mathematics

MDPI

Article

Modeling Residential Energy Consumption Patterns with Machine Learning Methods Based on a Case Study in Brazil

Lucas Henriques [1,2], Cecilia Castro [1,*], Felipe Prata [2], Víctor Leiva [3,*] and René Venegas [4]

[1] Centre of Mathematics, Universidade do Minho, 4710-057 Braga, Portugal; lucasdestefano2@hotmail.com
[2] Instituto Federal de Alagoas, Maceió 57035-350, Alagoas, Brazil; felipepratalima@gmail.com
[3] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile
[4] Doctorate Program in Intelligent Industry, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile; rene.venegas@pucv.cl
* Correspondence: cecilia@math.uminho.pt (C.C.); victorleivasanchez@gmail.com or victor.leiva@pucv.cl (V.L.)

Abstract: Developing efficient energy conservation and strategies is relevant in the context of climate change and rising energy demands. The objective of this study is to model and predict the electrical power consumption patterns in Brazilian households, considering the thresholds for energy use. Our methodology utilizes advanced machine learning methods, such as agglomerative hierarchical clustering, k-means clustering, and self-organizing maps, to identify such patterns. Gradient boosting, chosen for its robustness and accuracy, is used as a benchmark to evaluate the performance of these methods. Our methodology reveals consumption patterns from the perspectives of both users and energy providers, assessing the corresponding effectiveness according to stakeholder needs. Consequently, the methodology provides a comprehensive empirical framework that supports strategic decision making in the management of energy consumption. Our findings demonstrate that k-means clustering outperforms other methods, offering a more precise classification of consumption patterns. This finding aids in the development of targeted energy policies and enhances resource management strategies. The present research shows the applicability of advanced analytical methods in specific contexts, showing their potential to shape future energy policies and practices.

Keywords: artificial intelligence; consumption profiles; energy management; multi-class classification; pattern recognition; residential energy use

MSC: 68T10; 68T05

Citation: Henriques, L.; Castro, C.; Prata, F.; Leiva, V.; Venegas, R. Modeling Residential Energy Consumption Patterns with Machine Learning Methods Based on a Case Study in Brazil. *Mathematics* **2024**, *12*, 1961. https://doi.org/10.3390/math12131961

Academic Editor: Vassilis C. Gerogiannis

Received: 12 May 2024
Revised: 10 June 2024
Accepted: 15 June 2024
Published: 25 June 2024

1. Introduction

In the current digital era, the revolution of data has transformed numerous sectors. This revolution has driven not only technological advancements but also spurred a critical need for extracting valuable information [1,2]. Consequently, sophisticated methods of data analytics have emerged as a result of such a revolution. These methods are applied in fields ranging from medical and educational studies [3–5] to the optimization of industrial processes and global supply chains [6]. Such technological advancements enhance infrastructure reliability through defect detection and data privacy [7,8].

Machine learning methods have played a crucial role in healthcare by providing predictive insights that can save lives and improve treatment efficacy [9,10]. Additionally, methods such as smart meter data-based algorithms for optimal phase load balancing highly improve operational efficiency in distribution networks [11], reflecting broader trends in energy systems optimization [12,13].

Detecting energy consumption patterns is vital for developing efficient energy management strategies and designing future-oriented energy network architectures. This detection becomes more pressing in light of climate change and the continuous increase in energy demand [14].

Understanding the consumption profiles of both electrical power consumers and suppliers is relevant to achieving energy efficiency. This understanding facilitates the identification of opportunities to reduce unnecessary consumption and optimize the use of energy, leading to a sustainable management of energy resources [15]. The study presented in [16] demonstrated how detailed data analytics can facilitate the transition to sustainable and efficient energy systems. In the Brazilian context, marked by a strong economy and considerable demographic diversity, the aforementioned understanding of the consumption profiles is exacerbated. Specific policies, such as legal consumption limits for billing purposes, add complexity to the analysis of energy consumption in Brazil.

Clustering is effective for grouping variables identifying patterns that are used in various fields. In the study of energy consumption patterns, clustering has been effective [17–20]. However, the application of clustering methods requires adaptation to local contexts, which can present challenges and opportunities. The Brazilian scenario, with its diverse demography and economy, offers a distinct context where traditional methods may need tailoring for accurate analysis. Despite the extensive body of research in the field, important gaps persist, particularly in the application of clustering methods to specific contexts such as Brazilian households. These gaps highlight the need for research that is not only context-specific but also adaptable to the unique challenges presented by different cultures, economies, and geographies worldwide. Our study addresses these gaps. Therefore, the objective of the present study is to model and predict the electrical power consumption patterns in Brazilian households, considering the thresholds for energy use. We utilize machine learning methods particularly focused on clustering for energy consumption analysis considering gradient boosting (GB) as a benchmarking tool [21,22].

We select GB for its robustness in handling heterogeneous data and ability to capture complex interactions, making it a reliable benchmark for evaluating the performance of clustering algorithms. This selection ensures a comparison and validation of clustering methods, establishing a foundation for identifying consumption patterns. GB is suited for our study due to strategic reasons. Firstly, its versatility and predictive performance allow it to model complex relationships within the data, which is crucial for benchmarking algorithms that need to uncover subtle patterns. Also, GB incorporates methods that prevent overfitting, ensuring that the benchmark remains reliable even with complex and noisy datasets, thereby maintaining the validity of our clustering performance comparisons. Furthermore, GB is highly scalable and can efficiently handle large datasets, making it an ideal choice for comprehensive evaluations without prohibitive computational costs.

While other methods such as K-nearest neighbors, neural networks, random forests (RFs), and support vector machines (SVM) have their strengths, they were not used for specific reasons. RF handles diverse data and avoids overfitting, but GB improves model performance, making it more effective for our needs. SVM, utilized for supervised classification, could benchmark clustering effectiveness, but its use in classification makes it less appropriate for the unsupervised nature of clustering compared to GB. Neural networks, although capable of modelling complex patterns, are computationally intensive and challenging to interpret, adding overhead to the benchmarking process compared to GB. K-nearest neighbors could validate clustering results by comparing the proximity of data points to their nearest neighbors and evaluating cluster cohesion, but its susceptibility to noise and outliers makes it less robust compared to GB. These methods serve distinct roles, but GB is selected for its precision and reliability in benchmarking. This selection ensures that the benchmark is both robust and effective in validating clustering algorithms, providing us with confidence to identify accurate consumption patterns.

Tailored to Brazilian households, our research employs a methodology with various algorithms, including agglomerative hierarchical clustering (AHC), k-means (KM), and self-organizing maps (SOM), alongside advanced preprocessing techniques. By applying this methodology to a dataset of monthly electricity consumption from a random sample of Brazilian households over a year, we aim to uncover underlying consumption patterns. In our data analytics, we also use principal component analysis (PCA) and silhouette analysis.

PCA is helpful for dimensionality reduction and might serve to reduce the data to key components and then apply clustering on this reduced space [23]. Silhouette analysis is utilized as an internal metric for cluster quality, measuring cohesion and separation of the clusters.

Clustering identifies distinct patterns of energy use, categorizing households into low, medium, and high consumption groups. Our main conclusion is that KM offers the most accurate results for energy management, improving the precision of energy policies. Our methodology provides insights into energy management within Brazilian homes and contributes to the literature on clustering in energy consumption analysis. Our contribution goes beyond the specific context of Brazil. By successfully applying and evaluating clustering methods in this context, our study serves as a model that can be adapted and applied to other regions, each with their distinct challenges.

The rest of this article is organized as follows. Section 2 describes the clustering methods employed in our analysis, namely AHC, KM, and SOM. Then, in Section 3, we outline our data collection process and specify the preprocessing techniques utilized to prepare data for clustering. In Section 4, the clustering methods are used to show their respective results. This section discusses the load profiles obtained through each clustering method, along with their pros and cons. Section 5 concludes the article with our key findings and their implications, providing directions for potential future research.

2. Methodology

In this section, we discuss our methodology for generating and analyzing residential electrical load profiles, focusing on three clustering algorithms: AHC, KM, and SOM. Additionally, as mentioned, GB is used to benchmark the performance and robustness of these algorithms, following advanced benchmarking principles [24]. Each algorithm, including GB, is detailed with its operational principles, strengths, and limitations.

2.1. Background on Clusters Algorithm for Profile Classification

Clustering [25] is an unsupervised machine learning method that classifies data based on shared attributes, revealing inherent structures within complex datasets. In analysis of profiles, clustering categorizes them into distinct groups, where a cluster contains profiles with high similarity, whereas profiles in different clusters display low similarity.

When analyzing profile data, clustering methods are classified as: partition-based, hierarchy-based, and model-based. Partition-based methods produce clusters with a central point, minimizing the distance between data points and this center [26]. Hierarchical methods deal each data point as a single cluster and progressively merge the closest clusters, leading to a dendrogram of clusters [27]. Model-based methods select a specific structure—grounded in statistical or neural network methods—and tailor the data to best fit this model [28].

Cluster validity indices gauge the effectiveness of clustering algorithms by evaluating their robustness [29]. Depending on the clustering method used, the optimal number of clusters can differ. Determining the number of clusters may require subjective judgment, aligning with the objectives or needs of the stakeholders [30].

Within the energy consumption research and data mining domain, clustering methods were introduced for discerning energy consumption patterns. However, a universal best-practice algorithm for energy consumption analysis remains elusive. Noteworthy among these methods are AHC, KM clustering, and SOM [31], which are relevant for energy consumption analysis due to their ability to uncover hidden patterns in complex data without requiring predefined categories. The simplicity and efficiency of KM in handling large datasets make it ideal for initial segmentation. The unique topological structure of SOM offers intuitive visualizations of energy consumption, facilitating deeper insights. The hierarchical approach of AHC is invaluable for understanding the relationships between consumption patterns, offering a detailed cluster hierarchy.

The distinction between supervised and unsupervised methods is crucial in the present context. Energy consumption data often lack clear labels, making unsupervised methods like KM, SOM, and AHC more suited for identifying structures within the data. This identification allows for the discovery of natural groupings based on consumption behaviors, which can inform effective energy management strategies. In contrast, supervised methods would require predefined categories based on known outcomes, limiting their utility in scenarios where one wishes to explore and understand unknown patterns. Thus, the utilization of unsupervised methods is a strategic choice for advancing our comprehension of energy usage dynamics, laying the groundwork for adaptive energy solutions.

2.2. k-Means Clustering

This article focuses on a widely used partition-based or flat clustering algorithm that uses centroids named KM. The KM algorithm employs an expectation-maximization approach and its number of clusters, represented by k, is predetermined. In the expectation step, each data point is assigned to the nearest centroid forming clusters, with these centroids being initially chosen at random. Following this, in the maximization step, centroids are recalibrated by averaging all the data points within the corresponding cluster. These two steps are iterated until the algorithm converges, signified when centroids retain their positions from one iteration to the next. The KM algorithm stands out due to its simplicity and efficiency, finding applications in various clustering scenarios and big-data [25,32]. It gauges similarity by measuring clustering distances; that is, when two entities are closer, they are more similar. As highlighted in [33], the KM algorithm has the capability to distinguish distinct electricity consumption patterns and of identifying households with analogous consumption. This capability renders KM especially suitable for mining insights from smart meter datasets.

Upon initiation, the KM algorithm establishes a preset number of clusters, k say. The dataset is represented as X, consisting of N data vectors x_i, with i ranging from 1 to N, each of dimension p. Having randomized the centroid positions, the algorithm then calculates the Euclidean distance (ED) between each data vector x_i and generic centroid vector c_j, which is also p-dimensional. Each point x_i is affiliated with generic cluster C_j that is closest to centroid c_j. Next, the centroid of each cluster c_j is updated by computing the average of m data points x_i belonging to cluster C_j. This centroid is formulated as

$$c_j = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad x_i \in C_j, j \in \{1,\ldots,N\}, m \in \{1,\ldots,M\}.$$

After obtaining centroid c_j, the distances between each data point x_i and the newly adjusted centroids are recalculated, leading to potential reassignments of data points to clusters. This recalculation iterates until the clustering is stable, meaning no data point switches its cluster membership between consecutive iterations, or until an iteration limit is reached. A challenge with the KM algorithm is ascertaining the best number of clusters, k say [25]. The Elbow method, which is widely used for determining the optimal number of clusters, involves visualizing the total within-cluster sum of squared errors (SSE) given by

$$\text{SSE} = \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - c_j \|^2 . \tag{1}$$

The SSE expressed in (1) is plotted against a range of cluster numbers to identify the point where increasing the number of clusters no longer provides important gains in reducing the SSE. At first, as the cluster count grows, SSE witnesses a marked decrease. Nonetheless, as the number of clusters continues to rise, the rate of SSE decline slows, leading to the formation of an "elbow". This "elbow" point is typically considered the optimal number of clusters [25]. The process of how clusters are formed and centroids adjusted is described in Algorithm 1.

Algorithm 1 KM clustering method.

Require: Dataset X, number of clusters k
Ensure: Clustered data
1: Select k initial centroids randomly from X.
2: **while** centroids change **do**
3: **for** each point $x_i \in X$ **do**
4: Assign x_i to nearest centroid c_j.
5: **end for**
6: **for** each centroid c_j **do**
7: Update c_j to be the mean of all points assigned to C_j.
8: **end for**
9: **end while**
10: Form clusters C_j after centroids are adjusted.

While the KM method is renowned for its simplicity and broad applicability, making it a popular choice for various applications requiring quick and straightforward clustering solutions, AHC adopts a fundamentally different and more intricate approach. AHC builds clusters by starting from individual data points and progressively merging them into larger nested structures, offering a deep dive into the data-inherent hierarchies.

2.3. Agglomerative Hierarchical Clustering

AHC stands out by its approach to cluster formation, where observations are combined into nested clusters, visualized through a dendrogram. This tree-like diagram illustrates the step-by-step clustering process and highlights the similarity levels at which observations merge [34]. One key distinction of AHC compared to partitioning methods like KM lies in its flexibility regarding the number of clusters. Unlike KM, which requires presetting this number, AHC allows users to cut the dendrogram at different levels, offering adaptability in defining the number of clusters. Such adaptability is beneficial in electrical load clustering, where the optimal number of clusters might not be evident beforehand [35]. The mentioned flexibility positions AHC as an alternative to KM in scenarios requiring hierarchical insights into data structures [34,35]. The AHC algorithm commences by viewing each observation as an individual cluster and then progressively merges these clusters based on their similarities. In determining which clusters to merge, the AHC algorithm frequently employs the Ward minimum variance method, due to its simplicity, which seeks to minimize the total within-cluster variance [36]. Unlike other linkage methods that might focus on the distances between cluster centroids or furthermost points, the Ward linkage method emphasizes the minimization of variance within clusters. According to [37], the total within-cluster variance is analogous to the loss of information incurred when grouping objects into clusters. The AHC algorithm aims to minimize this loss when forming clusters.

The AHC algorithm starts with the observation x_i, for $i \in \{1, \ldots, N\}$, as an individual cluster C_j. By using the Ward linkage method, the distance between clusters C_j and C_l is calculated as

$$\text{Dist}(C_j, C_l) = \text{SSE}(C_j \cup C_l) - \text{SSE}(C_j) - \text{SSE}(C_l), \tag{2}$$

where $C_j \cup C_l$ denotes the union set of clusters C_j and C_l. Since each data point starts in its own cluster, the initial distance is zero and gradually increases as clusters are merged. Note that the distance defined in (2) can also be represented as the squared distance between the centers of the two clusters, $c_j \in C_j$ and $c_l \in C_l$, as

$$\text{Dist}(C_j, C_l) = \left(\frac{N_{C_j} N_{C_l}}{N_{C_j} + N_{C_l}} \right) \|c_j - c_l\|^2,$$

where c_j and c_l are the centers or centroids of the respective clusters, N_{C_j} and N_{C_l} are the numbers of elements in those clusters C_j and C_l, respectively, and $\| \cdot \|$ denotes the norm of the vector.

An $N \times N$ distance matrix is defined to represent the inter-cluster similarities. Each element of the matrix corresponds to the distance between two clusters stated as

$$H = \begin{pmatrix} 0 & \text{Dist}(C_1, C_2) & \dots & \text{Dist}(C_1, C_N) \\ \text{Dist}(C_2, C_1) & 0 & \dots & \text{Dist}(C_2, C_N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Dist}(C_N, C_1) & \text{Dist}(C_N, C_2) & \dots & 0 \end{pmatrix}.$$

The two clusters with the smallest distance are then merged. Hence, the distance matrix is updated, and the process of merging continues iteratively until a single cluster encompassing all the data points is obtained. This iterative merging process of hierarchical clustering provides a clear and structured approach to clustering, as described in Algorithm 2.

Algorithm 2 AHC method.

Require: Dataset X
Ensure: Dendrogram representing hierarchical clusters
1: Treat each data point in X as a single cluster.
2: Compute the pairwise distance matrix H for all clusters.
3: **while** there is more than one cluster **do**
4: **for** each cluster pair (C_j, C_l) **do**
5: Calculate $\text{Dist}(C_j, C_l)$ using the Ward linkage method.
6: **end for**
7: Merge the two clusters with the smallest $\text{Dist}(C_j, C_l)$.
8: Update the distance matrix H.
9: **end while**
10: Create a dendrogram from the cluster merging process.

Transitioning from traditional clustering methods to advanced neural network-based approaches brings to the forefront another powerful tool: SOMs.

2.4. Self-Organizing Maps

SOMs are a type of artificial neural network that use unsupervised learning to transform high-dimensional data into a low-dimensional discretized representation known as a map [38]. Unlike hierarchical clustering, which relies on merging clusters based on distances, SOMs focus on learning from input patterns to produce a spatial representation that maintains topological relationships. SOMs can identify the similarity relationships between input variables, making them suitable for dimensionality reduction and clustering. In the context of electrical load profiles, a SOM can identify and learn important features, patterns, regularities, or correlations in the input load data and represent them in a topological map [39]. The SOM algorithm uses a neighborhood structure among the clusters, where data points close to each other are placed in the same or neighboring clusters [40].

The algorithm initializes with a grid of M artificial neurons, typically arranged in a two-dimensional lattice corresponding to the SOM grid. Each neuron j in the SOM grid has a weight vector w_j associated with this neuron. The weight w_j has the same dimensionality p as the input data x_i, where $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, M\}$. The SOM algorithm iteratively adjusts the weights of the neurons to match the input data. An input vector x_i is randomly selected from the dataset, and the neuron whose weight vector w_j is most similar to the input vector is identified as the best matching unit (BMU). This similarity is measured using a distance metric named the BMU index and defined as

$$b = \arg\min_j \|x_i - w_j\|. \tag{3}$$

Once the BMU is identified, the weight vectors w_j of neurons within a neighborhood S around the BMU are updated to become more similar to the input vector.

The update rule is given by

$$w_j(t+1) = w_j(t) + \varepsilon(t)h_{jb}(t)(x_i - w_j(t)), \tag{4}$$

where $\varepsilon(t)$ is the learning rate and $h_{jb}(t)$ is the neighborhood function, which determines the degree of update and is defined as

$$h_{jb}(t) = \exp\left(-\frac{d(j,b)^2}{2\sigma(t)^2}\right), \tag{5}$$

where $d(j,b)$ is the distance between neuron j and the BMU index b, with $\sigma(t)$ being the neighborhood radius. Note that $d(j,b)$ measures the topological or geometric distance between neuron j and b in the lattice structure. This distance is used to determine the influence of the BMU on its neighboring neurons during the update process. The update is strongest for the neuron that wins (that is, this neuron is the BMU) and decreases as the distance from the BMU increases. This iterative process continues until a specified number of iterations is reached or the algorithm converges. To allow gradual convergence, both the learning rate and neighborhood function are reduced over time, expressed as $\varepsilon(t) = \varepsilon_0 \exp(-t\lambda)$ and $\sigma(t) = \sigma_0 \exp(-t\gamma)$, where λ and γ represent the learning and neighborhood decay rates, respectively. Hence, the SOM results in a low-dimensional map where similar input vectors are positioned close to each other, and dissimilar vectors are further apart. This map can be visually inspected and used as a powerful tool for data exploration, clustering, and visualization. Algorithm 3 describes the steps of a SOM.

Algorithm 3 SOM method.

Require: Dataset X, grid size M
Ensure: Topological map of input data
1: Initialize a grid of M neurons with random weights.
2: **for** each iteration t **do**
3: Select an input vector x_i from X randomly.
4: Determine the BMU b by using the formula presented in (3).
5: Define a neighborhood S around b.
6: **for** each neuron j in the neighborhood S **do**
7: Update the weight w_j using the rule stated in (4) with the learning rate $\varepsilon(t)$ and neighborhood function given in (5).
8: **end for**
9: **end for**
10: Generate the topological map.

2.5. Gradient Boosting as a Benchmarking Tool

GB is a machine learning method primarily used for regression and classification tasks. It builds models in a stage-wise fashion, optimizing for mean squared error. GB constructs additive models by sequentially fitting a simple base model, typically a decision tree, to the current pseudo-residuals [41,42]. Although GB is primarily used for classification and regression, we employ it here as a benchmarking tool to evaluate the performance of clustering algorithms, providing a reliable reference for comparison. As mentioned in the introduction, we select GB for its robustness in handling heterogeneous data and its ability to capture complex interactions. Its versatility and predictive performance allow it to model complex relationships within the data, which is crucial for uncovering subtle patterns. Additionally, it includes techniques to prevent overfitting, ensuring reliable benchmarks even with complex and noisy datasets. Furthermore, it is scalable and efficiently handles large datasets, making it ideal for comprehensive evaluations without prohibitive computational costs. These attributes make GB a reliable benchmark for evaluating clustering algorithms. Other benchmarking methods such as RF, SVM, and neural networks have their strengths but were not chosen for specific reasons indicated below.

RF handles diverse data and avoids overfitting, but GB gradient-based optimization enhances model performance. SVM, suitable for supervised classification, is less appropriate for the unsupervised nature of clustering. Neural networks, while capable of modeling complex patterns, are computationally intensive and difficult to interpret. This strategic selection ensures that the benchmark is both robust and highly effective in validating clustering algorithms, providing confidence in identifying accurate consumption patterns.

The GB algorithm proceeds through several steps. In the following, F represents the predictive function that is iteratively adjusted during the boosting process, x_i is the feature vector for observation i, and y_i is the target value for observation i. The GB model is initialized with a constant value calculated as

$$F_0 = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma), \tag{6}$$

where $L(y_i, \gamma)$ is the loss function. In each iteration $t \in \{1, \ldots, T\}$, residuals are stated as

$$r_{it} = -\left(\frac{\partial L(y_i, F_{t-1}(x_i))}{\partial F(x_i)} \right)_{F(x)=F_{t-1}(x)}, \quad i \in \{1, \ldots, N\}, \tag{7}$$

for each data vector x_i. Then, a base learner $h_t(x)$ is fitted to the residuals resulting in

$$h_t(x) = \arg\min_{h} \sum_{i=1}^{N} (r_{it} - h(x_i))^2. \tag{8}$$

Next, the multiplier γ_t is determined by

$$\gamma_t = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, F_{t-1}(x_i) + \gamma h_t(x_i)). \tag{9}$$

Lastly, the model is updated as

$$F_t(x_i) = F_{t-1}(x_i) + \eta \gamma_t h_t(x_i), \tag{10}$$

where η is the learning rate. The performance of GB is influenced by several hyperparameters, such as the number of boosting stages, tree depth, and learning rate. The proper tuning of these hyperparameters is crucial to ensure the reliability and robustness of the benchmarking results [43]. Algorithm 4 provides an overview of the GB process, illustrating how these hyperparameters are utilized and adjusted.

Algorithm 4 GB method.

Require: Dataset $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, number of iterations T, learning rate η
Ensure: Predictive model $F(x)$
1: Initialize the model with a constant value stated as in (6).
2: **for** $t = 1$ to T **do**
3: Compute pseudo-residuals r_{it} associated with each x_i, for $i \in \{1, \ldots, N\}$, as defined in (7).
4: Fit a base learner $h_t(x)$ to the calculated residuals by employing the expression given in (8).
5: Determine the multiplier γ_t as presented in (9).
6: Update the model $F_t(x_i)$ considering the formula established in (10).
7: **end for**
8: Formulate the predictive model.

To measure the performance of various clustering methods against the GB benchmark, we use metrics like precision, recall, and silhouette scores. These metrics enable a thorough evaluation of clustering quality, providing insights into the effectiveness and reliability of the methods. By establishing a clear quantifiable standard for successful clustering, GB enhances our understanding of the best methods for analyzing energy consumption data, guiding future research and applications in the field [44,45].

3. Data and Methods

This section outlines the methodologies and data employed in our analysis. It integrates the clustering methods utilized with GB for benchmarking purposes. Additionally, this section discusses the dataset comprising the energy consumption patterns of Brazilian households, detailing in data collection, processing, and preparation for analysis. Emphasis is placed on ensuring data quality and relevance, facilitating a comprehensive understanding of household energy consumption behaviors in Brazil.

3.1. Data

A survey was applied to $N = 383$ randomly selected households across Brazil, collecting data on monthly electrical power consumption as well as characteristics of the households and occupants. The period under study covered from January to December 2022 and this was specifically chosen to capture the evolving dynamics of household energy consumption in the aftermath of the COVID-19 pandemic—a time marked by important shifts in living and working habits that potentially altered energy use patterns. Deliberately including all four seasons within this timeframe aimed to encapsulate the impact of seasonal variations on energy consumption comprehensively. Although the analysis does not segregate results by season, covering the entire year, it ensures that the study reflects the influence of climatic changes, such as variations in temperature and daylight, which are known to affect energy needs for heating, cooling, and lighting. By examining data from 2022, our study not only provides insights into post-pandemic consumption behaviors but also accounts for the intrinsic seasonal factors that shape energy use, offering a holistic view of the current state of household energy consumption in Brazil. The collected load data represent the history of monthly energy consumption, expressed in kilowatt–hour (kWh), over one year from January to December 2022, as mentioned. This results in a dimensionality of twelve, with each dimension corresponding to each month, providing a comprehensive view of household energy usage patterns across seasons and months.

The dataset under study consists of elements $l_{i,j}$, each representing a monthly consumption data point in kWh for a household in row i and month in column j, with $j \in \{1, \ldots, 12\}$. Therefore, the consumption of household i is represented as a row matrix, and the consumption of all households in month m_j as a column matrix, given by $d_i = [l_{i,1}, \ldots, l_{i,12}]$ and $m_j^\top = [l_{1,j}, \ldots, l_{N,j}]$, where m_1 represents January (JAN), m_2 represents February (FEB), up to m_{12} for December (DEC). Eventually, the collected consumption for all 12 months for the $N = 383$ households constitutes the load dataset as an $N \times 12$ matrix given by

$$L = \begin{bmatrix} l_{1,1} & \cdots & l_{1,12} \\ \vdots & \ddots & \vdots \\ l_{N,1} & \cdots & l_{N,12} \end{bmatrix}.$$

3.2. Preprocessing Stage

The analysis of the dataset reveals that monthly energy consumption distributions generally exhibit a right-skewed non-normal pattern, as shown in the histogram of Figure 1. This skewness, reflecting variability in household sizes and usage patterns, is a common feature in household energy consumption data, highlighting the diverse range of energy consumption behaviors among households. Similar patterns are observed in the yearly average consumption data, maintaining the right-skewed distribution. The analysis covered thirteen scenarios: including each of the twelve months, m_j say, and the yearly averages of each dwelling, d_i namely. Monthly distributions for m_j show a consistent right-skewed non-normal pattern, characterized by a pronounced right tail. The yearly average consumption for d_i reflects this skewness, maintaining the characteristic shape of the distribution. Considering the distribution characteristics and nature of our data, we evaluate several methods for handling missing data, including mean and median imputations, as well as interpolation.

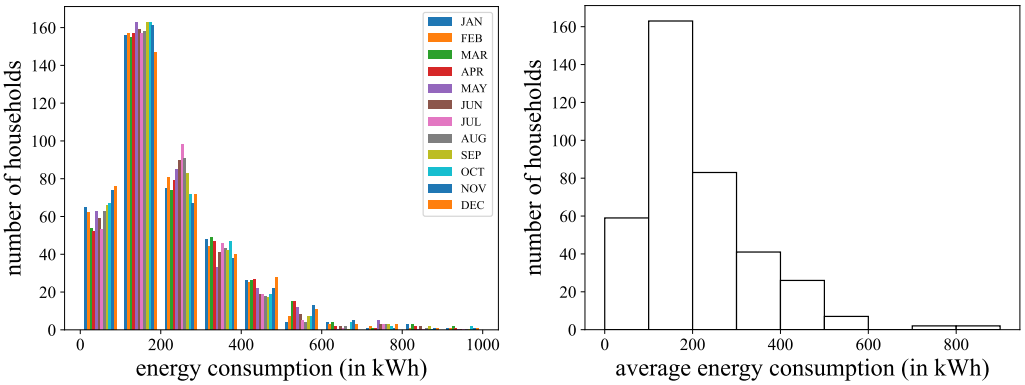

Figure 1. Bar-plot of monthly Brazilian energy consumption (**left**) and histogram of average yearly Brazilian energy consumption (**right**).

Given the skewness of the data distribution and presence of outliers, median imputation was selected as the most suitable method for maintaining the integrity of its central tendency without the influence of extreme values. The imputation step involves applying dimensionality reduction, which streamlined the dataset for efficient processing and analysis, as detailed in subsequent sections.

Rigorous preprocessing is fundamental for successful data analytics and the effective application of machine learning algorithms. The preprocessing stage involves eliminating anomalies and noise, calibrating the dataset to more accurately represent its inherent structure. Such calibration is important, especially in scenarios where external factors or regulatory constraints play an important role. One such external influence in the context of our study arises from the Brazilian law. The legislation mandates that if measured consumption falls below a certain legal threshold, energy companies must use this threshold value for billing purposes. Specifically, for consumers on a single-phase system, the threshold is set at 30 kWh, whereas for a three-phase system, this threshold is 100 kWh.

We recall the objective of our research is to model and predict the electrical power consumption patterns in Brazilian households, considering the thresholds for energy use. We must identify the factors influencing household energy consumption data. Then, it becomes clear that such data can distort our findings. This is because such threshold values do not genuinely represent household true energy consumption patterns, rendering them irrelevance and noisy for our dataset. To address and identify such distortions, we calculate two statistics: the average yearly consumption $\overline{d_i}$ for dwelling i across monthly load profile $l_{i,j}$ and its respective coefficients of variation (CV(d_i)), which are determined as

$$\overline{d_i} = \frac{1}{12}\sum_{j=1}^{12} l_{i,j}, \quad \mathrm{CV}(d_i) = \frac{\sqrt{\frac{1}{11}\sum_{j=1}^{12}(l_{i,j} - \overline{d_i})^2}}{\overline{d_i}}, \quad i \in \{1, \ldots, N\}.$$

Upon analyzing these statistics, it is evident that households with consumption equal to the statutory values exhibit CV(d_i) = 0, indicating no consumption variability. Table 1 lists all households identified through CV(d_i) = 0, which is subsequently removed from the dataset. We were able to identify potential duplicate data by recognizing pairs of households with an identical CV. It is pertinent to note that such duplications often occur when datasets from diverse sources are merged. As shown in Table 2, pairs of households sharing the same CV were considered duplicates. To eliminate redundancy, one entry from each pair of duplicates was removed from the dataset. Regarding outlier removal, in [46], it is argued that a genuine data point can contain critical information, and so it should not be discarded recklessly. Given that outliers may emerge in real-world load profiles, some clustering methods demonstrating that robustness to asymmetrically distributed data might temper their impact [35]. We discuss these methods in the subsequent sections, particularly in the context of electrical load profile clustering research.

Table 1. Households with identical electrical consumption in Brazil for the indicated month.

Household \ Month	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
d_{27}	100	100	100	100	100	100	100	100	100	100	100	100
d_{52}	30	30	30	30	30	30	30	30	30	30	30	30
d_{54}	30	30	30	30	30	30	30	30	30	30	30	30
d_{78}	100	100	100	100	100	100	100	100	100	100	100	100
d_{350}	30	30	30	30	30	30	30	30	30	30	30	30
d_{360}	30	30	30	30	30	30	30	30	30	30	30	30
d_{362}	30	30	30	30	30	30	30	30	30	30	30	30
d_{372}	30	30	30	30	30	30	30	30	30	30	30	30

Table 2. Households with duplicated electrical consumption in Brazil for the indicated month.

Household \ Month	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
d_{57}	342	362	329	322	276	298	272	0	0	0	0	0
d_{58}	342	362	329	322	276	298	272	0	0	0	0	0
d_{328}	214	226	244	286	249	225	234	205	243	238	218	285
d_{329}	214	226	244	286	249	225	234	205	243	238	218	285

3.3. Preprocessing Stage

Consequently, after eliminating 10 rows of inaccurate and duplicate data, the dataset now comprises $N = 373$ rows. After addressing distortions in the consumption data, the subsequent step involved handling missing data, a common issue within billing consumption datasets due to record gaps from energy companies. In our dataset, comprising data from 383 dwellings, missing data were identified in 40 dwellings (10.72%), accounting for 140 data points (3.13%) out of a total of 4476.

According to [47], missing data can be classified into three types: (i) missing completely at random, when the missing observations are unrelated to both observed and unobserved measurements; (ii) missing at random, when the probability of a missing value is only related to the available data; and (iii) missing not at random, which depends both on observed data and unobserved data. In this dataset, the missing data may be characterized as missing at random since the absence of monthly load profiles can be related to the observed and recorded load profiles.

To handle missing data, various imputation methods can be applied, including replacing the missing values with the mean, median, or inferred values [48]. Imputation approaches based on inference methods have a relatively high chance of predicting missing records close to their true values [49]. To determine the most suitable imputation method, three different approaches were implemented, and missing data were filled using the following methods: mean imputation, median imputation, and multivariate imputation by chained equations (MICE). In the mean imputation method, dwelling d_i with all profiles $l_{i,j} = 0$ is replaced by the average value $\overline{l_i}$ calculated using $\overline{l_i} = (1/r)\sum_{j=1}^{12} l_{i,j}$, for $l_{i,j} \neq 0$, where r represents the number of $l_{i,j} \neq 0$. In the median imputation method, dwelling d_i with all profiles $l_{i,j} = 0$ is replaced by the median value of the ordered monthly consumption where $l_{i,j} \neq 0$. Regarding the MICE method, it is worth noting that it operates under the assumption that the missing data are at random, and the imputed values are based on the observed data [50]. Initially, the algorithm assigns a single imputation, such as the mean, to each row with missing values. Subsequent imputations are performed based on random draws from the observed data, and this performance is repeated for a defined number of iterations or until convergence is achieved [51].

To compare the different imputation methods, we created two subsets: one containing only the dwellings with missing data ($N = 40$) and the other one comprising the complete dataset ($N = 373$). The box-plots for these imputed datasets are shown in Figure 2. These plots indicate that the dispersion among the methods within each group is quite similar, suggesting that there are no important differences in the distribution of the methods.

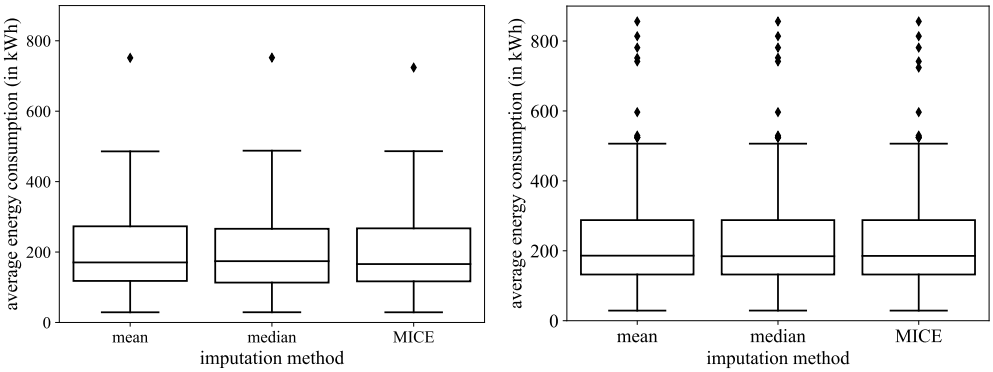

Figure 2. Box-plot for $N = 40$ (**left**) and $N = 373$ (**right**) dwellings with the listed method using Brazilian electrical consumption, where diamonds indicate outliers and lines within box the data median.

To further analyze the data, we calculated the yearly average $\overline{L}$ and its respective variance σ^2 for $\overline{d}_i$ using the same imputation method described by

$$\overline{L} = \frac{1}{373}\frac{1}{12}\sum_{i=1}^{373}\sum_{j=1}^{12} l_{i,j}, \quad \sigma^2 = \left(\frac{1}{372}\sum_{i=1}^{373}\left(\overline{d}_i - \overline{L}\right)^2\right)^{1/2}.$$

The descriptive statistics associated with each box-plot are detailed in Table 3. The analysis of the descriptive statistics reveals similar performances among the methods for both datasets ($N = 40$ and $N = 373$), with more notable differences observed in the medians of the smaller subset ($N = 40$). To assess whether the imputation methods uniformly impact the data, we applied the Friedman test, suitable for non-parametric data analysis.

Table 3. Descriptive statistics of the indicated imputation method, where Q_1 and Q_3 are the first and third quartiles, respectively, for the data of electrical load power consumption in Brazil.

Method	N	$\overline{L}$ (σ^2)	CV	$\overline{d}_{i_{\min}}$	Q_1	Median	Q_3	$\overline{d}_{i_{\max}}$
Mean	40	206.22 (136.44)	0.6616	28.90	117.94	170.50	273.08	751.44
	373	221.32 (129.54)	0.5853	28.90	132.17	186.08	287.75	856.08
Median	40	206.27 (137.54)	0.6668	29.00	113.27	174.08	266.04	752.08
	373	221.32 (129.65)	0.5858	29.00	132.17	184.42	287.75	856.08
MICE	40	204.44 (133.93)	0.6551	28.88	116.75	165.86	267.28	724.02
	373	221.12 (129.28)	0.5846	28.89	132.17	185.35	287.75	856.08

The Friedman test is selected for its ability to handle data that does not follow a normal distribution, comparing ranks instead of means. In this context, the null hypothesis is that the imputation methods do not exhibit significant differences in their effects on the data. The rejection of this hypothesis would indicate significant differences between the methods. Statistical analysis yielded a p-value of 0.80 for both datasets ($N = 40$ and $N = 373$), indicating the non-rejection of the null hypothesis and suggesting non-significant differences at 5% in the effects of the imputation methods.

As a p-value of 0.80 was obtained for the subset with $N = 40$, this leads to the non-rejection of the null hypothesis, implying that there is no significant difference in the effects of the imputation methods at 5% level. The same conclusion applies to the set $N = 373$, which also yielded a p-value of 0.80. Given the similar performance of the methods in terms of average consumption $\overline{d}_i$, an analysis of the impact of these imputation methods on the monthly consumption series of household i is needed.

Figure 3 shows the influence of each method on the monthly household consumption for selected samples. It becomes evident that the MICE method has a greater dispersion compared to imputation with the mean or median. Additionally, imputation with the mean or median, particularly with highly missing data, can create a misleading perception of stability in monthly energy consumption, as repeated monthly consumption values are atypical.

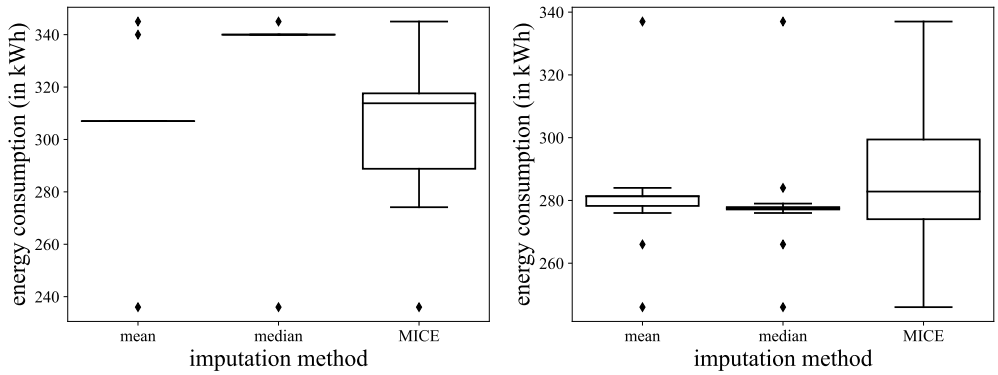

Figure 3. Box-plots for residences d_6 (**left**) and d_{299} (**right**) with the listed method using electrical consumption in Brazil, where diamonds indicate outliers and lines within box the data median.

In the highly missing dataset, 44.77% of the dwellings have at least two equal monthly consumption values, but never more than three. Even when equal monthly consumption occurs within a yearly period, the data still exhibit a distribution with a similar dispersion to that presented in the MICE method, as observed in the samples presented in Figure 4.

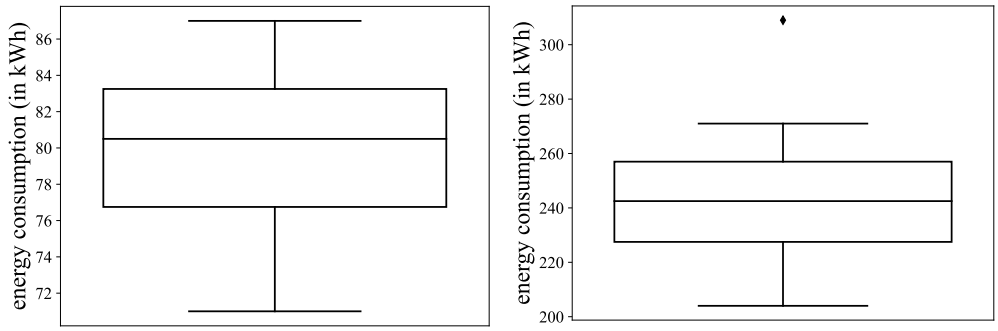

Figure 4. Box-plots for residences d_{339} (**left**) and d_{368} (**right**) using data of electrical consumption in Brazil, where diamond indicates outliers and lines within box the data median.

Despite the similar performance of the three imputation methods in terms of average household consumption $\overline{d_i}$, the MICE method was chosen for its ability to preserve the dispersion of the data distribution. This is particularly important for capturing the true variability in energy consumption patterns. The time series of all 373 dwellings, post-application of the MICE method, is shown in Figure 5, providing a comprehensive view of the adjusted consumption data.

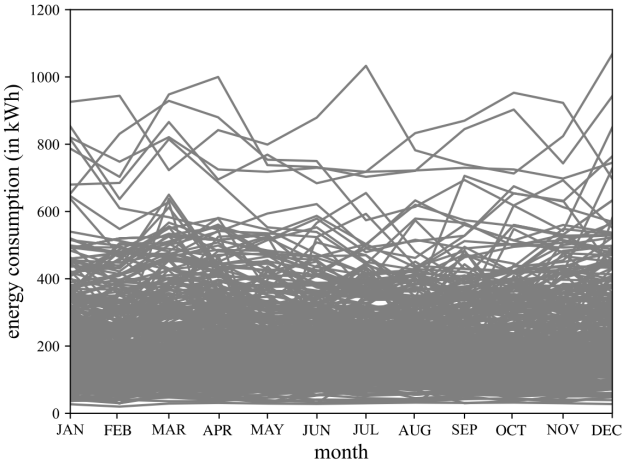

Figure 5. Time series plot of all 373 dwellings after applying the MICE method for data of electrical load power consumption in Brazil, where each line in gray is the monthly consumption of a dwelling, which for low consumptions are overlapping.

Having addressed the missing data, our analysis now turns to the critical step of reducing the data dimensionality. This step is essential to capture the main features of the dataset in an efficient way, enabling an effective analysis without being hampered by its volume or complexity. Reducing the dimensionality of data and extracting the main features are key aspects of clustering in data mining, as emphasized in [52]. Although there is no guarantee of preserving the cluster structure, PCA is an effective method for this purpose, widely used in various applications [53]. PCA, a method prevalent in multivariate statistics, transforms correlated variables into a smaller set of uncorrelated variables called principal components (PCs), effectively retaining most of the original set variance [54]. In PCA, the selection of PCs is based on their contribution to the dataset variance, with PC1 accounting for the most important variance and so on [55].

Our dataset includes twelve dimensions, m_j, for $j \in \{1, \ldots, 12\}$, each representing a month of energy consumption over a year. The high-dimensional nature of this dataset poses challenges for effective analysis and visualization. PCA is employed to streamline the analysis by converting this high-dimensional dataset into a more manageable format, enhancing the visualization of consumption trends and allowing for a more insightful interpretation of the inherent patterns [52].

Before implementing PCA, we assessed correlations in monthly consumption data using the Pearson correlation coefficient, calculated between month pairs m_j and m_l. Figure 6 shows a heatmap of these correlations, providing a visual representation of the relationships between different months.

Figure 6. Heatmap of Pearson coefficients for Brazilian electrical consumptions in the listed months.

The Pearson coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 no correlation, and 1 a perfect positive correlation. In our analysis, the Pearson coefficient fall between 0.81 and 0.94, indicating strong correlations among the monthly consumption data [56,57]. This supports the use of PCA to reduce the complexity of this interrelated data and highlight the consumption patterns.

It is essential to standardize the data to ensure that, each month, each energy consumption contributes equally to the analysis, regardless of its original scale or units. Variations in consumption across different months can be substantial due to seasonal changes and other factors. Without standardization, these variations could distort the PCA results, with misleading interpretations. By standardizing the data to have zero mean and one variance, we prevent any dimension from having a disproportionate effect on PCs. This ensures that PCA captures the patterns in the data, reflecting the true relationships and variations across the twelve dimensions of monthly energy consumption.

Thus, load $l_{i,j}$ is transformed to yield the standardized load $L_{i,j} = (l_{i,j} - \overline{m}_j)/\sigma_{m_j}$, where $\overline{m}_j$ denotes the mean, while σ_{m_j} stands for the standard deviation of the variable m_j. Subsequently, we calculate the covariance matrix C of these standardized data points as

$$C = \begin{bmatrix} \operatorname{cov}(m_1, m_1) & \dots & \operatorname{cov}(m_1, m_{12}) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}(m_{12}, m_1) & \dots & \operatorname{cov}(m_{12}, m_{12}) \end{bmatrix}$$

The eigenvalues ϱ of the matrix and the corresponding eigenvectors $\vec{v}$ can be obtained using $|C - \varrho I| = 0$ and $(C - \varrho I)\vec{v} = 0$, being I the identity matrix. Following this, we arrange the eigenvalues ϱ_j and their associated eigenvectors $\vec{v}_j = [v_{1,j}, \dots, v_{12,j}]$ in descending order, implying that $\varrho_1 > \cdots > \varrho_{12}$.

Considering the data reduction from a twelve-dimensional space to a two-dimensional one, the top two eigenvectors assist in data transformation, as shown in

$$\begin{bmatrix} L_{1,1} & \dots & L_{1,12} \\ \vdots & \ddots & \vdots \\ L_{373,1} & \dots & L_{373,12} \end{bmatrix} \begin{bmatrix} v_{1,1} & v_{1,2} \\ \vdots & \vdots \\ v_{12,1} & v_{12,2} \end{bmatrix} = \begin{bmatrix} x_{1,\mathrm{PC1}} & x_{1,\mathrm{PC2}} \\ \vdots & \vdots \\ x_{373,\mathrm{PC1}} & x_{373,\mathrm{PC2}} \end{bmatrix}.$$

Each dwelling in our study, d_i say, is represented by a coordinate pair $(x_{i,\mathrm{PC1}}, x_{i,\mathrm{PC2}})$ after applying PCA. This analysis condenses the dataset into a two-dimensional space, facilitating visualization and analysis. Figure 7 displays a scatterplot of the data points represented by these PC coordinates.

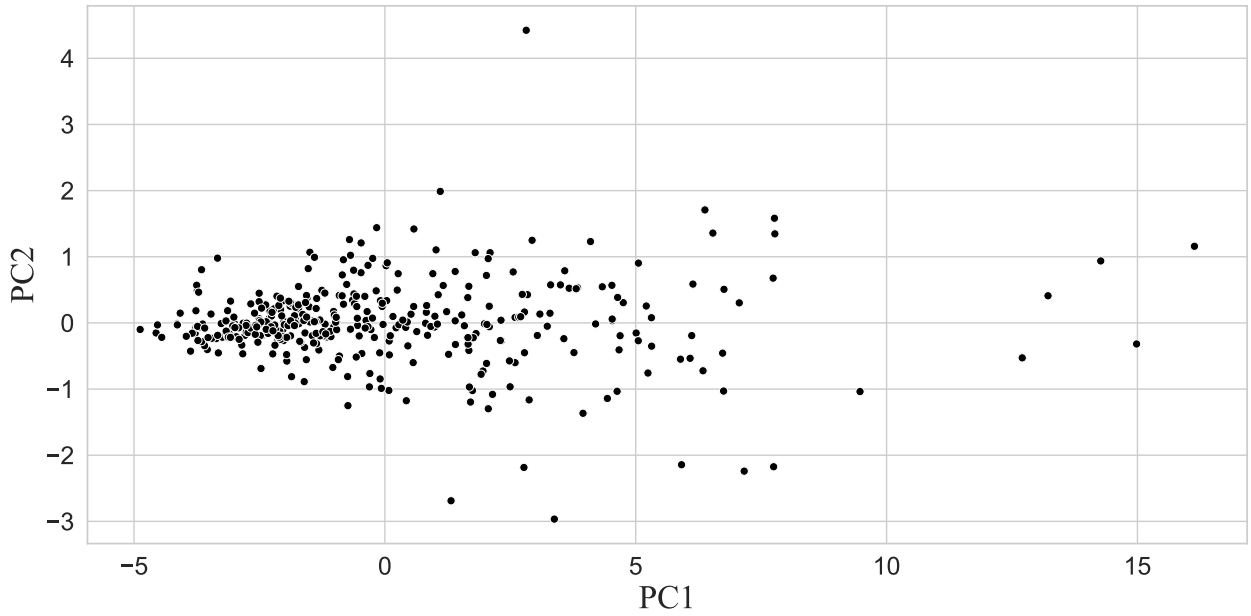

Figure 7. Scatter-plot of dwellings in the PCA-reduced space with two PCs for data on electrical load power consumption in Brazil.

The variance captured by the first two PCs is important, with PC1 accounting for 89.22% and PC2 for 3.40% of the total variance, cumulatively representing 92.62% of the dataset variance. This substantial accumulation is depicted in Figure 8(left), highlighting the effectiveness of PCA for our study. The primary aim of PCA is to derive uncorrelated PCs. The success of this objective is evidenced in the heatmap of the Pearson correlation coefficients of the PCs, as shown in Figure 8(right). This heatmap confirms the reduction in correlation among the newly derived PCs.

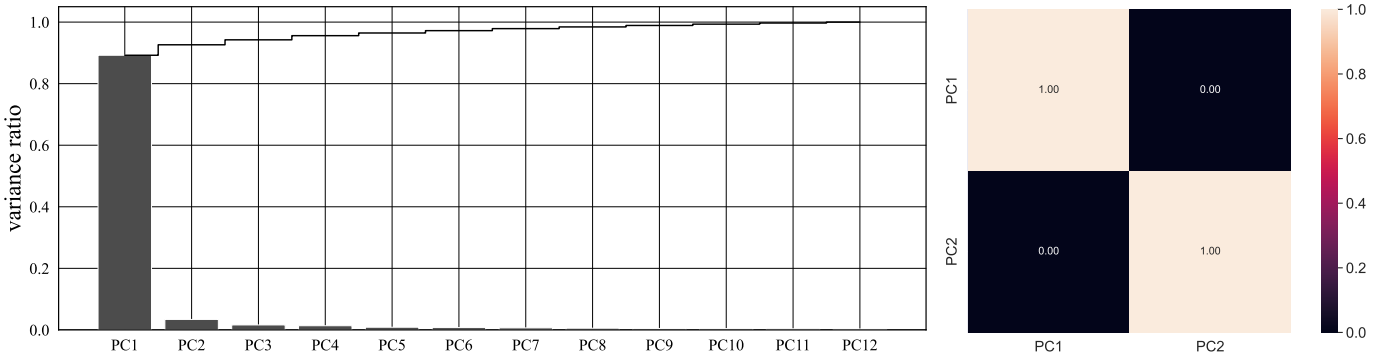

Figure 8. Bar-plot of variance ratio captured by the indicated PC (**left**) and heatmap of Pearson correlation among indicated PCs (**right**) for data of electrical load power consumption in Brazil.

With the completion of the dimensionality reduction through our PCA, we transition to the application of clustering methods. We utilize the PCs derived from PCA to conduct a clustering analysis on the reduced dataset. This analysis facilitates the efficient identification and categorization of distinct energy consumption profiles in Brazilian households.

4. Cluster Analysis Applied to the Electrical Load Dataset

In this section, we apply three clustering algorithms—KM, AHC, and SOM—within our Brazilian household electrical load dataset, using GB as a benchmark to assess their effectiveness. The analysis covers relevant evaluation metrics for unsupervised learning and highlights how these metrics apply to the chosen clustering method.

4.1. Background on Clustering Algorithms for Load Profile Classification

Cluster analysis is especially valuable in unsupervised learning scenarios, where data lack labels, and its objective is to unveil latent structures or groups within the dataset. Then, it is crucial to use an evaluation metric independent of observed labels to assess the effectiveness of clustering algorithms. A widely used metric in this context is the silhouette score, which evaluates the intra-cluster cohesion and inter-cluster separation of data points [58,59]. This score, ranging from -1 to 1, indicates well-defined and compact clusters when close to 1 [18]. The score for each data point is calculated using $S_{x_i} = (e_i - d_i)/\max(d_i, e_i)$, where d_i is the average distance between a data point x_i and all other points within the same cluster C_j, whereas e_i represents the smallest average distance from x_i to all points in any other cluster C_l, of which x_i is not a part. The expressions for d_i and e_i are defined as

$$d_i = \frac{\sum_{x_s \in C_j, s \neq i} \text{Dist}(x_i, x_s)}{N_{C_j} - 1}, \quad e_i = \min\left(\frac{\sum_{x_s \in C_l, l \neq s} \text{Dist}(x_i, x_s)}{N_{C_l}}\right),$$

where, as mentioned, N_{C_j} and N_{C_l} are the number of points in clusters C_l and C_l, respectively. The silhouette score provides a means to evaluate the clustering quality in the absence of observed labels. Data points with positive silhouette scores ($S_{x_i} > 0$) are typically well clustered, with values close to one indicating well-defined and compact clusters, and values close to zero indicating overlapping clusters. Negative scores ($S_{x_i} < 0$) represent possible misclustered points. The average silhouette width (ASW) is commonly utilized to analyze and select the optimal number of clusters, k say. The ASW states the mean silhouette score for all data points in the dataset, providing an overall measure of clustering quality. By calculating the ASW for the different values of k, we can determine the number of clusters that maximizes the ASW, so identifying the most appropriate clustering.

4.2. Load Profiles Generated with the k-Means Algorithm

Determining the optimal number of clusters k is a crucial step in the clustering process, as it affects the interpretability of the resulting clusters. For the KM algorithm, a combination of silhouette scores and the Elbow method can provide a more informed decision. We implemented KM clustering using the `scikit-learn` package of `Python` [60]. It was executed with the ED as a distance metric defined as

$$\mathrm{ED}(x_i, c_j) = \left(\sum_{s=1}^{p} \left(x_i(s) - c_j(s) \right)^2 \right)^{1/2},$$

where p represents the total number of dimensions (or features) of the data and, as mentioned, c_j is the centroid of cluster C_j.

As shown in Figure 9, the ASW decreases as the values of k increase. This decrease occurs because, as the number of clusters increases, the clusters tend to become smaller and less cohesive, often capturing more noise and less meaningful separation between data points. The evaluation of the silhouette score suggests that the optimal value of k can be 2, 3, or 4. To visualize the silhouette coefficients for each sample on a per-cluster basis, the `Yellowbrick` library [61] was used as a visual diagnostic tool to examine the density and separation of the clusters.

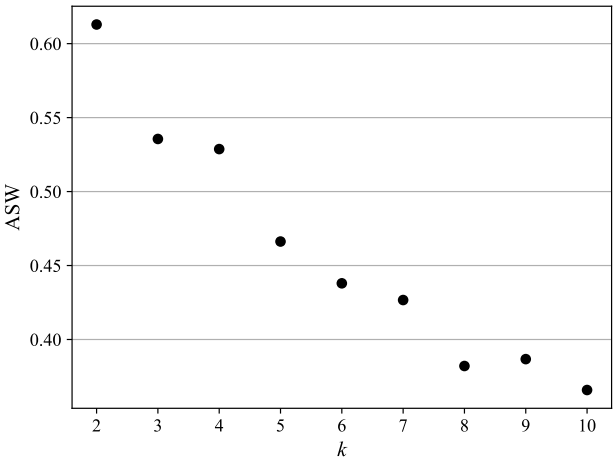

Figure 9. Plots of ASW for KM clustering with values of k varying from 2 to 10 for the data of electrical load power consumption in Brazil.

Figure 10(left) shows the silhouette-plot for $k = 2$, illustrating the distribution of silhouette coefficients within each cluster. Each bar represents a data point, with the width indicating the silhouette score, which measures how similar a point is to its own cluster compared to other clusters. The x-axis represents the silhouette scores, ranging from -1 to 1, and the y-axis indicates the cluster labels. The red dashed line represents the ASW for all clusters, which is 0.6129 in this case. The ASW provides a single value indicating the overall quality of the clustering for $k = 2$. Cluster 1 shows higher cohesion, with many points above the ASW line, whereas Cluster 2 displays lower cohesion, with most points below the ASW line. This suggests potential misclustering or a less clear definition in Cluster 2. Observing Figure 11(left), the greater cohesion within Cluster 1 compared to Cluster 2 observed becomes evident. This observation aligns with the findings from the silhouette-plot. The plot highlights distinct groupings and their spatial distribution. The PCA scatter-plot for $k = 2$ in Figure 11(left) reveals that the data in Cluster 1 (in black) are mainly located to the left of PC1 = 0, with some points to the right, showing less dispersion and symmetry around PC2 = 0. Conversely, the data points in Cluster 2 (in blue) are all positioned to the right of PC1 = 1, with a wider dispersion, indicating high variability within the cluster. This spatial distribution aligns with the silhouette-plot, confirming the higher cohesion of Cluster 1 and the lower cohesion of Cluster 2.

For $k = 3$, the results are shown in Figure 10(center), indicating an ASW of 0.5344, which is less than the ASW for $k = 2$, but still suggests a reasonable clustering structure. Notably, the silhouette scores for Clusters 1 and 2 are now predominantly above the ASW line, indicating improved cluster cohesion and separation for these clusters. The PCA scatter-plot for $k = 3$ in Figure 11(center) shows that data points in Cluster 1 (in black) are more confined to the left of PC1 = 0, reporting higher cohesion. Cluster 2 (in blue) shows moderate dispersion and is situated approximately between PC1 = -1 and PC1 = 4. Cluster 3 (in green) is further to the right with higher dispersion, indicating less cohesion compared to the other clusters. Observing PCA scatter-plots from Figure 11(left),(center), it is evident that, as k increases, the separation between clusters becomes clearer. Cluster 1 loses some points to Cluster 2, while Cluster 2 gains points from Cluster 1 and loses some to Cluster 3, reflecting the redistribution of points among the clusters.

Figure 10(right) illustrates silhouette scores for each sample within clusters using KM clustering. Cluster 1 (in black) has the highest area, indicating high cohesion. Cluster 2 (in blue) and Cluster 3 (in green) follow, with Cluster 4 (in orange) being the smallest. Thus, for $k = 4$, as depicted in Figure 10(right), Cluster 1 retains a similar structure compared to Cluster 1 from $k = 3$; see Figure 10(center). Clusters 3 and 4 emerge due to the split of Cluster 3 from $k = 3$. The ASW is now 0.5246, continuing to decrease, but still suggesting a reasonably good clustering structure. Most clusters have silhouette scores above the average threshold, indicating reasonable cohesion and separation. However, it is important to note that Cluster 4 is quite sparse, containing only a few data points, as shown in Figure 10(right). This limited size can pose challenges for future predictions.

Figure 11(right) visualizes the spatial separation of clusters for $k = 4$. Cluster 1 (in black) is concentrated around PC1 = -1 with less dispersion. Cluster 2 (in blue) is between PC1 = -1 and PC1 = 4 with moderate dispersion. Cluster 3 (in green) shows greater dispersion further to the right, and Cluster 4 (in orange) has the highest dispersion, extending even further right. This indicates a transfer of data points from Clusters 3 and 2. This transfer shows how increasing the number of clusters results in a refined separation and distribution of data among clusters. The dispersion patterns noted in the PCA plot align with the silhouette scores, confirming the variations in cohesion and separation across the clusters.

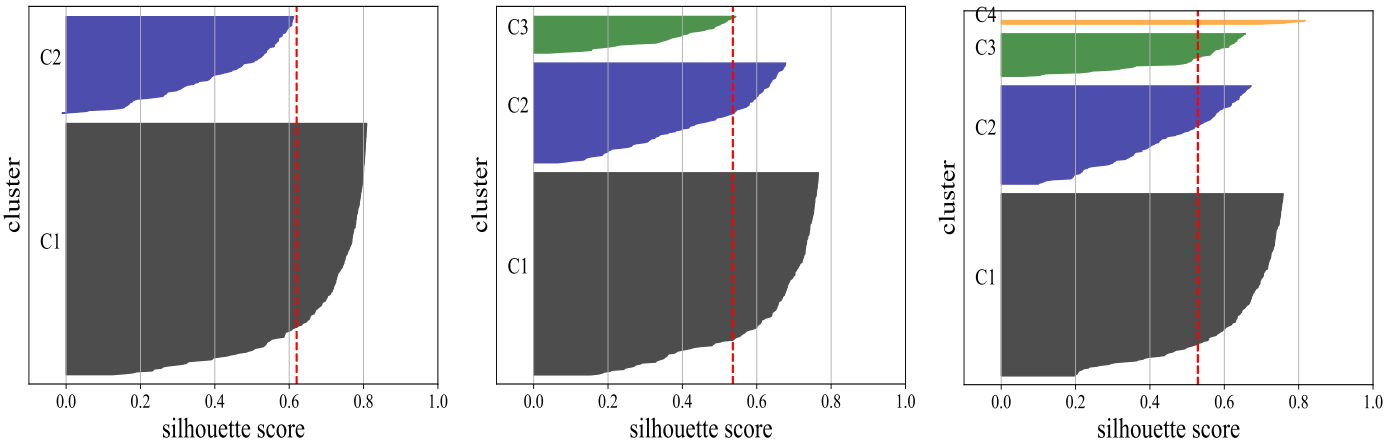

Figure 10. Silhouette-plot for $k = 2$ (**left**: C_1 in black and C_2 in blue), $k = 3$ (**center**: C_1 in black, C_2 in blue, and C_3 in green), and $k = 4$ (**right**: C_1 in black, C_2 in blue, C_3 in green, and C_4 in orange) for data of electrical load power consumption in Brazil, where the red dashed line represents the ASW.

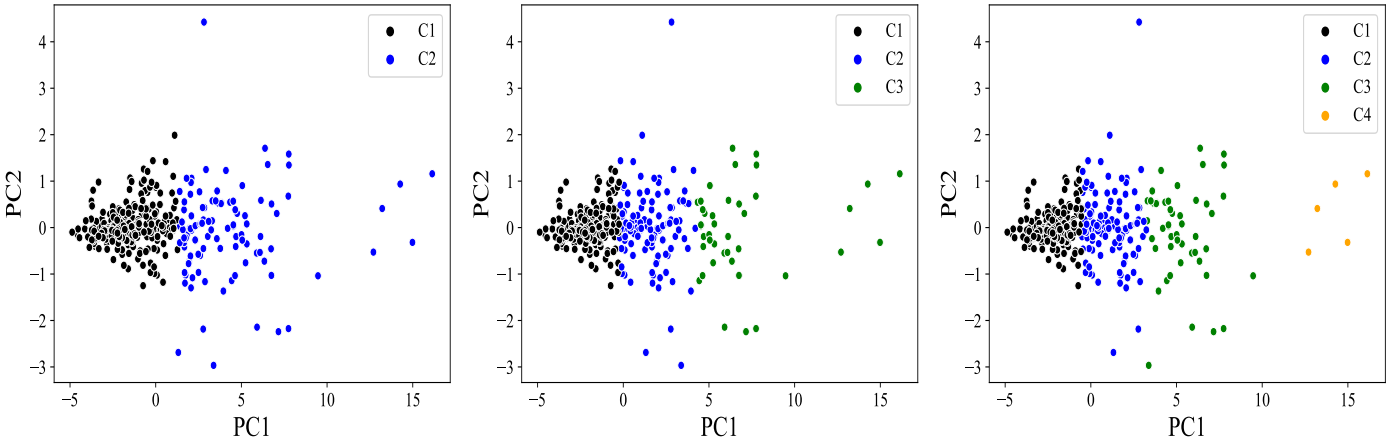

Figure 11. Scatter-plots of the PCs with $k = 2$ (**left**), $k = 3$ (**center**), and $k = 4$ (**right**) for the indicated cluster using data of electrical load power consumption in Brazil.

The Elbow method, which leverages the SSE, also known as inertia and defined in (1), determines the optimal number of clusters in the KM algorithm. As k increases, the inertia typically decreases. The optimal value of k is identified where this decline becomes pronounced, resembling an "elbow" [62].

A small range for values of k, such as $k = 1$ to $k = 10$, is chosen [63]. In our analysis, an elbow emerges at $k = 3$, suggesting that the optimal value of k for this dataset is $k = 3$, as shown in Figure 12. From this figure, we notice important inertia drops. For example, moving from $k = 2$ to $k = 3$ results in a drop of 620, with inertia from 1550 to 930. Beyond $k = 4$, reduction diminishes and inertia drops only by 190 between $k = 4$ and $k = 5$ (from 590 to 400). As this distance decrease, the data points are closer to each other within a cluster. Therefore, both the Elbow method and silhouette score analysis suggest that the optimal value of k for the KM algorithm applied to this dataset is $k = 3$ based on the inertia decrease rate.

The clustering results, as displayed in Figure 13, show the dispersion of data points in relation to their centroids, showing the relative positions and distances of data points within each cluster. Clusters C_2 and C_3 display a broader dispersion, while data points in Cluster C_1 seem more closely packed around its centroid.

The dispersion pattern is also reflected in the time series data for monthly household consumption. The time series for Cluster C_1, shown in Figure 14(left), exhibits a generally consistent temporal pattern interrupted by sporadic oscillations, showing consistent patterns with occasional fluctuations. In contrast, the time series data for Clusters C_2 and C_3, as depicted in Figure 14(center),(right), respectively, show more pronounced fluctuations throughout the monthly consumption timeframe, with the variations in Cluster C_3 being particularly extensive. Figure 14(center) illustrates moderate variability in energy use over time. Figure 14(right) depicts important fluctuations which are indicative of high variability in energy consumption.

Analyzing the yearly average consumption of households, $\overline{d_i}$, reveals distinct consumption patterns across the clusters. The division becomes even more pronounced when examining the scatter plot in Figure 13, which shows the dispersion of data points relative to their centroids for each cluster. This visualization effectively highlights the distinct energy consumption profiles among the clusters: a low consumption profile in C_1, medium in C_2, and high in C_3. The proximity of data points to their respective centroids in this scatter plot offers insights into the cohesiveness of each cluster and the variability within their consumption patterns, underscoring the diverse energy usage behaviors among Brazilian households.

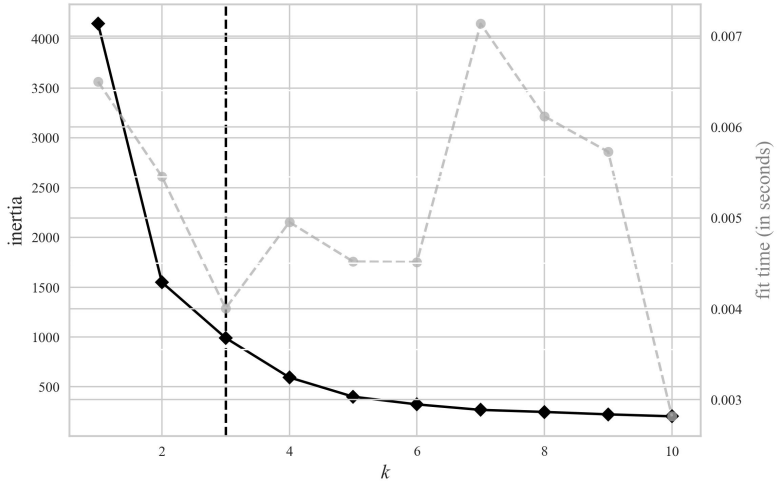

Figure 12. Plot of Elbow method using SSE for the electrical load data in Brazil, where the solid black line shows the SSE and the black dashed line the lowest SSE achieved at $k = 3$; whereas the gray dashed line represents the fit time (in seconds), which is the computational time required to fit the model for each number of clusters k.

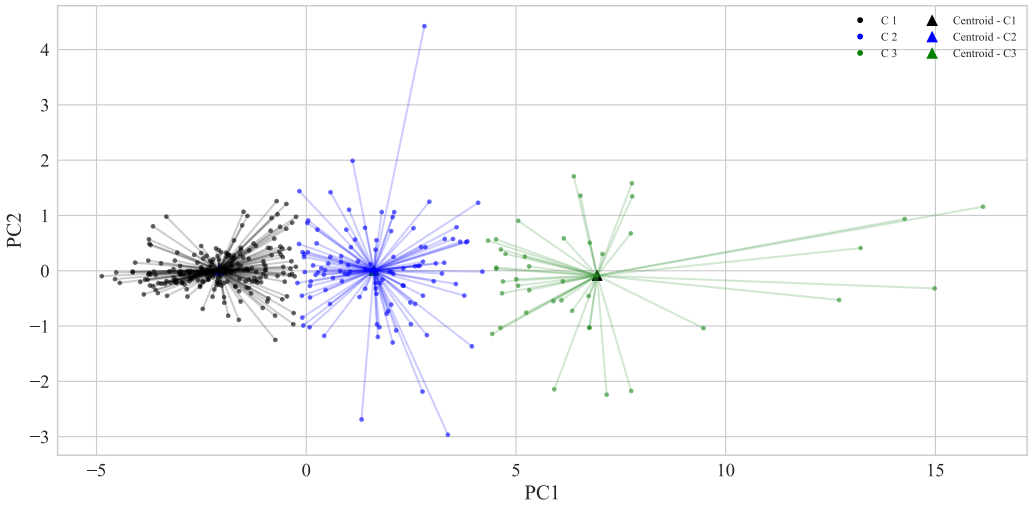

Figure 13. Scatter-plot of the indicated clusters and their centroids for the optimal value of k using data of electrical load power consumption in Brazil.

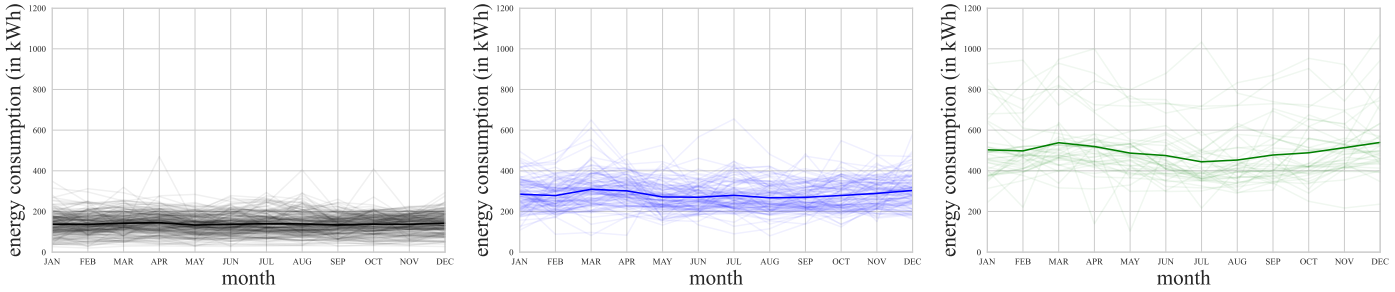

Figure 14. Time series for Clusters C_1 (**left**), C_2 (**center**), and C_3 (**right**) with data of electrical load power consumption in Brazil.

4.3. Load Profiles Using Agglomerative Hierarchical Clustering

The Ward linkage method, used in AHC, was conducted utilizing the `scikit-learn` package of `Python` [60]. The corresponding dendrogram of the AHC process is shown

in Figure 15 with the Ward linkage method. The x-axis denotes the data, while the y -axis illustrates the distance between them. Each vertical line represents a merge between clusters, with the height indicating the distance or dissimilarity at which this merge occurs.

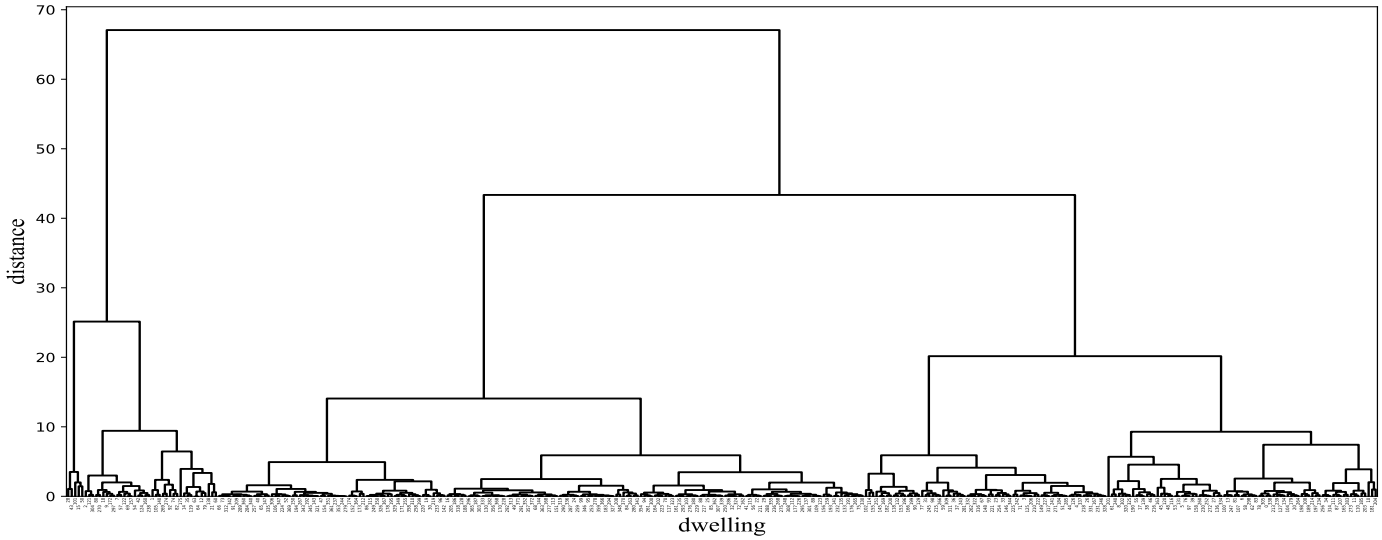

Figure 15. Dendrogram of the AHC process with the Ward linkage method for data of electrical load power consumption in Brazil.

From Figure 15, we visualize that, as the distance of vertical lines in the dendrogram increases, the distance between the corresponding clusters increases as well. This visualization aids in identifying the optimal number of clusters by examining the distances at which important merges occur, highlighting the hierarchical structure of data groupings.

At the first level, as shown in Figure 16(left), the dendrogram is divided into two clusters: C_1 in black on the right and C_2 in green on the left. At this cutting level, the majority of the data resides in a predominant Cluster C_1 with 330 data points, while a smaller Cluster C_2 contains 43 data points. At the second level, as shown in Figure 16(center), the data point division improves as only one cluster is split. Now, Cluster C_1 has 184 data points, C_2 has 146, and C_3 has 43 data points, showing a more balanced distribution. This level demonstrates an improvement in the division of data points among clusters. However, as shown in Figure 16(right), cutting at level 3 results in further division of the smallest cluster. The division creates an imbalance, as instead of a cluster with 43 data points, there is now a Cluster C_3 in green with 38 data points and a new Cluster C_4 in orange with only five data points, while C_1 and C_2 continue to have 184 and 146 data points, respectively. This level illustrates the effect of further segmentation on cluster balance, resulting in high imbalance by splitting the smallest cluster from the previous level.

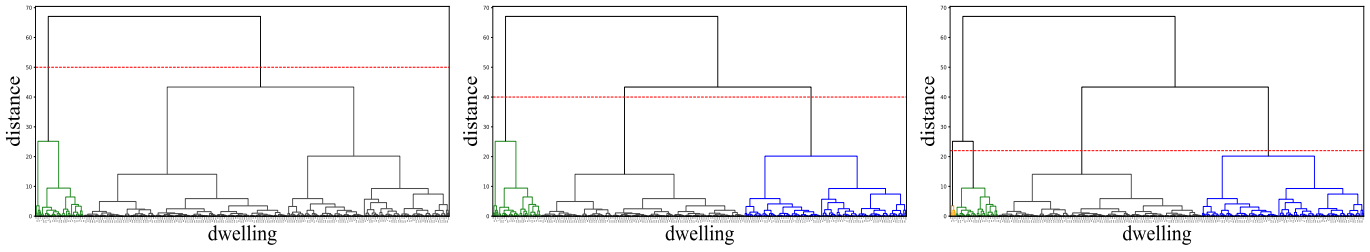

Figure 16. Dendrograms with initial cuts into two clusters (**left**: C_1 in black and C_2 in green), three clusters (**center**: C_1 in black, C_2 in blue, and C_3 in green), and four clusters (**right**: C_1 in black, C_2 in blue, C_3 in green, and C_4 in orange) for the data of electrical load power consumption in Brazil.

To determine the optimal dendrogram cut point, we analyze the cluster silhouette plots shown in Figure 17. Each plot displays the silhouette coefficients of clusters, highlighting

the optimal clustering structure at $k = 3$ and $k = 4$, with ASW values above a threshold of 0.5, suggesting strong cluster cohesion and separation. The plots show that all clusters are above the ASW line, and all ASW values are above the threshold of 0.5, with $k = 3$ and $k = 4$ yielding 0.5067 and 0.5097, respectively. For $k \geq 5$, all clusters result in poor ASW values, falling below the threshold.

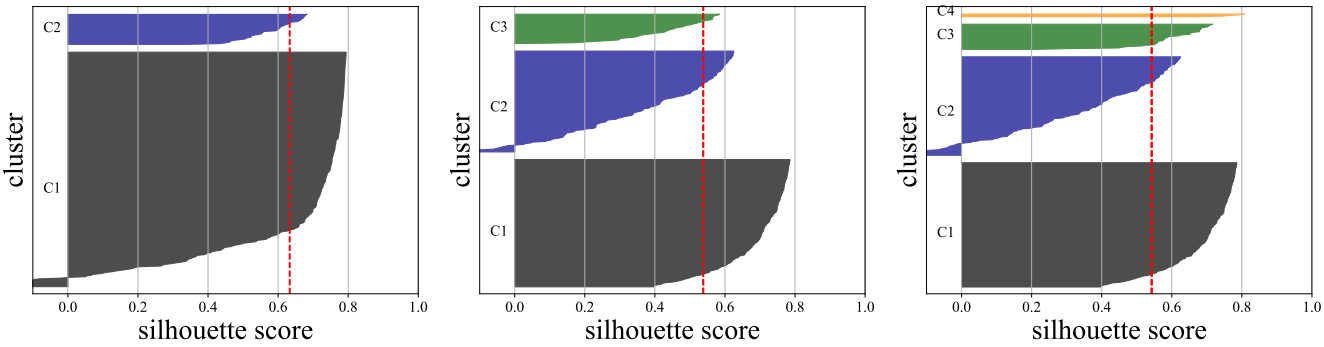

Figure 17. Silhouette-plots for hierarchical clustering with Ward linkage for $k = 2$ (**left**: C_1 in black and C_2 in blue), $k = 3$ (**center**: C_1 in black, C_2 in blue, and C_3 in green), and $k = 4$ (**right**: C_1 in black, C_2 in blue, C_3 in green, and C_4 in orange) for the data of electrical load power consumption in Brazil.

Upon analyzing the distribution of data points from Figure 17, it becomes evident that $k = 2$ and $k = 4$ represent two extreme possibilities. While $k = 2$ exhibits a cluster with a large concentration of data, $k = 4$ indicates a cluster with a small number of data points. Thus, among the analyzed cut levels, the dendrogram for $k = 3$, using the Ward linkage, presents better performance in terms of cluster distribution.

By examining the monthly consumption profiles obtained with $k = 3$, we can observe in Figure 18(left) the time series in each cluster. In Cluster C_1, we see an almost stable monthly average, while in the other clusters, there are dispersion, with C_3 being greater than C_2. This dispersion correlates with the number of data points and the compactness of the cluster; that is, as the cluster becomes more compact, the number of data points increases and the dispersion decreases.

The yearly average consumption presented in the box-plots of Figure 18(right) confirms that the variation observed in the monthly average is related to the compactness of the clusters. Cluster C_1, with more data points (N_{C_1} = 184), exhibits an interquartile range (IQR) of 50.15 kWh, while C_2 has 146 data points and an IQR of 91.81 kWh, with C_3 having 43 data points and an IQR of 95.54 kWh. Hence, as the number of data points decreases within the clusters and the IQR increases, the dispersion tends to increase. Here, IQR was used due to the potential presence of outliers.

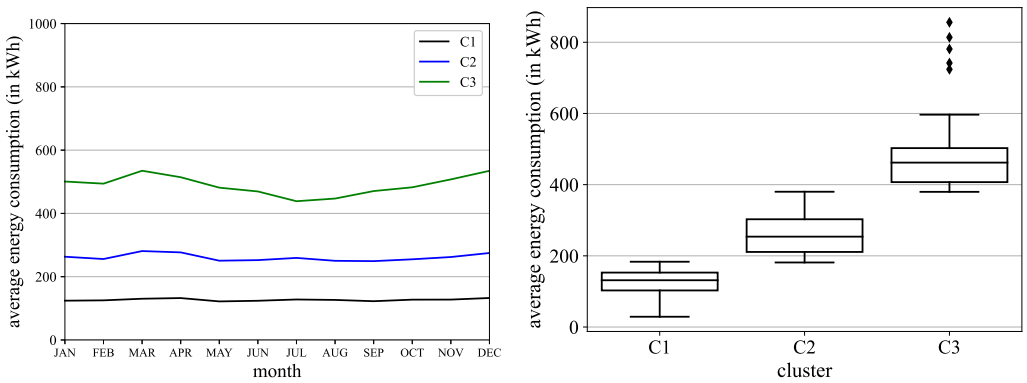

Figure 18. Plots of time series for monthly average consumption in each cluster with the Ward linkage method using $k = 3$ (**left**) and box-plots of yearly average consumption for the clusters (**right**), where diamonds indicate outliers and lines within the boxes the median of the data.

Therefore, three distinct profiles were identified: a low consumption profile in the first cluster C_1, with a maximum yearly average consumption of 183.41 kWh; a medium consumption profile in C_2, with a maximum value of 380.08 kWh; and a high consumption profile in C_3, with a maximum of 856.08 kWh. Thus, Figure 18(right) reveals distinct energy consumption profiles: low (Cluster C_1), medium (Cluster C_2), and high (Cluster C_3), with IQRs showing variability within clusters.

In summary, the AHC method utilizing the Ward linkage offered a comprehensive categorization of the data into discernible load profiles. This method has the ability to capture distinct consumption behaviors, underscoring its utility and relevance in understanding energy consumption patterns.

4.4. Load Profiles Using Self-Organizing Maps for Clustering

To cluster the load dataset using SOM, we begin by creating a neuron grid. According to the estimation proposed in [64], the number of neurons on the grid can be determined as approximately $5\sqrt{N}$. With $N = 373$, we obtain approximately 97 neurons. Since the grid is two-dimensional, we select a 10×10 neuron grid. The implementation of the SOM algorithm is achieved using the `MiniSom` library of `Python` [65].

The weight vector of each neuron is initialized randomly, and the ED is used as the distance metric between the neuron weight vector and the input vector. To determine the appropriate number of clusters, we evaluate the ASW for different values of k. As shown in Figure 19, the ASW values for $k = 2$ and $k = 3$ are above the threshold of 0.5, with values of 0.6057 and 0.5071, respectively. Figure 19 highlights the optimal clustering performance for $k = 3$ and $k = 4$, where all clusters surpass the threshold value of 0.5, indicating strong internal cohesion and separation from other clusters.

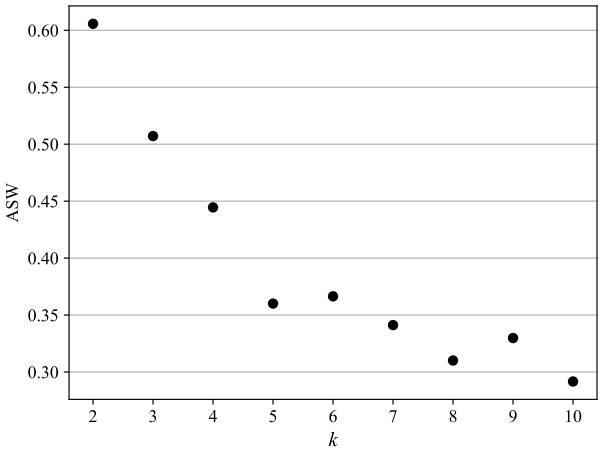

Figure 19. Plots of ASW for SOM clustering with k varying from 2 to 10 using the data of electrical load power consumption in Brazil.

To visualize the evolution of distances between neurons in each iteration, we utilize the U-matrix, which represents the distances between neurons by assigning colors to the cells. Darker colors indicate larger distances, while lighter colors indicate closer weight vectors. Initially, after the first iteration with the U-matrix, as shown in Figure 20(left), the neurons are mixed and not properly clustered. However, as the iterations progress, the weight vectors become more clustered, as displayed in Figure 20(right). The dark regions suggest dissimilar data points, while lighter or warmer regions indicate similar data points. Hence, in summary, Figure 20(left) illustrates initial neuron distances and cluster formations. Darker colors represent larger distances, indicating loosely connected or distinct clusters at the early stage of SOM training. Figure 20(right) shows a more defined clustering structure. Lighter colors indicate closer neurons, suggesting a clearer delineation of clusters as the SOM algorithm progresses.

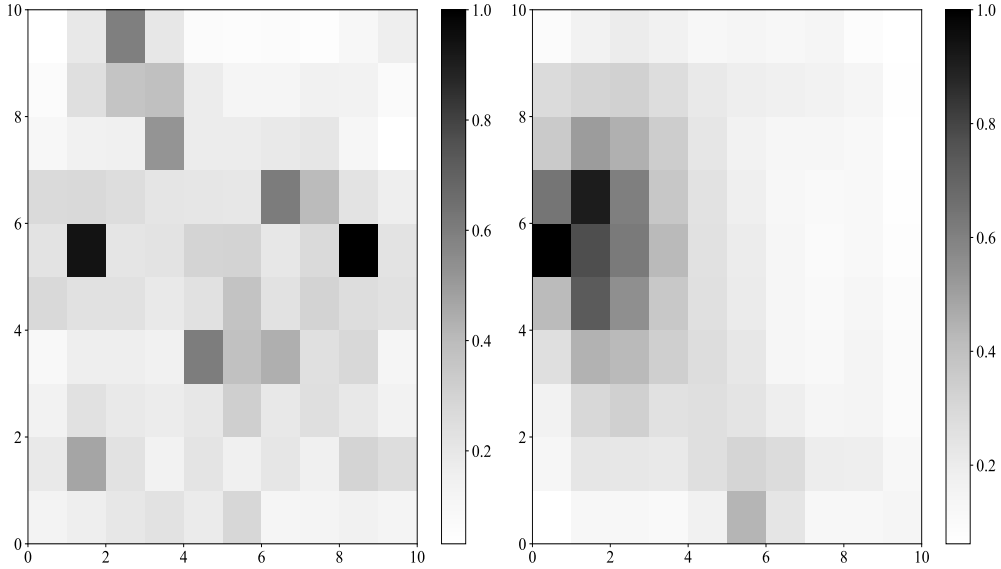

Figure 20. Plots of U-matrix after one iteration (**left**) and 4000 iterations (**right**) for the data of electrical load power consumption in Brazil.

By visualizing the BMUs marked with colors, we can identify the clusters. After the first iteration, as shown in Figure 21(left), the BMUs of each cluster exhibit distinct similarity patterns. However, as more iterations occur, the weight vectors become increasingly clustered, as displayed in Figure 21(right). This confirms the results obtained from the ASW analysis, indicating that $k = 3$ provides a good representation of the clusters. In summary, Figure 21(left) displays preliminary clustering patterns, while Figure 21(right) shows the refinement and clear definition of clusters over time with the SOM algorithm.

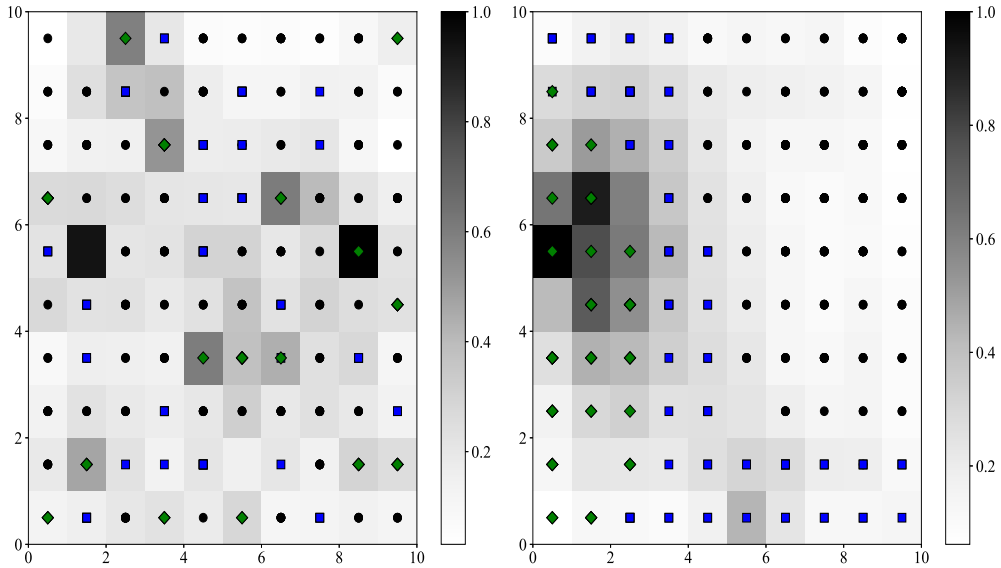

Figure 21. Plots of BMUs with cluster colors after one iteration (**left**) and 4000 iterations (**right**) for the data of electrical load power consumption in Brazil, where green diamonds are used for Cluster C3, blue squares for Cluster C2, and black dots for Cluster C1.

Assessing cluster quality through silhouette plots, as illustrated in Figure 22, reveals that, for $k = 2$, the silhouette values in the second cluster (C_2) fail to surpass the ASW threshold, presenting a limitation. Conversely, when $k = 3$, the silhouette plot shows all cluster silhouette values exceeding the threshold, suggesting an improved distribution of data points across the clusters.

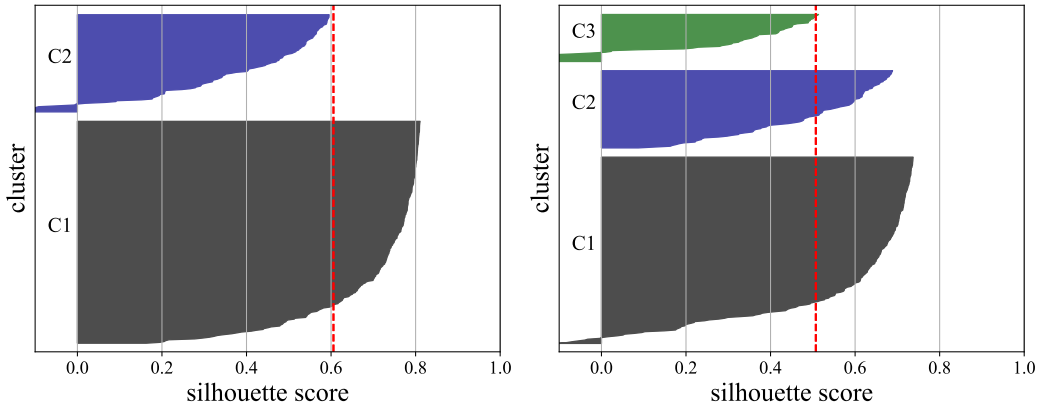

Figure 22. Silhouette-plots for the indicated cluster with $k = 2$ (**left**: C_1 in black and C_2 in blue) and $k = 3$ (**right**: C_1 in black, C_2 in blue, and C_3 in green) using the SOM algorithm for data of electrical load power consumption in Brazil.

By examining the monthly consumption profiles obtained with $k = 3$, we observe in Figure 23(left) the time series in each cluster. In Cluster C_1, we see an almost stable monthly average, while in the other clusters, there is dispersion, with C_3 being greater than C_2. This dispersion correlates with the number of data points and the compactness of the cluster; that is, as the cluster becomes more compact, the number of data points increases and the dispersion decreases.

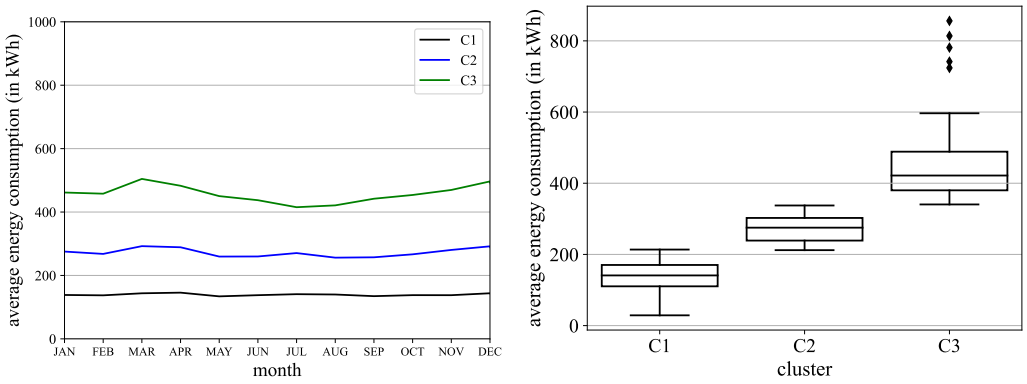

Figure 23. Plots of time series for monthly average consumption in each cluster with the Ward linkage method (**left**) and box-plots of yearly average consumption for the clusters (**right**) with $k = 3$ for data of electrical load power in Brazil, where diamonds indicate outliers and lines within the boxes the median of the data.

Upon analyzing the clusters using the yearly average consumption, as shown in the box-plots of Figure 23(right), we can identify three distinct consumption profiles:

- Cluster C_1 represents a low consumption class with a mean consumption of 139.32 kWh and a median consumption of 142.00 kWh.
- Cluster C_2 shows a medium consumption profile with a mean consumption of 272.24 kWh and a median consumption of 275.17 kWh.
- Cluster C_3 states a high consumption profile with a mean consumption of 457.73 kWh and a median consumption of 421.67 kWh.

Hence, Figure 23(right) illustrates the distribution within low, medium, and high consumption profiles. The difference between mean and median in Cluster C_3 is high, underscoring the impact of outliers on the mean consumption. The important difference between the mean and median in C_3 stems from the inclusion of five data points with consumption levels averaging between 600 kWh and 800 kWh, which impacts the mean but has little effect on the median.

In summary, employing SOM clustering with $k = 3$ yields distinct load profiles with different consumption patterns. Therefore, Clusters C_1, C_2, and C_3 represent low, medium, and high consumption profiles, respectively. The analysis of yearly and monthly average consumption further underscores the disparities among the clusters and provides valuable insights into the load characteristics of each cluster.

4.5. Discussion about the Clustering Methods Results

To evaluate the effectiveness of our clustering methods, we employ an analytical framework that integrates both unsupervised clustering methods and a supervised benchmarking tool. Figure 24 provides a visual overview of our analytical process, from data acquisition to the utilization of GB for benchmarking the clustering methods. This includes critical steps such as data preprocessing, dimensionality reduction, application of clustering algorithms, and rigorous performance evaluation against the GB benchmark.

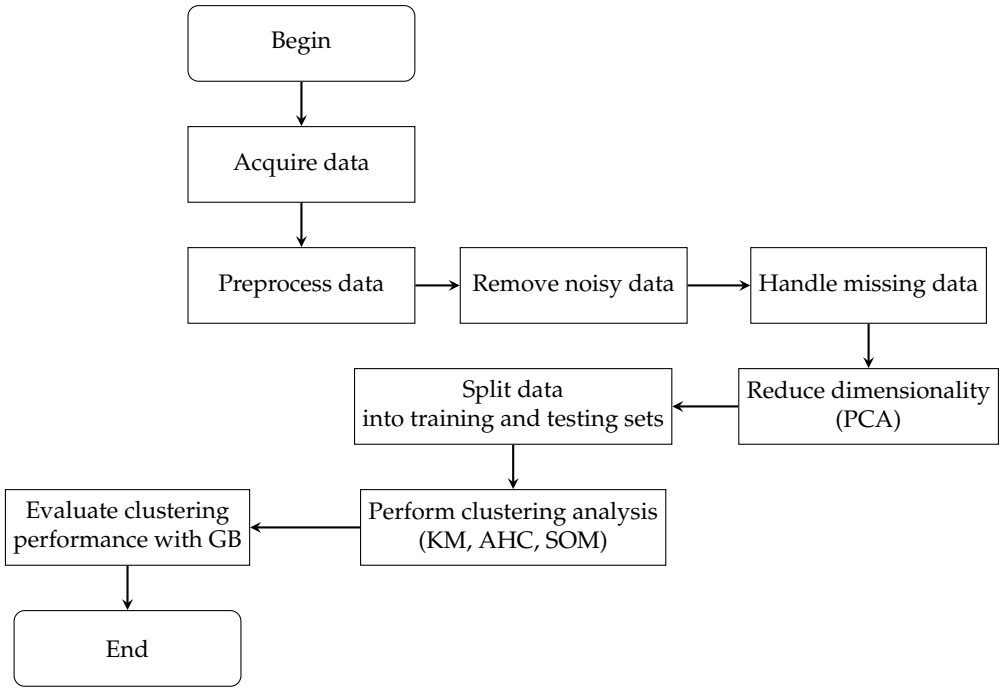

Figure 24. Flowchart of the data analysis process incorporating GB for benchmarking.

In our study, we performed clustering analysis with the ACH, KM, and SOM methods using GB as a performance baseline to enable a comprehensive comparison of these methods against a standard of predictive performance [66]. The use of GB to evaluate performance is based on its capability to handle heterogeneous data and capture complex interactions, allowing us to assess the effectiveness of each clustering algorithm. This ensures accurate billing for consumers and efficient resource management for energy providers.

To evaluate the clusters generated by KM, AHC, and SOM, we use precision (Prec) and recall (Rec) as metrics, comparing them against the patterns predicted by GB. Despite the challenges in defining true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) in unlabeled datasets, these metrics are crucial in our context. Precision measures the proportion of TP predictions among all positive predictions, helping to ensure that identified consumption patterns are accurate and not overestimated. Recall measures the proportion of TP predictions among all TPs, ensuring that the algorithm does not miss important consumption patterns. By using these metrics, we evaluate the performance of each clustering algorithm, balancing the need for accurate detection of consumption patterns with the risk of overestimation or underestimation, which is vital for both consumer satisfaction and resource management.

Specifically, precision and recall are defined, respectively, as $\text{Prec} = \text{TP}/(\text{TP} + \text{FP})$ and $\text{Rec} = \text{TP}/(\text{TP} + \text{FN})$. While metrics like normalized mutual information (NMI) and the rand index are well suited for evaluating unsupervised clustering methods, our choice of precision and recall is also justified by GB being a supervised method, where these metrics are highly relevant.

NMI measures similarity between predicted clusters and observed labels, considering the mutual information shared between them, while the rand index evaluates accuracy by examining all pairs of samples and determining how many pairs are correctly clustered together or separately. These metrics complement our evaluation by providing additional insights into clustering performance. Future work may explore the use of NMI and the rand index to provide a more comprehensive understanding of clustering effectiveness.

To accurately assess the performance of the clustering algorithms, the monthly consumption data from 373 dwellings were randomly split into two subsets: 70% forming the training set and the remaining 30% forming the test set. To ensure that this split maintained the proportionality of consumption categories—low (L), medium (M), and high (H)—present in the original dataset, we employed stratified sampling. This method guarantees that both subsets accurately reflect the variability of the complete dataset, providing a balanced foundation for evaluation. We then examined the accuracy within each consumption category to address the potential for aggregated accuracy metrics to obscure class-specific performance in multi-class problems. Our analysis, conducted with the `Scikit-learn` library and employing GB with a log-loss function, aims to clarify how effectively each algorithm categorizes dwellings into these consumption profiles.

The comparison of clustering methods in terms of precision and recall is given in Table 4, showing that the KM algorithm outperforms AHC and SOM, achieving a precision and recall of 92%, compared to 91% for both AHC and SOM. Despite the slight difference, this is important given the varying energy costs associated with different consumption profiles. The distinction between low and high consumption profiles, which correspond to lower and higher expenses, respectively, underscores the relevance of accurately categorizing dwellings. Such precision is not merely a technical achievement but has considerable financial implications for both consumers and energy suppliers.

Table 4. Metrics and advantages/disadvantages of the indicated clustering method for the data of electrical load power consumption in Brazil.

Method	Precision (Weighted Average)	Recall (Weighted Average)	Advantages	Disadvantages
KM	0.92	0.92	It is simple and computationally efficient.	It is sensitive to outliers and requires pre-set number of clusters.
AHC	0.91	0.91	It does not need for pre-set number of clusters and is good for small datasets.	it is computationally intensive and less efficient for large datasets.
SOM	0.91	0.91	It captures non-linear structures and has good visualization.	It needs careful parameter tuning and is computationally intensive.

Table 5 details the performance metrics by consumption profile, highlighting the precision and recall for each clustering method within the low, medium, and high consumption profiles. This detailed breakdown is crucial for understanding how each algorithm performs across different consumption levels.

The potential for cost discrepancies due to misclassification errors highlights the necessity of closely examining FPs and FNs within each consumption profile. These errors can result in incorrect billing and suboptimal energy management decisions. To provide a detailed analysis of these misclassifications, confusion matrices for each clustering algorithm are presented in Table 6. These matrices allow for an in-depth examination of how effectively each algorithm assigns dwellings to the correct consumption profile, emphasizing the practical impact of our study.

Table 5. Performance metrics by the indicated consumption profile for data of electrical load power consumption in Brazil.

Profile	Method	Precision	Recall	Comments
L	KM	0.95	0.94	It has slightly higher precision and good recall.
	AHC	0.92	1.00	It has perfect recall and slightly lower precision.
	SOM	0.94	0.97	It has high precision and recall.
M	KM	0.83	0.91	It has balanced performance.
	AHC	0.88	0.88	It has balanced precision and recall.
	SOM	0.83	0.86	It has consistent performance.
H	KM	1.00	0.85	It has perfect precision and lower recall
	AHC	1.00	0.64	It has perfect precision and highly lower recall.
	SOM	0.93	0.76	It has high precision and moderate recall

Table 6. Confusion matrix of the indicated algorithm for data of electrical load power consumption in Brazil.

		KM					**AHC**					**SOM**		
		Predicted					Predicted					Predicted		
		L	M	H			L	M	H			L	M	H
Observed	L	63	4	0	Observed	L	55	0	0	Observed	L	65	2	0
	M	3	29	0		M	5	38	0		M	3	24	1
	H	0	2	11		H	0	5	9		H	1	3	13

Examining the algorithms from a consumer perspective yields valuable insights. The SOM precision rate of 93% for high-profile dwellings implies that 7% of the dwellings classified by SOM as high-profile were misclassified. This is evident from the confusion matrix in Table 6, which shows a TP count of 13 and an FP count of 1 for SOM high-profile classification. This means that, while the algorithm identified 14 dwellings as high-profile, one of them was truly suited for the medium-profile. Similarly, evaluating the KM algorithm for its low-profile classification reveals a recall rate of 94%, indicating that 6% of the dwellings expected to be low-profile were incorrectly categorized as a higher profile. The confusion matrix in Table 6 details this with a TP count of 63 and an FN count of 4 for KM low-profile classification. Out of 67 dwellings expected to be in the low-profile, the algorithm accurately identified 63 but misclassified 4 as belonging to the medium-profile.

From the perspective of an energy company, the performance metrics unveil a distinct aspect. The AHC recall rate of 64% in classifying high-profile dwellings implies that 36% of dwellings expected in this category were incorrectly assigned to a lower one, as evidenced by the confusion matrix in Table 6. This misclassification, where 5 out of 14 high-profile dwellings are categorized into the medium-profile, underscores a potential underestimation of energy consumption. Conversely, an AHC precision of 92% in predicting low-profile dwellings suggests an 8% misclassification rate, where some dwellings are inaccurately categorized as lower than their observed profile. This discrepancy is critical for energy companies as it may result in revenue loss due to underbilling. Therefore, although KM, AHC, and SOM all exhibit high weighted average accuracies, their appropriateness varies depending on the stakeholder perspective. AHC is particularly precise from a consumer standpoint, effectively mitigating over-classification into costlier profiles. However, its tendency to underestimate high-profile predictions may concern energy companies. Conversely, KM and SOM offer more balanced performances, with KM slightly ahead due to its lower error rate in low-profile predictions and consistent accuracy across various measures. Thus, selecting a clustering algorithm for energy consumption analysis should be guided by the priorities of the stakeholders involved. Whether the focus is on minimizing overestimation for consumer protection or avoiding underestimation for accurate billing, the decision hinges on the requirements of the consumer or the energy provider. Table 7 compares the properties, advantages, and disadvantages of each clustering method discussed.

Table 7. Properties, advantages, and disadvantages for the indicated clustering method.

Method	Properties	Advantages	Disadvantages
KM	Iterative clustering based on centroids	It is simple and computational efficiency.	It is sensitive to outliers and requires pre-set number of clusters.
AHC	Hierarchical clustering, cluster merging	It does not require pre-set number of clusters and is good for small datasets.	It is computationally intensive and less efficient for large datasets.
SOM	Clustering based on neural networks	It captures nonlinear structures and has good visualization.	It requires careful parameter tuning and is computationally intensive.

In summary, our comparative analysis highlights the strengths and weaknesses of each clustering method. The KM method, with its simplicity and computational efficiency, is well suited for large datasets but is sensitive to outliers. The AHC method, while robust for small datasets and not requiring a pre-defined number of clusters, is computationally intensive and less efficient for larger datasets. The SOM method effectively captures nonlinear structures and provides good visualization capabilities but requires careful parameter tuning and is computationally intensive. The choice of clustering algorithm ultimately depends on the specific requirements of the application, including dataset size, desired clustering characteristics, and computational efficiency. Our study demonstrates the practical applicability of each method and provides a framework for selecting the appropriate clustering method based on the unique needs of energy consumption analysis.

5. Conclusions and Future Work

As we confront the persistent challenges of climate change and increasing energy demands, the urgency of innovative energy conservation and efficiency strategies becomes ever more critical. Developing a comprehensive understanding of energy consumption profiles from both consumer and supplier perspectives is vital for advancing these strategies. This understanding is essential for identifying opportunities to reduce and optimize energy usage, thereby promoting a more sustainable approach to energy resource management.

In Brazil, with its dynamic economy and diverse population, these challenges are relevant. Unique policies, such as consumption thresholds for billing, introduce complexity into the analysis of energy usage, obscuring true consumption patterns. This complexity shows the need for accurate data collection and the development of energy conservation and efficiency strategies tailored to the specific context of Brazilian energy consumption.

Our study utilized machine learning methods, specifically k-means, agglomerative hierarchical clustering, and self-organizing maps, to analyze energy consumption data in Brazilian households. The incorporation of gradient boosting as a benchmarking tool for these unsupervised learning methods facilitated a systematic evaluation of their efficacy. This benchmark introduced a novel dimension to our analysis, particularly in enabling a comparison of the clustering algorithms in effectively discerning consumption profiles.

The comparison highlighted in our analysis was particularly evident as we approached the results from both a consumer and an energy company perspective, yielding valuable insights. By categorizing "high-profile" dwellings as those with higher energy consumption and "low-profile" as those with lower consumption, we could observe the precision and recall rates of each algorithm more distinctly. For example, the precision rate of 93% for self-organizing maps in classifying high-profile dwellings underscores a notable rate of misclassification. This is further elucidated by our confusion matrix analysis, which shows how each clustering method performs in real-world scenarios.

Likewise, the recall rates of k-means in low-profile classifications and the performance of agglomerative hierarchical clustering in high-profile dwelling classifications unveiled important aspects concerning the potential underestimation or overestimation of energy consumption. These insights are pivotal for applications like accurate billing and consumer protection, where the comprehension of consumption behavior directly influences operational and strategic decisions.

Each utilized clustering method has shown its unique strengths in segmenting energy data into distinct profiles: k-means for its simplicity and efficiency, agglomerative hierarchical clustering for its balanced distribution capabilities, and self-organizing maps for its in-depth pattern recognition. However, our findings underscore that the selection of a clustering algorithm should be carefully aligned with the specific needs and perspectives of the stakeholders involved, whether it is minimizing overestimation for consumer protection or avoiding underestimation for accurate billing.

Our findings indicate that k-means offered the most accurate results for energy management applications, thereby enhancing the precision of energy policies. By providing a clearer identification of energy consumption patterns, our methodology supports the development of more targeted and effective energy conservation strategies. This improved classification directly informs policy-making processes by identifying specific areas where intervention is needed, promoting efficient and equitable resource allocation.

While precision and recall have been valuable metrics in our study, we acknowledge the limitations related to their reliance on labeled data, which might not always be reliable. Metrics like normalized mutual information and the rand index, which do not require labeled data, could provide complementary insights. Future work will explore these metrics to enhance our evaluation framework.

Beyond the methods explored in this study, the continuous advancements in deep learning, including formulations like large language models and latent mixture models, offer promising avenues for future analyses. These sophisticated models could further refine our understanding of energy consumption patterns, potentially unveiling more intricate relationships within the data that conventional methods might overlook.

Future work could also extend beyond the dataset used in the present study to compare energy consumption patterns across the different regions of Brazil, capturing the diverse socioeconomic and climatic conditions within the country. Additionally, analyzing the impact of the COVID-19 pandemic on energy usage could provide insights into how shifts in home office work and residential energy consumption behaviors have altered consumption patterns. Moreover, comparing Brazilian energy consumption with that of other countries in Latin America could highlight regional differences and similarities, offering a broader perspective on energy efficiency strategies suitable for the region.

Looking ahead, broadening our analysis to encompass additional data features, such as peak demand times and consumption variability, could enhance the profiling process further. Moreover, exploring other clustering algorithms or incorporating advanced deep learning methods, such as autoencoders, might provide deeper insights. Analyzing consumption data at more granular time intervals may also provide more detailed information for optimizing electricity demand management strategies.

The practical utilization of the consumption profiles identified in this study in initiatives like demand response programs and personalized energy-saving recommendations presents a promising avenue for future exploration. Such identifications hold the potential to substantially enhance load management practices, enabling more accurate demand forecasting and resource allocation. Overall, this study established a robust groundwork for further research in energy consumption analysis and hints at the prospect of extending these methodologies to other sectors that require detailed pattern analysis.

Author Contributions: Conceptualization, L.H., C.C. and F.P.; data curation, L.H., C.C. and F.P.; formal analysis, L.H., C.C., F.P., V.L. and R.V.; investigation, L.H., C.C., F.P. and V.L.; methodology, L.H., C.C., F.P., V.L. and R.V.; writing—original draft, L.H. and F.P.; writing—review and editing, C.C., V.L. and R.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by Portuguese funds through the CMAT—Research Centre of Mathematics of University of Minho, Portugal, within projects UIDB/00013/2020 (https://doi.org/10.54499/UIDB/00013/2020) and UIDP/00013/2020 (https://doi.org/10.54499/UIDP/00013/2020) (C.C.); and FONDECYT grant number 1200525 (V.L.) from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science, Technology, Knowledge, and Innovation.

Data Availability Statement: Data and codes are available from the authors upon request.

Acknowledgments: The authors would like to thank the editors and reviewers for their constructive comments, which led to improvements in the presentation of the article.

Conflicts of Interest: There are no conflicts of interest declared by the authors.

References

1. Rahman, M.Z.U.; Akbar, M.A.; Leiva, V.; Martin-Barreiro, C.; Imran, M.; Riaz, M.T.; Castro, C. An IoT-fuzzy intelligent approach for holistic management of COVID-19 patients. *Heliyon* **2024**, *10*, e22454. [CrossRef]
2. Cavalcante, T.; Ospina, R.; Leiva, V.; Martin-Barreiro, C.; Cabezas, X. Weibull regression and machine learning survival models: Methodology, comparison, and application to biomedical data related to cardiac surgery. *Biology* **2023**, *11*, 1394. [CrossRef] [PubMed]
3. Ospina, R.; Ferreira, A.G.O.; de Oliveira, H.M.; Leiva, V.; Castro, C. On the use of machine learning techniques and non-invasive indicators for classifying and predicting cardiac disorders. *Biomedicines* **2023**, *11*, 2604. [CrossRef]
4. Palacios, C.A.; Reyes-Suarez, J.A.; Bearzotti, L.A.; Leiva, V.; Marchant, C. Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy* **2021**, *23*, 485. [CrossRef]
5. Taylan, O.; Alkabaa, A.S.; Alqabbaa, H.S.; Pamukcu, E.; Leiva, V. Early prediction in classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods. *Biology* **2023**, *12*, 1179. [CrossRef] [PubMed]
6. Huerta, M.; Leiva, V.; Rojas, R.; Wanke, P.; Cabezas, X. A methodology for consolidation effects of inventory management with serially dependent random demand. *Processes* **2023**, *11*, 2008. [CrossRef]
7. Manchini, C.; Ospina, R.; Leiva, V.; Martin-Barreiro, C. A new approach to data differential privacy based on regression models under heteroscedasticity with applications to machine learning repository data. *Inf. Sci.* **2023**, *627*, 280–300. [CrossRef]
8. Li, D.; Yang, P.; Zou, Y. Optimizing Insulator Defect Detection with Improved DETR Models. *Mathematics* **2024**, *12*, 1507. [CrossRef]
9. Martin-Barreiro, C.; Cabezas, X.; Leiva, V.; de Santis, P.R.; Ramirez-Figueroa, J.A.; Delgado, E. Statistical characterization of vaccinated cases and deaths due to COVID-19: Methodology and case study in South America. *AIMS Math.* **2023**, *8*, 22693–22713. [CrossRef]
10. Alkadya, W.; ElBahnasy, K.; Leiva, V.; Gad, W. Classifying COVID-19 based on amino acids encoding with machine learning algorithms. *Chemom. Intell. Lab. Syst.* **2022**, *224*, 104535. [CrossRef]
11. Grigoras, G.; Neagu, B.C.; Gavrilaș, M.; Triștiu, I.; Bulac, C. Optimal phase load balancing in low voltage distribution networks using a smart meter data-based algorithm. *Mathematics* **2020**, *8*, 549. [CrossRef]
12. Delgado, E.; Cabezas, X.; Martin-Barreiro, C.; Leiva, V.; Rojas, F. An equity-based optimization model to solve the location problem for healthcare centers applied to hospital beds and COVID-19 vaccination. *Mathematics* **2022**, *10*, 1825. [CrossRef]
13. Ma, L.; Zhang, Y.; Leiva, V.; Liu, S.; Ma, T. A new clustering algorithm based on a radar scanning strategy with applications to machine learning data. *Expert Syst. Appl.* **2022**, *191*, 116143. [CrossRef]
14. Jahanger, A.; Awan, A.; Anwar, A.; Adebayo, T.S. Greening the Brazil, Russia, India, China and South Africa (BRICS) economies: Assessing the impact of electricity consumption, natural resources, and renewable energy on environmental footprint. *Nat. Resour. Forum* **2023**, *47*, 484–503. [CrossRef]
15. Michalakopoulos, V.; Sarmas, E.; Papias, I.; Skaloumpakas, P.; Marinakis, V.; Doukas, H. A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs. *Appl. Energy* **2024**, *361*, 122943. [CrossRef]
16. Lin, L.; Chen, C.; Wei, B.; Li, H.; Shi, J.; Zhang, J.; Huang, N. Residential electricity load scenario prediction based on transferable flow generation model. *J. Electr. Eng. Technol.* **2023**, *18*, 99–109. [CrossRef]
17. Zhang, X.; Ramírez-Mendiola, J.L.; Li, M.; Guo, L. Electricity consumption pattern analysis beyond traditional clustering methods: A novel self-adapting semi-supervised clustering method and application case study. *Appl. Energy* **2022**, *308*, 118335. [CrossRef]
18. Toussaint, W.; Moodley, D. Clustering residential electricity consumption data to create archetypes that capture household behaviour in South Africa. *S. Afr. Comput. J.* **2020**, *32*, 1–34. [CrossRef]
19. Abdalameer, A.K.; Alswaitti, M.; Alsudani, A.A.; Isa, N.A.M. A new validity clustering index-based on finding new centroid positions using the mean of clustered data to determine the optimum number of clusters. *Expert Syst. Appl.* **2022**, *191*, 116329. [CrossRef]
20. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2022**, *622*, 178–210. [CrossRef]
21. Touzani, S.; Granderson, J.; Fernandes, S. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy Build.* **2018**, *158*, 1533–1543. [CrossRef]
22. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **2013**, *7*, 21. [CrossRef] [PubMed]
23. Ramirez-Figueroa, J.A.; Martin-Barreiro, C.; Nieto, A.B.; Leiva, V.; Galindo-Villardón, M.P. A new principal component analysis by particle swarm optimization with an environmental application for data science. *Stoch. Environ. Res. Risk Assess.* **2021**, *35*, 1969–1984. [CrossRef]
24. Gupta, G.; Mathur, S.; Mathur, J.; Nayak, B.K. Blending of energy benchmarks models for residential buildings. *Energy Build.* **2023**, *292*, 113195. [CrossRef]

25. Liu, G.; Yang, J.; Hao, Y.; Zhang, Y. Big data-informed energy efficiency assessment of China industry sectors based on k-means clustering. *J. Clean. Prod.* **2018**, *183*, 304–314. [CrossRef]
26. Al-Wakeel, A.; Wu, J.; Jenkins, N. K-means based load estimation of domestic smart meter measurements. *Appl. Energy* **2017**, *194*, 333–342. [CrossRef]
27. Jafarzadegan, M.; Safi-Esfahani, F.; Beheshti, Z. Combining hierarchical clustering approaches using the PCA method. *Expert Syst. Appl.* **2019**, *137*, 1–10. [CrossRef]
28. Xu, D.; Tian, Y. A comprehensive survey of clustering algorithms. *Ann. Data Sci.* **2015**, *2*, 165–193. [CrossRef]
29. Yildiz, B.; Bilbao, J.I.; Dore, J.; Sproul, A.B. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Appl. Energy* **2017**, *208*, 402–427. [CrossRef]
30. Satre-Meloy, A.; Diakonova, M.; Grünewald, P. Cluster analysis and prediction of residential peak demand profiles using occupant activity data. *Appl. Energy* **2020**, *260*, 114246. [CrossRef]
31. Wei, Y.; Zhang, X.; Shi, Y.; Xia, L.; Pan, S.; Wu, J.; Han, M.; Zhao, X. A review of data-driven approaches for prediction and classification of building energy consumption. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1027–1047. [CrossRef]
32. Aykroyd, R.G.; Leiva, V.; Ruggeri, F. Recent developments of control charts, identification of big data sources and future trends of current research. *Technol. Forecast. Soc. Chang.* **2019**, *144*, 221–232. [CrossRef]
33. Wen, L.; Zhou, K.; Yang, S. A shape-based clustering method for pattern recognition of residential electricity consumption. *J. Clean. Prod.* **2019**, *212*, 475–488. [CrossRef]
34. Rajabi, A.; Eskandari, M.; Ghadi, M.J.; Li, L.; Zhang, J.; Siano, P. A comparative study of clustering techniques for electrical load pattern segmentation. *Renew. Sustain. Energy Rev.* **2020**, *120*, 109628. [CrossRef]
35. Si, C.; Xu, S.; Wan, C.; Chen, D.; Cui, W.; Zhao, J. Electric load clustering in smart grid: Methodologies, applications, and future trends. *J. Mod. Power Syst. Clean Energy* **2021**, *9*, 237–252. [CrossRef]
36. Randriamihamison, N.; Vialaneix, N.; Neuvial, P. Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. *J. Classif.* **2021**, *38*, 363–389. [CrossRef]
37. Ward, J.H., Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
38. Miljković, D. Brief review of self-organizing maps. In Proceedings of the 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija, Croatia, 22–26 May 2017; pp. 1061–1066.
39. Llanos, J.; Sáez, D.; Palma-Behnke, R.; Núñez, A.; Jiménez-Estévez, G. Load profile generator and load forecasting for a renewable based microgrid using self organizing maps and neural networks. In Proceedings of the International Joint Conference on Neural Networks, Brisbane, Australia, 10–15 June 2012; pp. 1–8.
40. Cottrell, M.; Olteanu, M.; Rossi, F.; Villa-Vialaneix, N.N. Self-organizing maps, theory and applications. *Rev. Investig. Oper.* **2018**, *39*, 1–22.
41. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
42. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
43. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
44. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
45. Di Persio, L.; Fraccarolo, N. Energy consumption forecasts by gradient boosting regression trees. *Mathematics* **2023**, *11*, 1068. [CrossRef]
46. Sainani, K.L. Dealing with non-normal data. *PM&R* **2012**, *4*, 1001–1005.
47. Emmanuel, T.; Maupong, T.; Mpoeleng, D.; Semong, T.; Mphago, B.; Tabona, O. A survey on missing data in machine learning. *J. Big Data* **2021**, *8*, 140. [CrossRef] [PubMed]
48. Fu, X.; Zeng, X.J.; Feng, P.; Cai, X. Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China. *Energy* **2018**, *165*, 76–89. [CrossRef]
49. Ashouri, M.; Haghighat, F.; Fung, B.C.M.; Lazrak, A.; Yoshino, H. Development of building energy saving advisory: A data mining approach. *Energy Build.* **2018**, *172*, 139–151. [CrossRef]
50. Azur, M.J.; Stuart, E.A.; Frangakis, C.; Leaf, P.J. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* **2011**, *20*, 40–49. [CrossRef]
51. Gibson, S.J.; Narendra, A.; Dainotti, M.G.; Bogdan, M.; Pollo, A.; Poliszczuk, A.; Rinaldi, E.; Liodakis, I. Using multivariate imputation by chained equations to predict redshifts of active galactic nuclei. *Front. Astron. Space Sci.* **2022**, *9*, 836215. [CrossRef]
52. Bosisio, A.; Berizzi, A.; Vicario, A.; Morotti, A.; Greco, B.; Iannarelli, G.; Le D. A method to analyzing and clustering aggregate customer load profiles based on PCA. In Proceedings of the 5th International Conference on Green Technology and Sustainable Development, Ho Chi Minh City, Vietnam, 27–28 November 2020; pp. 41–47.
53. Tasoulis, S.; Pavlidis, N.G.; Roos, T. Nonlinear dimensionality reduction for clustering. *Pattern Recognit.* **2020**, *107*, 107508. [CrossRef]
54. Aréchiga, A.; Barocio, E.; Ayon, J.J.; Garcia-Baleon, H.A. Comparison of dimensionality reduction techniques for clustering and visualization of load profiles. In Proceedings of the IEEE PES Transmission and Distribution Conference and Exposition-Latin America, Dallas, TX, USA, 3–5 May 2016; pp. 1–6.

55. Zhang, J.; Yang, X.; Shen, F.; Li, Y.; Xiao, H.; Qi, H.; Peng, H.; Deng, S. Principal component analysis of electricity consumption factors in China. *Energy Procedia* **2012**, *16*, 1913–1918. [CrossRef]
56. Akoglu, H. User's guide to correlation coefficients. *Turk. J. Emerg. Med.* **2018**, *18*, 91–93. [CrossRef] [PubMed]
57. Liu, Y.; Mu, Y.; Chen, K.; Li, Y.; Guo, J. Daily activity feature selection in smart homes based on Pearson correlation coefficient. *Neural Process. Lett.* **2020**, *51*, 1771–1787. [CrossRef]
58. Singh, S.; Yassine, A. Big data mining of energy time series for behavioral analytics and energy consumption forecasting. *Energies* **2018**, *11*, 452. [CrossRef]
59. Davies, D.L.; Bouldin, D.W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *2*, 224–227. [CrossRef]
60. 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.
61. Bengfort, B.; Bilbro R. Yellowbrick: Visualizing the scikit-learn model selection process. *J. Open Source Softw.* **2019**, *4*, 1075. [CrossRef]
62. Somu, N.; MR, G.R.; Ramamritham, K. A deep learning framework for building energy consumption forecast. *Renew. Sustain. Energy Rev.* **2021**, *137*, 110591. [CrossRef]
63. Guo, Z.; Zhou, K.; Zhang, X.; Yang, S.; Shao, Z. Data mining based framework for exploring household electricity consumption patterns: A case study in China context. *J. Clean. Prod.* **2018**, *195*, 773–785. [CrossRef]
64. Tian, J.; Azarian, M.H.; Pecht, M. Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. In Proceedings of the PHM Society European Conference, Nantes, France, 8–10 July 2014; Volume 2.
65. Vettigli, G. MiniSom: Minimalistic and NumPy-Based Implementation of the Self Organizing Map. 2018. Available online: https://github.com/JustGlowing/minisom/ (accessed on 4 June 2024).
66. Xie, B.; Zhu, C.; Zhao, L.; Zhang, J. A gradient boosting machine-based framework for electricity energy knowledge discovery. *Front. Environ. Sci.* **2022**, *10*, 1031095. [CrossRef]