A Proposed Intelligent Model with Optimization Algorithm for Clustering Energy Consumption in Public Buildings

Ahmed Abdelaziz[1], Vitor Santos[2], Miguel Sales Dias[3]
Nova Information Management School, Universidade Nova de Lisboa, 1070-312 Lisbon, Portugal[1, 2]
Information System Department, Higher Technological Institute, HTI, Cairo 44629, Egypt[1]
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, 1649-026 Lisbon, Portugal[3]

***Abstract*—Recently, intelligent applications gained a significant role in the energy management of public buildings due to their ability to enhance energy consumption performance. Energy management of these buildings represents a big challenge due to their unexpected energy consumption characteristics and the deficiency of design guidelines for energy efficiency and sustainability solutions. Therefore, an analysis of energy consumption patterns in public buildings becomes necessary. This reveals the significance of understanding and classifying energy consumption patterns in these buildings. This study seeks to find the optimal intelligent technique for classifying energy consumption of public buildings into levels (e.g., low, medium, and high), find the critical factors that influence energy consumption, and finally, find the scientific rules (If-Then rules) to help decision-makers for determining the energy consumption level in each building. To achieve the objectives of this study, correlation coefficient analysis was used to determine critical factors that influence on energy consumption of public buildings; two intelligent models were used to determine the number of clusters of energy consumption patterns which are Self Organizing Map (SOM) and Batch-SOM based on Principal Component Analysis (PCA). SOM outperforms Batch-SOM in terms of quantization error. The quantization error of SOM and Batch-SOM is 8.97 and 9.24, respectively. K-means with a genetic algorithm were used to predict cluster levels in each building. By analyzing cluster levels, If-Then rules have been extracted, so needs that decision-makers determine the most energy-consuming buildings. In addition, this study helps decision-makers in the energy field to rationalize the consumption of occupants of public buildings in the times that consume the most energy and change energy suppliers to those buildings.**

Keywords*—Energy consumption in public buildings; self-organizing map; K-means; genetic algorithm; principal component analysis

I. INTRODUCTION

The growing construction sector is struggling to cope with the increasing demand for energy despite efforts to develop sustainable buildings [1]. Therefore, improved energy efficiency and analysis of energy consumption patterns in buildings become necessary. This unveils the importance of understanding and classifying energy consumption patterns in buildings. For example, the more precise and pragmatic energy consumption profiles are computed, the better building energy quality evaluation becomes [2]. Energy consumption depends on various factors such as building characteristics, energy prices, and climate conditions, amongst others [3]. Therefore, aiming to classify the energy consumption of buildings requires advanced computational intelligent approaches, particularly adopting the latest trends in machine learning, such as deep learning techniques, which exploit familiarity from historical data and can support decision-makers in the energy domain, creating a basis for styling new power allocation dispositions, particularly for public buildings areas [4].

Energy consumption in public buildings merits particular attention since it accounts for a large share of final energy consumption if we look at 2019 figures from the OECD (Organization for Economic Cooperation and Development) countries, which reached 27% in the European Union [5]. For example, public buildings consume nearly one-third of all electricity in Portugal, increasing by 35% from 1995 to 2019 [6]. Understanding this consumption means solving a complex problem involving physical, technological, and performance characteristics of the dwelling, the status of the demography, socio-economic factors, climate and weather conditions, and the behavior of the building's occupants [7]. Therefore, academic research in European countries, notably Portugal, needs help understanding the energy consumption patterns of public buildings.

In the past, we can find several data mining and machine learning techniques that have been used for energy consumption classification. Among those, clustering is considered one of the most applied techniques [8]. Clustering comprises splitting objects with similar styles into various groups [9]. Researchers have provided many manuscripts on classifying energy consumption into discrete levels. For instance, Gouveia [10] discovered electricity consumption profiles in households through clusters by combining smart meters and door-to-door surveys. His study used hierarchical clustering to divide household profiles and obtained three clusters. Hernandez et al. [11] presented a study to classify daily load curves in industrial parks by using a self-organizing map and k-means to determine the number of clusters. Ford and Siraj [12] presented a fuzzy c-means clustering to classify smart meter electricity consumption data to similar groups. Hodes et al. [13] presented a study to classify residential houses with similar hourly electricity using the k-means algorithm. Azaza [14] presented a method to find the most responsible energy consumers in the peak hour by using

hierarchical clustering and a self-organizing map. Al-Jarrah et al. [15] presented a method to discover power consumption in buildings using multi-layered clustering. K-mean has been utilized to partition power consumption profiles. Then, the authors discover different patterns of power consumption profiles. Furthermore, Cai et al. [16] presented a hybrid method to divide the electricity consumption of an entire region into various levels by using k-means with particle swarm optimization. To extract behavior in daily electricity consumption in households, Nordahl et al. [17] utilized the centroids of the generated clusters.

Most research focuses on the total energy consumption in different buildings by reviewing the analyzed literature. However, other factors that affect energy consumption, such as the consumption behavior of the occupants of these buildings at peak time or during empty hours (00h00-02h00; 06h00-08h00; 22h00-00h00), were neglected. In this paper, we follow this literature trend, and we propose an intelligent computing model capable of automatically classifying energy consumption into discrete levels, such as low, medium, and high. In this model, we can discover the different consumption patterns of public buildings across the country and visualize such patterns at different levels of the geographical organization and during the year, showing the different districts, municipalities, and parishes in which, the energy consumption is low, medium or high, in a certain period, helping to direct the occupant's behavior in such public buildings. The contribution of our paper has four dimensions:

1) Development of a novel hybrid model for classifying energy consumption in buildings (with an application to public buildings): the SOM, PCA, K-means (KM), and Genetic algorithm (GA), referred to as the SPKG model.

2) Evaluation of the performance and precision of the proposed model is trained and tested with real big data of energy consumption of public buildings in Portugal, collected in the years 2018 and 2019 (81 260 public buildings of 238 Portuguese cities).

3) Correlation coefficient analysis, to understand the relationship between the factors influencing energy consumption in buildings and determine the optimal factors amongst them.

4) A clustering and classification model of energy consumption levels in buildings, featuring a comparison between SOM and Batch-SOM based on PCA, in terms of quantization error, to select the optimal model between them and determine the optimal number of clusters in energy consumption in buildings. In our approach, we use the PCA algorithm to optimize SOM's weights, which helps to enhance the SOM model's fitting ability. Moreover, GA was used to find the optimal initial centroids in KM. This last technique predicts the cluster label in each building.

The paper is organized as follows. Section II presents our related work. In Section III, we present research questions and methodology: intelligent computing model. Section IV presents experimental results and discussion. Finally, in Section V, we conclude and suggest lines for further work.

II. Related Work

Putting public buildings that use the same amount of energy into similar groups is a key part of figuring out how much better or worse one building performs compared to similar buildings, like peers in the same group. Therefore, it is imperative to correctly identify these divisions to help the decision-maker in energy on three essential points: rationalizing the occupants of public buildings that consume much energy, determining the required amount of energy expected in the coming years, and changing energy providers in public buildings.

The most common methods for analyzing energy consumption in buildings are the different types of clustering methods [17]. Previous research analyzed raw meter data and used that data to represent energy consumption patterns using traditional statistical methods such as regression analysis and others [18,19]. The two most used clustering methods are K-means and Hierarchical clustering, which provide the most energy for occupancy and load forecasting [20 - 22]. Other machine learning methods are used to predict power consumption and loads, such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and K-Shape and other clustering methods [21 - 25].

Some important studies focused on finding an optimal way to understand occupancy schedules and user demand in different buildings, using anomaly detection and clustering methods [8, 26, 27, 28]. In anomaly detection, occupancy behavior is often used to design strategies that fit dynamic needs, user conditions, and interior space [9]. In addition, it helps design future buildings with a strategy that conserves wasted energy [29].

Other studies focus on measuring electricity consumption in buildings with their various activities using different methods of machine learning (i.e., decision trees [30] and stochastic frontier analysis [31]). These studies used intelligent methods to determine the different forms of electrical loads. In addition, it has been applied to more than 3000 residential and non-residential buildings.

Literature efforts are being conducted to find an intelligent computing model for clustering energy consumption in buildings using different factors that depend on the state of those buildings at different times and discovering the energy consumption patterns of occupants in such buildings [6]. Identifying and clustering the energy load patterns of occupants in public buildings based on such consumption profiles can be beneficial to stakeholders who aim to improve the energy efficiency of buildings effectively. K-means clustering is one of the methods used in the analyzed literature. However, it shows several issues. For example, K-means cannot group data where the groups are of varying volume and density [32]. Secondly, centroids can be pulled by outliers [33]. Finally, K-mean assumes that all variables have the same variance [34]. Consequently, our work tries to find a more accurate clustering method to overcome the limitations of the K-means clustering approach.

By analyzing previous works, we noticed the inability of these studies to find data that represents the occupants'

behavior of buildings at different times. Also, some papers use traditional statistical models like regression analysis and common clustering methods without paying attention to how well these models divide energy use into similar groups. The inaccurate classification of energy consumption leads to several ways to mislead the decision-maker: (1) the inability to find buildings that have high energy consumption; (2) the lack of anticipation of the energy required to cover the needs of public buildings adequately, and finally, the inability to identify the best energy providers. An energy consumption dataset was collected from Portuguese public buildings in 2018 and 2019 to remedy these shortcomings. This dataset was used to train and test a hybrid intelligent computing model to cluster energy consumption in public buildings. We believe that the decision-maker in the energy field can rely on this model to make sound decisions regarding energy consumption in public buildings and energy providers. In addition, there is a clear difference between the data used in this study and the data used in previous studies in terms of data quality and size, as the quality of detailed data on electricity consumption at various times of the day and the large size of data compared to previous studies. Moreover, in the preprocessing section, recent hybrid intelligent techniques such as isolation forest and interpolation methods were used that were not used in the same form and accuracy in previous studies. Furthermore, public buildings with high energy consumption have been determined in detail compared to previous works. Finally, recent hybrid intelligent techniques such as KM with GA were used to predict cluster labels. All these features make this study distinct from the rest of the previous studies.

III. Research Questions and Methodology

To properly frame our research, we raised the following research question:

- RQ1: What types of data sources and critical factors can be adopted to profile the energy consumption of buildings?
- RQ2: Which intelligent computing technique(s) can be adapted to identify the number of clusters in the given energy consumption dataset?
- RQ3: What general rules can be extracted to help the decision-maker rationalize energy consumption for public buildings?
- RQ4: What are the different and essential patterns discovered in the given energy consumption dataset?

To tackle the raised research question, we propose a hybrid approach (see Fig. 1), with a mixture of machine learning and optimization techniques, namely, (SOM [9]), (PCA [4]), KM, and GA [4], referred to as the SPKG model, able to discover different energy consumption patterns in buildings, with a proof of concept of its application to public buildings in Portugal.

In this section, we describe in detail our proposed model, which is composed of four main phases, as depicted in Fig. 1, namely:

- Data Collection: Our collected data includes energy consumption and building characteristics, such as (but not limited to): unique energy point of delivery ID, address of such a point of delivery, contracted electrical power, electricity consumption, and billing data with the month of consumption. The objective of this phase is to ensure that the units of measurement are consistent, that the sampling rates are adequate, that the time series is the same and synchronized over time, and that there were no structural changes during the data collection period.
- Data Analysis and Pre-Processing: In this phase, we analyze the data in detail and, if needed, transform it to expose its information content better. We adopt different mathematical techniques, namely, outlier removal with Isolation Forest (ISF) [35] and polynomial interpolation [31].
- Feature Engineering: In this phase, we find the optimal variables used to discover energy consumption patterns in (public) buildings, adopting a coefficients analysis approach [35].
- Clustering Analysis: In this phase, we fine-tune, apply, and evaluate our SPKG hybrid machine learning model, which can find clusters (each cluster corresponds to an energy consumption profile of buildings), and cluster the energy consumption profile in a particular building. Intelligent computing techniques, such as SOM and KM, are assessed and compared for automatic cluster discovery and the definition and classification of energy consumption behavior in (public) buildings.
- Clustering Results: In this phase, we tried to find three important results: generate energy consumption rules, determine the final number of clusters, and determine municipalities and Portuguese building activities that consume high, medium, and low energy consumption.

A. Data Collection

The data used in this study consists of the energy consumed in public buildings in Portugal, with the following characteristics: monthly data collected during the years of 2018 and 2019 in 77 996 buildings of various public sectors and 238 cities, reaching 2 775 082 records. After removing the records related to public lighting (since it is outside the scope of our study) and removing buildings that do not contain consumption data for the full observed period of 24 months, the number of records used in this study reached 1 222 695, corresponding to 26 624 public buildings.

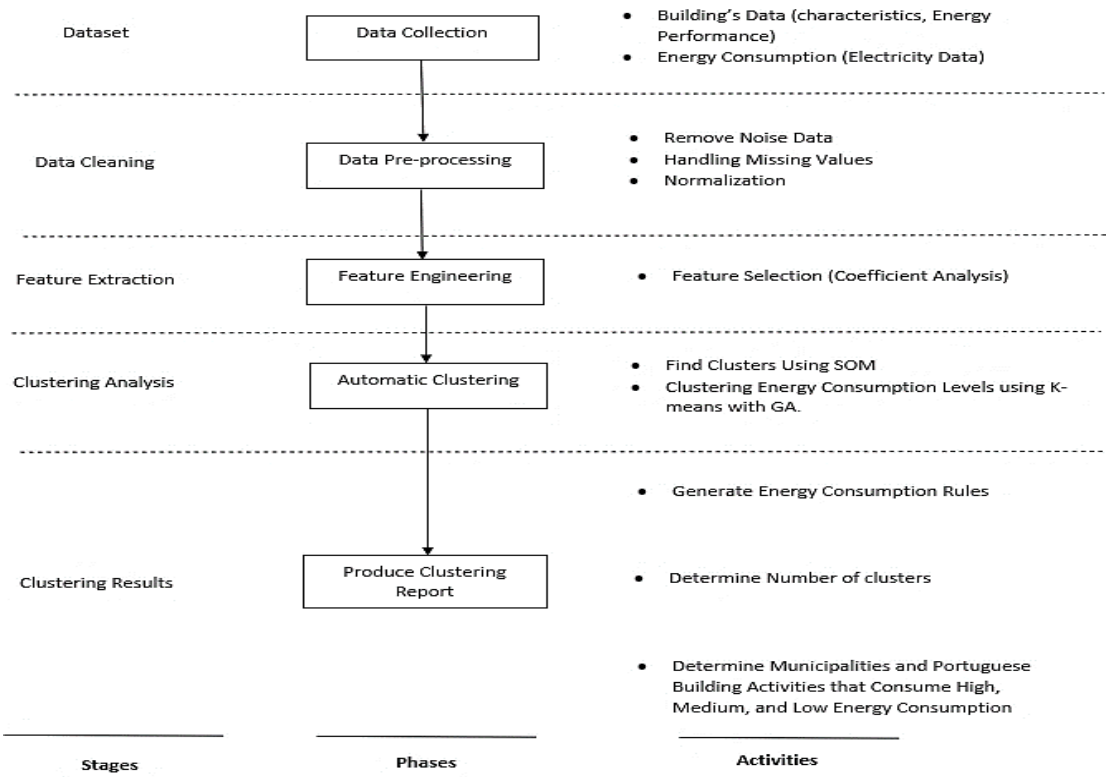

Fig. 1. Our proposed SPKG model for discovering energy consumption in public buildings

As mentioned, the dataset used in this study consists of two parts: the building characteristics and the actual energy consumption in these buildings (see Table I). Building characteristics include several attributes, namely:

- Unique energy is the point of the delivery ID of each building.
- Details of each building.

Energy consumption data, in our data set, includes:

- Actual active energy consumption in public buildings.
- Super empty: Active energy in the period 02h00-06h00 AM.
- Empty: Active Energy in the periods 00h00-02h00, 06h00-08h00, and 22h00-00h00.
- Outside empty: Lighting and plug loads that cannot be turned off.
- Peak: Active Energy in the periods 09h00-10h30 and 18h00-20h30.
- Full: Active Energy in the periods 08h00-09h00, 10h30-18h00, and 20h30-22h00.
- Total energy consumption: Active and Reactive Energy, where reactive energy is electrical energy that is stocked rather than transformed to some other form of energy and thus not "used" or "consumed."

TABLE I. DATASET DIMENSIONS OF ENERGY CONSUMPTION IN PUBLIC BUILDINGS

Dataset Dimensions	Attribute Name	Description
Characteristics of buildings	Unique Energy Point Delivery ID	The ID of each public building
	Business Partner	Identification of the institution that owns or rents the building.
	Building Address	Address of each building
	Municipality	City Location of each building
	Installation Type	Details of the electrical installation of each building
	Contracted Power	Power in MW has been agreed upon with the operator for each building.
	Year/Month	Consumption date
Energy consumption (Active Energy (KWh))	Simple	Total of active energy
	Super Empty	Active Energy (02h00-06h00)
	Empty	Active Energy (00h00-02h00; 06h00-08h00; 22h00-00h00)
	Outside Empty	Lighting and plug loads that cannot be turned off
	Peak	Active Energy (09h00-10h30; 18h00-20h30)
	Full	Active Energy (08h00-09h00; 10h30-18h00; 20h30-22h00)
	Total	Total of energy consumption (Active plus Reactive Energy)

B. Data Preprocessing

Outlier detection and missing value imputation are the two primary processes in the data preprocessing for missing data uti lizing the isolation forest and interpolation method. Here is a g eneral description of the procedure [36 - 42]:

Step 1: Outlier Detection using Isolation Forest

- Determine which features (columns) are missing data.
- For each characteristic, distinguish between the entire d ata (rows without missing values) and the partial data (r ows with missing values).
- Using the whole data for each feature, isolate outliers us ing the isolation forest algorithm. A well-liked approach for anomaly identification called isolatio n forest isolates outliers by building random forests and calculating the typical number of splits required to isola te a data point.
- Establish a threshold to help you spot outliers. This may depend on the number of splits or a predetermined cuto ff point.

Step 2: Missing Value Imputation

- Use interpolation techniques to impute the missing valu es for the features that have missing data. Interpolation i s a method that calculates the missing values from the d ata points already there.
- There are numerous interpolation techniques, including linear interpolation, polynomial interpolation, and methods tailored to time series, including forward-fill and backward-fill.

Step 3: Combine Outlier Detection and Imputation

- We chose to keep outliers in the data after identifying th em with the isolation forest.
- Use the selected polynomial interpolation technique to f ill in the data gaps for the missing values.

C. Feature Selection

This section aims to find the critical variables or factors in our energy consumption dataset. To overcome this problem, we used the T-test correlation coefficient. This statistical technique is used in literature to detect if two factors/variables are significant [43]. It can be helpful in our study. In our dataset, looking at pairwise correlations between the various variables (or factors) may propose a causal relation between two factors that we can investigate further. Eq. (1) computes the T-test value by assuming no correlation with ρ = 0, where, P refers to that; there is no relationship between variables [44].

$$t = r\sqrt{\frac{n-2}{1-r^2}} \quad (1)$$

In (1), *n* refers to the instances, and *r* represents the correlation coefficient of the energy consumption dataset. The importance of relevance is expressed in probability levels: *p* (e.g., significant at p = 0.05). The degree of freedom for entering the t-distribution is *n* – 2. If the *t* value is less than the critical value (CV) at a 0.05 significant level, the factor is not essential and is avoided [44].

In Algorithm 1, we build the correlation coefficients using the training dataset. In Steps 1 to 4, we calculate the correlation coefficients between the proposed factors. Step 6 to step 7 computes significant values by using the T-test. Finally, step 8 to step 10 finds the final list of energy consumption factors.

Algorithm 1: Feature Selection Algorithm

Input: S(F_1, F_2, ……., F_k, F_c) // a training data set
Output: S_{best} // the selected feature set
1. begin.
2. **For** I to k **do**
3. r = compute correlation coefficient (F_i, F_c)
4. **End**
// let P = 0.05 significant level
let P = 0 // assuming there is no significant correlation
5. **For** I to k **do**
6. t = compute significant values (r,p) for F_i // Eq.4
7. **If** t > CV // critical value
8. S_{list} = CV
9. S_{best} = S_{list}
10. **End**
11. **End**
12. **Return** S_{best}

D. Finding the Number of Clusters

To determine the optimal number of clusters in energy consumption data, we used three literature methods: Self-Organizing Map (SOM), the Elbow method, and the Bouldin & Davis method [14, 15]. These methods have been used in prior studies to find the optimal number of clusters, notably in energy consumption in buildings.

1) Self–Organizing Map: SOM is a specific class of neural networks utilized broadly as a clustering and visualization instrument in exploratory information analysis [45]. The main objective of SOM is to convert a complex high-dimensional discrete input space into a less low-dimensional discrete yield space by keeping the topology within the information but not the real separations [46, 47]. An unsupervised learning calculation employs a basic heuristic strategy for finding covered-up non-linear structures in high dimensional information [46].

The SOM method is adopted because it deals with big data accurately and effectively [45]. Contrary to the other methods mentioned, it better deals with small and medium-sized data [46]. Therefore, the SOM method determined the number of clusters in the energy consumption dataset. SOM is composed of three main processes: competition, cooperation, and adaptation [45].

The SOM network is composed of two layers, the input, and the output layer, as shown in Fig. 2. Each input variable is shown using an m-dimensional input vector [48]. In the output layer, the number of nodes indicates the most extreme number of clusters and impacts the precision and generalization capability of the SOM [45]. The arrangement of the SOM

begins with the initialization of the weight vectors [46]. Then, weights are joined that interface the input nodes to the output nodes and are overhauled through learning. Finally, to discover the best match unit (BMU), the spaces between an input (x) and the weight vectors (wi) of the SOM are calculated by using various measurement methods, such as [47,49]:

- Manhattan distance.
- Chebyshev distance.
- Euclidean distance.
- Mahala Nobis distance
- Vector product, among other methods.

Euclidean distance is an approved measure in most scientific papers [47], as shown in Eq. (2):

$$di\,(t) = \|x\,(t) - wi\,(t)\| \quad (2)$$

At the finish of the propinquity matching method (Determine the similarity between points in the dataset), the most excellent matching unit c at repetition t is identified by the minimum distance [45].

$$c\,(t) = \arg\min_i \|di\,(t)\| \quad (3)$$

By analyzing the weight vector wi(t) of the winning neuron, i at iteration t, the overhauled weight vector wi(t + 1) at iteration (t + 1) is determined by Using a discrete-time formalism in Eq. (4) [47].

$$wi(t+1) = wi(t) + \alpha(t)\,[x\,(t) - wi\,(t)] \quad (4)$$

The weights (α) adjustment rate diminishes away from the winning node regarding the Spatio-temporal decay function [46].

$$h_{ci}\,(t) = \exp\left(-(d*d)^{ci} / 2\sigma^2\,(t)\right) \quad (5)$$

where,

- d is the lateral distance between the winning neuron c and the excited neuron i.
- Sigma is the effective width or radius of the neighborhood at iteration i.

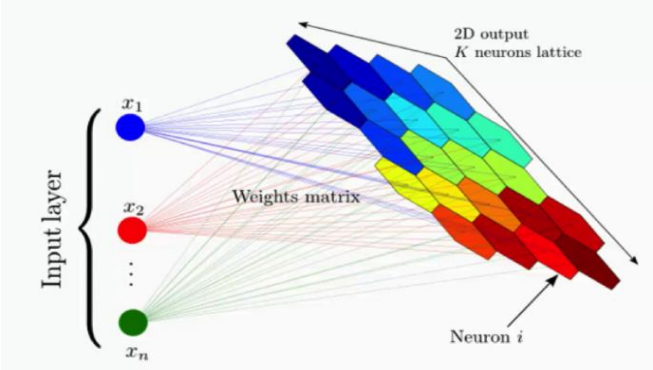

Fig. 2. Structure of SOM [45]

In Algorithm 2, we implemented a SOM network based on energy consumption data to determine the optimal number of clusters. Firstly, we have identified the lattice space of the 10 × 10, set weights based on random weights and PCA weights, set iterations from 100 to 1000. Secondly, pick the random points in energy consumption data, then find the best match point based on Eq. (5), set learning rate = 0.5, set neighborhood function = triangle, compute neighborhood distance weight matrix and modify SOM weight matrix, and finally, repeat from the step of picking random point (z) until the maximum number of iterations is reached.

Algorithm 2: Main Idea of the SOM Network Training

Input: ECD ← the energy consumption data.
Output: USOM ← U-matrix of SOM network.
1. β← initialize lattice nodes.
2. Ω← initialize weight vectors.
3. N ← Iteration count.
4. **For** i ← 1 to N do
5. z ← picks a random point in ECD.
6. c ← β closest to z.
7. move the weight vector of c closer to z.
8. move the weight vectors of the neighbors of c slightly closer to z.
9. **End**
Return USOM

2) Elbow method: We can plot the curve indicating the average inner per cluster sum of squared error (SSE) distance vs the number of clusters to discover a visual "elbow", the ideal number of clusters. The average inner whole of squares is the average distance between focuses interior of a cluster [11], as shown in Eq. (6).

$$_{k} = \sum_{r=1}^{k} \left(\frac{1}{n_r} + D_r \right) \quad (6)$$

Where:

- k is the number of clusters,
- n_r is the number of points in cluster r.
- D_r is the sum of distances between all points in a cluster.

3) Bouldin and davis method: In Davis and Bouldin (DB), the score is characterized as the average similitude degree of each cluster with its most identical cluster. The similitude is the proportion of within-cluster separations to between-cluster separations. In this way, clusters that are more distant separated, and less scattered will result in a distant better score. The least score is zero, with lower values indicating superior clustering [13], as shown in Eq. (7) and (8) [16].

$$DB(c) = \frac{1}{k} \sum_{i=1}^{k} \left(max_{j \le k, j \ne i}\, D_{ij} \right), \; k = |c| \quad (7)$$

D_{ij} is the "within-to-between cluster distance ratio" for the ith and jth clusters.

$$D_{ij} = \frac{di^{-} + dj^{-}}{dij} \quad (8)$$

where, di^{-} is the average distance between every data point in cluster i and its centroid, similar for dj^{-}. d_{ij} is the Euclidean distance between the centroids of the two clusters.

E. K-Means with GA

GA is a research process inspired by Charles Darwin's theory of naturalist evolution. It is a process to select the fittest individuals to reproduce to create offspring of the next generation. GA is good at dealing with multiple points and is good in noisy environments; therefore, it quickly helps implement any fitness function such as Euclidean distance in the energy consumption dataset. GA was used to find the optimal centroids in KM to speed up convergence between energy consumption points through three fitness functions which are Euclidean distance (ED), Manhattan distance (MD), and Cosine distance (CD), as shown in formulas (2, 9, 10) [47]. Moreover, it helps to improve the accuracy of KM in our study.

MD indicates the sum of the absolute values of the differences of the coordinates. For example, if X = (E, M) and Y = (B, K), the MD between X and Y is:

$$\text{MD} = |E - B| + |M - K| \quad (9)$$

CD calculates the cosine of the angle between vectors X and Y as shown below:

$$\text{CD} = \frac{X \cdot Y}{\|X\| \, \|Y\|} \quad (10)$$

Where:

- $\|X\|$ = Euclidean norm of vector, $X = (X_1, X_2, \ldots\ldots\ldots\ldots, X_n)$.
- $\|Y\|$ = Euclidean norm of a vector, $Y = (Y_1, Y_2, \ldots\ldots\ldots\ldots, Y_n)$.

KM aims to group identical data points as one cluster and detect underlying patterns. It has many challenges. First, determine the optimal number of previously determined clusters utilizing SOM. Second, determine the optimal centroid placement in each cluster utilizing GA. Thus, KM has been used to predict cluster labels in each building in ECD. In Algorithm 3, we constructed the improved KM using SOM and GA as inputs. From step 1 to step 4, improved KM tries to find the new centroid positions in each cluster for enhancing the accuracy of predicting the cluster label in each building in ECD.

Algorithm 3: Improved KM to predict cluster label in each building

Input: K = 3, // Specify the number of clusters using SOM
Initialize σ of centroids using GA.
Output: β ← predicting cluster label in each building in ECD
1. **Repeat**
2. Assign each point to its closest centroid.
3. Compute the new centroid of each cluster.
4. **Until** the centroid positions do not change.
Return β

IV. Experimental Results and Discussion

This section comprises four sections: data pre-processing, feature selection, finding the number of clusters, and finally, k-means with GA to produce energy consumption rules. We have used Python programming and the Scikit-learn library to implement the proposed algorithms.

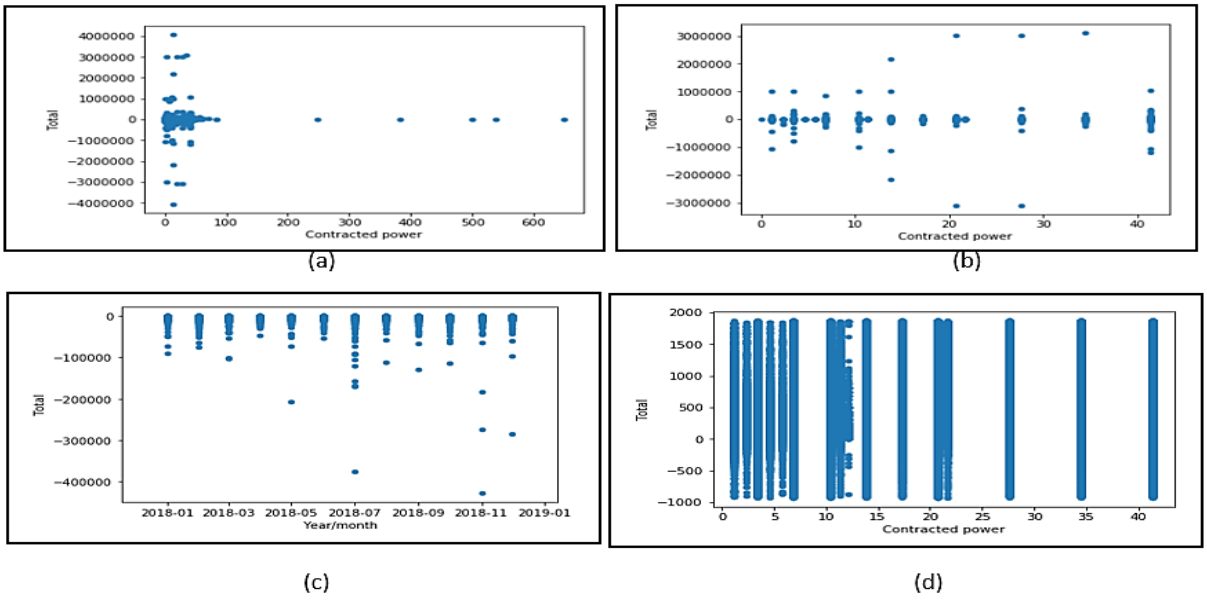

Fig. 3. Sample of data preprocessing.

A. Results of Data Preprocessing

Intelligent machine learning techniques always depend on the quality and efficiency of the dataset proposed in the study. Therefore, if the dataset provided is high quality and accurate, which helps to build and train an intelligent model with high efficiency. Furthermore, the energy consumption data is collected from a real-world environment. Therefore, it is unstructured and incomplete. Thus, we always need the pre-processing data stage to remove noise and outliers. Data pre-processing has two stages. Fig. 3 shows the steps for pre-processing the energy consumption dataset in the first stage. Initially, (a) the sample of the raw dataset was displayed in terms of contracted power (Xi) and total energy consumption (Yi); secondly, (b) Public buildings have been removed that

have several months less or more than 24 months, and public lighting buildings also have been removed because it is outside the scope of the study. Thirdly, (c) there are still public buildings that contain harmful and zero values. Fourthly, (d) outlier values have been removed using ISF, but harmful and zero values have also been removed.

After pre-processing, the final dataset was reached, which was relied upon to find the different patterns in energy consumption in public buildings. Fig. 4 shows the sample of the final data set between contracted power and total energy consumption.

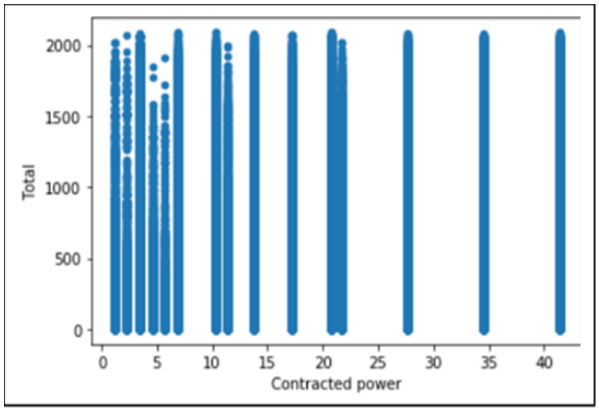

Fig. 4. Sample of Final Dataset

B. Results of Feature Selection

The aim of this section is to show the results of the T-test correlation coefficient and find the critical factors in the energy consumption dataset. Fig. 5 shows the relationships between energy consumption factors. We observed a relationship between contracted power with Full, Peak, Empty, outside empty, and total consumption, and there is also a relationship between Full and Peak. Moreover, there is a relationship between Empty and Outside Empty. Moreover, we can avoid the Super Empty factor because it contains null values in all the columns, and there is no relationship between it and all the other factors. Finally, there is a negative relationship between the Simple factor with Full, Peak, and Empty consumption.

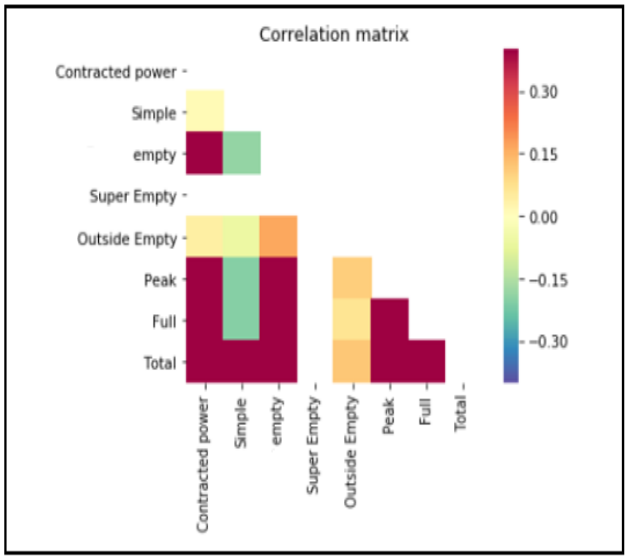

Fig. 5. The applied correlation coefficient in the energy consumption dataset.

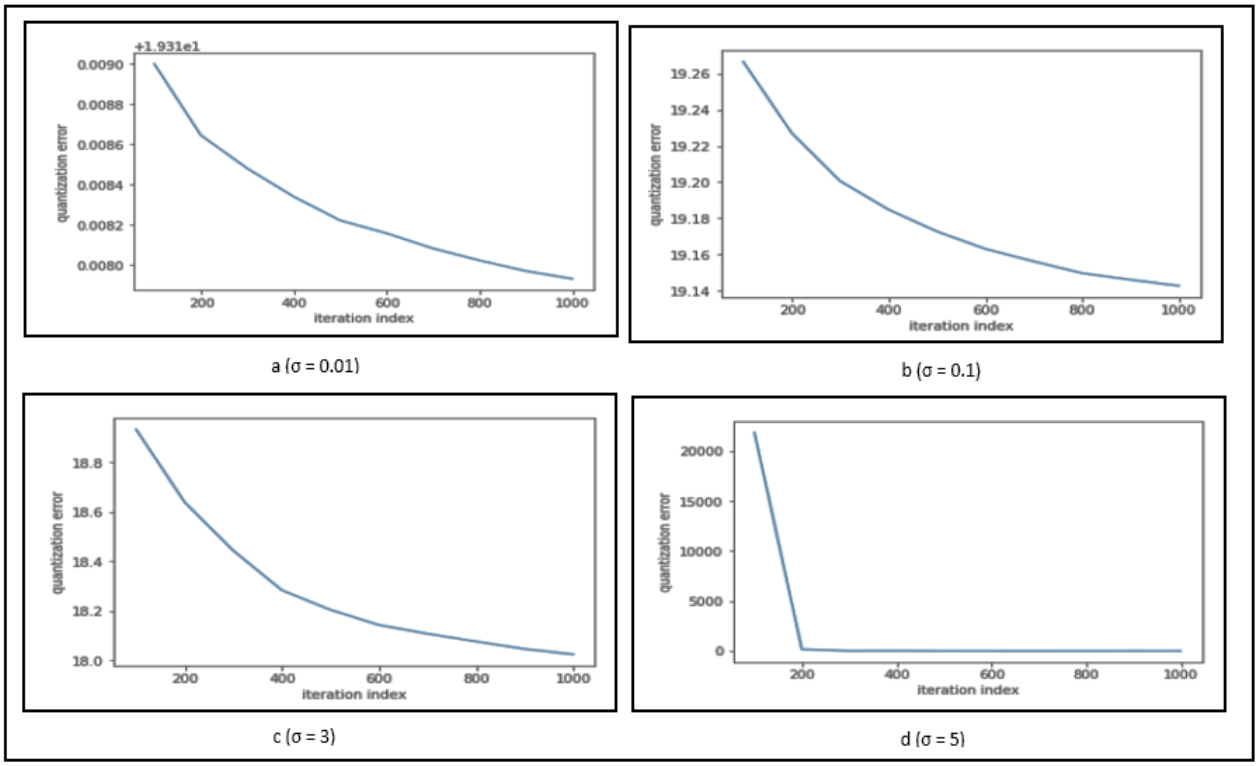

Fig. 6. q- error in random weights. Set iteration = 1000, (a) σ = 0.01, (b) σ = 0.1, (c) σ = 3 and (d) σ = 5.

C. Results of Finding Number of Clusters

This section shows the results of three methods to find an optimal number of clusters: self-organizing map, Elbow method, and Bouldin and Davis method. A comparison was made on the weights of the SOM network in two different ways, the first utilizing random weights and the second through PCA weights. We set the iterations = 1000 and, we set sigma = 0.01, 0.1, 3 and 5. By comparing random weights and PCA weights, PCA weights are better than random weights in terms of quantization error (q- error), especially in iteration = 1000 and sigma = 3, as shown in Table II and Fig. 6 and 7.

The q- error expresses the squared distance (usually the average Euclidean distance) between input data x and their corresponding so-called BMU. Thus, the QE reflects the average distance between each data vector (X) and its BMU, as shown in Eq. (11):

$$q - \text{error} = 1/N \sum_{i=1}^{N} \left\| X_i - (BMU_{(i)}) \right\| \quad (11)\ [47]$$

The q- error appeared within Table II and Fig. 6 and 7 are midpoints for all data patterns. A comparative assessment of how this quantization is changed permits us to recognize distinctive clusters, which is one of the primary purposes of utilizing these techniques.

The SOM network was trained in two different ways based on PCA weights: random training SOM (RTSOM) and batch SOM (BSOM). The batch overhaul does not require a learning rate function. Typically, profitable since it reduces the number of required parameters. PCA weights with RTSOM (PCAW-RTSOM) are better than PCA weights with BSOM (PCAW-BSOM) in terms of q- error. Q- error in PCAW-RTSOM and PCAW-BSOM is 8.97 and 9.24, respectively, as shown in Table III and Fig. 8.

TABLE II. A COMPARISON BETWEEN RANDOM WEIGHTS AND PCA WEIGHTS

SOM random weights			SOM PCA weights		
Iteration	Sigma	q- error	Iteration	Sigma	q- error
1000 0.01 19.32			1000 0.01		0.01
0.1		19.14	0.1		216.85
3		18.02	**3**		**13.14**
5		20.13	5		16.39

TABLE III. A COMPARISON BETWEEN PCAW-RTSOM AND PCAW-BSOM

Iteration	PCAW-RTSOM	PCAW-BSOM
	q- error	
100	15.26	13.86
200	14.65	13.92
300	10.69	14.14
400	10.00	12.72
500	9.50	14.91
600	10.22	10.05
700	9.77	11.81
800	9.47	11.22
900	9.32	12.31
1000	**8.97**	**9.24**

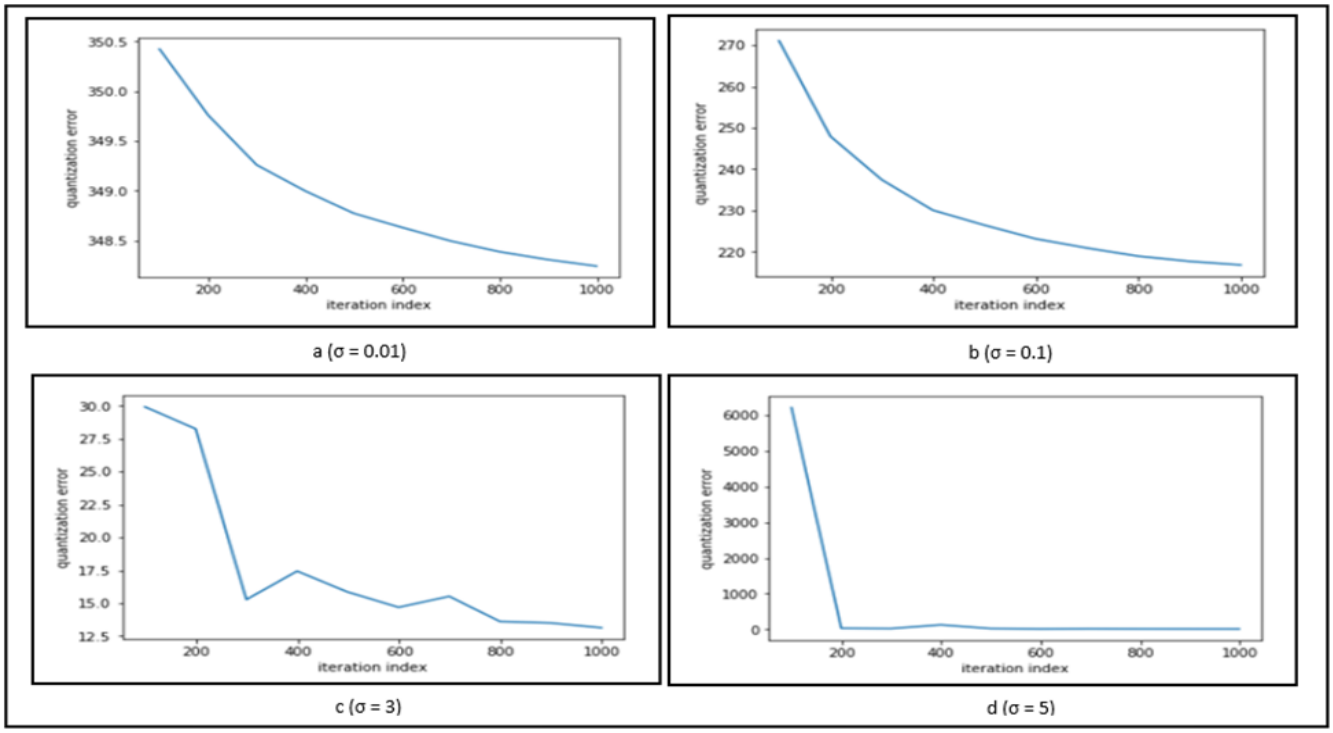

Fig. 7. q- error in PCA weights. Set iteration = 1000, (a) σ = 0.01, (b) σ = 0.1, (c) σ = 3 and (d) σ = 5.

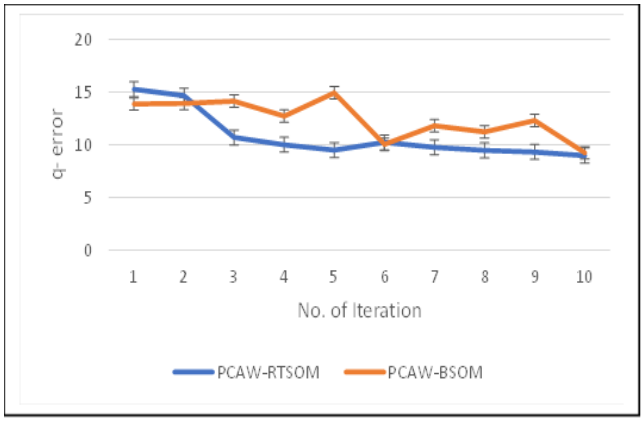

Fig. 8. A Comparison between PCAW-RTSOM and PCAW-BSOM.

Fig. 9 shows the visualization of U-matrix in PCAW-RTSOM and PCAW-BSOM. In that U-matrix, we may determine three light color areas (white color) that match the minimum values in the U-matrix and indicate three clusters in the energy consumption dataset. These areas are detached by dark blue, which matches the segregation between the clusters.

We have obtained three clusters by implementing Elbow and Bouldin & Davis method in our energy consumption dataset, as shown in Fig. 10 and 11.

The U-matrix in SOM shows the distances between the points (points represent the energy consumption dataset) on the SOM. The dark areas in that U-matrix show the areas of the map where the points are far away from each other so, which represents the segregation between the clusters, and the lighter areas show fewer distances between the points so, which means the number of clusters. The Elbow method is computed as the intermediate of the squared distances from the cluster centers of the clusters. Typically, the Euclidean formula is utilized. In Bouldin & Davis method, to obtain the intra-cluster scuttle, we compute the average distance between each vector within the cluster and its centroid, which computes the Euclidean method between the centroid of the cluster. Finally, we could determine three clusters (low, medium, high consumption) in the energy consumption dataset by analyzing the SOM network, Elbow method, and Davis-Bouldin method.

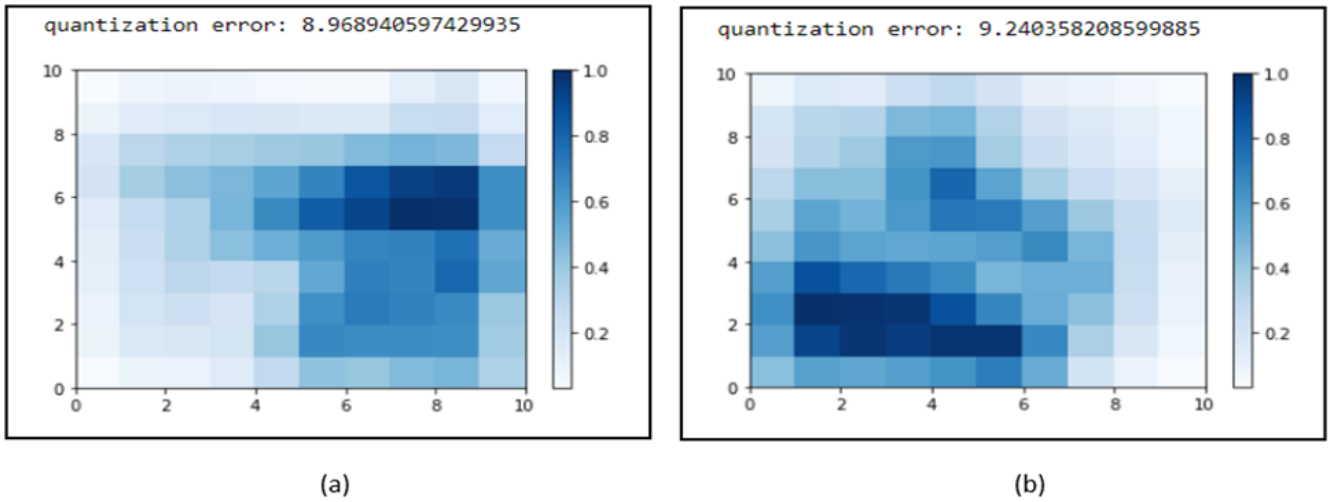

Fig. 9. A Comparison between PCAW-RTSOM and PCAW-BSOM

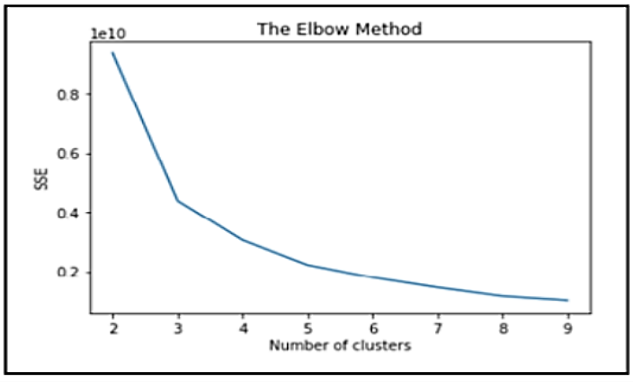

Fig. 10. Apply Elbow Method in Our Dataset

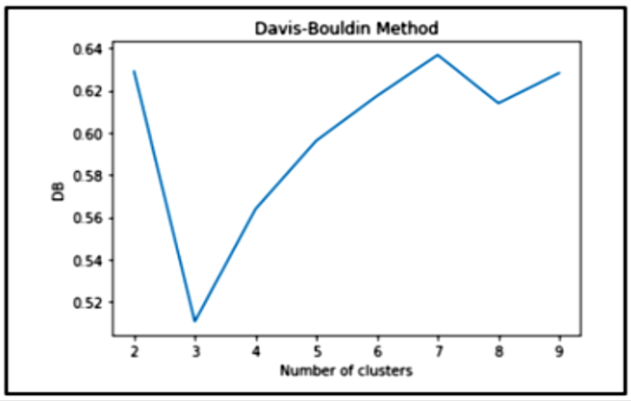

Fig. 11. Apply Davis-Bouldin Method in Our Dataset

D. K-Means with GA to Produce Energy Consumption Rules

In this section, we computed the distance between each cluster by two methods: The first method is K-means clustering with K-means++ initialization (KMCKI) and the second method is SPKG. GA has been implemented through the main parameters, as shown in Table IV. There are three methods to compute distances between clusters: ED, MD, and CD. We compared the performance between KMCKI and SPKG in terms of standard error (SE), as shown in formula 12, and standard deviation. CD with SPKG is better than all methods, as shown in Table V. Thus, this study relied on CD with SPKG to predict cluster labels in each building in ECD and detect underlying patterns.

$$SE = \frac{\text{STDEV}(\Omega)}{\sqrt{\text{COUNT}(\Omega)}} \quad (12) \qquad [11]$$

Where:

STDEV = Standard deviation

Ω = Distances between each center of clusters.

It is an important step to visualize big data analytics. The clustering outputs have been shown in different methods to facilitate decision-makers and stakeholders in the energy field in Portugal to take suitable decisions in energy consumption in public buildings. In addition, ECD has the significant factor of contracted power, which is very useful in understanding how much energy is consumed during different times in the day in each public building. Fig. 12 shows a sample analysis of the various visualizations that show the dimensions used in the ECD through CD with SPKG.

By analyzing the clustering results, several essential rules have been extracted to assist stakeholders in the energy sector in Portugal in identifying the different styles of public buildings, as shown in Table VI. Energy consumption rules help the decision-maker identify public buildings that need guidance for their occupants and change the energy suppliers for those buildings.

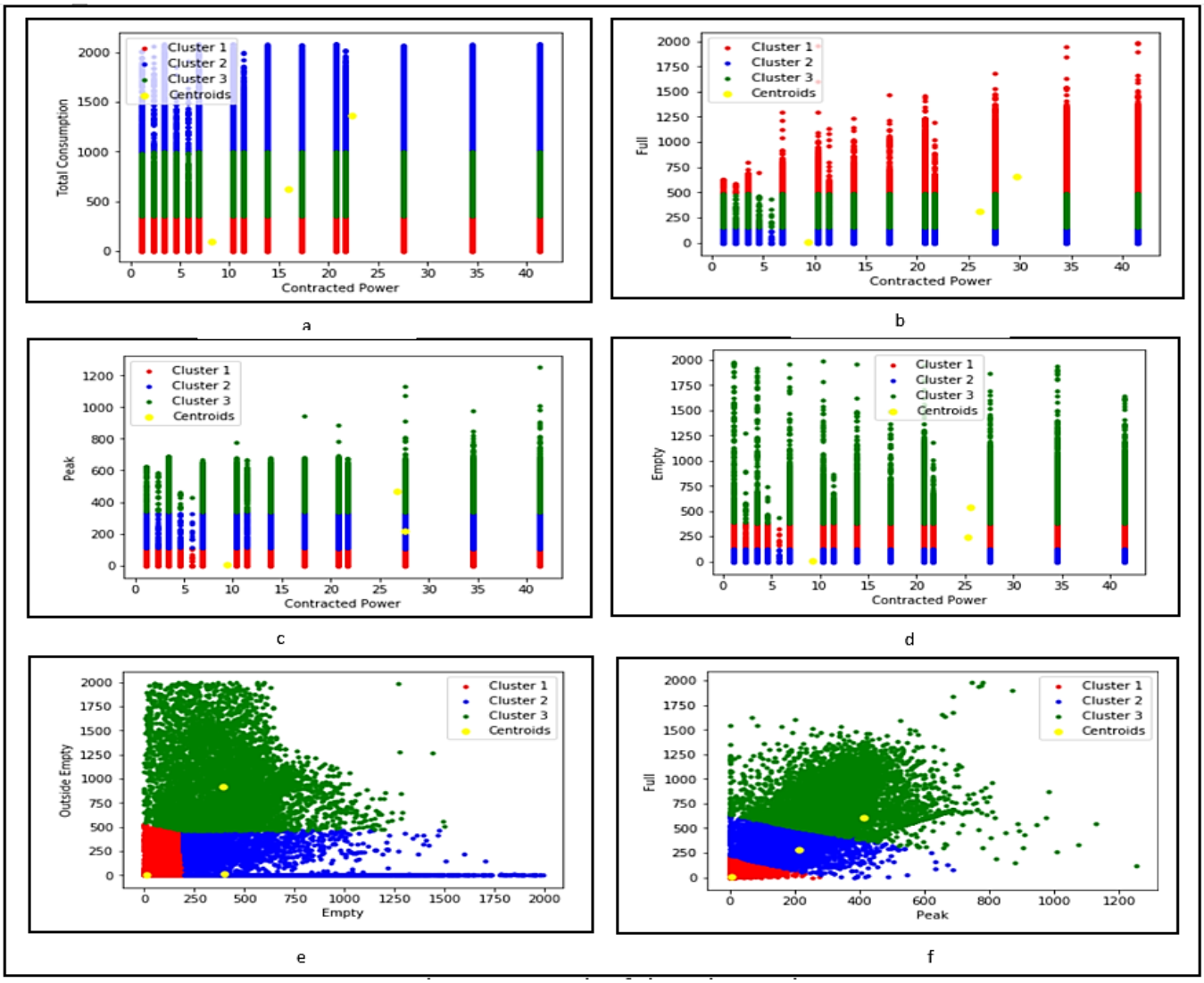

Fig. 12. Sample of clustering results.

TABLE IV. GA PARAMETERS

No	Parameters	Value
1	Population Size	ECD
2	Crossover Probability	0.5
3	Crossover type	Two points
4	Mutation Probability	0.6
5	Mutation type	Bit flip
6	Number of Iterations	100

TABLE V. A COMPARISON BETWEEN KMCKI AND SPKG IN TERMS OF SE AND STDEV

No	Method	SE	STDEV
1	ED with Kmeans++ (EDK)	93.19	465.99
2	MD with Kmeans++ (MDK)	184.14	920.73
3	CD with Kmeans++ (CDK)	0.004	0.021
4	ED with SPKG	88.49	442.45
5	MD with SPKG	174.94	874.71
6	CD with SPKG	**0.002**	**0.012**

TABLE VI. SAMPLE OF ENERGY CONSUMPTION RULES

No	Rules
1	Total<359 AND Full<157 Then cluster 1 (low energy consumption)
2	Total<359 AND Peak<111 Then cluster 1 (low energy consumption)
7	359<Total<992 AND 245<Outside empty<878 Then cluster 2 (medium energy consumption)
8	359<Total<992 AND 123<Empty<386 Then cluster 2 (medium energy consumption)
9	Total>=993 AND Full>=484 Then cluster 3 (high energy consumption)
10	Total>=993 AND Peak>=341 Then cluster 3 (high energy consumption)
14	157<Full<484 AND 111<Peak<341 Then cluster 2 (medium energy consumption)
15	Full>=484 AND Peak>=341 Then cluster 3 (high energy consumption)
16	Outside empty<245 AND Empty<123 Then cluster 1 (low energy consumption)
17	245< Outside empty <878 AND 123<Empty<386 Then cluster 2 (medium energy consumption)
18	Outside empty >=878 AND Empty>=386 Then cluster 3 (high energy consumption)
19	Total<359 AND Full<157 AND Peak<111 AND Outside empty<245 AND Empty<123 Then cluster 1 (low energy consumption)
21	Total>=993 AND Full>=484 AND Peak>=341 AND Outside empty>=878 AND Empty>=386 Then cluster 3 (high energy consumption)

Fig. 12 was better for detecting energy consumption levels; however, it could not determine the months in which energy consumption increases, as well as the months in which energy consumption decreases. Monthly consumption patterns show broader details of energy consumption by an occupant in public buildings. For ECD, the energy consumption levels were determined based on cluster label predictions, as shown in Fig. 13. Fig. 13 shows a noticeable increase in electricity consumption in January, February, November, and December. In addition, energy consumption levels decreased in June and July. It also helps the decision-maker identify the months of increased energy consumption for public buildings. Thus, the occupants of these buildings are guided promptly.

Fig. 13. Monthly energy consumption patterns captured in different clusters for the ECD.

Fig. 14 and Table VII show municipalities and Portuguese public buildings activities that contain the number of buildings that consume low energy at different times. Three municipalities contain public buildings that consume little energy in Fig. 14, such as 'LOULE', 'SANTA MARIA DA FEIRA', and 'LISBON'. In addition, Table VII shows Portuguese public buildings activities that consume little energy such as: 'INFRAESTRUTURAS PORTUGAL SA', 'GUARDA NACIONAL REPUBLICANA', and 'INSTITUTO SEGURANCA SOCIAL'.

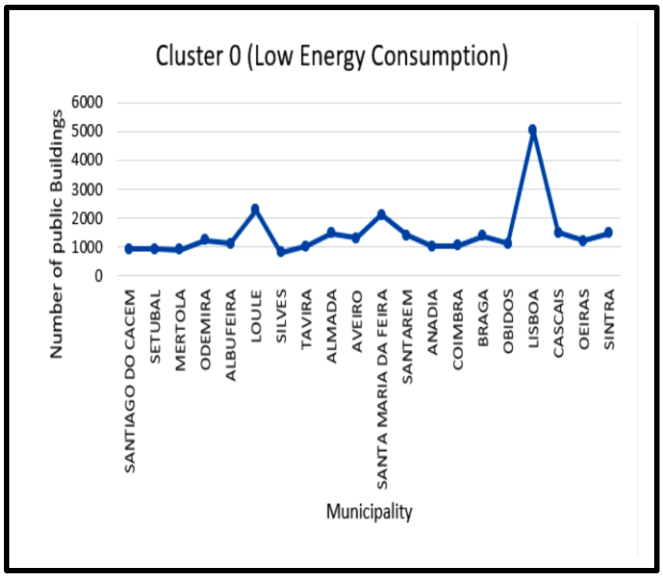

Fig. 14. Sample of Municipalities that Consume Low Energy Consumption

TABLE VII. SAMPLE OF PUBLIC BUILDINGS THAT CONSUME LOW ENERGY IN EACH MUNICIPALITY

Public buildings	Municipality	CACEM	SETUBAL	MERTOLA	ODEMIRA	ALBUFEIRA	LOULE	SILVES	TAVIRA	ALMADA	AVEIRO	SANTA MARIA DA FEIRA	SANTAREM	ANADIA	COIMBRA	BRAGA	OBIDOS	LISBOA	CASCAIS	OEIRAS	SINTRA
INFRAESTRUTURAS PORTUGAL SA		45	28	0	3	3	37	40	31	0	16	19	49	0	99	48	4	46	15	6	10
INSTITUTO SEGURANCA SOCIAL		3	0	21	7	6	12	12	0	13	0	23	0	0	0	0	0	0	0	62	5
ADMINISTRACAO REGIONAL SAUDE CENTRO IP		0	0	0	0	0	0	0	0	0	52	0	0	63	81	0	0	0	0	0	0
GUARDA NACIONAL REPUBLICANA		45	21	17	27	21	47	7	0	29	0	0	0	0	1	22	0	9	0	0	8

TABLE VIII. SAMPLE OF PUBLIC BUILDINGS THAT CONSUME MEDIUM ENERGY IN EACH MUNICIPALITY

Public building	Municipality	MONTIJO	CACEM	ALMODOV	MERTOLA	ODEMIRA	ALBUFEIRA	ALCOUTIM	CASTRO	LOULE	TAVIRA	SANTA MARIA FEIRA	MONTEMO	TORRES VEDRAS	BRAGANCA	VISEU	PORTO	VILA NOVA	LISBOA	CASCAIS	OEIRAS
IHRU INSTIT DA HABIT E REABILITACAO URBANA IP		0	105	0	0	0	0	0	0	23	0	0	0	0	0	0	2211	161	46	23	0
INFRAESTRUTURAS PORTUGAL SA		45	98	0	0	47	5	19	16	78	128	404	158	142	0	131	506	444	60	19	19
MUNICIPIO PORTO		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9797	0	0	0	0
MUNICIPIO OEIRAS		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8872

TABLE IX. SAMPLE OF PUBLIC BUILDINGS THAT CONSUME HIGH ENERGY IN EACH MUNICIPALITY

Public building	Municipality	SINTRA	ODEMIRA	ALBUFEIRA	LOULE	SILVES	ALMADA	AVEIRO	SANTA MARIA DA FEIRA	AZEMEIS	COIMBRA	SOURE	GUIMARAES	BRAGA	BARCELOS	MIRANDELA	LEIRIA	VILA NOVA DE GAIA	CASCAIS	LISBOA	OEIRAS
GUARDA NACIONAL REPUBLICANA		1	31	22	39	20	31	0	0	0	0	0	0	20	1	18	0	29	0	24	0
ADMINISTRACAO REGIONAL SAUDE CENTRO IP		0	0	0	0	0	0	70	0	0	115	54	0	0	0	0	106	0	0	0	0
ADMINISTRACAO REGIONAL SAUDE NORTE		0	0	0	0	0	0	0	272	42	0	0	50	102	53	0	0	165	0	0	0
AUTORIDADE TRIBUTARIA E ADUANEIRA		11	7	0	19	0	14	0	28	0	0	0	3	0	2	18	0	0	13	23	0
INSTITUTO SEGURANCA SOCIAL		17	14	16	9	17	0	0	3	0	0	2	0	0	20	0	0	16	0	0	5

Fig. 15 and Table VIII show the municipalities and Portuguese public buildings activities that contain the number of public buildings that consume energy on average between low and high consumption at different times. In Fig. 15, three municipalities contain public buildings that consume energy reasonably, such as 'PORTO', 'LISBOA', and 'OEIRAS'. In addition, Table VIII shows Portuguese public buildings activities that consume energy reasonably, such as: 'IHRU INSTIT DA HABIT E REABILITACAO URBANA IP', 'INFRAESTRUTURAS PORTUGAL SA', and 'MUNICIPIO PORTO'.

Fig. 16 and Table IX show the activities of municipalities and Portuguese public buildings containing the number of public buildings that consume high energy at different times. In

Fig. 16, four municipalities contain public buildings that consume high energy, such as: 'LOULE', 'SANTA MARIA DA FEIRA', 'VILA NOVA DE GAIA', and 'LISBOA'. In addition, Table IX shows Portuguese public buildings activities that consume high energy such as: 'GUARDA NACIONAL REPUBLICANA', 'ADMINISTRACAO REGIONAL SAUDE CENTRO IP', 'ADMINISTRACAO REGIONAL SAUDE NORTE', and 'AUTORIDADE TRIBUTARIA E ADUANEIRA'.

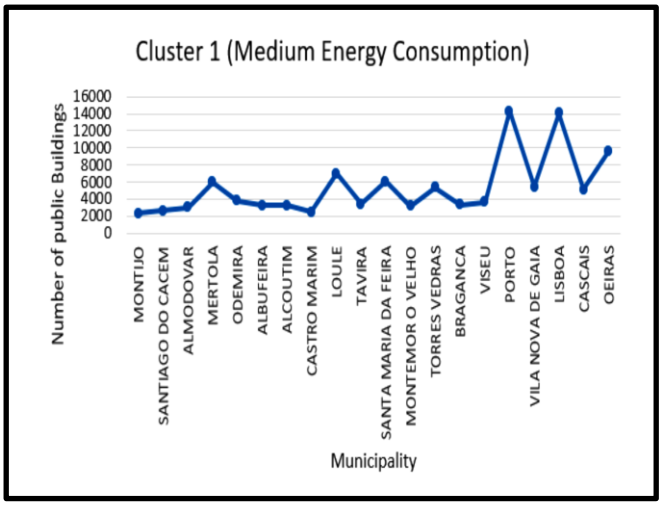

Fig. 15. Sample of Municipalities that Consume Medium Energy Consumption

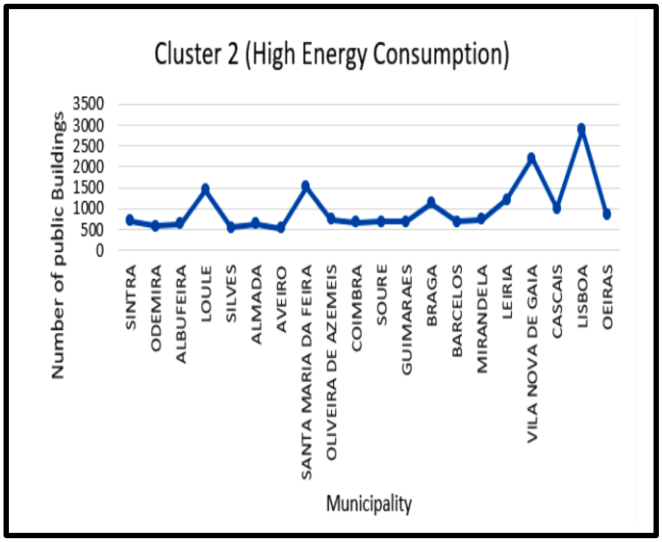

Fig. 16. Sample of Municipalities that Consume High Energy Consumption

This accurate analysis helps the decision-maker identify the municipalities and Portuguese public buildings activities that need to guide their consumers and change their energy providers.

By analyzing Fig. 14 to 16 and Tables VII to IX, municipalities such as 'LISBOA' and 'LOULE' contain public buildings with low, medium, and high energy consumption. In addition, there are Portuguese public buildings activities such as 'INFRAESTRUTURAS PORTUGAL SA' that consume low and medium energy. Therefore, we seek to find the distribution of the number of public buildings with different activities with low, medium, and high energy consumption over the different municipalities.

Tables VII, VIII, and IX show a sample of the public buildings located within each municipality. Knowing that each building has more than one location appears 24 times, distributed over 24 months over two years, 2018 and 2019.

By analyzing Fig. 17, the number of public buildings in these Municipalities increased in certain months in 2018 and 2019 as follows:

- LOULE: Aug-18, Oct-18, Jan-19, Feb-19, Mar-19, and Oct-19.
- SANTA MARIA DA FEIRA: Feb-18, Mar-18, Apr-18, May-18, Jun-18, Nov-18, Jan-19, and Feb-19.
- BRAGA: May-18, Aug-18, Oct-18, Jan-19, Mar-19, Apr-19, and May-19.
- VILA NOVA DE GAIA: Aug-18, Sep-18, Oct-18, Nov-18, and Jan-19 to Oct-19.
- LISBOA: Feb-18 to Nov-18, and Jan-19 to Dec-19

Regarding answering our research questions, and starting from RQ1, which aimed to collect public buildings energy consumption data in Portugal, and to find which where the critical factors in such dataset that could helped us in profiling such consumption, we were able to obtain aggregated monthly data for the years 2018 and 2019, regarding 77 996 buildings of various public sectors in 238 cities in Portugal, reaching 2 775 082 records. We concluded that all factors (variables) of the collected data are critical for the mentioned profiling, except for the super empty variable. Our RQ2 aimed to find the more appropriate intelligent computing techniques, for the preparation of the energy consumption dataset to proceed with further clustering analysis. Answering to this question, we adopted different mathematical techniques to that aim, namely, outlier removal with Isolation Forest and polynomial interpolation. With a dataset ready for clustering analysis, we raised RQ3, seeking first, to identify the number of clusters in the given energy consumption dataset, where we adopted literature techniques such as Self Organizing Map (SOM), the Elbow method, and the Davis – Bouldin method, and then to propose a novel and optimized hybrid model for classifying (labelling) energy consumption in buildings. This model includes a mix of different techniques, namely, SOM, Principal Component Analysis (PCA), K-means (KM), and Genetic algorithm (GA), is referred to as the SPKG model, and was applied successfully to our dataset, predicting the cluster label (low, medium, or high consumption) of each building. With a set of labelled buildings at hand, we turned our attention to RQ4, targeting to discover essential patterns and general rules in such labelled dataset, that could help the decision-maker to rationalize energy consumption. Therefore, we analysed the clustering results and came up with a set of rules that can help the characterization of energy consumption of a given public building in Portugal.

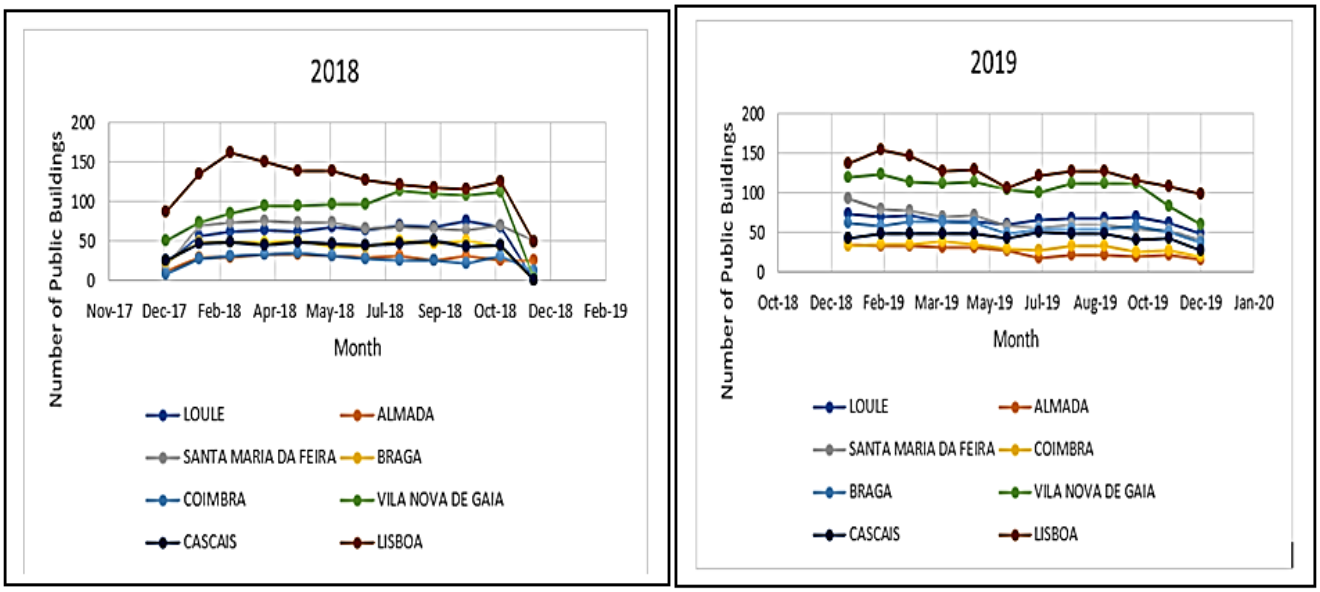

Fig. 17. Sample of Public Buildings in Different Municipalities that Consume High Energy in 2018 and 2019.

We have compared our results with state-of-the-art methods in the literature related to our work, in terms of using the K-means algorithm. M. Azaza study [14] and Al-Jarrah study [15]. The Standard Error (SE) of clustering in M. Azaza [14] and Al-Jarrah [15]m is 28.3 and 22.5, respectively. However, SE in our study is 0.002. Therefore, our study outperforms state-of-the-art methods in previous work, in terms of SE of the K-means algorithm.

V. Conclusion and Future Work

This paper presented a novel hybrid intelligent model for classifying the energy consumption level (low, medium, high) of buildings that was tested in a dataset of energy consumption of Portuguese public buildings. To frame our research, we raised four research questions that where properly answered. To understand our data, a correlation coefficient analysis was used to find the critical factors (variables) that influence energy consumption of public buildings and understand the relationship between those factors. In a data preparation step, an isolation forest was used to remove outliers in the dataset. Additionally, an interpolation method was used to find compensation values or estimate unknown values using related known values. As for our modelling approach, aiming at labelling the energy consumption level of each building, we first computed the number of clusters of energy consumption in the dataset, and SOM, the Elbow method, and Davis-Bouldin method all agreed in 3 as the figure for the found number of clusters (corresponding to low, medium, high consumption).

Then we used K-means with a Genetic Algorithm to predict the energy consumption cluster level of each building. This study provides contributions in four aspects. The first one considers factors that influence the energy consumption of buildings. The second one provides a novel model for classifying energy consumption of public buildings into levels (e.g., low, medium, and high). The third one provides analysis on real big data of the energy consumption of public buildings in Portugal, in the years of 2018 and 2019 (77 996 public buildings in 238 Portuguese cities). As an example, we were able to identify the municipalities that consume high energy levels. We have also identified monthly energy consumption patterns of buildings of the years of 2018 and 2019. The last aspect extracts proper scientific If-Then rules to help decision-makers rationalize the energy-consuming and determine the most energy-consuming public buildings, from a set of 3 values (low, medium, or high consumption).

Together, all these results may help the decision-maker to evaluate the public building's future energy requirements, and rationalize the occupants of those buildings, with the correct energy consumption behaviours.

As a recommendation for future work, we can think of using other techniques, such as statistical methods like multiple linear regression or logistic regression to find critical factors that influence energy consumption of public buildings. We could combine SOM and other optimization techniques (grey wolf, lion, and whale optimization), aiming find the optimal number of clusters of the energy consumption data of buildings. In addition, combining clustering and optimization techniques (grey wolf, lion, and whale optimization) could yield better prediction of cluster labels as for predicting the amount of energy consumption of buildings, this study follows the recent literature trend and suggests adopting machine learning approaches from the family of deep learning techniques, such as long short-term memory, convolutional neural networks, or deep forest.

Acknowledgment

This work has been supported by Portuguese funds through FCT-Fundação para a Ciência e Tecnologia, Instituto Público (IP), under the project FCT UIDB/04466/2020 by Information Sciences and Technologies and Architecture Research Center (ISTAR-IUL), and this work has also been supported by Information Management Research Center (MagIC)-Information Management School of NOVA University Lisbon.

REFERENCES

[1] T. A. Nguyen and M. Aiello, "Energy intelligent buildings based on user activity: a survey", Energy and Buildings, vol. 56, no. 1, pp. 244–257, 2013.

[2] M. Zhang and C. Y. Bai, "Exploring the influencing factors and decoupling state of residential energy consumption in Shandong", Journal of Cleaner Production, vol. 194, no. 1, pp. 253–262, 2018.

[3] N. Javaid, I. Ullah, M. Akbar, Z. Iqbal, F. Khan et al., "An intelligent load management system with renewable energy integration for smart homes", IEEE Access, vol. 5, no. 1, pp. 13587–13600, 2017.

[4] K. Li, C. Hu, G. Liu and W. Xue, "Building's electricity consumption prediction using optimized artificial neural networks and principal component analysis", Energy and Buildings, Vol. 108, no. 4, pp. 106–113, 2015.

[5] D. Zhao, M. Zhong, X. Zhang and X. Su, "Energy consumption predicting model of VRV (variable refrigerant volume) system in office buildings based on data mining", Energy, Vol. 102, no. 1, pp. 660–668, 2016.

[6] E. Agência, " Energy efficiency trends and policies in Portugal", Agência para a Energia, Vol. 1, no. 1, pp. 234–251, 2018.

[7] G. Shi, D. Liu and Q. Wei, "Energy consumption prediction of office buildings based on echo state networks", Neurocomputing, Vol. 126, no. 1, pp. 243- 264, 2016.

[8] S. Naji, A. Keivani, S. Shamshir, J. Alengaram, Z. Jumaat et al., "Estimating building energy consumption using extreme learning machine method", Energy, Vol. 97, no. 2, pp. 506–516, 2016.

[9] J. Massana, C. Pous, L. Burgas, J. Melendez and J. Colomer, "Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes", Energy and Buildings, Vol. 130, no. 4, pp. 519–531, 2016.

[10] J. P. Gouveia and J. Seixas, " Unravelling electricity consumption profiles in households through clusters: combining smart meters and door-to-door surveys ", Energy and Buildings, Vol. 116, no. 2, pp. 666 – 676, 2016.

[11] L. Hernández, C. Baladrón, J. Aguiar, B. Carro and A. Sánchez, " Classification and clustering of electricity demand patterns in industrial parks", Energies, vol. 5, no. 1, pp. 5215–5228, 2012.

[12] V. Ford and A. Siraj, " Clustering of smart meter data for disaggregation", In Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing, Austin, TX, USA, pp. 507–510, 2013.

[13] D. Rhodes, J. Cole, R. Upshaw, F. Edgar, E. Webber et al., "Clustering analysis of residential electricity demand profiles", Applied Energy, vol. 135, no. 4, 461–471, 2014.

[14] M. Azaza and F. Wallin, " Smart meter data clustering using consumption indicators: responsibility factor and consumption variability ", Energy Procedia, vol. 142, no. 4, pp. 2236 – 2242, 2017.

[15] Y. Al-Jarrah, Y. Al-Hammadi, D. Yoo and S. Muhaidat, "Multi-layered clustering for power consumption profiling in smart grids", IEEE Access, Vol. 5, no. 1, pp. 18459–18468, 2017.

[16] H. Cai, S. Shen, Q. Lin, X. Li and H. Xiao, "Predicting the energy consumption of residential buildings for regional electricity supply-side and demand-side management", IEEE Access, Vol. 7, no. 1, pp. 30386–30397,2019.

[17] C. Nordahl, V. Boeva, H. Grahn and P. Netz, "Profiling of household residents' electricity consumption behavior using clustering analysis", In Proceedings of the International Conference on Computational Science, Faro, Portugal, pp. 779–786, 2019.

[18] R. Granell, C. J. Axon and D. C. Wallom, "Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles", IEEE Transactions on Power Systems, Vol. 30, no. 1, pp. 3217–3224, 2015.

[19] M. Christ, N. Braun, J. Neuffer and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package)", Neurocomputing, Vol. 307, no. 4, pp. 72–77, 2018.

[20] C. Miller, Z. Nagy and A. Schlueter, "A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings", Renewable and Sustainable Energy Reviews, Vol. 81, no. 4, pp. 1365–1377, 2018.

[21] D. Hsu, "Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data", Applied Energy, Vol. 160, no. 1, pp. 153–163, 2016.

[22] A. Al-Wakeel, J. Wu and N. Jenkins, "K - means based load estimation of domestic smart meter measurements", Applied Energy, Vol. 194, no. 2, pp. 333– 342, 2017.

[23] A. S. Ahmad, M. Y. Hassan, M. P. Abdullah, H. A. Rahman, F. Hussin et al., "A review on applications of ANN and SVM for building electrical energy consumption forecasting", Renewable and Sustainable Energy Reviews, Vol. 33, no. 1, 102–109, 2014.

[24] D. Zhikuen, W. Zhan, T. Hu and H. Wang, "A comprehensive study on integrating clustering with regression for short-term forecasting of building energy consumption: case study of a green building", Buildings, Vol. 12, no. 10, pp. 1–20, 2022.

[25] Z. Chen, F. Xiao, F. Guo, F. Zhang, J. Yan et al., "Interpretable machine learning for building energy management: a state-of-the-art review", Advances in Applied Energy, Vol. 9, no. 1, pp. 1–19, 2023.

[26] T. Zhao, C. Zhang, T. Ujeed and L. Ma, "Methods on reflecting electricity consumption change characteristics and electricity consumption forecasting based on clustering algorithms and fuzzy matrices in buildings", Building Services Engineering Research and Technology, Vol. 43, no. 16, pp. 703–724, 2022.

[27] A. Galli, M. Savino, V. Moscato and A. Capozzoli," Bridging the gap between complexity and interpretability of a data analytics-based process for benchmarking energy performance of buildings", Expert System with Applications, Vol. 15, no. 1, pp. 388– 403, 2022.

[28] M. M. Ouf, H. B. Gunay and W. O'Brien, "A method to generate design sensitive occupant-related schedules for building performance simulations", Science and Technology for the Built Environment, Vol. 25, no. 1, pp. 221– 232, 2019.

[29] B. Dong, D. Yan, Z. Li, Y. Jin, X. Feng et al., "Modelling occupancy and behavior for better building design and operation—a critical review", Building Simulation, Vol. 11, no. 4, pp. 899–921, 2018.

[30] H. S. Park, M. Lee, H. Kang, T. Hong and J. Jeong, "Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques", Applied Energy, Vol. 173, no. 1, pp. 225–237, 2016.

[31] Z. Yang, J. Roth and R. K. Jain, "DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis", Energy and Buildings, Vol. 163, no. 1, pp. 58–69, 2018.

[32] K. Park and S. Son, "Novel load image profile-based electricity load clustering methodology". IEEE Access, vol. 7, no. 1, pp. 59048–59058, 2019.

[33] L. Wen, K. Zhou and A. Yang, "Shape-based clustering method for pattern recognition of residential electricity consumption", Journal of Cleaner Production, Vol. 212, no. 1, pp. 475–488, 2019.

[34] L. G. Swan and V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: a review of modeling techniques", Renewable Sustainability of Energy Review, vol. 13, no. 8, pp. 1819–1835, 2009.

[35] J. Kim, H. Naganathan, Y. Moon, O. Chong and S. Ariaratnam, "Applications of clustering and isolation forest techniques in real-time building energy-consumption data: application to LEED certified buildings", Journal of Energy Engineering, Vol. 143, no. 5, pp. 1 – 20, 2017.

[36] H. Yassine, G. Khalida, A. Abdullah, B. Faycal and A. Abbes, "Artificial intelligence-based anomaly detection of energy consumption in buildings: a review, current trends and new perspectives", Applied Energy, vol. 287, no. 1, pp. 1–26, 2021.

[37] S. Rodrigo and C. Marcelo, "Extended isolation forests for fault detection in small hydroelectric plants", Sustainability, vol. 12, no. 1, pp. 1–16, 2020.

[38] A. Daniel, G. Katarina, F. Hany, A. Miriam and B. Girma, "An ensemble learning framework for anomaly detection in building energy consumption", Energy and Buildings, vol. 144, no. 2, pp. 191–206, 2017.

[39] S. Jakob, T. Erik and L. Michael, "Anomaly detection forest", 24th European Conference on Artificial Intelligence, Belgium, Brussels, pp. 1–8, 2020.

[40] H. Sahand, C. Matias and J. Robert, "Extended isolation forest with randomly oriented hyperplanes", IEEE Transactions on Knowledge and Data Engineering, Vol. 1, no. 1, pp. 1–12, 2019.

[41] B. Elhadj, D. Belkacem, I. Bachir and A. Khadidja, "Numerical simulation of conjugate convection combined with the thermal conduction using a polynomial interpolation method", Advances in Mechanical Engineering, Vol. 9, no. 1, pp. 1–7, 2017.

[42] A. Abdelaziz, V. Santos and M. S. Dias, "Convolutional Neural Network With Genetic Algorithm for Predicting Energy Consumption in Public Buildings," *IEEE Access*, vol. 11, pp. 64049-64069, 2023.

[43] H. Zhao and F. Magoules, "Feature selection for predicting building energy consumption based on statistical learning method", Journal of Algorithms & Computational Technology, vol. 6, no. 1, pp. 59–77, 2012.

[44] M. Inga, "Feature selection for energy system modelling: identification of relevant time series information", Energy and AI, Vol. 4, no. 1, pp. 1–14, 2021.

[45] J. Lee, J. Kim and W. Ko, "Day-ahead electric load forecasting for the residential building with a small-size dataset based on a self-organizing map and a stacking ensemble learning method", Applied Sciences, Vol. 9, no. 1, pp. 1–19, 2019.

[46] A. E. Ioannou, D. Kofinas, A. Spyropoulou and C. Laspidou, "Data mining for household water consumption analysis using self-organizing maps", European Water, vol. 58, no. 2, pp. 443–448, 2017.

[47] Y.Long, M.Tang and H.Liao, "Renewable energy source technology selection considering the empathetic preferences of experts in a cognitive fuzzy social participatory allocation network", Technological Forecasting and Social Change, vol. 58, no. 1, pp. 421–432, 2021.

[48] A. Abdelaziz, V. Santos and S. Dias, "Machine learning techniques in the energy consumption of buildings: a systematic literature review using text mining and bibliometric analysis", Energies, Vol. 14, no. 1, pp. 1 - 25, 2021.

[49] T.Räsänen, J. Ruuskanen and M. Kolehmainen, "Reducing energy consumption by using self-organizing maps to create more personalized electricity use information", Applied Energy, Vol. 85, no. 4, pp. 830–840, 2008.