IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier –

Identification of Typical Time-Series House Consumption Profiles based on Unsupervised Learning Techniques

MARCELO FORTE[1, 2], CINDY P. GUZMAN[1], LUCAS PEREIRA[2], HUGO MORAIS[1]

[1]Department of Electrical and Computer Engineering, INESC-ID—Instituto de Engenharia de Sistemas e Computadores-Investigação e Desenvolvimento, Instituto Superior Técnico (IST), Universidade de Lisboa, Rua Alves Redol, 9, Lisboa, 1000-029 Lisbon, Portugal
[2]Interactive Technologies Institute (ITI), Laboratory for Robotics and Engineering Systems (LARSyS), Av. Rovisco Pais, 1, Lisboa, 1049-001, Lisbon, Portugal

Corresponding author: Hugo Morais (e-mail: hugo.morais@tecnico.ulisboa.pt).

This work has been developed under the U2DEMO project, funded by the European Union's Horizon Innovation Actions under grant agreement no. 101160684. Views and opinions expressed in this document are those of the authors only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. The work was also supported by Portuguese Foundation for Science and Technology (FCT) in the scope of the MIT-Portugal Programa, under grant (2022.15771.MIT), through national funds and by the project nº 56 - "ATE", financed by European Funds, namely "Recovery and Resilience Plan - Component 5: Agendas Mobilizadoras para a Inovação Empresarial", included in the NextGenerationEU funding program. The authors received funding from the Portuguese Foundation for Science and Technology under grant CEECIND/01179/2017 (L.P.), UIDB/50009/2020 (L.P.) and UIDB/50021/2020 (M.F., C.G., H.M.).

ABSTRACT Energy consumption profiles are a critical component of energy management strategies. With the recent widespread adoption of smart meters, there is a pressing need to develop robust methodologies for obtaining, characterizing, and visualizing these profiles. This paper presents a comparative analysis of various clustering methods for deriving consumer electricity profiles, aiming to identify the most suitable techniques for big time-series data and to evaluate how the granularity of the data influences the final profiles. Four approaches were considered for obtaining house electricity consumption profiles: K-means, Principal Component Analysis (PCA) combined with K-means, Self-Organizing Maps (SOM), and SOM combined with K-means. Additionally, an iterative study with the scores silhouette coefficient and Davies-Bouldin index validated the subjective indications from the elbow method. The results demonstrate that hybrid approaches offer superior performance compared to conventional clustering models, highlighting the effectiveness of combining PCA and SOM in improving model generalization and supporting more targeted energy management strategies. SOM combined with K-means provides the best clustering quality across all tested scenarios, making it the preferred method when computational resources are not a limiting factor. On the other hand, PCA combined with K-means highlights a good balance between meaningful profiles and the capacity to handle large datasets effectively.

INDEX TERMS Clustering, Consumption profile, Dimensionality reduction, Time-series data.

I. INTRODUCTION

THE increasing electrification of modern society is reshaping how energy is consumed and managed, driven by sustainability goals and technological advancements. Buildings are at the center of this transformation, accounting for approximately 28% of total final energy consumption worldwide in 2023 [1], while households, in particular, are becoming more energy-dependent due to the growing adoption of electric heating, cooling, and smart appliances. Enhancing energy efficiency and reducing carbon emissions in this sector remain critical priorities, requiring a deeper understanding of electricity consumption patterns to support more sustainable energy management strategies.

The widespread deployment of smart meters in residential, commercial, and industrial buildings has revolutionized electricity monitoring [2], [3]. Instead of relying on traditional subjective surveys, data collection has shifted to real-time, high-resolution measurements, enabling more accurate analysis of energy usage [4]. This big data has facilitated the development of advanced consumption profiling techniques, allowing the identification of behavioral patterns and the optimization of energy management strategies. For example,

with the data provided by smart meters, electricity providers can design personalized demand response (DR) programs promoting peak load reduction [5], improving system reliability, load forecasting [6], [7], grid planning, and designing of time-of-use tariffs tailored to typical consumption behaviors [8]. In the context of electric mobility, such data can enhance smart charging strategies by aligning EV charging with household routines and grid constraints [9], [10].

A fundamental concept in this context is *consumption profiles*, which aim to analyze and identify statistically representative electricity consumption patterns over different timeframes, such as a day (24 hours), a week, a season, or even a year. These profiles are typically derived from time-series data with resolutions ranging from one minute to one hour, being applied at various scales, including individual households, buildings, or even entire neighborhoods [11]. They provide crucial insights into typical consumption behaviors and how electricity use fluctuates over time.

As buildings and homes become increasingly integrated into the energy system, their role in achieving sustainability goals becomes even more significant. The adoption of demand response programs and cost-reflective electricity pricing has gained traction, encouraging consumers to adapt their consumption habits in response to grid conditions and price signals. Understanding how these consumption profiles are identified and classified is essential to support the transition to a more sustainable and intelligent energy system, where buildings and homes actively contribute to energy flexibility and decarbonization efforts.

A. RELATED WORKS

As verified in several studies [12]–[14], these consumption profiles are typically obtained using clustering methods. They serve as the foundation for numerous applications, including load forecasting [6], tariff design [8], and demand-side management [9]. For instance, Satre Meloy et al. [15] started clustering residential electricity consumption using K-means and Hierarchical clustering on cumulative consumption profiles of high temporal resolution (1 second) during evening peak hours. This approach identified two distinct demand patterns, linking them to occupant behaviors like cooking and appliance use. The main part of the methodology used predictive models, such as elastic net logistic regression and random forests, to further connect these clusters to activities, providing actionable insights for improving demand-side flexibility and reducing peak loads. On the other hand, Khan et al. [16] proposed a spatial-temporal framework for short-term electricity consumption forecasting in residential buildings. The model integrates K-means clustering with a hybrid distance metric to segment apartments based on consumption patterns into low, medium, and high clusters. Clustered data is then aggregated to reduce variability and enhance prediction accuracy. The forecasting component combines Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) in a stacked ensemble approach to optimize predictions. The results demonstrate superior performance compared to conventional machine learning (ML) and deep learning (DL) models, highlighting the effectiveness of clustering in improving model generalization and supporting more targeted energy management strategies.

However, most of these studies paid limited attention to the definition and selection of consumption profiles. Given their critical role in understanding consumption patterns and consumer behavior, it is essential to establish robust methodologies for obtaining, characterizing, and visualizing these profiles. With that in mind, Michalakopoulos et al. [17] presented a comprehensive ML framework for clustering residential electricity consumption profiles to optimize DR programs. The authors compared multiple clustering algorithms, including K-means, K-medoids, Hierarchical agglomerative clustering, and DBSCAN, using evaluation metrics such as the silhouette score, Davies–Bouldin index, and Calinski–Harabasz index to determine the optimal number of clusters. The framework also incorporates probabilistic classification using CatBoost and eXplainable Artificial Intelligence (xAI) techniques to enhance the interpretability of cluster assignments.

A critical challenge in this domain is translating raw big data from smart meters into meaningful consumption profiles [18]. Principal Component Analysis (PCA) [19] is often utilized to reduce the dimensionality of the data down to a few orthogonal components, which are then clustered to obtain the profiles. Duarte et al. [20] explored the impact of different data preparation techniques on creating consumption profiles for energy consumption analysis. Using real-world data from a university campus, the authors evaluated standardization, dimensionality reduction, and data enrichment methods. Techniques such as matrix normalization, standardization by rows, and PCA were applied along with K-means clustering to group load curves. Similarly, Bustamante et al. [21] studied individual residential electricity consumption in shared student housing using granular circuit-level data. The authors applied PCA to reduce the dimensionality and uncover key consumption variables, followed by K-means clustering to group residents into four categories based on their median daily consumption. Building on these approaches, Al-Jarrah et al. [22] proposed a multi-layered clustering framework for power consumption profiling in smart grids. This method first performs local clustering on smaller grids and then aggregates the results at a global level, reducing computational complexity and communication overhead while maintaining clustering accuracy. But some limitations remain. Notably, the framework relies solely on K-means, and the authors did not compare their approach with further clustering methods or evaluation metrics, limiting the understanding of whether the observed efficiency gains also lead to the best segmentation of power consumption behaviors. Nonetheless, these three studies highlighted how preprocessing techniques critically affect load analysis and subsequent energy management decisions.

Moreover, recent advances in AI and neural networks,

such as Self-Organizing Maps (SOMs) [23], offer innovative approaches to data clustering and dimensionality reduction [24], [25], as demonstrated by Rajabi et al. [14]. The authors compared five techniques for electrical load pattern segmentation: K-means, fuzzy C-means, Hierarchical clustering, SOMs, and Gaussian Mixture Model (GMM), alongside hybrid and adaptive approaches. The study emphasizes the importance of data preprocessing, including normalization and PCA, to improve clustering outcomes. Using metrics such as the Davies–Bouldin index, silhouette score, and Dunn index, the paper offers insights into the selection of clustering algorithms based on specific data characteristics and use cases, emphasizing the computational intensity of hierarchical and SOM methods.

As evidenced by the previously mentioned studies, despite its potential, SOM has been largely neglected in electricity consumption analysis [26], with prior studies presenting minor real-life applicability [27]. Furthermore, the aggregation and averaging of data frequently result in the loss of essential characteristics of load curves, complicating their interpretation and application due to the extensive amount of data provided by smart meters [28]. Additionally, determining the optimal temporal resolution of data for effective clustering is an ongoing concern, as key features may be lost at higher levels of aggregation, potentially affecting the outcome of the processes [29], [30]. This topic would benefit from further research.

Therefore, addressing these challenges is essential for developing robust methods to identify energy consumption profiles tailored to specific needs [31]. These profiles are vital for understanding consumer habits and behaviors, supporting demand-response programs, implementing profile-based tariffs, and enhancing forecasting models.

B. MAIN CONTRIBUTIONS AND PAPER ORGANIZATION

The present study addresses the critical points previously mentioned by presenting a comparative analysis of various methodologies to derive consumer electricity profiles. It benchmarks these methods using time-series data at different resolutions, aiming to establish a standard for processing large datasets while retaining the granularity needed to identify typical and extreme behaviors. Ultimately, this work seeks to provide a systematic and accessible approach to generating optimal profiles at a low computational cost. In particular, the main contributions can be listed as follows:

- A comprehensive and robust methodology based on unsupervised learning techniques that can be readily applied to various high-dimensional datasets across different regions, particularly when the objective is to identify typical profiles for characterizing household electricity consumption patterns;
- A benchmark analysis of various consumption profiling methods, specifically K-means, PCA with K-means, SOM, and SOM with K-means, to verify which yields the best profiles for different time resolutions (including 1-min, 5-min, 15-min, and 1-hour granularity). To achieve this, we present scores, namely the silhouette coefficient and Davies-Bouldin index, and perform a Fast Fourier Transform (FFT) study on the time-series data to find the most suitable time unit of analysis;
- Present insightful visual representations of the profiles to allow easy comparison of results for future studies, providing empirical input data for demand response programs and forecasting models.

The paper is organized as follows. Section II presents the proposed methodologies, along with a description of the dataset, dimensionality reduction, and clustering methods. Additionally, the evaluation scores are also described. Section III performs a detailed exposition of the obtained results, summarizing and commenting on the main findings. Finally, Section IV contains the conclusion and possible future work.

II. METHODOLOGY

The overview of the methodological approach is illustrated in Fig. 1, which represents the research flowchart of this work. A description of the datasets' characteristics is done in Section II-A. The data preprocessing steps are explained in Section II-B. Section II-C describes the main characteristics of the consumption profiling methods defined, and Section II-D presents the selected cluster validation techniques.

A. DATA DESCRIPTION AND ANALYSIS

Cluster analysis cannot be conducted without the availability of a dataset. Therefore, it is essential to use a suitable household energy consumption dataset. Kang et al. [4] provide an outstanding review of building electricity use profile models, offering the community a structured and carefully selected list of open datasets that can be used to foster data-driven research in this field.

This paper had access to private electricity consumption datasets from eighteen households from various municipalities in mainland Portugal, one house per city. Each dataset presents different periods of consumption data measurement, provided in power (W), all in a 1-minute resolution. Furthermore, it is worth noting that all homes belong to diverse social and economic backgrounds with differing family sizes and periods of measurement. This heterogeneity in the data enhances the robustness of the proposed methods, ensuring their effectiveness when applied to additional datasets. House 97 has the longest duration of data collection at 519 days, while house 91 has the shortest at 249 days, with at least one recorded instance of consumption data. Additionally, the datasets include information on the presence of photovoltaic systems for self-consumption, with five houses reporting this feature. Table 1 presents a summary of the datasets' characteristics.

B. STAGE 1: DATA PREPROCESSING

According to earlier research [15], data cleaning and preprocessing are two key processes in obtaining interpretable results from cluster analysis.

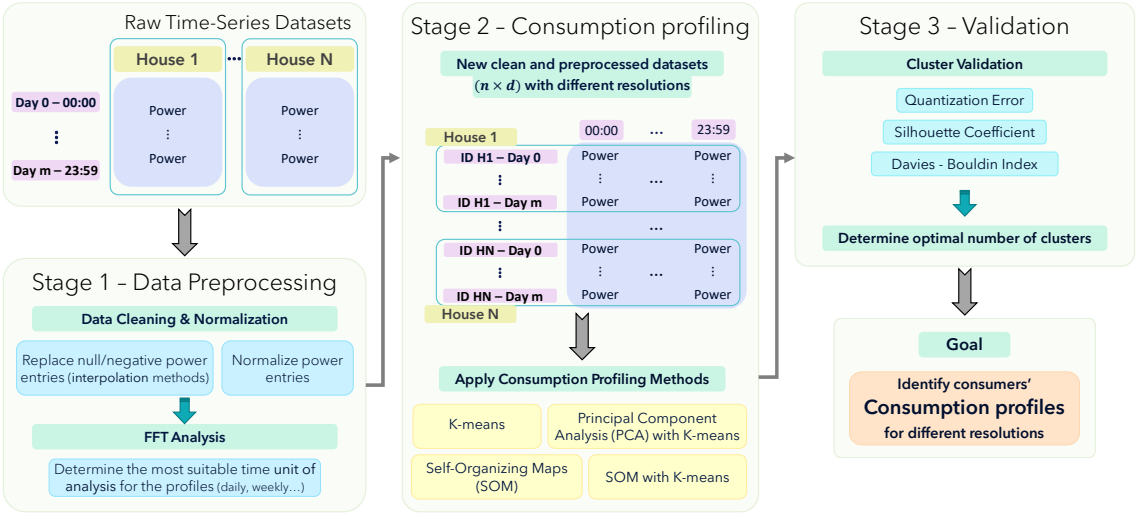


FIGURE 1. Overview of methodological approach.

1) Data Cleaning and Normalizing

Some entries in the dataset may lack power information or not include all consumption days. Interpolation using nearby entries can replace these absent values, assuming similar consumption patterns between nearby instants. When dealing with time-series data, removing specific rows that have missing entries may not be a practical solution, as this can result in gaps in daily, weekly, or seasonal profiles that do not accurately represent actual consumption behavior. Instead, it is advisable to consider the unit of analysis for the profiles and remove an entire day or week. This approach helps to avoid gaps and preserves the integrity of the data curves.

Some entries may also contain inaccurate information, such as abnormal consumption between periods of very low activity. These points, known as outliers, should be handled and eliminated using, for instance, thresholds for data removal and interpolation methods for replacement.

One of the most crucial steps in data analysis is normalizing the data before applying any dimensionality reduction or clustering methods, as these algorithms are sensitive to the scale of the data. Consequently, each value should range from 0 to 1 to ensure that each entry contributes equally to the distance calculation between data points, helping to improve accuracy [32].

2) Fast Fourier Transform Analysis

The Fast Fourier Transform (FFT) [33] is a mathematical algorithm used to convert data in the time domain (such as time-series signals) into the frequency domain, where it is represented as a combination of sinusoidal components with specific frequencies, amplitudes, and phases. It computes the Discrete Fourier transform (DFT) of a signal and its inverse (IDFT), with the most well-known FFT method being the Cooley-Tukey [34]. In the context of energy consumption data, FFT can help identify dominant periodic patterns, such as daily, weekly, or seasonal cycles, by revealing the frequencies at which the data exhibit the most significant variations [35]. This makes FFT a valuable tool for determining the optimal study period for consumption profiles, as it allows researchers to focus on the natural cycles of consumption behavior [36]. By identifying these dominant frequencies, FFT can help tailor the analysis to the intrinsic temporal structure of the data, ensuring that dimensionality reduction and profiling methods capture the most relevant and recurring patterns.

C. STAGE 2: CONSUMPTION PROFILING METHODS

Time-series data frequently belongs to the domain of big data, particularly when dealing with high-resolution data collected from smart meters. Consequently, it is increasingly important to consider this factor when defining and identifying consumption profiles. To this end, the present study introduces four distinct methodologies: **K-means**, **PCA with K-means**, **SOM**, and **SOM with K-means**.

K-means (for clustering) frequently appears in applications related to household consumption profiles, as mentioned in Section I-A. It serves as a basis for comparison with the remaining proposed methods. PCA (for dimensionality reduction) is also frequently employed in similar studies; therefore, we combined it with K-means in a hybrid approach. Furthermore, the review studies conducted by Aghabozorgi et al. [37] and Kang et al. [4] highlight that SOMs (for clustering and dimensionality reduction) have enormous potential, although rarely explored in this context.

In this study, by combining different approaches, the objective is to identify the most suitable techniques for big time-series data and how data granularity impacts the final profiles. Consequently, it is crucial to provide a brief introduction to these methods.

1) K-means Clustering

Cluster analysis (often known as **clustering**) corresponds to the general problem of partitioning a dataset into natural subgroups called clusters [38]. Objects within the same group should be as similar as possible (based on a similarity measure), while objects between different groups should be as dissimilar as possible.

Various methods for distinct strategies have been developed, with K-means [39] being one of the most widely applied due to its ease of use and broad applicability to diverse datasets. In the K-means algorithm, each cluster is characterized by a point (or multiple points in the case of time-series curves) known as the centroid.

For a dataset with n features, each centroid c_k can be represented as

$$c_k = [c_{k1}, c_{k2}, \ldots, c_{kn}]. \tag{1}$$

The algorithm assigns labels to the data based on its distance to these centroids, allocating each input entry to the nearest one. This process is iterative, with centroids being updated in each iteration. The most commonly employed similarity measure is the Euclidean distance [40]. However, in the context of time-series data, dynamic time wrapping (DTW) [41] emerges as a valuable alternative, as it can identify similar curves independent of the time steps, speed, or length.

2) Principal Component Analysis

Principal Component Analysis (PCA) [19] is a statistical technique employed for dimensionality reduction. It allows complex datasets to be summarized by identifying their most significant patterns of variation. Specifically, the principle of PCA is to convert a set of variables that may be correlated into a set of linearly independent features through orthogonal transforms called **principal components**. PCA transforms the original data into a new coordinate system, where the dimensions (principal components) are ranked based on the amount of variance they explain [38].

Given the input data $D \in R^{n \times d}$, the algorithm first centers it by subtracting the mean from each point $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$, resulting in matrix Z. Next, it computes the eigenvectors $U = (u_1\, u_2 \ldots u_d)$ and eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_d)$ of the covariance matrix $\sum$, defined by

$$\sum = \frac{1}{n}(Z^T Z). \tag{2}$$

Then, given the desired variance threshold α, PCA selects the smallest set of dimensions r that capture at least α fraction of the total variance (normally $\alpha = 0.9$ or higher). Finally, it computes the coordinates of each point in the new r-dimensional principal component subspace, to generate the new data matrix $A \in R^{n \times r}$

$$A = \{a_i \,|\, a_i = U_r^T x_i, \text{for } i = 1, \ldots, n\}. \tag{3}$$

In the context of energy consumption time-series analyses, PCA is particularly valuable for identifying key consumption patterns and reducing redundancy in high-dimensional datasets [42]. By isolating the most influential components (power values at seconds, minutes, or hours, according to the granularity), the method enables a focus on the underlying trends and behaviors of energy consumption profiles. This approach not only simplifies the analysis but also enhances the interpretability of clustering or profiling methods by ensuring that the most relevant features of the data are preserved [19].

3) Self-Organizing Maps

As proposed by Kohonen [23], a Self-Organization Map (SOM) consists of a single-layer unsupervised artificial neural network designed for information compression and clustering. SOMs project high-dimensional data onto a lower-dimensional grid, typically with two dimensions, preserving the topological relationships within the data. This means that similar points are mapped close to each other, thus forming clusters represented by each neuron.

The method begins by initializing the neurons' weights with random values or samples from the input data and defining the grid size. Then the training process consists of three steps [43]:

- **Competition**, when each neuron competes to capture the current input point. All neurons calculate a similarity distance measure (typically Euclidean distance), winning the most similar neuron to the input point (best matching unit, BMU);
- **Cooperation**, when the BMU identifies the topological neighborhood of excited neurons, according to a radius;
- **Adaptation**, when the BMU and the neighbors adapt their weights according to the current training epoch and learning rate.

The training is also characterized by the smoothing step: the neighborhood radius and the learning weight decrease with each iteration.

SOMs are particularly effective for identifying complex consumption behaviors that may not be evident through traditional clustering methods. They can handle non-linear relationships in the data and provide an intuitive visual representation of the clusters, making it easier to interpret and analyze the results. By using SOMs, researchers can uncover typical and atypical consumption profiles, aiding in tasks such as demand forecasting, tariff design, and load management.

D. STAGE 3: CLUSTERING VALIDATION TECHNIQUES

Since working with unsupervised methods, no ground truth is available; therefore, internal validation should be used to quantify the performance of the methods [38] and select the ideal number of clusters and SOM grid size. Three internal validation metrics, Quantization Error [44], silhouette coefficient [45], and Davies-Bouldin index [46] can be employed,

being the most suitable for this type of analysis based on the studies reviewed in Section I-A.

1) Quantization Error

The **Quantization Error** (QE) [44] is a critical aspect of SOMs. It is a statistical measure of variance and is associated with the final synaptic weight of the neurons after learning.

The QE mathematically expresses the squared distance (usually the average Euclidean distance) between input data x and their corresponding winning neurons. Thus, the QE reflects the average distance between each data point (x_i) and its BMU, defined as

$$QE = \frac{1}{n}\sum_{i=1}^{n} \| x_i - BMU_i \|, \tag{4}$$

with n as the number of samples. It plays an important role in assessing the quality of SOMs on the same dataset, particularly when determining the best size of the topological grid or the optimal model parameters, such as the learning rate and neighborhood radius. Generally, smaller QE values indicate better results.

2) Silhouette Coefficient

For each point x_i, the silhouette coefficient is

$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}, \tag{5}$$

where $\mu_{out}^{min}(x_i)$ is the mean of the distances from x_i to points in the closest cluster, and $\mu_{in}(x_i)$ is the mean distance from x_i to points in its own cluster. The total **silhouette coefficient** [45] is defined as the mean s_i value across all points, given by (6), where a value close to $+1$ indicates good clustering.

$$SC = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{6}$$

3) Davies-Bouldin Index

The Davies-Bouldin measure for a pair of clusters C_i and C_j is defined as

$$DB_{ij} = \frac{\sigma_{\mu_i} + \sigma_{\mu_j}}{\delta(\mu_i, \mu_j)}, \tag{7}$$

where μ_i denote the centroid of cluster i, $\sigma_{\mu_i} = \sqrt{var(C_i)}$ represents the dispersion of the points around the respective centroid (square root of the total variance) and $\delta(\mu_i, \mu_j)$ is the distance between the centroids.

The **Davies-Bouldin index** [46], for k clusters, is thus defined as

$$DB = \frac{1}{k} \cdot \sum_{i=1}^{k} \max_{i \neq j}\{DB_{ij}\}, \tag{8}$$

meaning that for each cluster C_i it is chosen the cluster C_j that returns the largest DB_{ij} ratio. Therefore, smaller DB values mean better clustering (clusters are well separated and each one is well represented by its centroid).

III. APPLICATION OF CLUSTERING METHODS IN CONSUMPTION PROFILES

This section describes the evaluation of different approaches to obtaining house consumption profiles. It includes two main subsections, focusing on the data preprocessing steps and the presentation of results from the analyzed datasets. In each subsection, the fundamental aspects of the applied methodology are critically and concisely discussed.

The code was written in Python using the Google Colab platform, and the *scikit-learn* library [47] for the preprocessing, PCA, clustering, and evaluation methods, and the SOM was implemented using MiniSOM [48]. Most parameters were left at default, while those modified are justified throughout the text. Specifically, we employed the 'full' SVD solver for PCA, with the variance threshold α defined as *n_components* = 0.9. For SOM, a learning rate of 0.5 and a neighborhood radius of 1.1 were selected, as these modifications yielded superior overall results for all methods.

A. DATA CLEANING AND NORMALIZATION

The initial stage in obtaining household consumption profiles is data preprocessing, as illustrated in Fig. 1. Each house includes a power consumption dataset, making it necessary to create a single dataframe comprising all the houses. The datasets were normalized using the MinMaxScaler approach, ensuring consumption values between 0 and 1. When working with consumption curves, the MinMaxScaler is preferable to the StandardScaler since it does not modify the shape of the curves [32]. The latter assumes a Gaussian distribution for the data, an invalid option for uncertain data such as residential consumption.

Furthermore, regarding missing data, all days with more than eight hours of missing data were excluded, corresponding to 137 days distributed across all houses. These days typically involved blackouts in the electrical grid, which caused the smart meters to turn off. Subsequently, on the remaining days with incomplete data, the backward fill (*bfill*) and forward fill (*ffill*) methods of the *Pandas* library were employed to replace NaN values by propagating the following or previous valid observation for the same instant. For instance, if there is a missing value on 12/12/2022 at 17:00:00, the method will try to replace that value with a valid observation from the day immediately before or after. Consequently, one ensures a complete and cleaned dataframe, with dimensions 7633×1440. Table 1 presents a summary of the cleaned dataset's characteristics, along with the original datasets.

B. FFT ANALYSIS AND FINAL PIVOT DATAFRAMES

With the data cleaned and organized, all that remains is to identify the best unit of analysis for the consumption profiles. By applying the FFT, it was clear that there is a daily pattern in the electricity consumption of the houses. Fig. 2 illustrates the outcome of the FFT application for the example of house 97 (highest number of days). With

TABLE 1. Summary of the main characteristics of the original and cleaned datasets.

		Before cleaning			After cleaning		
House ID	**With PV**	**Start Date**	**End Date**	**# of days**	**Start Date**	**End Date**	**# of days**
27		03/12/2021 15:31	31/03/2023 23:58	484	04/12/2021 00:00	31/03/2023 23:59	479
29	X	27/01/2022 12:28	31/03/2023 23:58	429	28/01/2022 00:00	31/03/2023 23:59	420
31		06/12/2021 13:13	31/03/2023 23:58	481	07/12/2021 00:00	31/03/2023 23:59	474
37	X	09/02/2022 11:47	31/03/2023 23:58	416	12/02/2022 00:00	31/03/2023 23:59	413
41	X	11/01/2022 16:17	31/03/2023 23:58	445	12/01/2022 00:00	31/03/2023 23:59	421
64		03/02/2022 10:59	17/11/2022 11:21	288	04/02/2022 00:00	16/11/2022 23:59	286
67		27/01/2022 14:33	31/03/2023 23:58	429	28/01/2022 00:00	31/03/2023 23:59	428
71		26/11/2021 10:00	31/03/2023 23:58	491	26/11/2021 00:00	31/03/2023 23:59	490
76		03/03/2022 15:16	31/03/2023 23:58	394	04/03/2022 00:00	31/03/2023 23:59	393
86	X	21/01/2022 09:59	31/03/2023 23:58	435	21/01/2022 00:00	31/03/2023 23:59	434
91		25/11/2021 12:40	31/07/2022 23:58	249	26/11/2021 00:00	31/07/2022 23:59	247
92	X	13/01/2022 13:49	31/03/2023 23:58	443	14/01/2022 00:00	31/03/2023 23:59	438
93		16/02/2022 10:56	31/03/2023 23:58	409	17/02/2022 00:00	31/03/2023 23:59	408
94		02/12/2021 15:31	31/03/2023 23:58	485	03/12/2021 00:00	31/03/2023 23:59	483
96		22/11/2021 16:29	19/02/2023 13:44	454	23/11/2021 00:00	18/02/2023 23:59	405
97		29/10/2021 16:18	31/03/2023 23:58	519	30/10/2021 00:00	31/03/2023 23:59	516
100		13/01/2022 17:43	31/03/2023 23:58	443	14/01/2022 00:00	31/03/2023 23:59	439
105		11/12/2021 10:44	31/03/2023 23:58	476	12/12/2021 00:00	31/03/2023 23:59	459

a period of approximately 1440 minutes for all houses, the fundamental frequency of the data corresponds to 24 hours. Consequently, the consumption data were arranged in a pivot table, where the columns represent the minutes of the day (00:00 − 23:59), and each row contains the house ID and the corresponding date of consumption data (recall Fig. 1).

After formatting the data according to the required configuration for the various methods, it was possible to create different dataframes with varying resolutions. For this study, we tested data with granularity of 1 minute, 5 minutes, 15 minutes, and 1 hour, resulting from averaging several intervals corresponding to these specified resolutions.

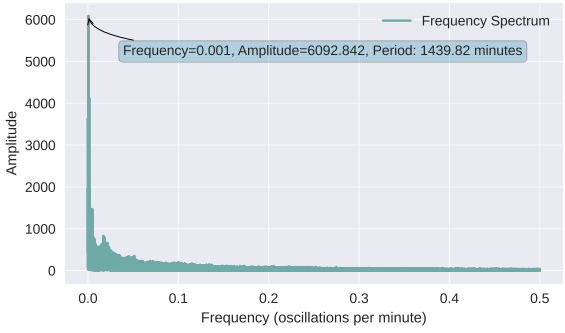


FIGURE 2. Spectral analysis of power consumption for House 97.

C. CONSUMPTION PROFILING RESULTS

Four approaches were considered for obtaining electricity consumption profiles: K-means, PCA combined with K-means, SOM, and SOM combined with K-means. The research aims to address two phases. First, we want to determine which method produces the best profiles for the original data resolution of 1 minute, comparing them with the standard K-means method, the most employed algorithm (remember Section I-A). After having well-defined results, the second phase of the investigation assesses the impact of data resolution on the behavior of the profiling methods.

This section is divided into six parts. The first four subsections present and explain each approach, followed by a critical comparison of the results based on the scores and profiles obtained. It concludes with a comprehensive analysis of the clusters corresponding to the best method.

1) K-means Clustering Results

As the most widely used method in the literature for obtaining consumption profiles, this approach involves applying the K-means technique to the pivot table of the daily consumption data. The method requires defining the number of clusters, k, in advance. To determine this, the elbow method [49] offered an initial suggestion for the most appropriate number of clusters for the data: between 8 and 12. Selecting only 6 clusters resulted in different behaviors being combined into the same clusters, leading to the loss of significant profiles. On the other hand, choosing 14 clusters did not reveal new profiles; instead, it merely resulted in the division of existing clusters, leading to redundant profiles concerning consumption behavior.

Additionally, we performed an iterative analysis using the silhouette and Davies-Bouldin metrics to validate the subjective insights gained from the elbow method. This analysis indicated that the optimal number of clusters within the 8-12 range would be $k = \{8, 10, 12\}$. These values served as the basis for the remaining approaches, described next.

Furthermore, the initialization of the cluster centroid was left at default (*k-means++*), which uses sampling based on an empirical probability distribution of the points [47]). Table 2 summarizes the optimal scores obtained for the K-means consumption profiling approach.

2) PCA with K-means Results

This approach involves applying PCA to the pivot table to reduce the dimensionality of the data. It is followed by the application of K-means to the newly reduced dataset while varying the number of clusters between $k = \{8, 10, 12\}$ according to the case study.

Starting with PCA, it was first necessary to determine the optimum number of components. By defining the variance retention threshold α as 0.9, the analysis determined the number of principal components to be 157. This means that the first 157 components effectively capture 90% of the variability in the data, significantly reducing its complexity. The pivot table decreased from 7633×1440 to 7633×157 after PCA, while preserving the essential patterns and structures for clustering. Table 2 reveals a summary of the optimal scores obtained for the PCA + K-means consumption profiling method.

3) SOM Results

The third technique includes applying a SOM as the primary source for obtaining clusters. It is necessary to define specific parameters to implement this method, such as the grid size, learning rate, and neighborhood radius. According to K-means results, it is clear that the SOM grid should be composed of 8, 10, or 12 neurons to enable comparison between methods. Therefore, $[2 \times 4]$, $[2 \times 5]$, and $[4 \times 3]$ grids were used to obtain the 8, 10, and 12 clusters, respectively.

A comprehensive preliminary analysis was performed on various grid sizes to identify the optimal learning rate and neighborhood radius values. An iterative tuning process determined that a learning rate of 0.5 and a neighborhood radius of 1.1 provided the best outcomes. These parameters resulted in the highest silhouette score and the lowest Davies-Bouldin index and quantization error. Furthermore, the initial weights were derived as random samples from the dataset rather than as random points (aiming to align with the initialization process of K-means), and trained for 10000 epochs. Table 2 reveals a summary of the optimal scores obtained for the SOM consumption profiling method.

4) SOM with K-means Results

Motivated by the SOM's capacity to reduce dimensionality (similar to the aim of PCA), this approach seeks to test whether the SOM can solely reduce dimensionality, while K-means focuses on identifying consumption profiles.

To determine the optimal number of neurons to utilize, a square grid was chosen to ensure an equal spatial distribution of the units. The rule $M \approx 5 \times \sqrt{n}$ was employed, where M represents the number of units and n indicates the number of samples tested. This policy is referenced commonly in the literature [50], [51]. Given that there are 7633 days, this calculation yields approximately 437 neurons, corresponding to a grid of roughly $[21 \times 21]$ units.

Additionally, a new iterative analysis was conducted to identify the preferred learning rate and neighborhood radius.

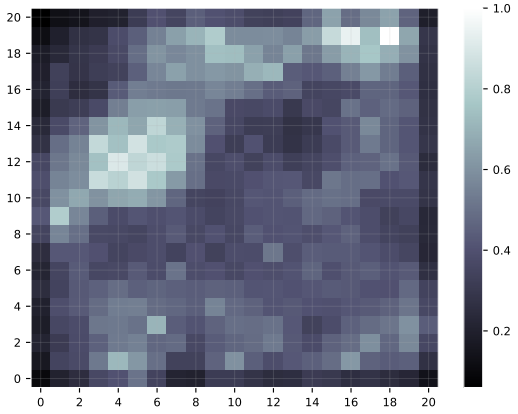


FIGURE 3. Unified distance matrix for the SOM + K-means approach, comprising [21x21] units.

This analysis yielded values of 0.5 and 1.1, respectively, identical to those obtained with the SOM approach (section III-C3). This outcome was anticipated, given that the dataset tested is the same. Table 2 reveals a summary of the optimal scores obtained for the SOM + K-means consumption profiling method.

Fig. 3 presents the Unified distance matrix (U-matrix) [24] of the $[21 \times 21]$ grid, where each cell represents a node of the SOM. The shading intensity corresponds to the distances between the weight vectors of neighboring nodes. Darker areas indicate smaller distances (regions that should represent clusters), and lighter areas reveal larger distances (indicating boundaries between clusters).

5) Comparison of Methods

Table 2 summarizes the obtained scores for the benchmark analysis of 1-minute resolution consumption data, comparing the performance of the profiling approaches K-means, PCA combined with K-means, SOM, and SOM combined with K-means across different metrics and cluster sizes. The metrics considered include the Silhouette Score, Davies-Bouldin Score, Computational Cost (in seconds), and Quantization Error (for SOM-based methods).

Analyzing Table 2, the results demonstrate a clear improvement in clustering quality when transitioning from basic K-means to PCA + K-means, SOM, and finally to SOM + K-means. Across all cluster sizes, SOM combined with K-means achieves the highest silhouette scores, indicating the best-defined and most well-separated clusters. For example, with 8 clusters, the silhouette score increases from 0.1081 (K-means) to 0.1690 (SOM + K-means), corresponding to a 56.3% improvement. Simultaneously, the Davies-Bouldin score decreases from 2.7413 to 2.6611, revealing that the clusters are both more compact and better separated, with a relative reduction of 2.9%.

While not achieving the same peak performance as the SOM + K-means, PCA + K-means provides a compelling trade-off between clustering quality and computational effi-

TABLE 2. Comparison of metrics for each approach, considering 1-minute resolution.

$k = 8$ **Clusters**	**K-means**	**PCA + K-means**	**SOM [2x4]**	**SOM [21x21] + K-means**
Silhouette Score	0.1081	0.1288	0.1406	0.1690
Davies-Bouldin Score	2.7413	2.7080	2.9248	2.6611
Computational Cost (s)	2.6245	0.3256	2.1712	49.4151
Quantization Error	-	-	3.6502	2.7652
$k = 10$ **Clusters**	**K-means**	**PCA + K-means**	**SOM [2x5]**	**SOM [21x21] + K-means**
Silhouette Score	0.1067	0.1172	0.1275	0.1423
Davies-Bouldin Score	2.6438	2.6651	2.9579	2.5433
Computational Cost (s)	1.9483	0.3706	2.4599	45.5021
Quantization Error	-	-	3.6145	2.7652
$k = 12$ **Clusters**	**K-means**	**PCA + K-means**	**SOM [4x3]**	**SOM [21x21] + K-means**
Silhouette Score	0.1043	0.1149	0.1182	0.1376
Davies-Bouldin Score	2.7850	2.6569	2.9880	2.6682
Computational Cost (s)	1.9126	0.5772	2.8392	43.9700
Quantization Error	-	-	3.5299	2.7652

ciency. The application of PCA reduces the dataset's dimensionality, significantly accelerating the clustering process. For example, with 8 clusters, PCA + K-means achieves a reasonable silhouette score of 0.1288 at a fraction of the computational cost (0.3256 seconds) compared to SOM + K-means (49.4151 seconds). This efficiency makes PCA + K-means particularly suitable for scenarios where computational resources are limited. SOM approach demonstrates mixed results. While it performs better than the basic K-means, it struggles to match the cluster quality of SOM + K-means, as evidenced by its higher Davies-Bouldin scores and quantization errors. For example, with 12 clusters, SOM achieves a Davies-Bouldin score of 2.9880, compared to 2.6682 for SOM + K-means. The higher quantization errors of SOM further emphasize the advantage of refining the initial clustering using K-means.

Nevertheless, it is crucial not only to consider the scores but also to compare the resulting profiles to confirm if the identified clusters align with the expected trends. A comprehensive examination revealed that the most meaningful profiles correspond to $k = 12$. Fig. 4, Fig. 5, Fig. 6, and Fig. 7 exhibit the distribution of daily consumption profiles across the identified clusters using percentile plots, for the K-means, PCA + K-means, SOM, and SOM + K-means approaches, respectively. Each subplot corresponds to a distinct cluster, with the x-axis representing the time of day (in hours) and the y-axis representing the normalized power consumption. The shaded regions depict the data's spread within each cluster, with the light blue area encompassing 80% of the data (between the 10th and 90th percentiles) and the darker teal area highlighting the interquartile range (IQR, between the 25th and 75th percentiles). The solid dark blue line represents the median (50th percentile) power consumption profile. The number of days assigned to each cluster is indicated in the top left corner of each subplot.

Table 3 presents the correspondence between the various profiles identified through the different methodologies.

TABLE 3. Cluster correspondence between the different consumption profiling approaches.

Cluster Description	**K-means**	**PCA + K-means**	**SOM**	**SOM + K-means**
Lunch Time Peak Consumption	1	1	1	1
Bimodal Early Morning and Evening Peaks	2	2	2	2
Evening and Nighttime High Consumption	3	3	3	3
Moderate Daytime Use with Dinner Time Peak	4	4	4	4
Flat Consumption Over 24 Hours	5	5	5	5
Nighttime Consumption Peaks	6	6	-	6
High Consumption During Working Hours	7	-	7	7
High Pre-Dawn Usage with Evening Peak	8	8	8	8
Consistent All-Day Use with Evening Peak	9	9	9	9
Overnight Peaks with Reduced Evening Use	10	10	10	10
Afternoon Peak Usage	-	-	-	11
Low Consumption with Late Evening Peak	12	12	12	12
Daytime and Late Evening High Consumption	-	7	6	-
Morning-to-Midday Peak Usage	11	11	11	-

Observing Fig. 7, one can see that the profiles display distinct consumption patterns over the 24 hours. For instance, some clusters have harsh evening peaks (e.g., clusters 4 and 12), while others exhibit flatter, more consistent consumption (e.g., clusters 5 and 9). However, despite the good distinction between clusters, there is an uneven distribution of days between the profiles: cluster 5 includes 3713 days, and cluster 6 only 68 days. This may indicate dominant consumption patterns in the dataset, with the method capable of grouping the most different consumptions into smaller clusters. On the other hand, PCA + K-means (Fig. 5) distributes the days more evenly: the most populous profile, cluster 5, has 2705 days, with the least typical cluster

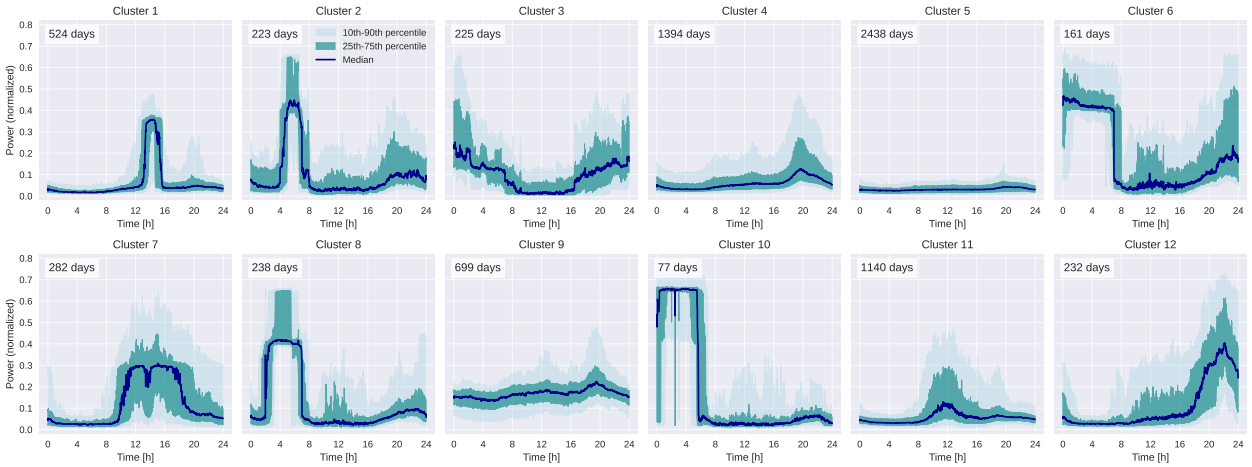


FIGURE 4. Results for 12 clusters, 1-min resolution using K-means approach.

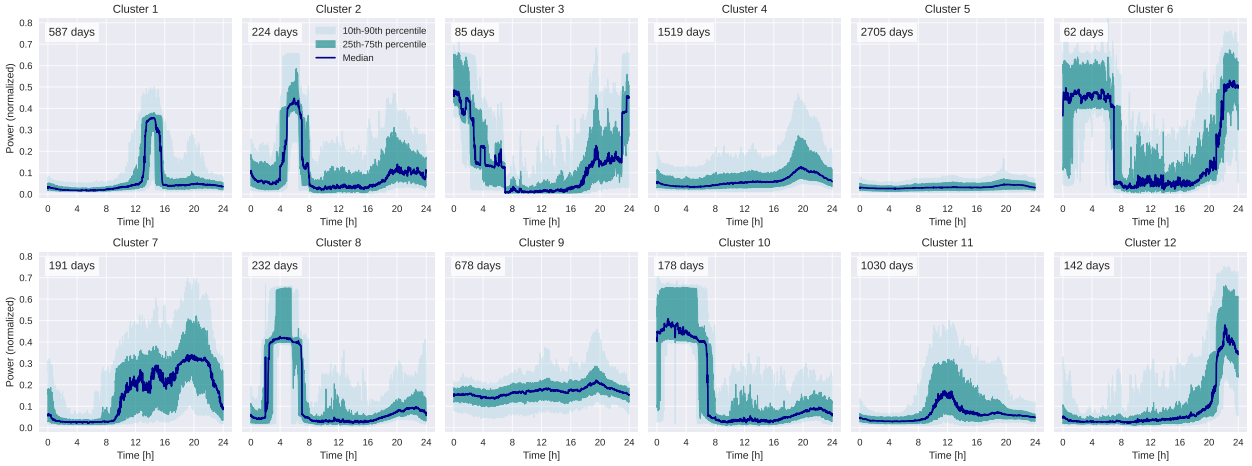


FIGURE 5. Results for 12 clusters, 1-min resolution using PCA + K-means approach.

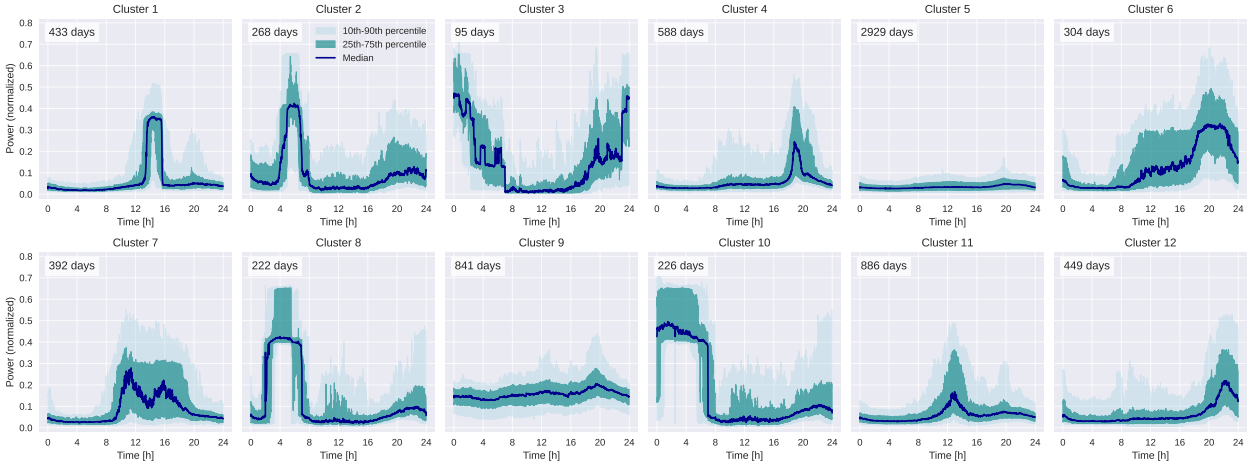


FIGURE 6. Results for 12 clusters, 1-min resolution using SOM approach.

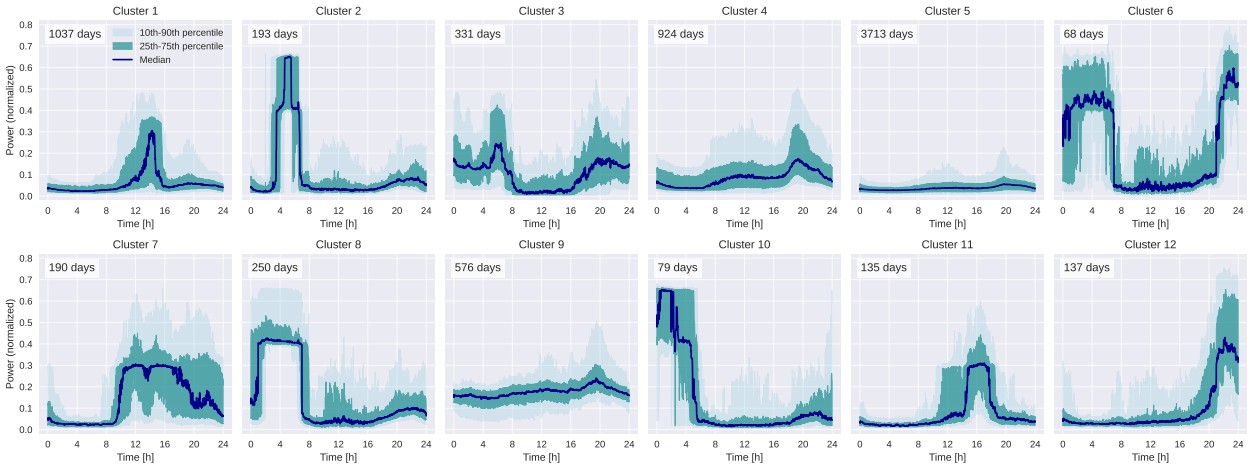


FIGURE 7. Results for 12 clusters, 1-min resolution using SOM + K-means approach.

having 62 days. Even so, the method discovered cluster 11, characterized by morning-to-midday peak usage, not present in SOM + K-means. K-means also grouped this cluster (cluster 11 in Fig. 4), similar to SOM (cluster 11 in Fig. 6). In addition, SOM and PCA + K-means also identified a medium-sized profile typical of daytime and late evening high consumption, which the remaining methods failed to achieve. Conversely, only SOM + K-means created a typical afternoon consumption cluster (refer to Table 3).

In general, SOM + K-means produces more diverse clusters, managing to group less frequent consumption patterns into distinct clusters. This behavior is likely due to combining the SOM's topology-preserving mapping and K-means' refinement. SOM captures some unique patterns but struggles to provide the same level of cluster separation as SOM + K-means. By contrast, K-means provides a less refined distribution of clusters, with a more consistent distribution of curves over the clusters. Nonetheless, the difference with PCA + K-means is not as noticeable as it might be in practice. The profiles found are virtually identical, revealing that PCA + K-means is an extremely valuable option for obtaining faster results with big data, yet obtaining better scores.

6) Analysis of the Identified Consumption Profiles

Since the SOM + K-means method achieved the best scores and captured the most significant behavior among the methods tested, it is now important to briefly analyze the characteristics of the consumption profiles obtained by this method. As mentioned above, each cluster contains several days from different households. Fig. 8 illustrates the distribution of the clusters among the houses under study, while Fig. 9 presents a bar plot displaying the monthly distribution of each identified cluster.

A visual examination of the data reveals the prevalence of specific profiles across the analyzed families, such as cluster 5. This profile accurately represents the consumption

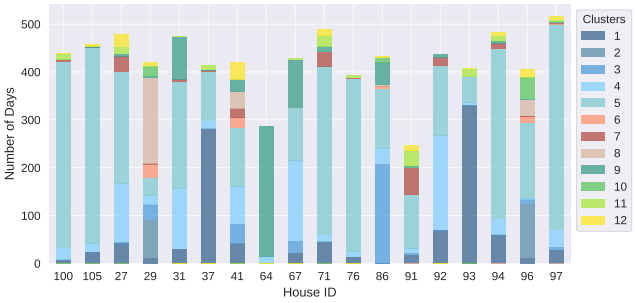


FIGURE 8. Distribution of days per house in each daily SOM + K-means consumption cluster.

patterns of houses 100, 105, 76, and 97, exhibiting relatively constant electricity usage throughout the day and night. This suggests energy-efficient routines or low occupancy, making these families ideal candidates for load shifting or automated flexibility programs, as their stable demand enables more accurate forecasting and control. Cluster 9 also demonstrates a consistent pattern but holds a higher average consumption. This profile includes a subtle peak in power usage during dinner time and primarily represents house 64.

Cluster 1 is another profile frequently observed across various households. In particular, houses 37 and 93 exhibit a notable predominance of days within this cluster, representing their most typical profile. Characterized by a distinct midday consumption peak, reaching a minimum at night, this pattern reflects typical lunchtime behavior, which could benefit from time-of-use tariffs that incentivize pre-heating or appliance usage during off-peak hours. Interestingly, this profile emerges in all households except 86, indicating that midday activity is widespread but not universal.

On the other hand, cluster 8 depicts the opposite behavior, with sustained high consumption during the night and lower daytime usage, with occasional peaks. This pattern is

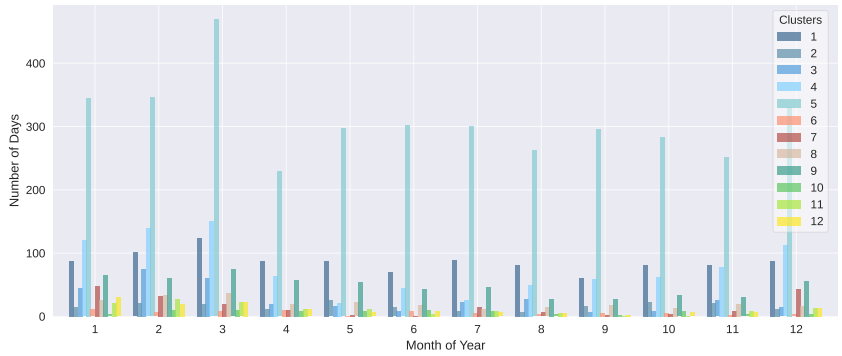


FIGURE 9. Monthly distribution for each daily SOM + K-means consumption cluster.

unusual and may indicate nighttime appliance use, charging of electric vehicles (EVs), or inefficient loads such as heating/cooling systems running overnight. House 29, which exhibits this behavior noticeably, could particularly benefit from incentive-based EV charging schemes, helping to shift demand to periods of excess renewable generation, such as during wind curtailment or in response to solar production declines [52]. Profiles 6 and 10 exhibit similar patterns, although they are less common and unpredictable.

It is also worth noting that some houses exhibit greater homogeneity in their consumption patterns than others. For instance, house 64 can be characterized by a relatively limited set of three profiles, indicating a more predictable and stable electricity demand. In contrast, house 96 displays a much more diverse range of twelve distinct clusters, reflecting highly variable daily patterns. As a result, this household may require adaptive or real-time management strategies, while house 64 could effectively use static management approaches due to its stable electricity usage, saving computational resources.

Additionally, one can identify different monthly consumption patterns among the various clusters in Fig. 9. Notably, profiles 4, 9, and 12 demonstrate a preference for the winter months, in contrast to clusters 1 and 5, which are distributed evenly throughout the year without significant fluctuations. Overall, the autumn and winter months (November to April) exhibit a consistent distribution across multiple clusters, while fewer clusters characterize the spring and summer months (from May to October).

To get a complete overview of the consumption behavior, Fig. 10 illustrates a lattice heatmap of the monthly distribution of daily electricity usage profiles across the different households. Each cell reveals how many days in a month a household's consumption pattern matched a specific cluster. The intensity of the colors indicates how frequently each cluster occurred, with darker colors representing a higher frequency of days associated with that cluster, confirming the distinct patterns among the various profiles.

Fig. 10 easily reveals that house 27 shifts its consumption between clusters 4 and 5, regardless of the month of the year, similarly to house 37, which alternates more frequently between clusters 1 and 5. On the other hand, house 41's behavior is distributed across several clusters throughout the year, with the winter months dominated by clusters 3 and 4, while cluster 5 is most typical in May, September, and October. This insight is relevant for optimizing energy systems and supporting sustainable practices by enabling demand response strategies, dynamic pricing models, or even forecasting consumption to improve energy production and distribution planning throughout the year [53].

Moreover, the identification of distinct and recurring consumption patterns can support the development and tuning of energy management algorithms, particularly in applications involving home energy storage. For instance, these typical daily household consumption profiles could guide battery control strategies to enhance self-consumption, reduce peak loads, and support grid services by determining optimal charging and discharging schedules [54].

D. IMPACT OF THE RESOLUTION ON THE CONSUMPTION PROFILES

Following the analysis of consumption profiles at a 1-minute resolution, the present section examines how the scores and clusters fluctuate in response to changes in data granularity, maintaining the four approaches considered previously. Table 4 summarizes the scores according to the 5-minute, 15-minute, and 1-hour resolutions for 12 clusters. For the PCA + K-means approach, the selected number of components was 70, 35, and 12, respectively.

Looking at the 5-minute resolution results in Table 4 and comparing them with those from the 1-minute resolution in Table 2, one can see improvements in the silhouette score of 7.9% for K-means, 11.66% for PCA + K-means, 9.2% for SOM, and 12.9% for SOM + K-means. The Davies-Bouldin score also decreases, with this reduction being more noticeable in K-means and PCA + K-means. This trend persists as the resolution decreases, with SOM + K-means reaching a silhouette score of 0.1981 and a Davies-Bouldin index of 1.8712 at 1-hour resolution. Concerning the computational cost, PCA + K-means consistently proves to be the

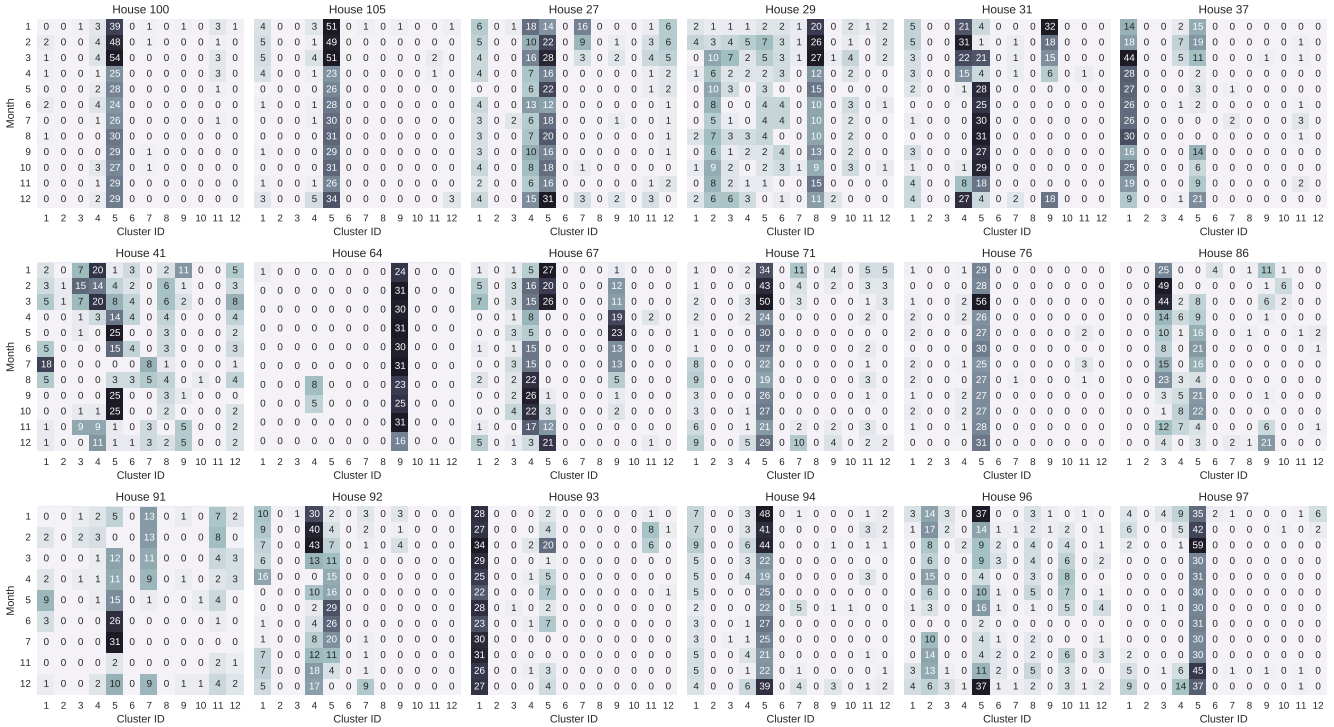


FIGURE 10. Lattice heatmap of the monthly distribution of daily SOM + K-means electricity consumption profiles across the different houses.

fastest approach across all resolutions due to the reduction in dimensionality, while SOM + K-means incurs a significant computational burden, being repeatedly the most expensive method. Despite this, SOM + K-means obtains the lowest quantization error at all resolutions, reinforcing its ability to capture finer details in the data compared to the SOM approach.

However, this improvement in scores may not reflect a better grouping process but rather a consequence of data smoothing. As temporal resolution decreases, short-term fluctuations and noise are attenuated, leading to similar curves and simplifying the clustering problem. At higher resolutions where more detail is preserved, such as 1 minute, clustering becomes more challenging, resulting in lower scores but potentially more meaningful distinctions between patterns.

In addition to the metrics evaluation, a systematic analysis of the clustering outputs across resolutions is needed to reveal how temporal granularity shapes the practical relevance of the resulting profiles. Fig. 11 illustrates how data resolution affects the behavior of cluster 4 (one of the most identifiable profiles) obtained using the SOM + K-means approach for 12 clusters. Similarly, Fig. 12 depicts the same outcome for cluster 3, one of the most asymmetrical consumption profiles.

Observing Fig. 11 and Fig. 12, one can verify that, at finer resolutions (e.g., 1 minute), the clusters capture detailed behavioral signatures such as short appliance usage, cooking peaks, or EV charging events. These nuances are critical for demand response or anomaly detection applications, which require high-precision control [9]. In contrast, these specific features are smoothed at coarser resolutions (e.g., 1 hour), resulting in broader and more uniform profiles that may be more suitable for tariff design or long-term load forecasting [6]. For instance, at the 1-minute resolution, cluster 3 reveals a set of peak patterns in the late evening and night that is lost at the 1-hour resolution.

Ultimately, this observation highlights the importance of aligning temporal resolution with the intended application, as the higher resolution may offer greater behavioral insight at the cost of lower clustering compactness and higher computational costs.

E. CASE STUDY: LISBON METROPOLITAN AREA

To further validate the generalizability of the proposed methodology, an additional analysis was conducted focusing exclusively on a subset of households located in the Lisbon Metropolitan Area. This region, being the most densely populated in Portugal, features diverse behavior and electricity consumption patterns, making it a suitable test case for evaluating the robustness of the clustering approach across different urban dynamics.

From the original dataset, consisting of eighteen households from various municipalities across mainland Portugal, we selected those located in the Lisbon region, corresponding to ten families. The same data preprocessing and

TABLE 4. Comparison of metrics for each approach, considering varying resolutions, for $k = 12$ clusters.

5-minute resolution	**K-means**	**PCA + K-means**	**SOM [4x3]**	**SOM [21x21] + K-means**
Silhouette Score	0.1126	0.1283	0.1291	0.1553
Davies-Bouldin Score	2.6133	2.4750	2.9940	2.5979
Computational Cost (s)	0.9834	0.1869	1.2286	16.8648
Quantization Error	-	-	1.4424	1.0978
15-minute resolution	**K-means**	**PCA + K-means**	**SOM [4x3]**	**SOM [21x21] + K-means**
Silhouette Score	0.1214	0.1499	0.1421	0.1596
Davies-Bouldin Score	2.3792	2.2667	2.4195	2.4306
Computational Cost (s)	0.2529	0.1816	0.6015	6.5450
Quantization Error	-	-	0.7698	0.5413
1-hour resolution	**K-means**	**PCA + K-means**	**SOM [4x3]**	**SOM [21x21] + K-means**
Silhouette Score	0.1455	0.1795	0.1671	0.1981
Davies-Bouldin Score	1.8968	1.8769	2.0480	1.8712
Computational Cost (s)	0.1702	0.0759	0.8714	3.4504
Quantization Error	-	-	0.2985	0.1827

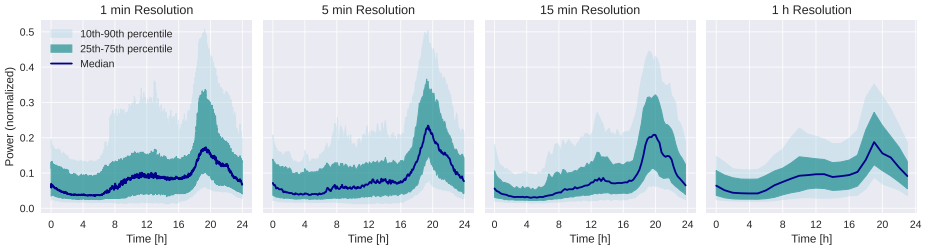


FIGURE 11. Comparison of SOM + K-means cluster 4 profiles across different temporal resolutions (1 min, 5 min, 15 min, and 1 h), for $k = 12$.

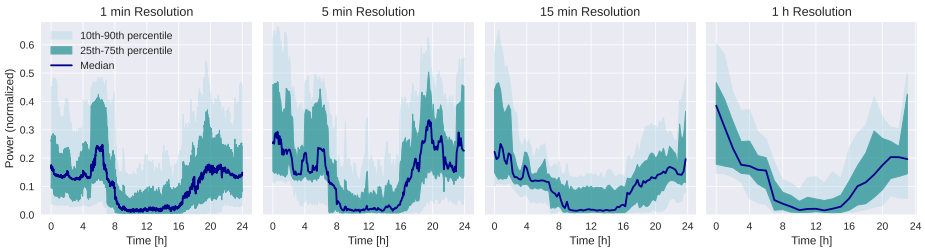


FIGURE 12. Comparison of SOM + K-means cluster 3 profiles across different temporal resolutions (1 min, 5 min, 15 min, and 1 h), for $k = 12$.

clustering pipeline described in Section II was applied to this smaller, region-specific group of houses. For this study, the SOM + K-means method was chosen due to the better overall results confirmed in Section III-C. The ideal number of clusters obtained was six profiles, represented in Fig. 13, with a silhouette score of 0.1761, a Davies-Bouldin score of 1.9762, a computational cost of 83.7322 seconds, and a quantization error of 2.4385.

Despite the reduction in sample size, the methodology successfully identified consistent and interpretable consumption profiles comparable to the results obtained when using all households' datasets. Specifically, the clusters identified are virtually identical to some of those found previously (recall Fig. 7), as expected.

This analysis demonstrates that the methodology is not limited to a specific geographic or demographic context and can be applied to any group of households. It also serves as a practical example of how the method can be tailored for local or municipal-level planning and decision-making, which is particularly relevant for Distribution System Operators (DSOs), urban planners, and energy analysts aiming to understand localized energy consumption behavior.

IV. CONCLUSIONS AND FUTURE WORK

Energy consumption profiles are a critical component of energy management strategies, particularly with the recent

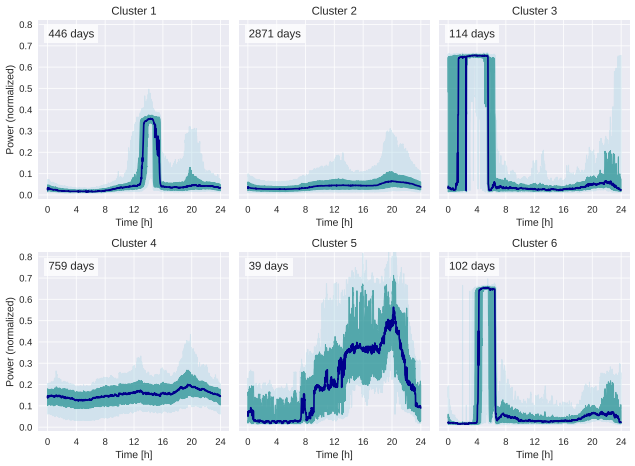


FIGURE 13. Representative daily consumption profiles for Lisbon Metropolitan Area households, using SOM + K-means approach.

widespread adoption of smart meters, which has created a pressing need to develop robust methodologies for obtaining, characterizing, and visualizing these profiles from big data. This study evaluated multiple clustering methods, including K-means, PCA + K-means, SOM, and SOM + K-means, using data resolutions ranging from 1 minute to 1 hour. By analyzing cluster profiles, quality metrics, and computational costs, we demonstrated that temporal resolution has a significant impact on clustering results, the efficiency of the process, and the interpretability of the findings. These results highlight the importance of aligning temporal resolution with the specific goals of energy data analysis and management.

Among the methods studied, hybrid approaches like SOM combined with K-means consistently delivered the best clustering quality across all tested scenarios, achieving the highest silhouette scores and lowest quantization errors. This makes SOM + K-means particularly well-suited for identifying nuanced patterns in high-resolution data. However, its significantly higher computational cost may limit its practical application in large datasets or resource-constrained scenarios. Conversely, PCA combined with K-means offers a good balance between performance and computational efficiency, effectively handling large datasets while maintaining meaningful profiles.

By compressing the data to its most essential features, PCA facilitated a more efficient and computationally feasible clustering process. The reduced dataset maintained the ability to reveal meaningful patterns, enabling K-means clustering to focus on the most relevant features. This dimensionality reduction is particularly critical in high-dimensional datasets, as it mitigates issues such as overfitting and the curse of dimensionality.

Using finer resolutions, such as 1-minute data, helps capture detailed consumption patterns and variability but comes with higher computational costs and lower scores due to the added noise and short-term fluctuations. On the other hand, lower resolutions, like 1 hour, flattened these fluctuations, resulting in improved clustering metrics such as the silhouette coefficient and Davies-Bouldin index. However, these improvements often reflect the loss of fine-grained details rather than inherently better clustering quality. The analysis of cluster 3 and cluster 4 across different granularities further demonstrated the impact of temporal resolution on clustering results. At 1-minute resolution, subtle variations in energy use were preserved, providing detailed insights into consumption patterns. Lower resolutions, such as 15-minute and 1-hour data, revealed more generic profiles, reducing variability within clusters and simplifying the interpretation of results. These findings emphasize the importance of tailoring temporal resolution to the specific goals of the study: finer granularity for detailed, short-term analyses and coarser resolutions for broader, long-term insights. Understanding these transformations provides a clearer view of the trade-offs between preserving detail and achieving computational efficiency.

In summary, this study highlights the critical role of temporal resolution in shaping clustering outcomes, computational requirements, and practical applications in energy data analysis. The results demonstrate that hybrid methods, particularly those combining PCA and SOM, effectively improve clustering performance and model generalization, making them valuable tools for large-scale time-series data applications.

Future research should explore the proposed methodology in this paper and seek to enhance it with new approaches, validation metrics, or even parameters. The SOM could improve by selecting more adapted initial weights for the neurons. Several studies in other domains have explored this topic [55], and applying similar methods to energy consumption data could bring significant benefits. Another aspect would be to create dynamic approaches that adjust the resolution based on the complexity of the data, leveraging the strengths of fine and coarse resolutions, thus reducing the impact on real use cases. Furthermore, exploring the generalizability of clustering models across different regions or periods remains an open challenge, especially nowadays, with consumption patterns changing due to electrification, energy efficiency measures, or the growing adoption of EVs [52]. Finally, integrating external factors, such as climate or photovoltaic production, could also improve the interpretability and value of the clustering results, allowing for more targeted energy management strategies, accurate forecasts, and sustainable energy solutions.

REFERENCES

[1] I. E. Agency, "World Energy Outlook 2024 – Analysis," Oct. 2024. [Online]. Available: https://www.iea.org/reports/world-energy-outlook-2024

[2] Z. He, J. Khazaei, and J. D. Freihaut, "Optimal integration of Vehicle to Building (V2B) and Building to Vehicle (B2V) technologies for commercial buildings," Sustainable Energy, Grids and Networks, vol. 32, p. 100921, Dec. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352467722001667

[3] B. Völker, A. Reinhardt, A. Faustine, and L. Pereira, "Watt's up at Home? Smart Meter Data Analytics from a Consumer-Centric Perspective," Energies, vol. 14, no. 3, p. 719, Jan. 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/3/719
[4] X. Kang, J. An, and D. Yan, "A systematic review of building electricity use profile models," Energy and Buildings, vol. 281, p. 112753, Feb. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778822009240
[5] S. Pelekis, A. Pipergias, E. Karakolis, S. Mouzakitis, F. Santori, M. Ghoreishi, and D. Askounis, "Targeted demand response for flexible energy communities using clustering techniques," Sustainable Energy, Grids and Networks, vol. 36, p. 101134, Dec. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S235246772300142X
[6] J. Amantegui, H. Morais, and L. Pereira, "Benchmark of Electricity Consumption Forecasting Methodologies Applied to Industrial Kitchens," Buildings, vol. 12, no. 12, p. 2231, Dec. 2022. [Online]. Available: https://www.mdpi.com/2075-5309/12/12/2231
[7] A. M. Tzortzis, S. Pelekis, E. Spiliotis, E. Karakolis, S. Mouzakitis, J. Psarras, and D. Askounis, "Transfer Learning for Day-Ahead Load Forecasting: A Case Study on European National Electricity Demand Time Series," Mathematics, vol. 12, no. 1, p. 19, Dec. 2023. [Online]. Available: https://www.mdpi.com/2227-7390/12/1/19
[8] H. C. Jeong and S.-K. Joo, "Personalized Electricity Tariff Recommendation Method for Residential Customers Lacking Historical Metering Data Incorporating Customer Profiles and Behavioral Changes," IEEE Access, vol. 12, pp. 73 426–73 435, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10520307/
[9] C. P. Guzmán, A. Lekidis, P. Pediaditis, P. M. Carvalho, and H. Morais, "Intelligent Participation of Electric Vehicles in Demand Response Programs," in 2023 International Conference on Smart Energy Systems and Technologies (SEST). Mugla, Turkiye: IEEE, Sep. 2023, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10257505/
[10] G. Kormpakis, A. Lekidis, E. Sarmas, I. Papias, F. Serepas, G. Stravodimos, and V. Marinakis, "Home Energy Management Systems: Challenges, Heterogeneity & Integration Architecture Towards A Smart City Ecosystem," in 2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC). Funchal, Portugal: IEEE, Jun. 2024, pp. 1–10. [Online]. Available: https://ieeexplore.ieee.org/document/10794289/
[11] N. Çankaya, "Deriving Power Consumption Models From Energy Bills for Optimal Sizing of Hybrid Power in Commercial Buildings," IEEE Access, vol. 12, pp. 115 042–115 054, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10637379/
[12] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," Applied Energy, vol. 141, pp. 190–199, Mar. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261914012963
[13] K. Li, Z. Ma, D. Robinson, and J. Ma, "Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering," Applied Energy, vol. 231, pp. 331–342, Dec. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261918313606
[14] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," Renewable and Sustainable Energy Reviews, vol. 120, p. 109628, Mar. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032119308354
[15] A. Satre-Meloy, M. Diakonova, and P. Grünewald, "Cluster analysis and prediction of residential peak demand profiles using occupant activity data," Applied Energy, vol. 260, p. 114246, Feb. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261919319336
[16] A.-N. Khan, N. Iqbal, A. Rizwan, R. Ahmad, and D.-H. Kim, "An Ensemble Energy Consumption Forecasting Model Based on Spatial-Temporal Clustering Analysis in Residential Buildings," Energies, vol. 14, no. 11, p. 3020, May 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/11/3020
[17] V. Michalakopoulos, E. Sarmas, I. Papias, P. Skaloumpakas, V. Marinakis, and H. Doukas, "A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs," Applied Energy, vol. 361, p. 122943, May 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S030626192400326X
[18] A. Ushakova and S. Jankin Mikhaylov, "Big data to the rescue? Challenges in analysing granular household electricity consumption in the United Kingdom," Energy Research & Social Science, vol. 64, p. 101428, Jun. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214629620300050
[19] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1-3, pp. 37–52, Aug. 1987. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0169743987800849
[20] O. G. Duarte, J. A. Rosero, and M. D. C. Pegalajar, "Data Preparation and Visualization of Electricity Consumption for Load Profiling," Energies, vol. 15, no. 20, p. 7557, Oct. 2022. [Online]. Available: https://www.mdpi.com/1996-1073/15/20/7557
[21] C. Bustamante, S. Bird, L. Legault, and S. E. Powers, "Energy Hogs and Misers: Magnitude and Variability of Individuals' Household Electricity Consumption," Sustainability, vol. 15, no. 5, p. 4171, Feb. 2023. [Online]. Available: https://www.mdpi.com/2071-1050/15/5/4171
[22] O. Y. Al-Jarrah, Y. Al-Hammadi, P. D. Yoo, and S. Muhaidat, "Multi-Layered Clustering for Power Consumption Profiling in Smart Grids," IEEE Access, vol. 5, pp. 18 459–18 468, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7947198/
[23] T. Kohonen, "The self-organizing map," Proceedings of the IEEE, vol. 78, no. 9, pp. 1464–1480, Sep. 1990. [Online]. Available: http://ieeexplore.ieee.org/document/58325/
[24] B. Brentan, G. Meirelles, E. Luvizotto, and J. Izquierdo, "Hybrid SOM+k-Means clustering to improve planning, operation and management in water distribution systems," Environmental Modelling & Software, vol. 106, pp. 77–88, Aug. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364815217301834
[25] A. Abdelaziz, V. Santos, M. S. Dias, and A. N. Mahmoud, "A hybrid model of self-organizing map and deep learning with genetic algorithm for managing energy consumption in public buildings," Journal of Cleaner Production, vol. 434, p. 140040, Jan. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959652623041987
[26] F. S. Villar, P. H. J. Nardelli, A. Narayanan, R. C. Moioli, H. Azzini, and L. C. P. Da Silva, "Noninvasive Detection of Appliance Utilization Patterns in Residential Electricity Demand," Energies, vol. 14, no. 6, p. 1563, Mar. 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/6/1563
[27] K. Jayanth Krishnan and K. Mitra, "A modified Kohonen map algorithm for clustering time series data," Expert Systems with Applications, vol. 201, p. 117249, Sep. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S095741742200625X
[28] S. Mitra, B. Chakraborty, and P. Mitra, "Smart meter data analytics applications for secure, reliable and robust grid system: Survey and future directions," Energy, vol. 289, p. 129920, Feb. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544223033145
[29] J. C. Hernandez, F. Sanchez-Sutil, A. Cano-Ortega, and C. R. Baier, "Influence of Data Sampling Frequency on Household Consumption Load Profile Features: A Case Study in Spain," Sensors, vol. 20, no. 21, p. 6034, Oct. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/21/6034
[30] W. Khan, J. Y. Liao, S. Walker, and W. Zeiler, "Impact assessment of varied data granularities from commercial buildings on exploration and learning mechanism," Applied Energy, vol. 319, p. 119281, Aug. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261922006389
[31] C. Guo, S. Ci, Y. Zhou, and Y. Yang, "A Survey of Energy Consumption Measurement in Embedded Systems," IEEE Access, vol. 9, pp. 60 516–60 530, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9406578/
[32] L. B. De Amorim, G. D. Cavalcanti, and R. M. Cruz, "The choice of scaling technique matters for classification performance," Applied Soft Computing, vol. 133, p. 109924, Jan. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494622009735
[33] M. T. Heideman, D. H. Johnson, and C. S. Burrus, "Gauss and the history of the fast fourier transform," Archive for History of Exact Sciences, vol. 34, no. 3, pp. 265–277, 1985. [Online]. Available: http://www.jstor.org/stable/41133773
[34] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," Mathematics of Computation, vol. 19, no. 90, pp. 297–301, 1965. [Online]. Available: https://www.ams.org/mcom/1965-19-090/S0025-5718-1965-0178586-1/
[35] H. Musbah, M. El-Hawary, and H. Aly, "Identifying Seasonality in Time Series by Applying Fast Fourier Transform," in 2019 IEEE Electrical Power and Energy Conference (EPEC). Montreal, QC, Canada: IEEE, Oct. 2019, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9074776/

[36] J. M. Abreu, F. Câmara Pereira, and P. Ferrão, "Using pattern recognition to identify habitual behavior in residential electricity consumption," Energy and Buildings, vol. 49, pp. 479–487, Jun. 2012. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778812001363

[37] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – A decade review," Information Systems, vol. 53, pp. 16–38, Oct. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306437915000733

[38] M. J. Zaki and W. Meira, Jr, Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd ed. Cambridge University Press, 2020.

[39] L. M. L. Cam and J. Neyman, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification. University of California, 1967, google-Books-ID: IC4Ku_7dBFUC.

[40] A. Singh, A. Yadav, and A. Rana, "K-means with three different distance metrics," International Journal of Computer Applications, vol. 67, no. 10, 2013.

[41] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, ser. AAAIWS'94. AAAI Press, 1994, p. 359–370. [Online]. Available: https://dl.acm.org/doi/10.5555/3000850.3000887

[42] L. Wen, K. Zhou, and S. Yang, "A shape-based clustering method for pattern recognition of residential electricity consumption," Journal of Cleaner Production, vol. 212, pp. 475–488, Mar. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959652618337697

[43] D. Miljkovic, "Brief review of self-organizing maps," in 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). Opatija, Croatia: IEEE, May 2017, pp. 1061–1066. [Online]. Available: http://ieeexplore.ieee.org/document/7973581/

[44] B. Dresp, J. M. Wandeto, and H. O. Nyongesa, "Using the quantization error from Self-Organizing Map (SOM) output for fast detection of critical variations in image time series," Des Données à la Décision - From Data to Decisions , vol. 18-2, no. 1, 2018. [Online]. Available: https://hal.science/hal-02182882

[45] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, Nov. 1987. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0377042787901257

[46] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979. [Online]. Available: http://ieeexplore.ieee.org/document/4766909/

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[48] G. Vettigli, "Minisom: minimalistic and numpy-based implementation of the self organizing map," 2018. [Online]. Available: https://github.com/JustGlowing/minisom/

[49] T. Kodinariya and P. Makwana, "Review on determining of cluster in k-means clustering," International Journal of Advance Research in Computer Science and Management Studies, vol. 1, pp. 90–95, 01 2013.

[50] J. Tian, M. H. Azarian, and M. Pecht, "Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm," PHM Society European Conference, vol. 2, no. 1, Jul. 2014. [Online]. Available: https://papers.phmsociety.org/index.php/phme/article/view/1554

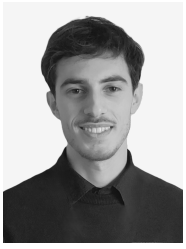
[51] S. Licen, A. Astel, and S. Tsakovski, "Self-organizing map algorithm for assessing spatial and temporal patterns of pollutants in environmental compartments: A review," Science of The Total Environment, vol. 878, p. 163084, Jun. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0048969723017035

[52] M. Forte, C. P. Guzman, A. Lekidis, and H. Morais, "Clustering Methodologies for Flexibility Characterization of Electric Vehicles Supply Equipment," Green Energy and Intelligent Transportation, p. 100304, Mar. 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2773153725000544

[53] A. Hernandez, R. Nieto, L. De Diego-Oton, M. C. Perez-Rubio, J. M. Villadangos-Carrizo, D. Pizarro, and J. Urena, "Detection of Anomalies in Daily Activities Using Data from Smart Meters," Sensors, vol. 24, no. 2, p. 515, Jan. 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/2/515

[54] A. Maturo, C. Vallianos, A. Buonomano, A. Athienitis, and B. Delcroix, "Clustering-driven design and predictive control of hybrid PV-battery storage systems for demand response in energy communities," Renewable Energy, p. 123390, May 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960148125010523

[55] V. P. Murugesan and Punniyamoorthy M., "Development of a New Means to Improve the Performance of Self-Organizing Maps:," International Journal of Data Analytics, vol. 3, no. 1, pp. 1–16, Aug. 2022. [Online]. Available: https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJDA.307065



MARCELO FORTE completed the MSc in Electrical and Computer Engineering in 2023 by Universidade de Lisboa - Instituto Superior Técnico, and the BSc in Electrical and Computer Engineering in 2021 also by Universidade de Lisboa - Instituto Superior Técnico. He has received 4 awards and/or honours for his academic journey. He is a Researcher at INESC-ID: Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa since 2023, participating in multiple Horizon Europe funded projects. Currently, he is pursuing a Ph.D. degree, working in the areas of Engineering and Technology with an emphasis in Data Science, Machine Learning, Electric Vehicles, and Renewable Energy.

CINDY P. GUZMAN received her B.S. degree in electrical engineering from the Universidad de la Costa, Barranquilla, Colombia, in 2011, and her M.Sc. and Ph.D. degrees in electrical engineering from São Paulo State University (UNESP), Ilha Solteira, Brazil, in 2015 and 2021, respectively. She worked as a Postdoctoral Researcher with the School of Electrical and Computer Engineering, University of Campinas (UNICAMP) from August 2021 to November 2022, contributing to projects focused on sustainable microgrids and electromobility. Since November 2022, she have been working as a postdoctoral researcher at INESC-ID Lisboa in the R&D projects: EV4EU - Electric Vehicle Management for Carbon Neutrality and AHEAD - (AI-informed Holistic Electric Vehicles Integration Approaches for Distribution Grids) developing strategies for optimal EV charging station placement, grid resource management, and sustainable mobility solutions.

LUCAS PEREIRA received his Ph.D. in Computer Science from the University of Madeira, Portugal, in 2016. Since then, he has been an Assistant Researcher at ITI/LARSyS of Instituto Superior Técnico - University of Lisbon, leading the Further Energy and Environment Research Laboratory (FEELab). His research applies data science, machine learning, and human-computer interaction to develop practical solutions for future energy systems and sustainable built environments. His work is characterized by the real-world deployment and evaluation of monitoring technologies and software systems, ensuring practical impact in energy and sustainability domains.

HUGO MORAIS is a senior researcher at the Portuguese R&D Institute INESC-ID and an Associate Professor at Instituto Superior Técnico, University of Lisbon. With a background in Electrical Engineering (specializing in Power Systems) and a Ph.D. in Electrical and Computer Engineering (2012), he has contributed to over 40 academic and industrial research projects, ranging from national to French and European-funded initiatives. At INESC-ID he is the coordinator of three ongoing Horizon Europe projects and is actively involved in the other four. His research primarily focuses on advancing smart grid technologies and tools. He has been actively involved in the development of strategies for electric mobility integration in power systems. Hugo Morais (IEEE Senior Member, 2018) has received 26 awards and authored over 250 scientific papers, including more than 90 published in international peer-reviewed journals, with more than 8,000 citations. He also serves as an editor for several leading scientific journals, including Energies, Electricity Journal, and Frontiers in Energy Research.