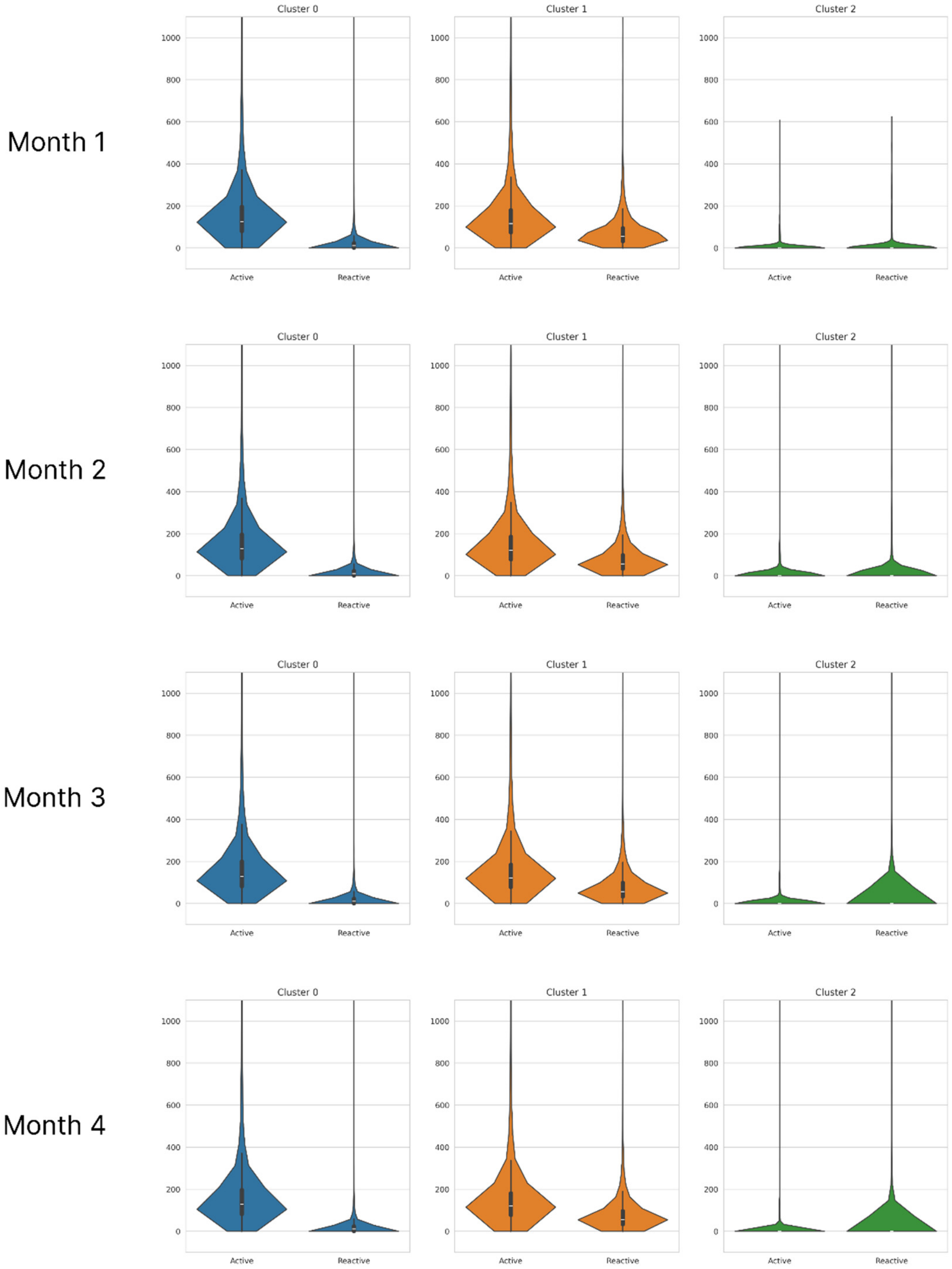

Figure 7. Statistical exploration of the consumption magnitude of users, when grouped by the three power factor clusters. For each month, the distribution of magnitudes of active and reactive consumption is shown as a violin plot discriminated by clusters. Each color represents the same cluster in all plots, and cluster colors and numbers (0, 1, 2) also coincide with the ones shown in Figure 6.

4.6. User Segmentation and Anomaly Detection

Based on the different behaviors found for users in both active and reactive energy (which can be divided into high, medium and low consumption considering absolute measurements, and into normal and atypical profile curves considering normalized measurements), a more detailed characterization of users can be built by combining these different properties. To achieve a more specific identification of user behavior, it is necessary to consider two different scenarios: one associated only with active energy measurements, and another associated with active and reactive energy measurements. This is required because the number of users with active energy measurements alone is greater than the number of users with measurements of both types of energy (as stated in Table 1).

Considering only active energy measurements, it is possible to classify users into six groups: high measurements with normal behavior, high measurements with anomalous behavior, medium measurements with normal behavior, medium measurements with anomalous behavior, low measurements with normal behavior, and low measurements with anomalous behavior. To illustrate this fine-grain classification of users, we take month 3 as an example (however, this general idea is applicable to all of them). By performing the separation on the active energy data of month 3, Table 3 is obtained, in which the relative weight of each one of the groups is determined. The discrepancy in users with the values shown in Table 1 is due to users whose consumption was zero in all times.

Table 3. Users discriminated by active energy consumption behavior (both in magnitude and according to their curve geometry) in month 3. The last column shows the relative proportions of normal and anomalous values in each consumption level.

	Normal	**Anomalous**	**Proportion**
Low consumption	38,276	21,088	64.5%/35.5%
Medium consumption	831	336	71.2%/28.8%
High consumption	61	19	76.3%/23.7%

Figure 8 shows some examples of consumption profiles identified within each of the six groups. Although all the plots show frequent and high amplitude variations, with steep valleys and peaks, the profiles labeled as normal (in blue) tend to exhibit this fluctuating behavior throughout the day, rising and falling more or less constantly, while the profiles labeled as anomalous (in red) show some much flatter regions followed by abrupt variations, and the location of these flat areas appears not to be consistent. Although the magnitude of the consumption variations can vary greatly from case to case, the main difference appears to lie in whether these changes are constant throughout the day, or whether they occur more suddenly and infrequently.

When considering users with both types of measurements, a much finer separation can be established, dividing the users both by their magnitudes in active and reactive energy consumption, and determining whether they fit within the normal behavior for both types of energy, for only one or for neither. The results of this separation can be seen in Table 4, where the rows correspond to active measurements and the columns to reactive measurements. In this case, users with zero consumption were also discarded.

Some examples of profile curves for this scenario are shown in Figure 9. In each case, both profiles are shown, active ones as continuous lines and reactive ones as dotted lines, and the colors indicate whether these profiles are considered normal (in blue) or anomalous (in red). In most users, the geometries of both profiles are very similar in terms of shape. This is especially notable when the magnitudes of both profiles are similar, and a little less so when they differ greatly. Visually, the distinction between the groups is quite clear, particularly between groups with a large difference in magnitude (high active and medium reactive, high active and low reactive), groups with a moderate difference (high active and high reactive, medium active and low reactive), groups with more similar magnitudes (medium active and medium reactive, low active and low reactive), and groups where reactive is greater (low active and medium reactive). For profile geometries, the differentiation between normal profiles with constant variations and atypical profiles with flat areas is clearly maintained in active energy, but in reactive energy it is a little more difficult to see, given the lower magnitude of their variations compared to active energy. These smaller variations could explain why reactive profiles with some flatter regions or less frequent variations are more likely to be classified as normal.

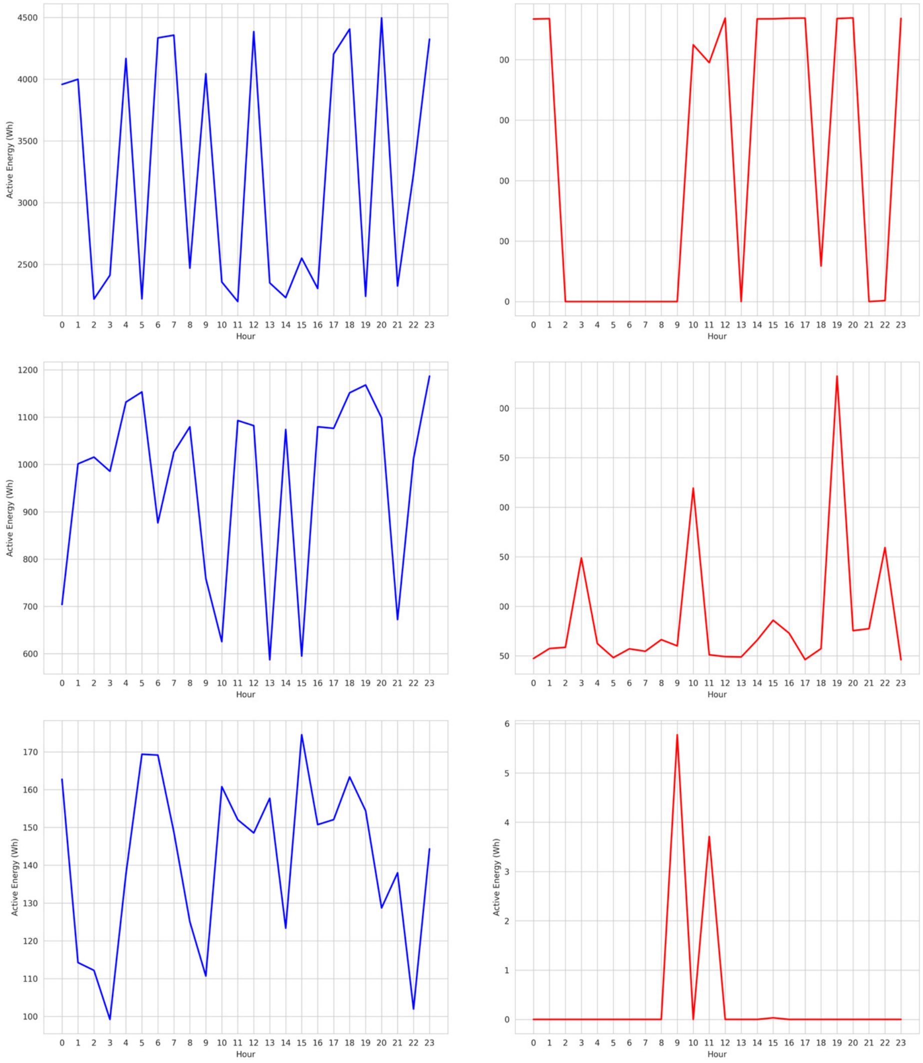

Figure 8. Examples of profile curves associated with users in each one of the six groups defined in Table 3. From top to bottom, high, medium and low consumption magnitudes; from left to right, normal and anomalous behaviors.

Table 4. Users discriminated by active and reactive energy consumption behavior in month 3. In each cell, users are discriminated by the normality (or anomalousness) of their curve geometries for both types of measurement.

		Reactive		
		Low Consumption	**Medium Consumption**	**High Consumption**
Active	Low consumption	Both normal: 14,621	Both normal: 125	Both normal: 1
		One normal: 8340	One normal: 35	One normal: 0
		Anomalous: 2320	Anomalous: 21	Anomalous: 0

Table 4. *Cont.*

		Reactive		
		Low Consumption	**Medium Consumption**	**High Consumption**
Active	Medium consumption	Both normal: 153	Both normal: 170	Both normal: 2
		One normal: 85	One normal: 44	One normal: 1
		Anomalous: 37	Anomalous: 26	Anomalous: 0
	High consumption	Both normal: 3	Both normal: 8	Both normal: 14
		One normal: 2	One normal: 2	One normal: 2
		Anomalous: 4	Anomalous: 0	Anomalous: 2

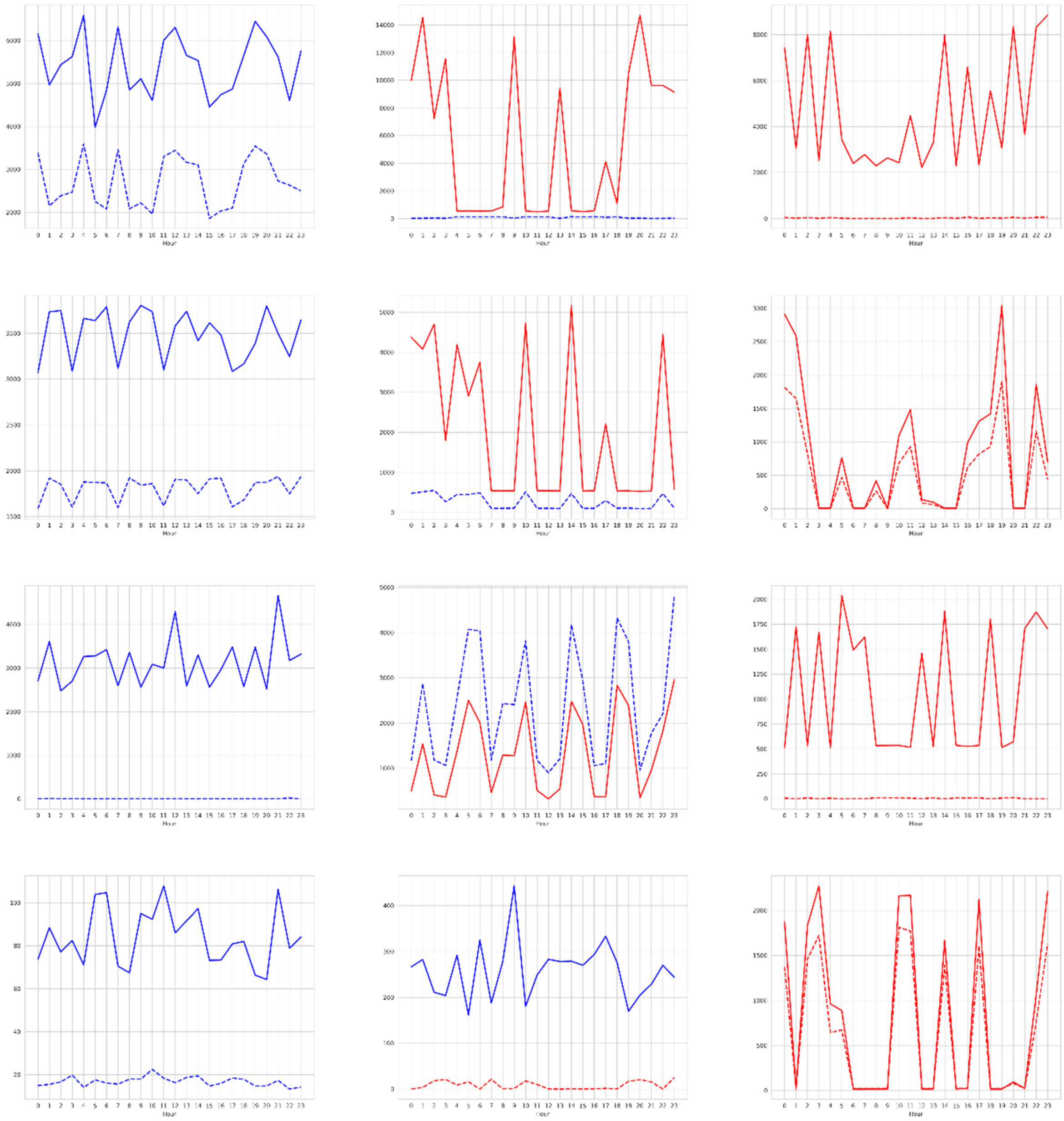

Figure 9. Examples of profile curves associated groups of users defined in Table 4. Left column, users with both normal active and reactive profiles. Middle column, users with one normal profile and one anomalous profile. Right column, users with both anomalous active and reactive profiles.

In summary, anomalous users can be identified in two ways. If only active power is analyzed, anomalous users can be identified by analyzing whether their curves vary consistently throughout the day or whether they have flat areas surrounded by valleys or peaks; the latter could be signaled as anomalous behavior. When looking at both active and reactive power consumption, curves with flat areas are still more likely to be associated with anomalous users, although this is less noticeable for reactive power. The frequency of anomalous users does not appear to be closely related to the magnitude of their consumption, appearing in similar proportions for the different levels. Regarding the geometry of consumption curves, anomalous users tend to exhibit curves with similar geometries, particularly in the flatter areas or those with minimal variation. In contrast, users without anomalous consumption typically display slightly greater discrepancies (though still minimal) in the location of peaks and valleys or the slopes of their changes. Using these clues, grid operators can identify users with these rare consumption profiles and define strategies to analyze these users in more detail and determine the cause of their consumption variations.

5. Conclusions

In this work, we present the application of a series of clustering methods to a data set of active and reactive energy measurements of end users taken by AMI meters located in Colombia. The data set covers a period of four months in 2023, and the results of these methods are compared through the use of the Davies–Bouldin metric. This allows us to identify interesting patterns that reveal the behavior of users with respect to their consumption. In terms of consumption magnitude, the most effective methods for user segmentation are consistently k-means and bisecting k-means. When considering the geometry of consumption patterns throughout the day, the leading methods are HDBSCAN and OPTICS. The segmentations found when considering active and reactive energy allow users to be associated with a small number of well-differentiated clusters, depending on their monthly consumption magnitudes for both active energy (low around 150 kWh, medium between 850 and 1300 kWh, and high around 3000 kWh per month or more) and reactive energy (low until 36 kVArh per month, medium between 90 and 720 kVArh per month, and high above 1800 kVArh). In addition to this, an analysis of the proportion of active energy over total energy (also known as the power factor) is performed, showing a behavior clearly divided into three clusters of users: a group with a clear predominance of active energy (centered near 90%), a group with a clear predominance of reactive energy (centered near 10%), and a group where both types of consumption are similar, with a slight advantage of active energy (centered near 64%). The number of clusters remained constant across all months, which suggests a clear delineation between these groups, independent of seasonal variations.

As users are distinguished by both the magnitude and the distribution of their consumption throughout the day, they can also be classified based on the separations provided by some of the clustering methods, particularly the identification of anomalies as data points that cannot be allocated to any cluster. This allows for the further differentiation of users based on four characteristics: their level of active consumption, their level of reactive consumption, the abnormality of their active profile, and the abnormality of their reactive profile. Anomalous behaviors tend to be represented in profiles with both sharp changes and flat areas throughout the day, particularly in active profiles. This result can serve as the basis for a simple procedure to identify users showcasing anomalous behavior in real time and with a relatively low computational cost. In combination with the segmentation of user behaviors, this strategy can help network operators in two clear ways: to identify users that generate a larger amount of reactive energy and thus to develop effective compensation strategies or to adapt tariff schemes for reactive energy generation, and to focus efforts in specific users to explain abnormal variations and to address the underlying problems that may cause them.

This proposal offers a straightforward and efficient method for characterizing user behavior, independent of external data sources. It emphasizes direct analysis, focusing solely on consumption measurements. This approach is applicable to any smart grid with AMI meters capable of measuring active and reactive energy at the user level. While the results obtained for the data set used are applicable to a specific location and time period, the geographic and socioeconomic characteristics of the users who are part of the sample could allow these results to be extrapolated to other regions of Colombia or to countries with similar geography and economic level in Latin America. Future lines of work with this proposal may include new sources of user information that could allow relating the results obtained with categories such as customer types (residential, commercial, industrial) or with other types of variables such as geographic location (e.g., elevation or temperature) or socioeconomic conditions (e.g., income). Another interesting possibility may be directed to combine this proposal with methodologies focused on other levels of the electrical grid, such as electrical hubs or substations, to comprehensively understand the overall distribution of energy in the network.

Author Contributions: Conceptualization, J.R.G.; methodology, O.A.B.-B.; investigation, O.A.B.-B.; writing—original draft preparation, O.A.B.-B.; writing—review and editing, O.A.B.-B. and J.R.G.; project administration, J.R.G.; funding acquisition, J.R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Electrical Machines and Drives (EM&D) research group from Universidad Nacional de Colombia, Project: "Red de cooperación de soluciones energéticas para comunidades", code: 59384, and Center of Excellence in Scientific Computing: CoE-SciCo, Code: 52625.

Data Availability Statement: Additional data are available on request by contacting the corresponding author of this manuscript.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Mohassel, R.R.; Fung, A.S.; Mohammadi, F.; Raahemifar, K. A survey on advanced metering infrastructure and its application in smart grids. In Proceedings of the 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), Toronto, ON, Canada, 4–7 May 2014; pp. 1–8.
2. Ghosal, A.; Conti, M. Key management systems for smart grid advanced metering infrastructure: A survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 2831–2848. [CrossRef]
3. Li, W.T.; Yuen, C.; Hassan, N.U.; Tushar, W.; Wen, C.K.; Wood, K.L.; Hu, K.; Liu, X. Demand response management for residential smart grid: From theory to practice. *IEEE Access* **2015**, *3*, 2431–2440. [CrossRef]
4. Assad, U.; Hassan, M.A.S.; Farooq, U.; Kabir, A.; Khan, M.Z.; Bukhari, S.S.H.; Jaffri, Z.U.A.; Olah, J.; Popp, J. Smart grid, demand response and optimization: A critical review of computational methods. *Energies* **2022**, *15*, 2003. [CrossRef]
5. Nazir, A.; Shaikh, A.K.; Shah, A.S.; Khalil, A. Forecasting energy consumption demand of customers in smart grid using Temporal Fusion Transformer (TFT). *Results Eng.* **2023**, *17*, 100888. [CrossRef]
6. Ahmad, A.; Javaid, N.; Mateen, A.; Awais, M.; Khan, Z.A. Short-term load forecasting in smart grids: An intelligent modular approach. *Energies* **2019**, *12*, 164. [CrossRef]
7. Lilliu, F.; Vinyals, M.; Denysiuk, R.; Recupero, D.R. A novel payment scheme for trading renewable energy in smart grid. In Proceedings of the Tenth ACM International Conference on Future Energy Systems, Phoenix, AZ, USA, 25–28 June 2019; pp. 111–115.
8. Duarte, J.E.; Rosero-Garcia, J.; Duarte, O. Analysis of Variability in Electric Power Consumption: A Methodology for Setting Time-Differentiated Tariffs. *Energies* **2024**, *17*, 842. [CrossRef]

9. Dileep, G.J.R.E. A survey on smart grid technologies and applications. *Renew. Energy* **2020**, *146*, 2589–2625. [CrossRef]
10. Ponce-Jara, M.A.; Ruiz, E.; Gil, R.; Sancristóbal, E.; Pérez-Molina, C.; Castro, M. Smart Grid: Assessment of the past and present in developed and developing countries. *Energy Strategy Rev.* **2017**, *18*, 38–52. [CrossRef]
11. Sarkar, M.N.I.; Meegahapola, L.G.; Datta, M. Reactive power management in renewable rich power grids: A review of grid-codes, renewable generators, support devices, control strategies and optimization algorithms. *IEEE Access* **2018**, *6*, 41458–41489. [CrossRef]
12. Stanelytė, D.; Radziukynas, V. Analysis of voltage and reactive power algorithms in low voltage networks. *Energies* **2022**, *15*, 1843. [CrossRef]
13. Águila Téllez, A.; López, G.; Isaac, I.; González, J.W. Optimal reactive power compensation in electrical distribution systems with distributed resources. *Rev. Heliyon* **2018**, *4*, 746. [CrossRef] [PubMed]
14. Montoya, O.D.; Gil-González, W. Dynamic active and reactive power compensation in distribution networks with batteries: A day-ahead economic dispatch approach. *Comput. Electr. Eng.* **2020**, *85*, 106710. [CrossRef]
15. Andrade, I.; Pena, R.; Blasco-Gimenez, R.; Riedemann, J.; Jara, W.; Pesce, C. An active/reactive power control strategy for renewable generation systems. *Electronics* **2021**, *10*, 1061. [CrossRef]
16. Stanelyte, D.; Radziukynas, V. Review of voltage and reactive power control algorithms in electrical distribution networks. *Energies* **2019**, *13*, 58. [CrossRef]
17. Shaheen, A.M.; Spea, S.R.; Farrag, S.M.; Abido, M.A. A review of meta-heuristic algorithms for reactive power planning problem. *Ain Shams Eng. J.* **2018**, *9*, 215–231. [CrossRef]
18. Rajabi, A.; Li, L.; Zhang, J.; Zhu, J.; Ghavidel, S.; Ghadi, M.J. A review on clustering of residential electricity customers and its applications. In Proceedings of the 2017 20th International Conference on Electrical Machines and Systems (ICEMS), Sydney, Australia, 11–14 August 2017; pp. 1–6.
19. Ruff, L.; Kauffmann, J.R.; Vandermeulen, R.A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T.G.; Müller, K.R. A unifying review of deep and shallow anomaly detection. *Proc. IEEE* **2021**, *109*, 756–795. [CrossRef]
20. Coman, C.M.; Florescu, A.; Oancea, C.D. Improving the efficiency and sustainability of power systems using distributed power factor correction methods. *Sustainability* **2020**, *12*, 3134. [CrossRef]
21. Wahab, K.; Rahal, M.; Achkar, R. Economic improvement of power factor correction: A case study. *J. Power Energy Eng.* **2021**, *9*, 1–11. [CrossRef]
22. Milton, M.A.; Pedro, C.O.; Xavier, S.G.; Guillermo, E.E. Characterization and classification of daily electricity consumption profiles: Shape factors and k-means clustering technique. In Proceedings of the E3S Web of Conferences, 3rd International Conference on Power and Renewable Energy, Berlin, Germany, 21–24 September 2018; Volume 64, p. 08004.
23. Motlagh, O.; Berry, A.; O'Neil, L. Clustering of residential electricity customers using load time series. *Appl. Energy* **2019**, *237*, 11–24. [CrossRef]
24. Aggarwal, C.C.; Reddy, C.K. Chapter I: An Introduction to Cluster Analysis. In *Data Clustering. Algorithms and Applications*; CRC Data mining and Knowledge Discovery Series; Chapman & Hall: Boca Raton, FL, USA, 2024; ISBN 978-1-4665-5821-2.
25. Viegas, J.L.; Vieira, S.M.; Melício, R.; Mendes, V.M.F.; Sousa, J.M. Classification of new electricity customers based on surveys and smart metering data. *Energy* **2016**, *107*, 804–817. [CrossRef]
26. Amri, Y.; Fadhilah, A.L.; Setiani, N.; Rani, S. Analysis clustering of electricity usage profile using k-means algorithm. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Kerala, India, 5–8 June 2016; Volume 105, p. 012020.
27. Okereke, G.E.; Bali, M.C.; Okwueze, C.N.; Ukekwe, E.C.; Echezona, S.C.; Ugwu, C.I. K-means clustering of electricity consumers using time-domain features from smart meter data. *J. Electr. Syst. Inf. Technol.* **2023**, *10*, 2. [CrossRef]
28. Tureczek, A.; Nielsen, P.S.; Madsen, H. Electricity consumption clustering using smart meter data. *Energies* **2018**, *11*, 859. [CrossRef]
29. Banerjee, S.; Choudhary, A.; Pal, S. Empirical evaluation of k-means, bisecting k-means, fuzzy c-means and genetic k-means clustering algorithms. In Proceedings of the 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Dhaka, Bangladesh, 19–20 December 2015; pp. 168–172.
30. Rohilla, V.; Chakraborty, S.; Singh, M.S. Data clustering using bisecting k-means. In Proceedings of the 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 18–19 October 2019; pp. 80–83.
31. Wang, K.; Yang, R.; Liu, C.; Samarasinghalage, T.; Zang, Y. Extracting Electricity Patterns from High-dimensional Data: A comparison of K-Means and DBSCAN algorithms. In Proceedings of the IOP Conference Series: Earth and Environmental Science, 2022, Medan, Indonesia, 29 October 2022; Volume 1101, p. 022007.
32. Zhang, L.; Deng, S.; Li, S. Analysis of power consumer behavior based on the complementation of K-means, DBSCAN. In Proceedings of the 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, China, 26–28 November 2017; pp. 1–5.

33. Liu, X.; Ding, Y.; Tang, H.; Xiao, F. A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy Build.* **2021**, *231*, 110601. [CrossRef]
34. Wang, F.; Li, K.; Duić, N.; Mi, Z.; Hodge, B.M.; Shafie-Khah, M.; Catalão, J.P. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. *Energy Convers. Manag.* **2018**, *171*, 839–854. [CrossRef]
35. Miraftabzadeh, S.M.; Colombo, C.G.; Longo, M.; Foiadelli, F. K-means and alternative clustering methods in modern power systems. *IEEE Access* **2023**, *11*, 119596–119633. [CrossRef]
36. Wang, P.; Govindarasu, M. Anomaly detection for power system generation control based on hierarchical, DBSCAN. In Proceedings of the 2018 North American Power Symposium (NAPS), Fargo, ND, USA, 9–11 September 2018; pp. 1–5.
37. Hurst, W.; Montañez CA, C.; Shone, N. Time-pattern profiling from smart meter data to detect outliers in energy consumption. *IoT* **2020**, *1*, 6. [CrossRef]
38. Chaudhry, M.; Shafi, I.; Mahnoor, M.; Vargas DL, R.; Thompson, E.B.; Ashraf, I. A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. *Symmetry* **2023**, *15*, 1679. [CrossRef]
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.