Unsupervised Clustering for Pattern Recognition of Heating Energy Demand in Buildings Connected to District-Heating Network

Mikel Lumbreras
ENEDI Research Group, Department of Energy Engineering, Faculty of Engineering of Bilbao
University of the Basque Country UPV/EHU
Bilbao, Spain
mikel.lumbreras@ehu.eus

Koldobika Martin-Escudero
ENEDI Research Group, Department of Energy Engineering, Faculty of Engineering of Bilbao
University of the Basque Country UPV/EHU
Bilbao, Spain
koldobika.martin@ehu.eus

Gonzalo Diarce
ENEDI Research Group, Department of Energy Engineering, Faculty of Engineering of Bilbao
University of the Basque Country UPV/EHU
Bilbao, Spain
gonzalo.diarce@ehu.eus

Roberto Garay-Martinez
Building Technologies
TECNALIA, Basque Research and Technology Alliance (BRTA)
Bilbao, Spain
roberto.garay@tecnalia.com

Ruben Mulero
Building Technologies
TECNALIA, Basque Research and Technology Alliance (BRTA)
Bilbao, Spain
ruben.mulero@tecnalia.com

Abstract — This paper presents a novel framework for the identification of different consumption patterns of heating loads of buildings. The approach to analyzing the consumption data is carried out by a combination of unsupervised clustering models. Density based clustering is used for outlier detection in the original dataset and K-means for pattern recognition. The proposed framework is then applied to a real building connected to the district heating in Tartu (Estonia). Three main day-types are identified for the building as an outcome of the clustering process, with different patterns throughout these days. More than 60% of the analyzed Cluster Validation Indexes studied in this paper show that classifying the daily demand profiles in three clusters is the optimal classification.

Keywords — Unsupervised Clustering, Pattern Recognition, Heating Energy Demand, Data-Driven Model, District-Heating Networks

I. INTRODUCTION

The building sector is responsible for the consumption of approximately the 40% of the primary energy in the European Union (EU) [1]. Increasing the energy efficiency in buildings has become an important objective for the EU by means of different directives.

District-Heating (DH) networks are systems in which energy is distributed from a centralized source to many consumers. DH networks are very efficient systems covering around 13% of the total heat loads in EU buildings [2]. With the progressive implementation of the so-called 4th Generation District-Heating (4GDH) [3], heat is supplied at very low temperatures, commonly with the injection of heat from low-grade renewable energy sources (RES).

With the above-mentioned issues, networks are pushed to their limit to improve efficiency and sustainability, requiring higher accuracy in the characterization of the heat loads in buildings. In this process, the identification of different energy usage patterns is a key step.

Smart energy meters are being massively implemented in buildings [4], with remote reading capabilities for high frequency data (hourly and sub-hourly measurements of different variables in the system). Key variables are flow & return temperature and energy consumption.

In contrast to electricity data, data quality level has not been available for heat loads until recent times. Therefore, few studies regarding machine-learning (ML) techniques applied to heating energy can be found. However, literature from ML applied to electricity can be used as reference for this type of analysis.

Unsupervised clustering techniques aim to group or organize data that are similar without prior knowledge of groups´ definitions. The main or most used clustering algorithms are the partitioning algorithms, such as K-means, and hierarchical clustering. Other clustering algorithms have also been developed for specific objectives: fuzzy clustering, density clustering or model based clustering.

Regarding ML techniques applied to electricity demand data, [5] found that k-means was the most effective algorithm to investigate electricity load patterns in a data large data set with 1910 residential and 1919 non-residential buildings. Three fundamental clusters were obtained from this analysis. Moreover, [6] studied daily electricity usage pattern of three office buildings by means of a combination of clustering techniques and applied for anomaly detection.

Therefore, a gap is identified in literature, since very few references can be found on unsupervised clustering techniques applied to heating demand in buildings.

This paper explores the applicability of clustering processes to heating energy profiles of buildings connected to DH networks. For that purpose, a multistep methodology combining different clustering algorithms presented in Section II is proposed.

II. METHODOLOGY

This paper explores the use of unsupervised clustering techniques for pattern recognition of heating energy demand in buildings connected to a DH network. The multistep methodology proposed is illustrated in Fig. 1 and applied to a specific building, but replicable for any building.

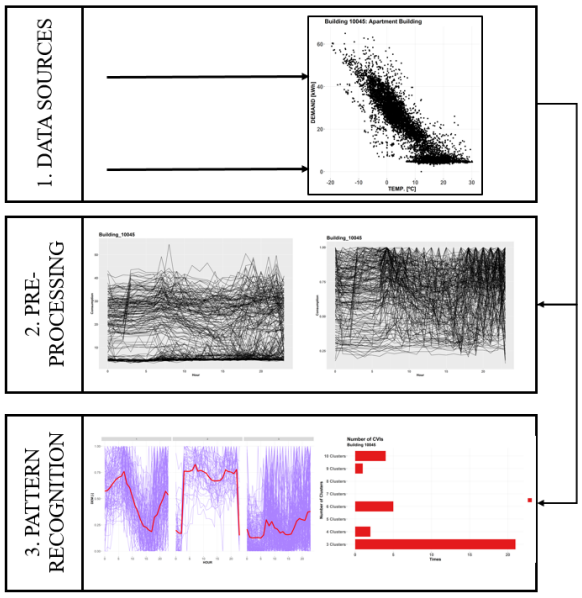

Fig. 1. Overview of the general methodology.

Firstly, Section A presents the data sources used in this study and Section B explains the pre-processing activities carried out with the raw data, including the outlier detection algorithm. Finally, Section C presents the unsupervised clustering algorithm proposed for the identification of the patterns of use of the energy and how the identified clusters are evaluated by means of different Cluster Validation Indexes (CVIs). The software used for the implementation of the different algorithms used throughout the study is R [7].

The following subsections will provide details of each phase.

A. Data Sources

The starting point of this study is the collection of the data sets used. For this purpose, data from a real building is used. This building is located in Tartu (Estonia) and this dwelling is currently connected to the DH in Tartu, which is managed by [8].

The smart energy meter is situated in the substation of the building, which divides the primary and secondary side of the network. The energy meter installed in the buildings is the Multical® 603 from Kasmtrup [9]. The accuracy of this device is always higher than the one fixed in EN-1434-1:2015 and the measuring error remains below 5% in all the variables read. Heat energy consumption is saved as a cumulated variable and is read hourly. Consequently, the hourly energy use is calculated as the measured reading in that hour minus the measured value in the previous hour.

On the other hand, climatic data is obtained from the Physics Institute of the University of Tartu [10]. This weather station collects data with a 15-minute frequency. This data is resampled from a 15-minute frequency to hourly intervals, obtaining 8760 readings representing each hour of the year.

Therefore, these two sources are coupled by matching calendar variables of both sources, obtaining a data set with hourly values. This hourly data set is used for outlier detection methodology that is explained in the following section. This dataset is conformed by hourly readings of heat consumption, hourly outdoor temperature, solar irradiance and calendar variables. Furthermore, these hourly data is ordered into daily profiles for pattern recognition (Section C), obtaining a dataset with 24 columns, corresponding each of the column with the measure of the column with the hour of the day and each row will contain daily data.

B. Outlier Detection & Other Pre-processing activities

The raw data set obtained from the previous step may include reading errors and outliers that can disrupt and hide the real consumption patterns in the building. Reading errors are directly removed from the original data set.

For the identification of outliers, density based clustering is proposed, by means of DBSCAN algorithm [11]. The objective of this algorithm is to identify high-density observations that are closely together and the points that are identified in low-density areas are considered outliers. The central concept of this algorithm is that the neighborhood of a given radius has to contain at least a minimum number of observations. In this algorithm, two parameters have to be defined: eps and MinPts. The parameters eps present the initial radius and MinPts is the minimal number of points in the region denotes by eps. The algorithm starts with the selection of a random point (core point) and checks its radius. If the number of points inside this area is higher than MinPts, these points are labelled also as core points and the algorithm will start a new cluster. The next step is to randomly select another point that has not been visited in previous steps and apply the same procedure. After all the points are processed, the points that are not assigned to any cluster are labelled as noise. The main difficulty of this algorithm is the optimization of these two parameters. According to [12], MinPts is initialized as the dimension of the dataset plus one. An overly small eps can cause that values that are not outliers are considered as outliers and an overly high eps value cause that the outliers are not identified. For the optimization of this step, calculation of K-nearest neighbors' distances is carried out and the elbow of the ascending ordered distances correspond the optimal value. The elbow of that curve is calculated by means of the second derivative of the k-NN distances.

The other main pre-processing step is the normalization of the data set. Since the objective of this study is the recognition of the patterns of use, the real value of the demand is not so interesting as the temporal variation of the profile of the demand. The objective of normalizing data is to reduce all the values of the demand into the same range, in this case between 0 and 1. Thus, clean data set (without outliers) is normalized by the following equation:

$$q_{nor}(t) = q(t)/q_{max}(t) \quad (1)$$

Where $q_{nor}(t)$ and $q(t)$ correspond with normalized and actual heating demand at time t (t = 1,2…, 24), respectively and $q_{max}(t)$ is the daily maximum heating load.

C. Patter Recognition & Cluster Validation Index (CVI)

Consumption patterns are daily loads or a fraction of daily load that are repeated over time. These consumption patterns may be repeated over various day within a heating season. The

consumption patterns can be used to understand the heating load of a building and the patterns of use of the heating demand of the users, without additional information.

For the identification of days with similar heating load profiles, K-means clustering algorithm is used.

This algorithm starts with the definition of K, the number of clusters in which the data set is aimed to be clustered. This variable is introduced at the beginning of the algorithm by the author. Then, random K initial centroids are created and the distance between each daily feature and the cluster center is calculated. The cluster center is updated by calculating the mean value of all the daily consumptions in that cluster. This last step is repeated until the centers do not change.

There is no initial conditions that can determine how many clusters are found in the heat consumption in the building. So, this step is repeated for K = {3, 4,…, 10}. K = 2 is skipped in order to avoid the weekday/weekend identification. Thus, for a specific building, eight different clustering process are carried out.

Among the eight clustering processes, the optimal process is the one that better recognize the consumption patterns of the heating energy demand in the building. For this purpose, clustering validation indexes (CVIs) are used. In this study, more than 30 different CVIs are evaluated, including the most common CVIs, such as, Silhouette Index, Dunn Index or Davies Bouldin Index. Each of the indexes used in the study present their own evaluation way, but like general norm, these indexes evaluate the intra-cluster distance (distance between observations inside a cluster) and inter-cluster (distance between observations from different clusters). A low intra-cluster distance and high inter-cluster distance mean that the identified clusters are separated and compact.

III. Results

A. Description of the Case Building

The building used as the Case Building for showing the results is referred as Building 10045 to avoid the disclosure of usage patterns of specific users. The chosen building is currently used as residential apartments and it is part of a larger data set composed by 43 buildings. For this study, hourly data for entire 2019 is used. In Fig. 2, hourly heat consumption is shown against the outdoor temperature for this building.

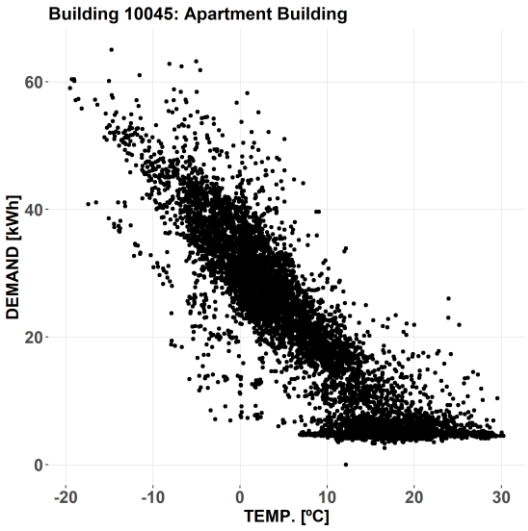

Fig. 2. Heat Consumption vs Outdoor Temperature. Raw Data

B. Outlier Detection

In this process, DBSCAN algorithm is used to identify the outliers. It is observed that outdoor temperature is the climatic variable that most affect the heat consumption, in other words, the climatic with highest correlation with heat consumption. Thus, the density based clustering algorithm is applied with these two dimensions: (i) Hourly heat consumption [kWh] and (ii) Outdoor temperature in [ºC]. As previously commented, MinPts is initialized as the dimensions of the data set plus one. This way, MinPts is initialized as three. 3-NN distance is then calculated and the ordered distance is shown in Fig. 3. The elbow of this function is also marked in Fig. 3.

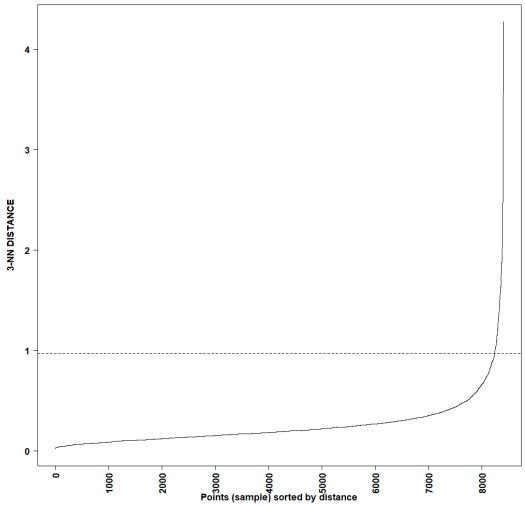

Fig. 3. Outlier Drection knee

The optimal eps is 0.9697 and as it can be observed in Fig. 3, very few outliers (around 2.5% of the total dataset) are identified as such. More than 8200 points are considered as good data. This way, the identified outliers are removed from the original dataset, before continuing with the rest of the framework.

C. Pattern Recognition & CVIs

After removing the outliers, the clean data set is now organized by days and normalized following the Eq. (2). K-means is applied in this step to the normalized data. Using too

few clusters could not be useful to discover the patterns in the building, while using too many clusters could result in insignificant differences across some of the patterns. Therefore, the optimal number of clusters to be analyzed was selected to be from 3 to 10. K-means is applied with the Euclidean distances.

Fig. 4 shows the results from clustering with K = 3 in the clustering algorithms, whereas Fig. 5 presents the results of the unsupervised clustering with K = 6. In these images, each of the plot represents a cluster and in each of the clusters, each of the line, correspond with the daily profile of the demand in this cluster. Then, the red line in each of the cluster is the mean heating demand profile of all the days included in the corresponding cluster. This red line of each of the clusters eases the visualization of the different consumption patterns in the building.

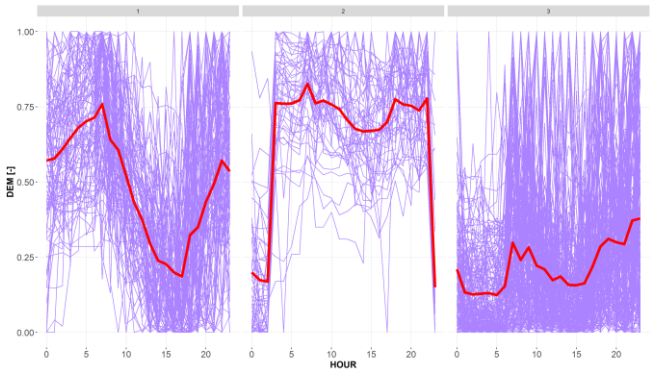

Fig. 4. Three daily patterns of heating demand.

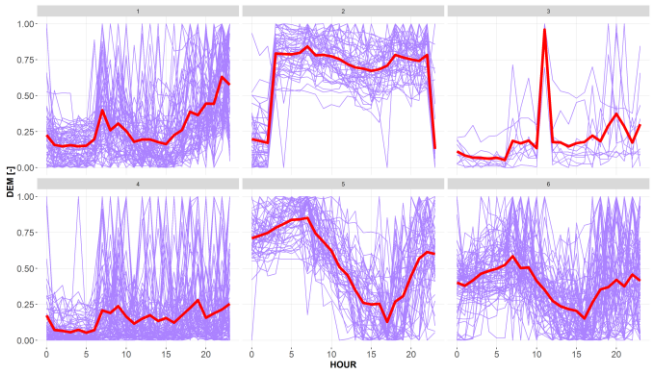

Fig. 5. Six daily patterns of heating demand

Fig. 4 & Fig. 5 are chosen to be presented because the CVIs conclude that these are the optimal clustering values, as summarized in Table I. More than 30 different CVIs have been evaluated and Table I shows the number of indexes that say that classification of the daily profiles is the optimal for each number of clusters. As commented previously, each CVI has its own equation for the cluster validation and this is why, it is possible to obtain different results. In this case, more than 60% CVIs conclude that the optimal classification is with K=3, followed by around 15% of the results concluding that the optimal classification results come from K=6.

TABLE I. NUMBER OF CVIS FOR EACH CLUSTERING PROCESS

Nº of Clusters	**Nº CVIs**
K = 3 Clusters	21
K = 4 Clusters	2
K = 5 Clusters	0
K = 6 Clusters	5
K = 7 Clusters	0
K = 8 Clusters	0
K = 9 Clusters	1
K = 10 Clusters	4

Although as a first approach, results from Fig. 4 & Fig. 5 seems to be very different, actually, they are not. Similar patterns are recognized in both cases:

- When K=6, Clusters 5 and 6 are very similar shape. The same applies to Clusters 1 and 4.
- Clusters 5 & 6 (K=6) are similar to Cluster 1 (K=3). The same applies to Clusters 1 & 4 (K=6) against Cluster 3 (K=3)
- Cluster 2 presents similar consumption patterns for both cluster numbers.

The unique and main difference between these two classifications is Cluster 3 (K=6). Only a few days are classified in this cluster and thus, the importance of this cluster in the overall image of the consumption patterns in the building is considered low.

Therefore, three main consumption profiles are identified:

- In the first cluster, the demand is gradually increasing in the first hours of the day, possibly caused by the domestic hot water consumption of the morning. From 7am onwards, the demand is reduced caused by the reduction of the space-heating demand in the hours when there is little occupancy in the building. These off-peak hours reach the minimum demand at 17pm and from then on, demand increases.
- The demand profile of the second cluster present a very strong increase around 3am until 5am approximately. The demand along the day remains relatively constant and at high values. This is probably caused by the very cold temperatures that Tartu (Estonia) usually present in winter. These low temperatures make necessary to consume space-heating demand throughout the day. Finally, at 23pm a drastic reduction of the demand is shown. Thus, from 23pm to 3am, a night setback is identified in which the set-point temperature is reduced because it is expected that users of the building will be sleeping.
- The demand profile of the third cluster is the most stable one. The days included in this cluster do not present the night setback from the second cluster. Besides, relative peak demands are presented at 7/8am coinciding with the hours when it is expected to be a high demand for domestic hot water. Then, the demand continues to increase as long as the outdoor temperature decreases. Unlike cluster 2, the demand of

this cluster show great correlation with the outdoor temperatures and other climatic variables affecting the demand.

Finally, Fig. 6 summarizes the optimal clustering profiles, showing the cluster of each of the day in the calendar. It is observed that the most determining calendar factor is the seasonal variation. Cells in purple are incomplete days caused by outlier removal or reading errors).

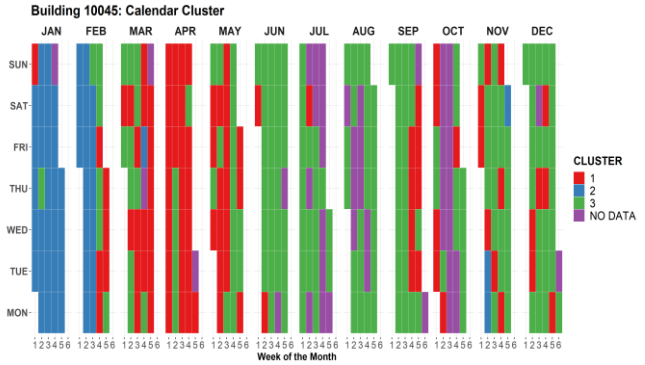

Fig. 6. Annual calendar with the identified clusters (K=3)

IV. Discussion

This paper explores the use of clustering methods for patterns recognition of heating energy demand in buildings connected to a DH network. Density based clustering is used outlier detection and K-means algorithm is used for the identification of different consumption patterns.

Clustering is an unsupervised technique that provides classification of different variables that can be used for the identification of demand patterns. However, this technique does not provide unique results and require interpreting these results by experts on building physics. As an example, both K = 3 and K = 6 clustering show good classification results but CVIs return that the optimal cluster classification is obtained with K = 3. This makes necessary to evaluate different CVIs and approach to the optimal results by means of a statistical analysis. In general, the proposed framework is difficult to be replicable in all the cases.

V. Conclusions & Further work

The proposed framework in this paper provides a general solution to identify representative heating demand usage patterns and discover deeper insightful knowledge behind the patterns. This study increases the interpretability of clustering results and application value of discovered knowledge about the consumption patterns of the analyzed building.

The future work will be focused on analyzing the real causes of the identified consumption patterns, by means of correlation between demand and climatic and calendar variables. Moreover, the proposed framework is expected to be applied to other buildings connected to the same DH network, so that results may be obtained.

Acknowledgment

The authors would like to thank Fortum Tartu for providing data from the substations for academic purposes.

The authors would like to acknowledge the Spanish Ministry of Science and Innovation (MICINN) for funding through the Sweet-TES research project (RTI2018-099557-B-C22).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 768567.

References

[1] Luis Pérez-Lombard, José Ortiz, Christine Pout, A review on buildings energy consumption information, Energy and Buildings, Volume 40, Issue 3, 2008, Pages 394-398, IS0SN 0378-7788,

[2] Henrik Lund, Renewable energy strateg1ies for sustainable development, Energy, Volume 32, Issue 6, 2007, 4Pages 912-919, ISSN 0360-5442,

[3] Haoran Li, Natasa Nord, Transition to the 4th generation district heating - possibilities, bottlenecks, and challenges, Energy Procedia, Volume 149, 2018, Pages 483-498, ISSN 1876-6102,

[4] Liu X, Golab W, Golab W, Ilyas IF. Benchmarking smart meter data analytics. In Proc of the 18th international conference on extending database technology: 2015. p. 385-96.

[5] June Young Park, Xiya Yang, Clayton Miller, Pandarasamy Arjunan, Zoltan Nagy, Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset, Applied Energy, Volume 236, 2019, Pages 1280-1295, ISSN 0306-2619.

[6] Xue Liu, Yong Ding, Hao Tang, Feng Xiao, A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data, Energy and Buildings, Volume 231, 2021, 110601, ISSN 0378-7788.

[7] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[8] Fortum Tartu, https://www.fortumtartu.ee/ (Accesed in 2021).

[9] Karmstrup, https://www.kamstrup.com/es-es, Multical® 603 (Accessed in January 2021).

[10] University of Tartu, Institute of Physics, Laboratory of Environmental Physics, http://meteo.physic.ut.ee/?lang=en (Accessed in 2020).

[11] M. Ester, H. Kriegel, X. Xu, D.- Miinchen, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, In: Proceedings of the 2nd ACM SIGKDD, Portland, Oregon; 1996. pp. 226–231.

[12] M. Hahsler M. Piekenbrock S. Arya D. Mount R, Package 'dbscan' 2020.