Energy Reports 14 (2025) 1207–1221

Contents lists available at ScienceDirect

Energy Reports

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

Research paper

Whale optimization-based clustering for buildings' energy consumption profiles

Viorica Rozina Chifu, Tudor Cioara *, Ionela Danci, Cristina Bianca Pop, Ionut Anghel

Computer Science Department, Technical University of Cluj-Napoca, Cluj-Napoca, Romania

ARTICLE INFO

Keywords:
Energy peaks
Energy profiles clustering
Energy valley
Magnitude
Variation
Whale optimization

ABSTRACT

The smart grids require advanced building demand-side management solutions to balance variable renewable production. Energy consumption profile segmentation can increase the efficiency of the processes, enabling grid operators to provide more personalized requests to different customer needs and behaviors increasing engagement and participant satisfaction. However, the complexity of clustering energy profiles is amplified by the diversity of consumer behavior and the need to adapt segmentation over time as consumption patterns evolve. Establishing the number of consumer energy profile clusters in advance is challenging, as most clustering algorithms are sensitive to initial parameter settings, which can affect their performance. In this paper, we propose a new clustering method based on the Whale Optimization Algorithm (WOA), which addresses the problem of dynamicity and variability of energy data. The whale individual is defined as a set of active centroids described by a compressive set of features extracted based on peak and valley periods focused on magnitude, variation, and efficiency. To promote compact and well-separated clusters of energy profiles the Calinski-Harabasz index was used as a fitness function. Population initialization is performed using K-Means++ algorithm which ensures a diversified initial distribution of solutions in the solution space. The evolution of solutions is ensured through the mechanisms of the WOA algorithm, which allow the gradual updating of candidate solutions to accelerate convergence and produces high-quality solutions. The efficiency of the method was evaluated on a real dataset of daily energy consumption of buildings on a university campus. The experimental results show that the WOA method outperforms the selected state-of-the-art methods for comparison. WOA obtained the lowest Davies–Bouldin index (0.527), the highest Silhouette score (0.484) and the lowest Ball–Hall coefficient (1.851), indicating superior segmentation in terms of both internal cohesion and inter-cluster separability.

1. Introduction

The integration of renewable energy sources into the smart grid makes the system less predictable and more complex to manage (Khalid, 2024). These sources are characterized by variable and intermittent production driven by weather conditions requiring sophisticated management and coordination solutions for efficient integration. The peak production of renewable energy often does not coincide with peak demand (Khalid, 2024). Furthermore, deploying these sources at the edge of the grid puts local communities and microgrids at risk of blackouts or load shedding. In this context, demand-side management is important for balancing energy production with the demand for increasing grid stability (Panda et al., 2023). The consumers are encouraged to adjust their demand to better match the renewable availability consuming more when production is high and consuming less during periods of low production. To make these programs more effective the grid operators use tools like consumer segmentation to provide more personalized requests to different customer needs and behaviors aiming to increase engagement and participant satisfaction (Hayn et al., 2014). However, with the large-scale deployment of IoT energy metering significant amounts of data at high granularity becomes available on consumers' energy profiles (Wang et al., 2023). Therefore, the grid operators are going one step further by incorporating profile segmentation in demand response programs focusing on identifying energy usage patterns, behavioral tendencies, and other contextual factors based on energy data (Bartusch and Alvehag, 2014). The energy customers can be clustered based on when and how they use energy, allowing for more advanced pattern matching and better identification of flexible

* Corresponding author.
E-mail addresses: viorica.chifu@cs.utcluj.ro (V.R. Chifu), tudor.cioara@cs.utcluj.ro (T. Cioara), danci.vi.ionela@student.utcluj.ro (I. Danci), cristina.pop@cs.utcluj.ro (C.B. Pop), ionut.anghel@cs.utcluj.ro (I. Anghel).

https://doi.org/10.1016/j.egyr.2025.07.034
Received 17 February 2025; Received in revised form 12 June 2025; Accepted 1 July 2025
Available online 23 July 2025

consumers who can easily shift their load in response to match the renewable energy availability.

In general, clustering algorithms are used to segment consumer profiles by grouping them based on identified energy consumption patterns. Various clustering methods have been applied in this context, including partitional, hierarchical, probabilistic, deep learning, and bio-inspired clustering techniques (Rajabi et al., 2020). However, the complexity of consumer segmentation is amplified by the diversity of their behavior and the need to adapt segmentation as consumption patterns evolve. Partitioning algorithms such as K-means (Nystrup et al., 2021; Jeong et al., 2021; Ofetotse et al., 2021; Palaniappan et al., 2024; Henriques et al., 2024; Wen et al., 2024) and K-medoids (Michalakopoulos et al., 2024) are efficient but require a predefined number of clusters, which poses a challenge due to the variability and complexity of energy data. This makes it difficult to determine the number of consumer energy profile clusters in advance. Hierarchical algorithms (Ahir and Chakraborty, 2022; Arias-Requejo et al., 2023) offer greater flexibility by grouping energy profiles in a hierarchy of clusters. However, they become inefficient for large data sets such as those generated by thousands of energy consumers. Probabilistic algorithms, such as Mixture of Gaussians (GMM) (Kaur and Gabrijelčič, 2022), assume a specific data distribution, often Gaussian, which may not adequately reflect the true variability of energy consumption. Additionally, they are sensitive to initial parameter settings, which can impact their performance. Deep learning-based clustering methods (Sun et al., 2019; Eskandarnia et al., 2022; Kumar and Mallipeddi, 2024; Wang et al., 2024), such as autoencoders, can learn complex representations of consumption behavior but require large computational resources and large training time. Hybrid methods (Sandoval Guzmán et al., 2024; Zhang et al., 2022) combine different techniques to handle noise and anomalies, providing more robust solutions, but introduce additional complexity and often come with higher implementation costs due to the integration of multiple algorithms. In this context, bio-inspired metaheuristics have recently been adopted as promising solutions for energy data clustering, either for determining the optimal number of clusters in the partitioning algorithms or for solving the clustering problem modeled as an optimization problem. However, despite the recent advantages there are open research gaps that limit their adoption for energy profile segmentation. Energy load profiles display complex dynamic behavior, including high dimensionality, non-linear tendencies, and non-homogeneous profile data challenging the representation of metaheuristic individuals and fitness function. Many approaches opt for dimensionality reduction, however the number of features considered is rather small and can lead to insufficient data representation failing to capture the data variability and the peaks and valleys dynamics in energy profiles. Moreover, encoding of centroids and individuals is complex as it needs to maintain energy profile variability and handle temporal dynamics for obtaining high-quality clusters. Finally, the quality of the clustering process is influenced by the diversity, and quality of the initial population of individuals so it is important to start with diverse, high-quality individuals.

In this paper, - a WOA-based clustering for energy consumption profiles of buildings capable of handling data variability and temporal dynamicity is proposed. A whale individual is represented as a set of centroids, each with an activation flag indicating whether the cluster associated with that centroid is active. Centroids are described by a compressive set of features extracted based on peak and valley periods focused on magnitude, variation, and efficiency, allowing to capture of variability and dynamics of the energy consumption profiles. Moreover, it eliminates the need to specify a predefined number of clusters enabling their dynamic determination for data at hand. As a fitness function, the Calinski-Harabasz (CH) index is used which measures the cluster's compactness and separability. This choice is justified by its ability to balance both essential criteria for efficient clustering, promoting compact and well-separated clusters of energy profiles. The initial population of individuals is generated with the K-Means++ algorithm, which ensures the diversity and uniform distribution of individuals in the solution space. This approach prevents stagnation, speeds up convergence, and reduces the likelihood of weak clusters forming early, ensuring a well-diversified population. The updating of individuals in the population follows the specific mechanism of the WOA algorithm, which balances the exploitation of existing solutions with the exploration of new solutions, accelerating the convergence to an optimal solution. Moreover, it is calibrated to progressively improve the fitness value, thus ensuring a fast and efficient convergence of the algorithm.

The novel contributions of this work can be summarized as follows:

- Development of a novel WOA-based clustering algorithm applied in the context of segmentation of daily energy consumption profiles. The algorithm simultaneously optimizes the number and position of centroids, eliminating the need to specify the number of clusters in advance, a common limitation of classical methods-
- Defining a new representation of energy profiles by extracting automatically derived features from the energy consumption profile, using moving average filtering to identify peak and valley intervals. These features capture information about the magnitude, median, local variation and duration of energy consumption in peak and valley intervals, providing a more expressive search space adapted to the real dynamics of the energy consumption data.
- Integrating K-Means++ initialization into WOA algorithm, to generate a diversified and competitive starting population, thus accelerating convergence and avoiding stagnation in local minimum, an aspect rarely addressed in the bio-inspired optimization methods applied to clustering.
- Defining a modular architecture, easily extendable to meet the requirements imposed by dynamic contexts, where consumption behaviors change over time (i.e. concept drift). Because the method optimizes the cluster configuration at each execution, without depending on a pre-trained static model, it can be integrated into a sliding window framework, where the clustering process is periodically restarted based on the most recent available data. This feature offers high potential for application in smart energy systems, where continuous adaptation is essential. Comparative evaluation of the WOA method against reference algorithms such as K-Means, DBSCAN, Agglomerative Clustering, Time Series Clustering with Variational Recurrent Auto-encoders (VRAE), Deep Autoencoder with K-Means, using four performance metrics: Davies-Bouldin Index, Dunn Index, Silhouette Score and Ball-Hall Index. Experimental results show that WOA method obtains the best values for most of these metrics, which confirms its superiority in the clustering process of energy consumption profiles compared to the state-of-the-art algorithms included in the analysis.
- Wilcoxon statistical tests, to highlight the superiority of the proposed method compared to state-of-the-art methods
- An error budget analysis, to quantify the impact of the variation of each adjustable parameter on the performance of the algorithm.

The rest of the paper is structured as follows: Section 2 reviews the literature on energy profile clustering approaches; Section 3 presents our daily energy clustering approach; Section 4 describes the evaluation results, while Section 5 discusses the performance of WOA-based clustering in terms of global and local fitness evolution and population diversity as well as the algorithm computational overhead. Section 6 concludes the paper and presents future work.

2. Related work

The clustering methods for energy consumption profiles presented in this section have been classified according to the methodology employed. Each methodology is suited for different types of load profile data and clustering objectives (see Table 1).

Partitioning-based solutions divide a set of load profiles into a

Table 1
Clustering Methods Used for Energy Profile Segmentation.

Clustering type	Algorithms	Features	Advantages	Disadvantages
Partitioning	K-Means	Energy data (Nystrup et al., 2021; Jeong et al., 2021; Ofetotse et al., 2021; Palaniappan et al., 2024; Henriques et al., 2024), temporal and frequency statistics (Nystrup et al., 2021), extreme data points (Jeong et al., 2021), home configuration (Ofetotse et al., 2021)	Fast and efficient for large energy datasets	Predefined no. of clusters, sensitive to initial centroids, noise, and outliers.
	Fuzzy C-Means	Socio-demographic information (Wen et al., 2024)	Multiple cluster membership, less sensitive to noise and outliers	Predefined no. of clusters, slow convergence, sensitivity to initial centroids, computationally expensive
Hierarchical Clustering	Agglomerative	Energy data (Michalakopoulos et al., 2024; Ahir and Chakraborty, 2022; Arias-Requejo et al., 2023), daily energy average and peaks (Michalakopoulos et al., 2024), weather variables (Arias-Requejo et al., 2023)	Does not require a no. of clusters	Computationally expensive, poor scalability
Probabilistic Clustering	Gaussian mixture model	Daily average, standard deviation, seasonal scores (Kaur and Gabrijelčič, 2022)	capture data complexity, increased flexibility	Sensitive to initializations, computationally expensive
Deep Learning Clustering	Autoencoders	Energy data (Sun et al., 2019; Eskandarnia et al., 2022; Kumar and Mallipeddi, 2024; Wang et al., 2024)	Detects complex patterns, automates feature extraction	Requires big data, computationally expensive
Hybrid Clustering	DBSCAN	Weather variables (Sandoval Guzmán et al., 2024)	High-quality clustering	Increased complexity
	DLDA and AP	Peaks (Zhang et al., 2022)		
	Association rules and K-Medoid	Temporal (Funde et al., 2019)		
	Markov model and adaptive K-Mean	Density peaks (Wang et al., 2016)		
Bio-inspired Clustering	PSO	Energy data (Cuevas et al., 2019)	Less sensitive to noise and outliers; capture complex patterns	Influenced by the algorithm's adjustable parameters, predefined no. of cluster
	Our Solution	*Energy peaks and valleys*	*Dynamically adjusted no. of clusters, high-quality clustering*	*Influenced by the algorithm's adjustable parameters*

predefined number of clusters based on similarity, ensuring that each profile is assigned to exactly one cluster. In most cases, K-Means or its variants are used with different types of features. However, a major disadvantage of the K-Means algorithm is that it is sensitive to the initial choice of cluster centers, which may lead to convergence to suboptimal solutions, and it requires prior specification of the number of clusters. Nystrup et al. (2021) apply the K-Means to cluster load profiles based on wavelet coefficients. Even if the algorithm is fast and efficient, it is sensitive to variability in load profiles, particularly temporal or seasonal fluctuations. Jeong et al. (2021) consider extreme points in load profiles, selected based on demographic characteristics of residential customers, such as floor area, number of household members, income, age, and level of education. Although promising, this approach has certain limitations as it can miss subtle variations in load profiles caused by unusual events. In addition, the use of a narrow set of demographic factors may reduce the accuracy of the results, therefore lifestyle or household composition could improve the clustering quality. Ofetotse et al (Ofetotse et al., 2021). combine K-Means analysis with feature selection based on the silhouette score for households' segmentation based on their energy consumption. The analysis identifies four distinct groups of households, differentiated by housing type, energy consumption, and number of household appliances providing a detailed understanding of consumption patterns for effective energy-saving strategies and demand management. K-Means with different classifiers are used by Palaniappan et al. (2024) for classifying energy consumption patterns showing promising results for linear regression and support vector machine. Henriques et al. (2024) use different clustering methods, including K-Means, agglomerative hierarchical clustering, and self-organizing maps, to discover distinct patterns of energy consumption. Principal component analysis is applied to reduce the data dimensionality, and the silhouette score assesses the quality of the clusters. The results indicate that K-Means provides the most accurate results, grouping households according to low, medium, and high consumption. Wen et al., (2024) use Robust-learning Fuzzy C-Means to cluster clients based on weekly load patterns. Symmetric uncertainty and Pareto analysis are used to extract relevant features used to train a deep learning network, which predicts the likelihood of households adopting demand management services. However, the algorithm performance can be affected by noise sensitivity, uneven data distribution, and suboptimal choice of the number of clusters.

In contrast, hierarchical techniques do not require the number of clusters beforehand. A downside of these solutions is that identifying the optimal number of clusters may be subjective and can be affected by the unique traits of the data set. Michalakopoulos et al. (2024) conducted a segmentation of energy customers using Agglomerative Hierarchical Clustering (HAC), K-means, K-medoids, and DBSCAN algorithms. The performance, as evaluated by indices such as Davies–Bouldin, Calinski–Harabasz, and Silhouette Score, shows that K-means, HAC, and DBSCAN achieve similar results for four out of the seven identified clusters. However, the remaining three clusters require further refinement. Ahir and Chakraborty (2022) apply agglomerative hierarchical clustering to analyze the energy usage patterns of residential consumers using six load profile characteristics: base load, discretionary load, morning use, pre-peak use, peak use, and post-peak use. The solution emphasizes the importance of understanding consumption habits for making informed decisions and identifying representative load profiles. However, hierarchical clustering is computationally expensive for large datasets and rigid, as any change requires recomputing the entire tree. Arias-Requejo et.al (Arias-Requejo et al., 2023) propose a data-driven methodology to predict the energy demand of HVAC systems in smart buildings, considering the influence of weather conditions on energy consumption. Hierarchical Agglomerative Clustering is used to identify clusters of similar environmental conditions affecting the building energy consumption and create tailored prediction models.

A more flexible approach to modeling clusters is given by the probabilistic clustering methods that use probability distributions and assign data points to clusters based on likelihood. However, they feature a high computational complexity. Kaur and Gabrijelčič (2022) use for energy profile clustering characteristics such as relative average consumption and standard deviation for different periods of the day as well as seasonal variation scores. The Gaussian mixture model identifies clusters with similar profiles. However, determining the optimal number of

clusters is a complicated and resource-consuming process. Sun et al. (2019) propose a new probabilistic method for basic demand estimation, which combines a clustering algorithm based on deep learning with quantile regression forest models. The clustering algorithm uses deep learning to identify daily load patterns, while quantile regression forest models estimate demand during dynamic usage events. Although promising, the method has high computational complexity and requires high-quality data to provide accurate clustering results.

Deep learning clustering methods apply advanced machine learning techniques to segment load and power consumption profiles; however, the accuracy of segmentation depends on high-quality data. Eskandarnia et al. (2022) propose a deep learning framework for load profiling in smart meter analysis, that integrates an autoencoder to reduce the dimensionality of smart meter readings and improve the clustering quality. Similarly, Kumar and Mallipeddi (2024) introduce an advanced deep learning-based framework that combines autoencoders with the K-Means for the segmentation of load patterns. The encoder is used to reduce the dimensionality of the data, generating compact representations and minimizing reconstruction losses. The generated clusters are refined by recomputing the associated loss and updating the embeddings. Wang et al. (2024) propose a deep clustering algorithm based on Particle Swarm Optimization (PSO) for analyzing electricity load curve clusters that is used to determine the optimal hyperparameters for a convolutional autoencoder. This optimized autoencoder, combined with a deep clustering model developed using a confidence-based sample selection strategy, enhances the accuracy and efficiency of data grouping, thereby improving the overall quality and effectiveness of the clustering process. Hybrid clustering methods combine different types of clustering solutions leveraging on the strengths of each and addressing their limitations. Sandoval Guzmán et al. (2024) propose a hybrid method that combines a tensor decomposition algorithm with DBSCAN to improve the clustering results; however, the solution features a high complexity. Zhang et al., (2022) combines deep linear discriminant analysis with affinity propagation clustering to improve the classification of electricity consumption patterns in households. The affinity propagation is used to create a load dictionary that better captures daily consumption variations. The proposed method not only overcomes the limitations of traditional clustering approaches but also allows for a more detailed exploration of consumption habits and correlations with household characteristics, with the potential to be applied in various sectors and regions. Funde et al. (2019) proposes an energy consumer behavior analysis method that combines reason-based association rules with the K-Medoid algorithm. Association rules are used to discover meaningful patterns and relationships in time series of energy data within a defined time window. The K-Medoid algorithm performance can significantly depend on the initial choice of medoids, which can lead to suboptimal solutions. Wang et al. (2016) model energy consumption dynamics using a time-based Markov model which considers that consumption behaviors depend on the current state and evolve predictably over time. To reduce communication and storage costs, the symbolic aggregate approximation technique is used, which discretizes and compresses consumption data, transforming load curves into a series of symbols. For segmenting consumers based on similar behaviors, adaptive K-Means is combined with the CFSFDP technique, providing an efficient and robust solution for data clustering.

Recently, some approaches used bio-inspired algorithms, to determine the optimal number of clusters or define the clustering problem as an optimization problem to find the best clustering solution. Lakshmi et al. (2018) and Balavand et al. (2018) use the Crow Search Algorithm to generate the initial centroids in the K-means algorithm. Anter et al. (2019) apply the Crow Search Algorithm to generate the initial centroids for the Fast Fuzzy C-means algorithm bringing significant improvements in the performance of the clustering (i.e., fast convergence and the high quality of results). Soppari and Chandra (2020) use WOA to identify the optimal set of initial centroids in the Fuzzy K-means algorithm used to determine the watermarked regions. This approach helps improve the accuracy and efficiency of the clustering process by ensuring an optimal distribution of centroids from the start. Cuevas et al. (2019) use the WOA to cluster unlabeled data, treating the clustering problem as an optimization problem solved with this algorithm. A major drawback of this method is the need to specify the number of clusters in advance. An incorrect choice of the number of clusters can result in the formation of irrelevant or poorly defined clusters. Nasiri and Modarres Khiyaban (2018) combine the PSO with dimensionality reduction techniques to cluster electric charge profiles. To determine the optimal number of clusters the Improved Visual Assessment of Cluster Tendency algorithm is applied.

The method proposed in this paper advances the state of the art by addressing some of the identified clustering limitations. The number of clusters and centroids is dynamically updated in each iteration, eliminating the need for an initial estimate and reducing the risk of determining a suboptimal number. The centroids and individuals are codified using features derived based on peaks and valleys to capture more accurately the variability and dynamics of energy consumption profiles. The method provides effective exploration and exploitation mechanisms, facilitating the identification of the most relevant and consistent clusters by leveraging the CH index as a fitness function, which measures cluster compactness and separability. Furthermore, by starting from a high-quality initial population generated with the K-Means++ algorithm, our method accelerates the convergence of the algorithm. Although the main purpose of the proposed method is to segment energy consumption profiles based on users' daily behavior, its architecture allows for application in contexts where consumption patterns evolve over time. Recent literature has addressed the concept's drift through complex models, such as (Azeem et al., 2025), where the authors use an adaptive ensemble of LSTM networks with dual attention and dynamic feature selection to adapt the prediction to behavioral changes. Although efficient, this approach involves high computational complexity and requires constant access to labeled data. In contrast, our method provides an unsupervised and efficient solution that can be periodically re-executed on updated data, without the need for a previously trained model. Thus, changes in consumption behavior can be naturally reflected in the obtained clustering structure, without explicit interventions or costly adaptations.

3. Daily energy profile clustering

For segmenting energy consumption profiles, an unsupervised clustering method based on the Whale Optimization (WOA) algorithm is proposed. The choice of this evolutionary strategy is motivated by its ability to efficiently explore highly complex solution spaces, characterized by high variability and irregular data distribution. In contrast to conventional methods that aggregate consumption values based on fixed intervals and predefined weights, the presented approach directly analyzes the energy consumption profiles and allows for the automatic segmentation of them according to real consumption behavioral patterns. The method includes an automatic feature extraction process, applied to each consumption profile. These features are calculated based on the daily energy consumption profile and include the duration and magnitude of peak and valley periods, average values, and other indicators derived from the consumption curve. Thus, each profile is represented by a numerical vector that synthesizes the relevant information for the clustering process. The process of identifying the optimal clustering configuration is performed using WOA. An individual in the algorithm population represents a possible clustering solution, defined by a set of centroids in the extracted feature space. The positions and number of centroids are iteratively adjusted, according to a fitness function based on the Calinski–Harabasz index, which measures the internal cohesion of the clusters and the separability between them. The goal is to identify that configuration that maximizes the quality of the energy consumption profiles segmentation.

Fig. 1 illustrates the architecture of the WOA based method,

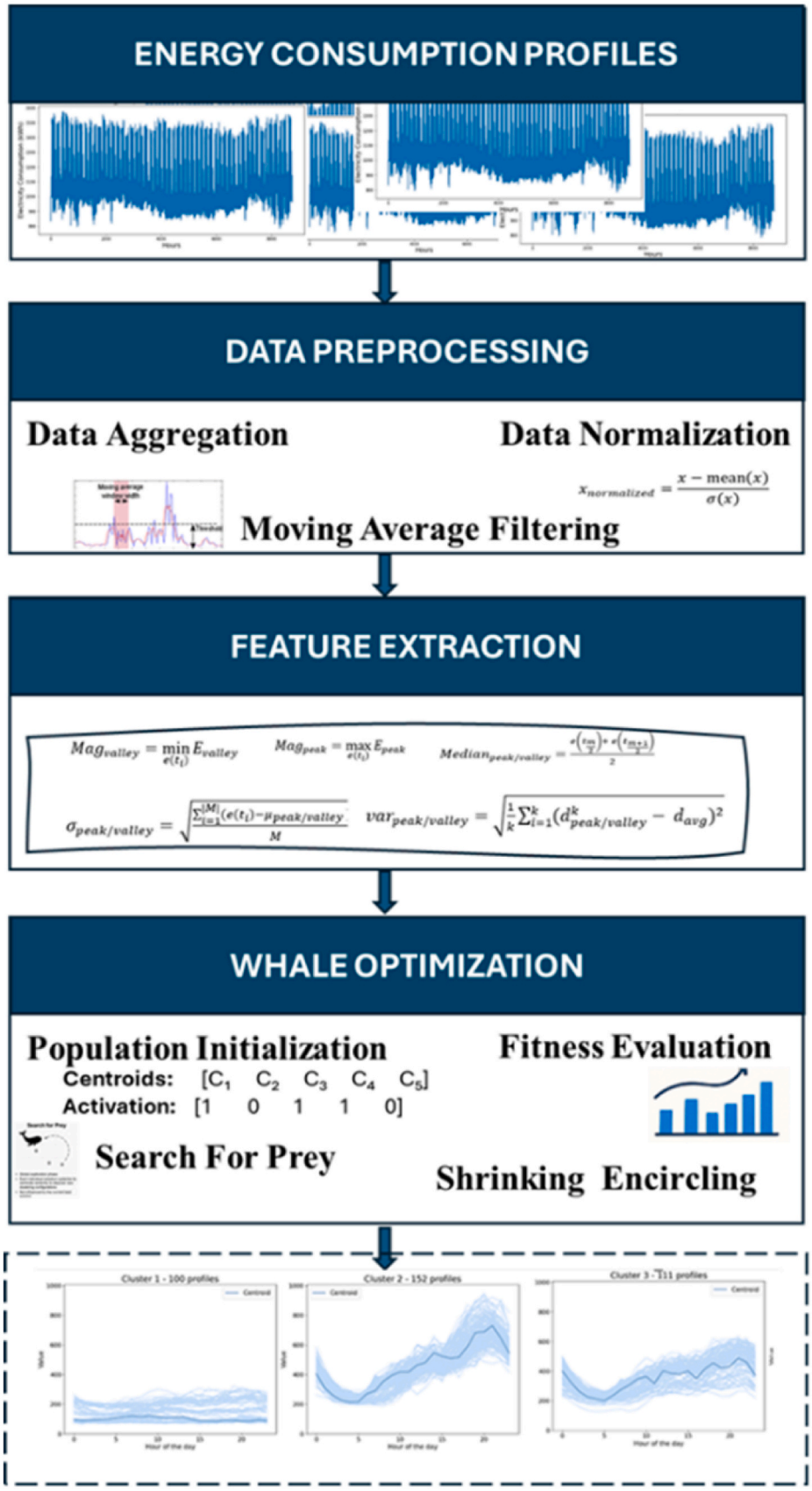

Fig. 1. Method Architecture.

structured in three main stages: data pre-processing, feature extraction and clustering configuration optimization. In pre-processing stage, the initial data, recorded every 15 min, are aggregated at the hourly level. Subsequently, the values are normalized, and a moving average filter is applied to reduce noise and highlight the overall structure of the curve. This smoothing facilitates the subsequent identification of peak and valley consumption periods, which are used in the feature extraction process. In the feature extraction stage, each consumption profile is transformed into a numerical vector describing the structure of the daily energy consumption profile. The features are automatically extracted based on the identification of peek and valley consumption periods, and include information such as the duration, magnitude and distribution of these periods over time. The result is a vector representation that synthesizes the shape and dynamics of the curve for each energy consumption profile. In the clustering configuration optimization stage, WOA algorithm is applied on the set of features, to determine the optimal clustering configuration.

3.1. Peaks and valleys derived features

A daily energy profile is formally defined as:

$$E_{day} = \{e(t_n) \,|n = \{0, 1, ..T = 24\}\} \tag{1}$$

where $e(t_n)$ is the energy value associated with the time slot t_n of the 24 h of the day.

The peaks and valleys in energy demand are defined as intervals where the energy consumption is consistently higher or lower than the average. A peak corresponds to an interval of high energy consumption due to increased activity (e.g. intensively using appliances), and a valley corresponds to a period of significantly reduced energy usage (e.g., late at night).

To facilitate the identification of peak and valley consumption periods, a moving average filter was applied, with a smoothing window of size w. The moving average at time slot i is computed as the arithmetic mean of the energy values of the time series representing the energy consumption profile from index $i-w$ to i:

$$\widehat{e(t_i)} = \quad \frac{1}{w+1} \sum_{n=i-w}^{i} e(t_n) \tag{2}$$

The time slots u at the beginning of the interval ($u < w$) require special attention because the time window required for calculating the moving average is not always completely available. To handle these situations, a weighted moving average method was used, which is calculated as follows:

$$\widehat{e(t_j)} = \frac{\sum_{i=0}^{n} w_i * e\left(t_{j-i}\right)}{\sum_{i=0}^{n} w_i} \tag{3}$$

where: $\widehat{e(t_j)}$ is the estimated energy value of the time-weighted moving average, $e\left(t_{j-i}\right)$ are the energy values of the time series at $j.j-1,\ldots.j-n$, w_i are the weights assigned to these values, and typically $w_i = n-i+1$ for a decreasing linear weighting, with w_0 being the largest share and w_n the smallest.

For the energy values thresholds to identify the peaks and valleys periods -a dynamic threshold-based approach has been adopted. The threshold values are determined based on the mean (μ) and standard deviation (σ) for the moving average:

$$\mu = \frac{1}{|T|} \sum_{i=1}^{|T|} \widehat{e(t_i)} \tag{4}$$

$$\sigma = \sqrt{\frac{1}{|T|} \sum_{i=1}^{|T|} \left(\widehat{e(t_i)} - \mu\right)^2} \tag{5}$$

A peak is defined as consecutive time slots (m) when the daily profile energy consumption values exceed the sum of the mean μ and the standard deviation σ:

$$E_{peak} = \{e(t_i) \,| \quad e(t_i) > \tau_{peaks} = \mu + \sigma\} \tag{6}$$

Similarly, a valley is identified when the daily profile consumption is less than the difference between the mean and the standard deviation.

$$E_{valley} = \{e(t_i) \,| \quad e(t_i) < \tau_{valley} = \mu - \sigma\} \tag{7}$$

Therefore, the length or duration (m) of the interval of an energy peak or valley continues until the value of the moving average returns below or above the thresholds.

On the energy peaks and valleys identified, the following features have been extracted to provide an in-depth understanding of daily energy consumption dynamics: magnitude, median, and total consumption of each peak and valley, total consumption, average frequency of peaks and valleys periods, variation in consumption during peak and valley periods, load factor, duration of peaks and valleys, variation in the durations, width of peaks or valleys periods.

The magnitude of peaks and valleys shows how high or low energy consumption values can reach, highlighting the variability and extremes:

$$Mag_{peak} = \max_{e(t_i)} E_{peak} \tag{8}$$

$$Mag_{valley} = \min_{e(t_i)} E_{valley} \tag{9}$$

The median value of the peak or valley periods is the energy value in the middle of the period ($m/2$) in case of an odd number of energy values. For an even number of time slots, it is computed as the arithmetic mean of the two energy values associated with the middle timeslots in the interval:

$$Median_{peak/valley} = \frac{e\left(t_{\frac{m}{2}}\right) + \quad e\left(t_{\frac{m+1}{2}}\right)}{2} \tag{10}$$

The total energy consumption of peak or valley periods provides an overall measure of the amount of energy used and is determined as:

$$Total_{peak/valley} = \sum_{i=1}^{m} e(t_i) \tag{11}$$

The consumption variation during peaks and valleys measures consumption fluctuations in energy usage. It is computed as the standard deviation of consumption values during peak and valley periods:

$$\sigma_{peak/valley} = \sqrt{\frac{\sum_{i=1}^{|M|} \left(e(t_i) - \mu_{peak/valley} \quad\right)^2}{M}} \tag{12}$$

where $\mu_{peak/valley}$ is the average consumption during the peak or valley periods.

The peak and valley load factor compares the average energy consumption $\mu_{peak/valley}$ periods with the maximum recorded energy during these periods:

$$LF_{peak/valley} = \frac{Mag_{peak/valley}}{\mu_{peak/valley}} \tag{13}$$

Duration variation measures the fluctuations in the length of peak and valley periods observed in a daily consumption profile:

$$\text{var}_{peak/valley} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(d^{k}_{peak/valley} - \quad d_{avg})^{2}} \tag{14}$$

where $d^{k}_{peak/valley}$ is the length of each peak or valley period, k is the total number of peak/valley periods in the daily profile and d_{avg} is the average duration of the peaks and valleys identified.

The clustering of the daily energy consumption profiles was performed, based on the average values of the extracted features computed for each daily energy consumption pattern.

In addition, the ratio between the sum of peak periods durations and the valley durations was also used as a feature. This report provides an insight into the proportion of activity during peak and peak consumption periods over the observed interval.

$$Ratio = \frac{\sum_{i=1}^{k_p} Total_{peak}}{\sum_{i=1}^{k_v} Total_{valley}} \quad , \quad k = k_p + k_v \tag{15}$$

where N_p is the number of peak periods and N_s is the number of valley periods.In the final stage, the data were normalized by applying Z-score to ensure that all features have the same scale, with a normalized - distribution having a mean of 0 and a standard deviation of 1.

3.2. Clustering using WOA

The clustering - a set of z daily energy consumption profiles denoted by:

$$ECP = \{E_{day,1}, \ldots, E_{day,z}\} \tag{16}$$

is formulated as an optimization problem, aiming to partition the set of profiles into a number k clusters, based on their similarity. Each daily profile is characterized by the set of features introduced in Section 3.1:

$$f(E_{day}) = \{f_1(E_{day}), \quad f_2(E_{day}) \ldots f_r(E_{day})\} \tag{17}$$

The optimization problem was approached using the Whale Optimization Algorithm (WOA), a metaheuristic inspired by the hunting behavior of humpback whales, specifically their ingenious bubble-net feeding method. This cooperative hunting strategy involves a hierarchical leadership structure with three main phases: searching for prey, encircling prey, and attacking the prey using the bubble-net method. The search for prey corresponds to the exploration stage of the search space while attacking the prey corresponds to the exploitation phase.

Within the WOA, an individual in the population of whales is modeled as a vector of potential cluster centroids:

$$Pop = \{\left[\overrightarrow{centroid}, \quad \overrightarrow{activate}\right] | I = (centroid_i, \quad activate_i), centroid_i \in ECP \wedge i = 1, \ldots, c\} \tag{18}$$

where c is the maximum number of clusters, $centroid_i$ is the centroid of cluster C_i and $activate_i$ is activation status corresponding to cluster C_i. The maximum number of clusters is computed as:

$$c = Round(\sqrt{z}) \tag{19}$$

where z represents the number of energy consumption profiles that should be clustered. The centroid of a cluster is a daily energy consumption profile and, therefore, is modeled using the same set of features as defined in relation (17). The activation status specifies if a cluster is activated or not and is defined as:

$$activate_i = \begin{cases} 1, if flag_i \geq \tau \\ 0, otherwise \end{cases} \tag{20}$$

where τ is a predefined threshold and $flag_i$ is the activation flag computed as the centroid diameter distance. The centroid diameter distance evaluates the spread or dispersion of the data points within a cluster, relative to the centroid of that cluster.

$$flag_i(C_i \quad) = \frac{1}{|C_i|}\sum_{E_{day} \in C_i} \quad d(E_{day}, centroid_i) \tag{21}$$

To compute the threshold value, τ, - a formula that dynamically adjusts the threshold based on the current population of individuals is defined:

$$\tau = \overrightarrow{mean} \quad _{activate} + \alpha * \sigma \tag{22}$$

where $\overrightarrow{mean_{activate}}$ is the mean of activation flags in the population of individuals, σ is the standard deviation of activation flags in the population and α is a scaling factor.

The fitness function evaluates the quality of an individual using the Calinski-Harabasz index. It measures the clustering effectiveness by assessing both the internal cohesion of the clusters as well as the clear external separation among them:

$$fitness(I) = \frac{BCSS}{No_{cl} - 1} * \frac{N - k}{WCSS} \tag{23}$$

In this formula, No_{cl} is the number of clusters, z is the total number of energy consumption profiles that need to be clustered, $BCSS$ is the Between-Cluster Sum of Squares, and $WCSS$ is the Within-Cluster Sum of Squares. The Within-Cluster Sum of Squares (WCSS) is calculated as the sum of the distances between the energy consumption profiles belonging to each cluster and the centroids of each cluster, while the Between-Cluster Sum of Squares (BCSS) is computed as the weighted sum of distances between each cluster centroid and the overall centroid ($centroid_{overall}$) of all energy consumption profiles:

$$WCSS = \sum_{i=1}^{No_{cl}} \sum_{E_{day} \in C_i} d(E_{day}, centroid_i) \tag{24}$$

$$BCSS = \sum_{i=1}^{No_{cl}} |C_i| * d(centroid_i, centroid_{overall}) \tag{25}$$

The $K-means++$ method (Ikotun et al., 2023) was used to generate the individuals from the initial population of whales, ensuring that all individuals have a good chance of representing a good solution, thus contributing to the overall effectiveness of the optimization process. The first sub-set of q centroids, $0 < q < c$ are randomly selected from the set of daily energy consumption profiles and added to the vector of centroids, $\overrightarrow{centroid}$. Then for each energy consumption profile, $E_{day} \in ECP$ that has not yet been chosen - the minimum distance to the already selected centroids is computed if the distance is small, the similarity between the energy consumption profile and the cluster's centroid is high, otherwise it is low. New centroid $q+1$ is selected using a probability based on the distance to the existing centroids and the steps are repeated until all c clusters of the $\overrightarrow{centroid}$ vector, are selected. For each centroid the activation status is randomly assigned.

- The WOA phases of searching for prey, shrinking encircling, and spiral updating position have been adapted for our representation of daily energy profiles and clustering process.

To update the position of an individual only the vector of centroids from the representation of an individual is considered. In the search for prey phase, each individual updates its position based on its current position, the position of an individual randomly selected from the current population, and two coefficient vectors $\vec{A}$ and $\vec{D}$:

$$\overrightarrow{centroid(t+1)} = \overrightarrow{centroid_{random}(t)} - \vec{A} * \vec{D} \tag{26}$$

$$\vec{D} = \quad \vec{C} * \overrightarrow{centroid_{random}(t)} - \quad \overrightarrow{centroid(t)} \tag{27}$$

where $\overrightarrow{centroid(t)}$ is the vector of centroids from the current population of individuals, $\overrightarrow{centroid_{random}(t)}$ is a vector of centroids randomly selected

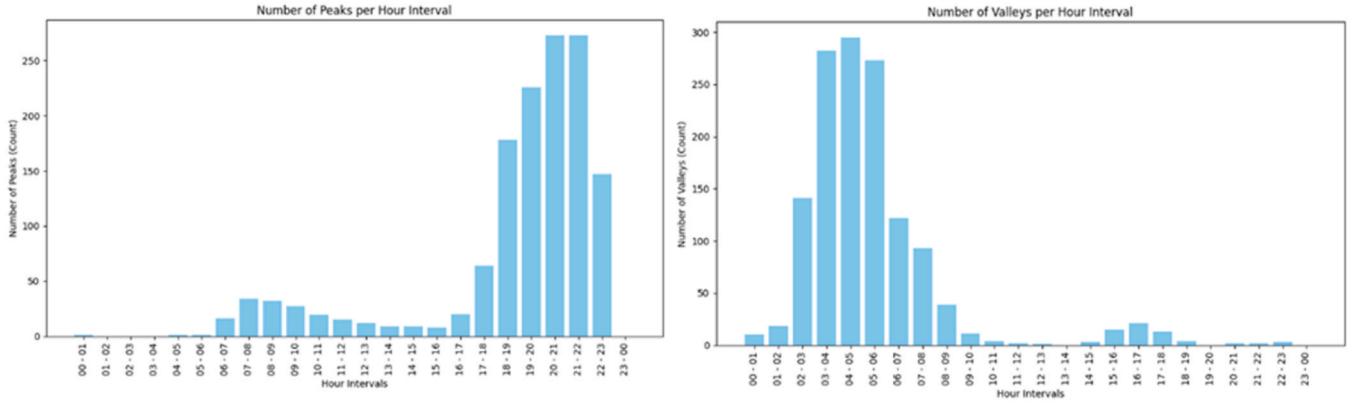

Fig. 2. The frequency distribution of peaks and valleys in energy consumption data.

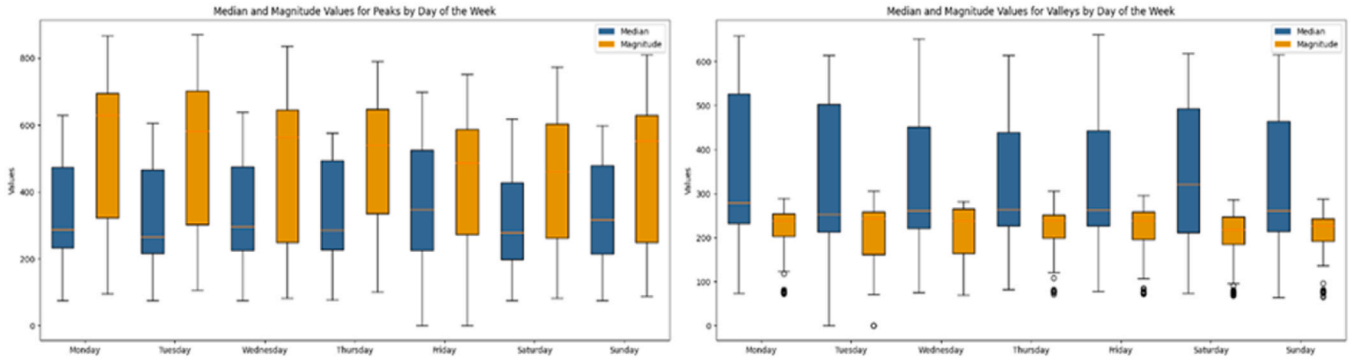

Fig. 3. Distribution of the median and the magnitude values of peaks and valleys over a week.

from the current population and $\overrightarrow{centroid(t+1)}$ is the vector of centroids of the new generated population. $\overrightarrow{A}$ and $\overrightarrow{C}$ are coefficient vectors defined as:

$$\overrightarrow{A} = 2 * a * \overrightarrow{r} - a \tag{28}$$

$$\overrightarrow{C} = 2 * \overrightarrow{r} \tag{29}$$

$\overrightarrow{r}$ is a vector with randomly generated values in the interval $(0,\ 1]$ and a is a value that linearly decreases from 2 to 0 throughout the iterations.

In the shrinking encircling phase, the individuals update their position relative to the best individual $\overrightarrow{centroid_{best}(t)}$, determined so far:

$$\overrightarrow{centroid(t+1)} = \overrightarrow{centroid_{best}(t)} - \overrightarrow{A} * \overrightarrow{D} \tag{30}$$

$$\overrightarrow{D} = \overrightarrow{C} * \overrightarrow{centroid_{best}(t)} - \overrightarrow{centroid(t)} \tag{31}$$

Table 2
Parameters configurations considered for WOA-based clustering.

Config. #	$iter_{max}$	*popSize*	*b*	*α*	*Fitness*
1	100	40	0.9	0.5	1789.3
2	100	50	0.5	0.5	1722.1
3	100	40	0.5	0.5	1721.9
4	150	60	0.1	0.	1717.6
5	150	50	0.1	0.5	1716.2
6	150	40	0.1	0.5	1716.2
7	100	50	0.1	0.5	1715.7
8	150	40	0.5	0.5	1710.6
9	100	40	0.1	0.5	1702.9
10	100	60	0.1	0.5	1688.7

In the phase of spiral updating position, the position of all individuals in the population is updated using helix-shaped movement mechanism:

$$\overrightarrow{centroid(t+1)} = \overrightarrow{D'} * e^{b*l} * \cos(2 * \pi * l) + \overrightarrow{centroid_{best}(t)} \tag{32}$$

$$\overrightarrow{D'} = \overrightarrow{centroid_{best}(t)} - \overrightarrow{centroid(t)} \tag{33}$$

where l represents a random number in the interval $[-1,\ 1]$, e is the Euler's constant, b is the constant for defining the shape of the logarithmic spiral.

The WOA-based clustering algorithm (Algorithm 1) consists of two main phases: the initialization phase (lines 1–3) and the iterative phase (lines 5–25). In the initialization phase, the initial population of whale individuals is generated using the K-Means++ algorithm. A set of clusters is formed by identifying the active centroids in the individual and assigning the energy consumption profiles to the cluster whose centroid is closest (lines 3–6). The fitness value is calculated to identify the best individual in the population (line 7). In each iteration, the WOA variables are initialized, as well as the coefficient vectors A and C. The algorithm updates each whale individual in the population by applying the three strategies: searching for prey (17−23), forming circles, and updating the spiral position (24−25). The associated clusters are updated based on the individual's new centroids, and the vector of clusters corresponding to the individual is added to CLUSTERS_SET. At the end of each iteration, the best individual is updated based on the fitness value calculated according to CLUSTERS_SET. The algorithm returns the set of clusters that corresponds to the best individual identified by the algorithm.

ALGORITHM 1. : WOA for clustering energy consumption profiles

Inputs: EPC - set of energy profiles, $iter_{max}$ - the maximum number of iterations, $popSize$ - the population size, b - the constant value used in defining the logarithmic spiral shape, k - the maximum number of clusters
Outputs: $CLUSTERS$ - the optimal set of clusters
Begin

$Fitness = \emptyset, \overrightarrow{A} = \emptyset, \vec{C} = \emptyset, CLUSTERS = \emptyset, \ CLUSTERS\_SET = \emptyset$
$\overrightarrow{centroid} = K_{Means++}(EPC)$
$Pop = [\overrightarrow{centroid}, \overrightarrow{activate}]$
foreach I in Pop do
  $CLUSTERS = ASSIGN - PROFILES(EPC, I)$
  $CLUSTERS\_SET = CLUSTERS\_SET \cup CLUSTERS$
endFor
$I_{best} = SELECT - BEST - FITNESS(Pop, CLUSTERS\_SET)$
$iter = 0$
while $(iter <= iter_{max})$
  $CLUSTERS\_SET = \emptyset$
  $l = RANDOM\ (-1, 1),\ p = RANDOM\ (0, 1),\ \vec{r} = RANDOM\ (0,1)$
  $a = LINEAR - DECREASE\ (2,0)$
  $\vec{A} = UPDATE(\vec{r}, a), \vec{C} = UPDATE(\vec{r})$
  foreach I in Pop do
    if $(p < 0.5)$ *then*
      if $(|A| < 1)$ *then*
        $\vec{D} = CALCULATE(I, \vec{C}, I_{best})$
        $I = UPDATE - POSITION\ (\vec{A}, \vec{D}, I_{best})$
      else
        $\vec{D} = CALCULATE(I, \vec{C}, I_{random})$
        $I = UPDATE - POSITION\ (\vec{A}, \vec{D}, I_{random})$
      endIf
        $\overrightarrow{D'} = CALCULATE(I, I_{best})$
        $I = UPDATE - POSITION\ (\overrightarrow{D'}, b, l, I_{best})$
    endIf
  $CLUSTERS = ASSIGN - PROFILES(ECP, I)$
  $CLUSTERS\_SET = CLUSTERS\_SET \cup CLUSTERS$
  endFor
  $I_{best} = UPDATE - BEST - FITNESS(Pop,\ CLUSTERS\_SET)$
  $iter = iter + 1$
endWhile
$CLUSTERS = ASSIGN - PROFILES(ECP,\ I_{best})$
return CLUSTERS
End

4. Evaluation results

To evaluate WOA-based clustering a dataset with 8568 data points representing hourly energy data of student campus buildings for a year. Fig. 2 illustrates the frequency distribution of peaks and valleys in energy consumption, grouped by hourly intervals. The most common consumption peaks occur between 18:00 and 21:00, indicating an increase in student activities during this time, probably related to domestic activities and other evening activities such as studying. The periods of minimum consumption are concentrated between 02:00 and 06:00, when most students are asleep and energy-intensive activities are reduced. After 06:00, consumption increases gradually, with the resumption of daily activities.

Fig. 3 shows the distribution of median and the magnitude values of energy consumption peaks and valleys for each day of the week. For peak periods, median energy consumption values are relatively constant throughout the week, with slight variability. However, the magnitude values vary more, indicating a fluctuation in peak intensity between days. This suggests that although average daily consumption remains approximately stable, consumption peaks may vary in intensity from day to day. In the case of valley periods, the median values also remain approximately constant, but the magnitude is considerably smaller. Also, the presence of outliers is observed, suggesting significant decreases in energy consumption during those periods, probably due to special events or conditions that affected consumption.

To evaluate the algorithm's behavior, different parameter configurations were investigated. The number of iterations and the population size were varied within the limits of 50–300 and 10–50, respectively.

These intervals were chosen because, while a larger number of iterations may improve the final solution, it can also significantly increase execution time. Parameter b varies between 0 and 1 with higher values favoring efficient exploration, while lower values indicate more

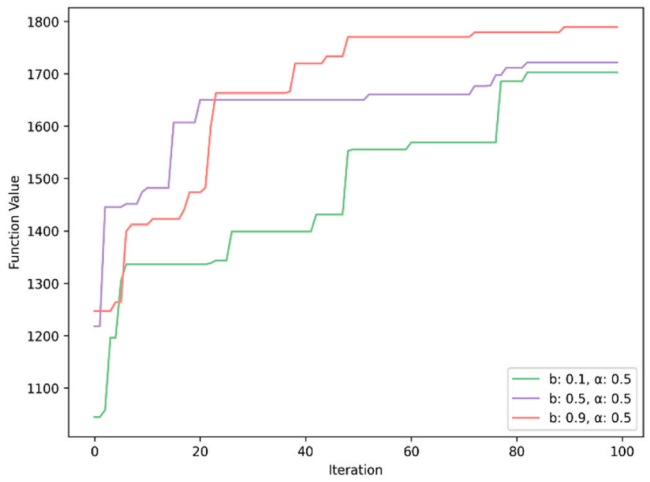

Fig. 4. The fitness evolution for the case where the number of iterations is 100, the population size is 40, b and α varies.

intensive exploitation. Scaling factor α has a similar range where a value of 1 imposes strict restrictions on cluster activation and smaller values provide increased flexibility.

Table 2 shows the configurations of the adjustable parameters for which the best values of the fitness function were obtained. For 100 iterations and a population size of 40, the high fitness values are obtained in about 20–30 iterations. Comparatively, for a population of 30, although the solutions converge, the increase in fitness values is slower, indicating that this size does not allow efficient exploration of the solution space. For larger populations such as 50 and 60, the quality of the solutions is good, but the increased population size leads to higher execution time without providing significant advantages over the 40 population. Within the population of 40, the combination of b = 0.9 and α = 0.5 generated the best final fitness values. The optimal configuration was set at 100 iterations, population size of 40, b= 0.5 and α= 0.5, providing an optimal balance between fast convergence and the quality of the obtained solutions.

The graph in Fig. 4 illustrates the evolution of the fitness value for three different settings of b and α parameters, during 100 iterations, with a population of 40. Although the purple (b=0.5, α=0.5) and green (b=0.1, α=0.5) lines converge slightly faster than the red line (b=0.9, α=0.5), the difference in the number of iterations required for convergence is insignificant. However, the red line reaches the highest fitness value, indicating that this configuration produces higher-quality solutions, which is the reason why it was chosen as optimal.

Based on the optimal configuration of parameters, - the clustering performance of the WOA algorithm was evaluated, focusing on the quality of the generated clusters. To assess the separability and compactness of the clusters - validation metrics such as the Davies-Bouldin Index, the Dunn Index, the Silhouette Score, and the Ball-Hall Coefficient were used. In addition, advanced visualization techniques were applied to analyze the similarity between grouped consumption profiles.

The results showed that the 377 daily consumption profiles were grouped into four distinct clusters (see Fig 5). The three main clusters included 103, 116, and 114 profiles, which showed similar features, during peak and valley periods. The centroids of these clusters provide a clear representation of the overall energy consumption pattern. The fourth cluster, composed of only 14 profiles, shows a distinct

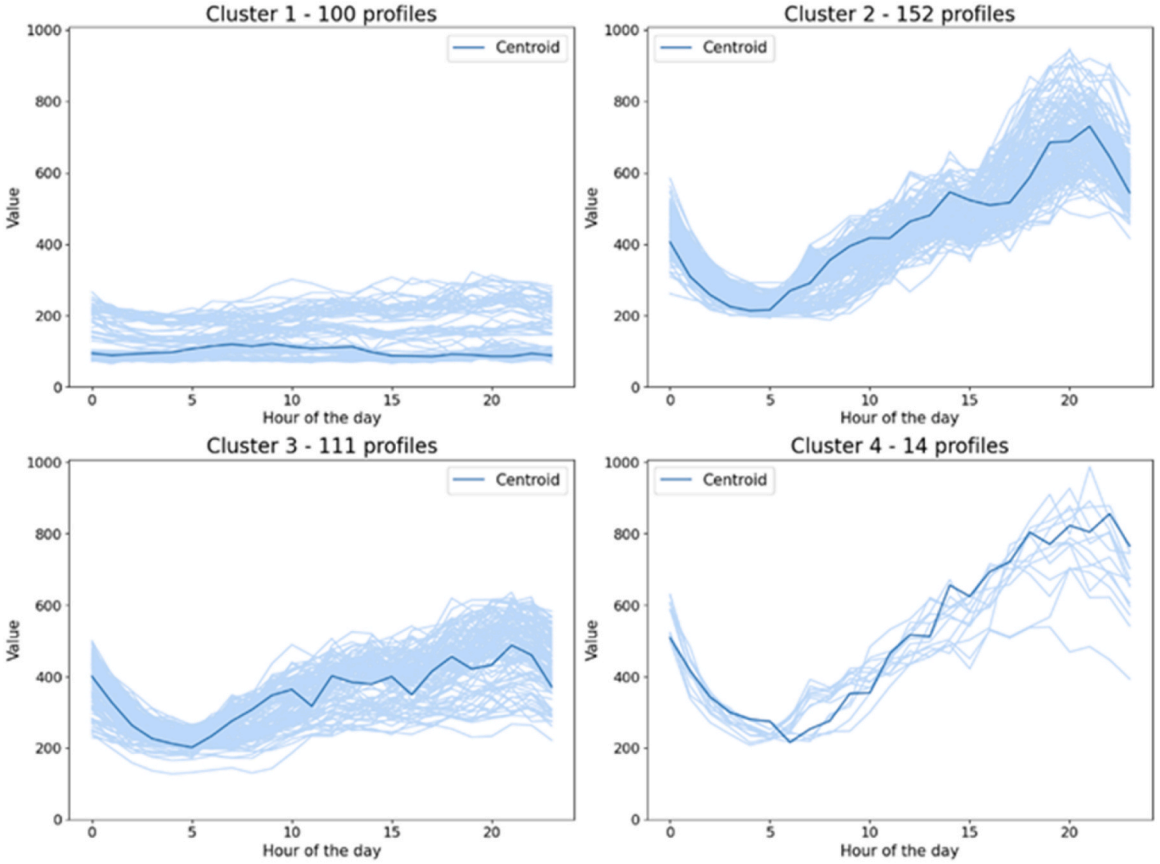

Fig. 5. The clusters resulting from running the WOA clustering algorithm.

Table 3
Energy profiles clustering performance metrics.

Agorithm	DB Index	Dunn Index	Silhouette	Ball-Hall	Time Complexity	Sensitivity to Noise
WOA	0.527	0.032	0.484	1.851	Medium	Low
K-Means	1.146	0.012	0.296	2.068	Low	High
DBScan	2.606	0.012	0.194	4.028	Medium	Reduced
Agglomerative Clustering	0.677	0.040	0.456	1.856	Medium	High
TimeSeries Clustering with VRAE	9.918	0.001	−0.057	5.446	High	High
Autoencoder with KMeans	2.848	0.008	0.266	4.164	High	Medium

Table 4
P-values obtained from applying the Wilcoxon test to compare the performance of the WOA method with reference algorithms, on four internal validation metrics.

Metrics	WOA vs Kmeans	WOA vs DBSCAN	WOA vs Agglomerative	WOA vs Autoencoder + Kmeans	WOA vs Variational Recurrent Autoencoder
Silhouette score	0.000001707	0.000001708	0.004341475	0.000001715	0.000001708
Dunn index	0.000001733	0.000001733	0.000261245	0.000001733	0.000001733
Davies-Bouldin index	0.000001721	0.000001715	0.000674563	0.000001708	0.000001707
Ball-Hall coefficient	0.00000173	0.000001729	0.004985803	0.00000173	0.000001729

consumption behavior, indicating that these profiles differ significantly from the others. This separation indicates the uniqueness of these profiles and suggests the presence of atypical consumption behaviors that do not align with the general trends observed in the rest of the data set.

The performance of the proposed method was evaluated through a comparative analysis with a selection of clustering algorithms frequently used in research literature. The comparison included both conventional methods, such as K-Means, DBSCAN and Agglomerative Clustering, and modern deep learning-based approaches, such as TimeSeries Clustering with Variational Recurrent Auto-encoders (VRAE) (https://statics.teams.cdn.office.net/evergreen-assets/safelinks/1/atp-safelinks.html) and Deep Autoencoder with K-Means (https://github.com/tschechlovdev/AutoEncoder_KMeans). The results obtained are summarized in Table 3.

In the case of K-Means, the Elbow method was used to determine the optimal number of clusters. Notably the number of clusters identified by our algorithm corresponds to that determined by the Elbow method for K-Means. This suggests that WOA effectively identifies the optimal

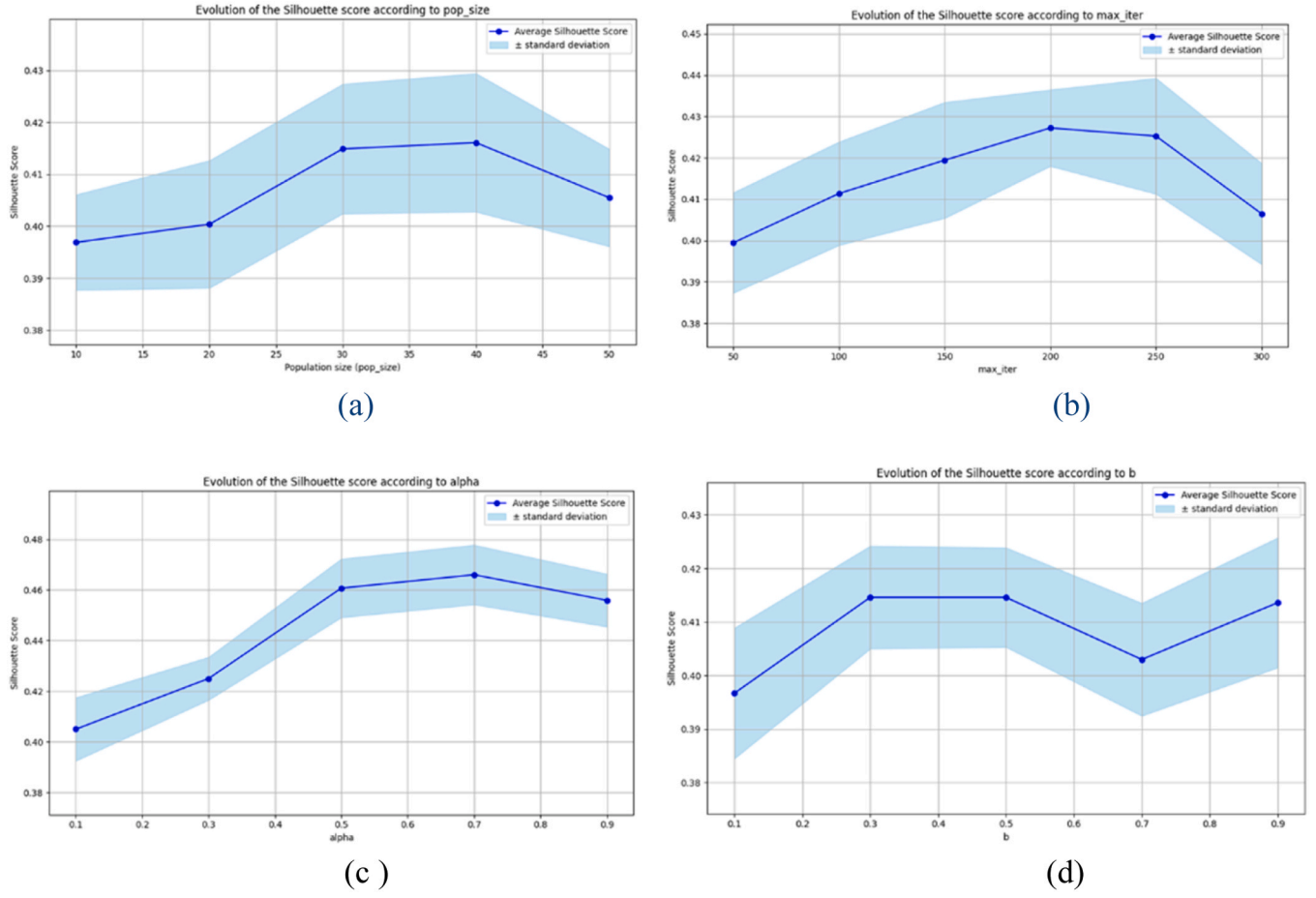

Fig. 6. Evolution of the Silhouette based on: (a) population size (pop_size), (b) maximum number of iterations (iter_max), (c) exploration coefficient (α), (d) coefficient for logarithmic spiral shape (b).

number of clusters, validating its efficiency against a well-established method.

The comparative analysis conducted between the WOA algorithm and methods selected from the state-of-the-art highlights that WOA achieves the best results. Specifically, WOA records the best values for DB Index (0.527), Silhouette Score (0.484), and Ball-Hall Index (1.851), which suggests a clear separation of clusters and high internal cohesion. It also stands out for its low sensitivity to noise, a crucial aspect in practical applications. Although the Agglomerative Clustering algorithm scores the highest Dunn score (0.040), it is surpassed by WOA in most other metrics. The complexity of WOA is medium, which strikes a reasonable balance between performance and computational complexity.

To evaluate the statistical significance of the differences between the performances of the analyzed algorithms, the Wilcoxon signed-rank test was applied to the values of the four validation metrics (i.e. Silhouette, Davies–Bouldin, Ball–Hall and Dunn) obtained by the WOA algorithm, in comparison with K-Means, DBSCAN, Agglomerative Clustering, Deep Autoencoder with K-Means, Time Series Clustering with Variational Recurrent Autoencoders (VRAE) and Deep Autoencoder with K-Means. Each algorithm was run 30 times, and for each execution the values of the validation metrics were recorded. Table 4 presents the p-values corresponding to each combination of algorithms and validation metrics. In all cases, the p-values are below the 0.05 significance threshold, confirming that WOA forms clusters of energy consumption profiles that are better separated from each other and more homogeneous within each cluster compared to classical and machine learning-based methods.

Finally, an error budget analysis was performed to assess the sensitivity of the WOA method to the variation of its configuration parameters. Each adjustable parameter wase individually modified while the others were kept constant. For each parameter, the algorithm was run 30 times, and for each run the Silhouette score was recorded. The Silhouette score was chosen as the reference metric, as it simultaneously reflects internal cohesion and separability between clusters, providing a synthetic measure of the quality of segmentation. The results presented in Fig. 6 show the evolution of the mean value of the Silhouette score, together with the standard deviation associated with each configuration. It is observed that the performance of the WOA algorithm increases up to a certain threshold for each adjustable parameter, after which it decreases slightly. In addition, the blue areas around the curves, which indicate the standard deviation, are narrow, suggesting a low variability between runs and, implicitly, a high stability of the algorithm. These results indicate that WOA provides good and consistent results in grouping energy consumption profiles, even under conditions of moderate variations in the adjustable parameters, provided that the ranges in which they vary are chosen appropriately.

Table 5 summarizes, for each adjustable parameter analyzed, the maximum variation of the standard deviation associated with the Silhouette score. Overall, the observed standard deviations remain low ($\sigma < 0.015$) in all tested configurations, which indicates a high consistency of the method with respect to changes in the values of the algorithmic parameters.

5. Discussion and limitations

The performance of WOA-based clustering is analyzed from the perspective of the global fitness evolution, population diversity, execution time, and the computational resources used. Fig. 7 illustrates the fitness evolution over 100 iterations showing that the algorithm is refining its solutions and converging towards the global optimum. The global fitness evolution shows rapid growth in the first 20 iterations, indicating efficient exploration of promising solutions. After this stage, the growth rate of fitness slows down, indicating that the algorithm continues to improve solutions, but at a slower rate as it approaches the optimum. In the latter stages, fitness stabilizes, which reflects the algorithm convergence, and short periods of stagnation highlight efforts to refine the solutions. However, subsequent increases suggest that the algorithm succeeds in overcoming local minima, discovering even better solutions.

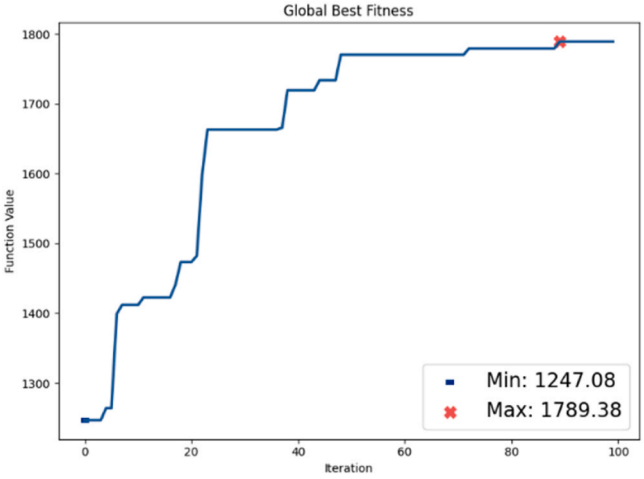

Fig. 7. WOA-based clustering global fitness evolution.

The population diversity (see Fig 8) varies up to iteration 46, suggesting active exploration of the solution space. After iteration 50, diversity stabilizes, indicating a transition to exploitation, focusing on improving the best-identified solution. The percentages of exploration and exploitation show an alternation between the search for new solutions and the optimization of existing ones. As iterations progress, exploration decreases, and the algorithm focuses on exploiting already discovered solutions.

Although the proposed algorithm generates high-quality results in terms of clustering energy consumption profiles, it faces a significant challenge related to high computational overheads. This overhead derives from two main sources: the evaluation of the fitness function and the absence of a parallelization strategy in the process of updating the individuals in the population. Computing of the fitness function based on CH index introduces substantial computational complexity. This process requires intensive resource allocation, thus increasing the total execution time of the algorithm. Moreover, sequentially updating the individuals in the population without using parallelization aggravates the computational overhead. This limitation negatively influences the scalability of the algorithm on large datasets.

To better assess and understand the impact of these factors on the performance of WOA based clustering the execution time was monitored depending on the volume of energetic data as well as memory and CPU usage. Fig. 9 illustrates the evolution of execution time for varying the number of daily energy profiles indicating an almost linear increase in the execution time as the size of the data set increases. However, it is noted that the execution time becomes significantly higher as the number of energy consumption profiles increases. This suggests that although the algorithm is scalable, its efficiency is limited, and execution times become quite high even for a relatively small number of profiles.

To analyze the impact on memory and CPU usage, - two experiments were performed. In the first experiment, the population size was varied,

Table 5
Analysis of the sensitivity of the Silhouette score to the variation of the WOA algorithm parameters.

Parameter	Tested Values	$\sigma_{max}(\textbf{\textit{Silhouette}})$
popSize	[50−300]	0.01331
$iter_{max}$	[10–50]	0.01401
α	(0,1)	0.01247
b	(0,1)	0.01219

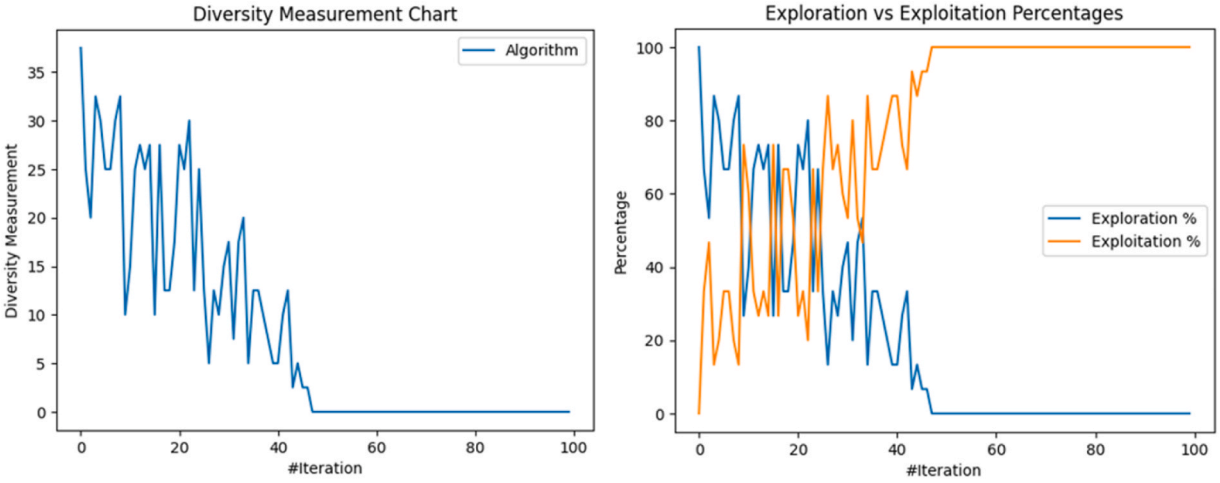

Fig. 8. Diversity evolution and the search space exploration vs exploitation percentages.

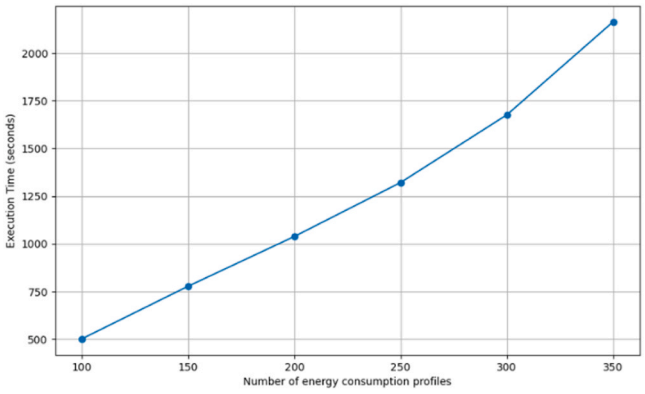

Fig. 9. Execution time increasing number of energy consumption profiles.

keeping constant the number of profiles, the number of iterations, and the values for α and b. The number of profiles was set to 350, $iter_{max}$ to 100, b to 0.9, and α to 0,5. In the second experiment, - the number of profiles was varied, keeping the population size, number of iterations, α and b constant. The optimal values used in this case were popSize= 40, $iter_{max}$= 100, b= 0.9 and α= 0,5. The obtained results presented in Fig. 10 show that memory usage remains constant regardless of variations in population size or number of consumption profiles, showing minimal fluctuations. This behavior suggests that the algorithm is independent of these parameters, benefiting from efficient data structures and libraries such as the least recently used cache, which optimizes memory consumption by keeping only recently used elements, and NumPy, which minimizes memory overhead. In contrast, CPU usage varies significantly, especially during the initialization and evolution stages. As the population size increases, CPU utilization can reach values of up to 90 % during the evolution phase, which indicates the need for more computing power to handle an increased number of whale individuals. The number of iterations also influences CPU usage: the more iterations, the longer CPU usage stays at high levels. However, despite these spikes, the global average CPU utilization stabilizes around 70–80 % regardless of population size or profiles, but significant increases in CPU usage are seen when the number of iterations is higher.

Although computationally the proposed method has a stable behavior in terms of memory usage and a reasonable scalability with respect to processor usage, a possible limitation is the assumption that data distribution remains constant during the optimization process. In real applications, consumer behavior may undergo significant changes over time (e.g., seasonality, changes in routine), a phenomenon known as concept drift. In its current form, the proposed method assumes a stationary data set (i.e. statistics such as the mean, variance, and covariance of consumption are assumed to not change over time) and uses a static feature extraction, performed only once before running the algorithm. Thus, no explicit mechanism for adapting to structural changes in the data is included. However, the architecture of the method

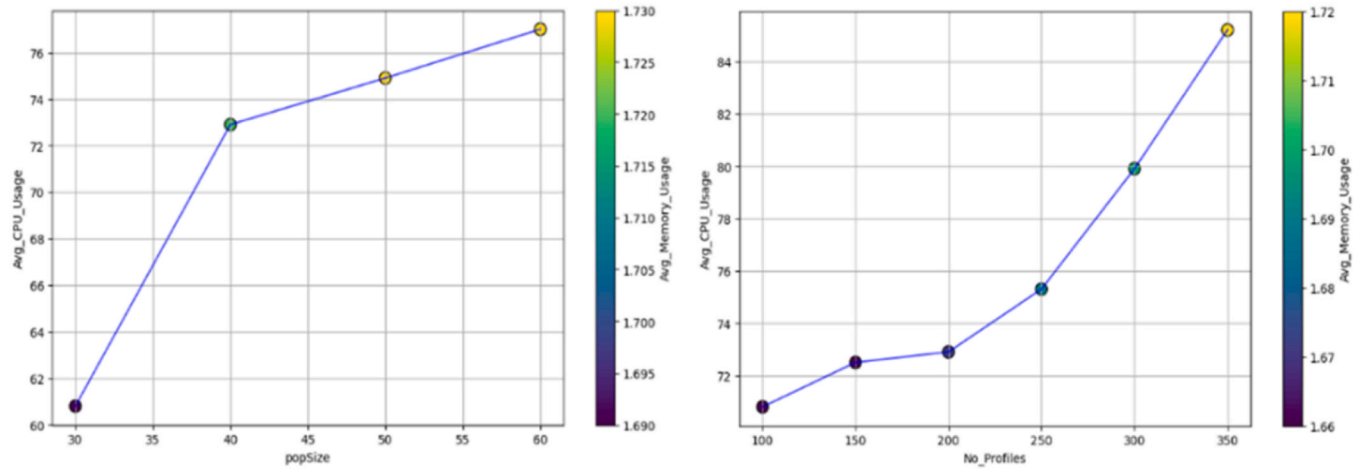

Fig. 10. CPU and Memory Usage when varying the population size and the number of profiles to be clustered.

allows for extension in this direction, by introducing data updating and feature adaptation mechanisms that consider variations in consumption behavior over time.

The first extension direction would be to use a sliding window strategy, in which the feature extraction process and the WOA algorithm are repeated periodically, at fixed intervals (e.g., once a day or once a week). For each time window, the algorithm is run on the new data, and some individuals (i.e. individuals with high fitness scores) from the previous population can be considered as starting points in the current optimization. In this way, the method can gradually detect changes in the data structure and update the cluster configuration accordingly.

The second direction of extension aims to replace the manual and fixed feature extraction process with an incrementally trained autoencoder model. Instead of features being extracted only once before optimization, the autoencoder automatically learns a set of numerical features that describe the shape of the daily consumption curve. These features are generated directly by the model, without human intervention, and are automatically updated as new data becomes available. By progressively training the autoencoder on new available data, the resulting features constantly reflect current consumption behavior. In this configuration, the WOA algorithm no longer optimizes using manually extracted features, but uses automatically learned features that constantly reflect current consumption behavior. By combining this approach with periodic re-runs of the optimization on time windows, the method can react to changes in the data distribution, without requiring labels or manual interventions.

6. Conclusions

In this paper, - an innovative approach for clustering energy consumption profiles, using Whale Optimization Algorithm was proposed. The daily consumption profiles are preprocessed by means of a moving average filtering method to identify peak and valley periods and to derive features, such as the magnitude, duration and variability of peaks and valley periods. These features were used to capture the dynamicity and variability of energy time series while encoding the whale individuals and clusters centroids. To improve the speed of convergence and the quality of the obtained clusters, - the population of individuals was initialized with K-Means++ algorithm. The quality of the clustering solutions was evaluated using a fitness function based on the CH index, due to its ability to evaluate the clusters' compactness and separability. The evaluation of the clustering results for energy consumption profiles is promising, showing strong cluster separation with a Dunn index of 0.032 and a Silhouette index of 0.484. Cluster compactness is also satisfactory, as indicated by the Davies-Bouldin index (0.527) and the Ball-Hall index (1.85). These results outperformed the performances of the K-means and DBSCAN algorithms and are comparable to those obtained by hierarchical clustering showing the efficiency of WOA-based clustering. The representation of an individual as a set of active centroids, described by features extracted from peak and valley periods, combined with the balanced exploration and exploitation mechanism of the WOA algorithm, underlines the high efficiency of our method in clustering energy consumption profiles. The fitness evolution showed a rapid increase in the first 20 iterations, followed by a plateau, showing the rapid convergence of the clustering solution with an effective balance between the exploration and exploitation phases.

The proposed method provides a framework for practical applications in the energy field, especially in the context of demand management and consumption forecasting. Its ability to perform unsupervised segmentation, without requiring a predefined number of clusters and without training on labeled sets, makes it suitable for real-world scenarios where we have heterogeneous consumption patterns that constantly change. By automatically extracting features that describe the behavior of daily energy consumption profiles and by automatically determining the optimal number of clusters, the algorithm can identify energy usage patterns that can be correlated with various consumer categories. These patterns can be used as input for adaptive demand-side management policies, dynamic resource allocation, personalized tariffs or automatic actions to reduce energy consumption during peak periods. In addition, the obtained segmentation of energy consumption profiles can be used to train specialized forecasting models for each cluster, which contributes to increasing the prediction accuracy. Thus, the method indirectly contributes to reducing load peaks, optimizing the use of energy infrastructure, and increasing the grid's capacity to deal with unforeseen fluctuations in demand.

For future work, the computational overhead of WOA-based clustering can be addressed by investigating ways in which it can have its execution parallelized. The parallelization of the fitness function evaluation and the individual updating process may significantly contribute to the improvement of execution time.

CRediT authorship contribution statement

Ionut Anghel: Writing – review & editing, Visualization, Investigation. **Tudor Cioara:** Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Viorica Rozina Chifu:** Writing – original draft, Validation, Methodology, Conceptualization. **Cristina Bianca Pop:** Writing – original draft, Formal analysis, Conceptualization. **Ionela Danci:** Writing – original draft, Visualization, Validation, Software.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially supported by the European Commission under the Horizon Europe Framework Program through the DEDALUS Project (Grant 101103998). Additional support was provided by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI - UEFISCDI, project number PN-IV-P8-8.1-PRE-HE-ORG-2023–0111, within PNCDI IV, as well as by the project "Research on Innovative Solutions for Smart Buildings, Efficiently Integrated with the Energy System" (N-C-CDI 3998).

Data availability

Data will be made available on request.

References

⟨https://github.com/tschechlovdev/AutoEncoder_KMeans⟩.
⟨https://statics.teams.cdn.office.net/evergreen-assets/safelinks/1/atp-safelinks.html⟩.
Ahir, Rajesh K., Chakraborty, Basab, 2022. A novel cluster-specific analysis framework for demand-side management and net metering using smart meter data. ISSN 2352-4677 Sustain. Energy Grids Netw. 31, 100771. https://doi.org/10.1016/j.segan.2022.100771.
Anter, Ahmed M., Hassenian, Aboul Ella, Oliva, Diego, 2019. An improved fast fuzzy c-means using crow search optimization algorithm for crop identification in agricultural. ISSN 0957-4174 Expert Syst. Appl. 118, 340–354. https://doi.org/10.1016/j.eswa.2018.10.009.
Arias-Requejo, Desirée, Pulido, Belarmino, Keane, Marcus M., Alonso-González, Carlos J., 2023. Clustering and deep-learning for energy consumption forecast in smart buildings, 1 IEEE Access (99), 1. https://doi.org/10.1109/ACCESS.2023.3331329.
Azeem, Abdul, Ismail, Idris, Sheeraz Mohani, Syed, Usman Danyaro, Kamaluddeen, Hussain, Umair, Shabbir, Shahroz, Zaman Bin, Rahimi, 2025. Jusoh, Mitigating concept drift challenges in evolving smart grids: an adaptive ensemble LSTM for enhanced load forecasting. Energy Rep. 13, 1369–1383.
Balavand, Alireza, Kashan, Ali Husseinzadeh, Saghaei, Abbas, 2018. Automatic clustering based on crow search algorithm-kmeans (CSA-Kmeans) and data envelopment analysis (DEA). Int J. Comput. Intell. Syst. 11, 1322–1337. https://doi.org/10.2991/ijcis.11.1.98.
Bartusch, Cajsa, Alvehag, Karin, 2014. Further exploring the potential of residential demand response programs in electricity distribution. ISSN 0306-2619 Appl. Energy 125, 39–59. https://doi.org/10.1016/j.apenergy.2014.03.054.

Cuevas, Erik, Barocio, Emilio, Conde, Arturo, 2019. Clustering Representative Electricity Load Data Using a Particle Swarm Optimization Algorithm. In: Metaheuristics Algorithms in Power Systems. Studies in Computational Intelligence, 822. Springer, Cham. https://doi.org/10.1007/978-3-030-11593-7_8.

Eskandarnia, Elham, Al-Ammal, Hesham M., Ksantini, Riadh, 2022. An embedded deep-clustering-based load profiling framework. ISSN 2210-6707 Sustain. Cities Soc. 78 (2022), 103618. https://doi.org/10.1016/j.scs.2021.103618.

Funde, Nitesh A., Dhabu, Meera M., Paramasivam, Aarthi, Deshpande, Parag S., 2019. Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data. ISSN 2210-6707 Sustain. Cities Soc. 46, 101415. https://doi.org/10.1016/j.scs.2018.12.043.

Hayn, Marian, Bertsch, Valentin, Fichtner, Wolf, 2014. Electricity load profiles in Europe: the importance of household segmentation. ISSN 2214-6296 Energy Res. Soc. Sci. 3, 30–45. https://doi.org/10.1016/j.erss.2014.07.002.

Henriques, Lucas, Castro, Cecilia, Prata, Felipe, Leiva, Víctor, Venegas, René, 2024. Modeling residential energy consumption patterns with machine learning methods based on a case study in Brazil. Mathematics 12, 1961. https://doi.org/10.3390/math12131961.

Ikotun, Abiodun M., Ezugwu, Absalom E., Abualigah, Laith, Abuhaija, Belal, Heming, Jia, 2023. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. ISSN 0020-0255 Inf. Sci. 622, 178–210. https://doi.org/10.1016/j.ins.2022.11.139.

Jeong, Hyun Cheol, Jang, Minseok, Kim, Taegon, Joo, Sung-Kwan, 2021. Clustering of load profiles of residential customers using extreme points and demographic characteristics. Electronics 10 (3), 290. https://doi.org/10.3390/electronics10030290.

Kaur, Ramanpreet, Gabrijelčič, Du.šan, 2022. Behavior segmentation of electricity consumption patterns: a cluster analytical approach. ISSN 0950-7051 Knowl. Based Syst. 251 (2022), 109236. https://doi.org/10.1016/j.knosys.2022.109236.

Khalid, Muhammad, 2024. Smart grids and renewable energy systems: Perspectives and grid integration challenges. ISSN 2211-467X Energy Strategy Rev. 51, 101299. https://doi.org/10.1016/j.esr.2024.101299.

Kumar, Abhimanyu, Mallipeddi, Rammohan, 2024. A deep clustering framework for load pattern segmentation. ISSN 2352-4677 Sustain. Energy Grids Netw. 38, 101319. https://doi.org/10.1016/j.segan.2024.101319.

Lakshmi, K., Visalakshi, N.K., Shanthi, S., 2018. Data clustering using Kmeans based on crow search algorithm. Sadhana 43, 190. https://doi.org/10.1007/s12046-018-0962-3S.

Michalakopoulos, Vasilis, Sarmas, Elissaios, Papias, Ioannis, Skaloumpakas, Panagiotis, Marinakis, Vangelis, Doukas, Haris, 2024. A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs. ISSN 0306-2619 Appl. Energy 361, 122943. https://doi.org/10.1016/j.apenergy.2024.122943.

Nasiri, Jhila, Modarres Khiyaban, Farzin, 2018. A whale optimization algorithm (WOA) approach for clustering. Cogent Math. Stat. 5 (1).

Nystrup, Peter, Madsen, Henrik, Blomgren, Emma M.V., de Zotti, Giulia, 2021. Clustering commercial and industrial load patterns for long-term energy planning. ISSN 2666-9552 Smart Energy 2, 2021. https://doi.org/10.1016/j.segy.2021.100010.

Ofetotse, Eng L., Essah, Emmanuel A., Yao, Runming, 2021. Evaluating the determinants of household electricity consumption using cluster analysis. ISSN 2352-7102 J. Build. Eng. 43, 102487. https://doi.org/10.1016/j.jobe.2021.102487.

Palaniappan, Somasundaram, Karuppannan, Sundararaju, Velusamy, Durgadevi, 2024. Categorization of Indian residential consumers electrical energy consumption pattern using clustering and classification techniques. ISSN 0360-5442 Energy 289, 129992. https://doi.org/10.1016/j.energy.2023.129992.

Panda, Subhasis, Mohanty, Sarthak, Rout, Pravat Kumar, Sahu, Binod Kumar, Parida, Shubhranshu Mohan, Samanta, Indu Sekhar, Bajaj, Mohit, Piecha, Marian, Blazek, Vojtech, Prokop, Lukas, 2023. A comprehensive review on demand side management and market design for renewable energy support and integration. ISSN 2352-4847 Energy Rep. 10, 2228–2250. https://doi.org/10.1016/j.egyr.2023.09.049.

Rajabi, Amin, Eskandari, Mohsen, Ghadi, Mojtaba Jabbari, Li, Li, Zhang, Jiangfeng, Siano, Pierluigi, 2020. A comparative study of clustering techniques for electrical load pattern segmentation. ISSN 1364-0321 Renew. Sustain. Energy Rev. 120, 109628. https://doi.org/10.1016/j.rser.2019.109628.

Sandoval Guzmán, Betsy, Barocio Espejo, Emilio, Elser, Miriam, Korba, Petr, Segundo Sevilla, Felix Rafael, 2024. A hybrid clustering approach for electrical load profiles considering weather conditions based on matrix-tensor decomposition. ISSN 2352-4677 Sustain. Energy Grids Netw. 38, 101326. https://doi.org/10.1016/j.segan.2024.101326.

Soppari, Kavitha, Chandra, N.Subhash, 2020. Development of improved whale optimization-based FCM clustering for image watermarking. ISSN 1574-0137 Comput. Sci. Rev. 37, 100287. https://doi.org/10.1016/j.cosrev.2020.100287.

Sun, Mingyang, Wang, Yi, Teng, Fei, Ye, Yujian, Strbac, Goran, Kang, Chongqing, 2019. Clustering-Based Residential Baseline Estimation: A Probabilistic Perspective. IEEE Trans. Smart Grid 10 (6), 6014–6028. https://doi.org/10.1109/TSG.2019.2895333.

Wang, Jianxiao, Gao, Feng, Zhou, Yangze, Guo, Qinglai, Tan, Chin-Woo, Song, Jie, Wang, Yi, 2023. Data sharing in energy systems. ISSN 2666-7924 Adv. Appl. Energy 10, 100132. https://doi.org/10.1016/j.adapen.2023.100132.

Wang, Li, Yang, Yumeng, Xu, Lili, Ren, Ziyu, Fan, Shurui, Zhang, Yong, 2024. A particle swarm optimization-based deep clustering algorithm for power load curve analysis. ISSN 2210-6502 Swarm Evolut. Comput. 89 (2024), 101650. https://doi.org/10.1016/j.swevo.2024.101650.

Wang, Yi, Chen, Qixin, Kang, Chongqing, Xia, Qing, 2016. Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications. Sept. 2016 IEEE Trans. Smart Grid 7 (5), 2437–2447. https://doi.org/10.1109/TSG.2016.2548565.

Wen, Hanguan, Liu, Xiufeng, Yang, Ming, Lei, Bo, Xu, Cheng, Chen, Zhe, 2024. A novel approach for identifying customer groups for personalized demand-side management services using household socio-demographic data. ISSN 0360-5442 Energy 286, 129593. https://doi.org/10.1016/j.energy.2023.129593.

Zhang, Xiaohai, Ramírez-Mendiola, José Luis, Li, Mingtao, Guo, Liejin, 2022. Electricity consumption pattern analysis beyond traditional clustering methods: a novel self-adapting semi-supervised clustering method and application case study. ISSN 0306-2619 Appl. Energy 308, 118335. https://doi.org/10.1016/j.apenergy.2021.118335.