Journal of Building Engineering 65 (2023) 105732

Contents lists available at ScienceDirect

Journal of Building Engineering

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

Unsupervised recognition and prediction of daily patterns in heating loads in buildings

Mikel Lumbreras [a,*], Gonzalo Diarce [a], Koldobika Martin [a], Roberto Garay-Martinez [b], Beñat Arregi [c]

[a] ENEDI Research Group, Energy Engineering Department, Faculty of Engineering of Bilbao, University of the Basque Country (UPV/EHU), Pza. Ingeniero Torres Quevedo 1, Bilbao, 48013, Spain
[b] Institute of Technology, Faculty of Engineering, University of Deusto, Av. Universidades, 24, 48007, Bilbao, Spain
[c] TECNALIA, Basque Research and Technology Alliance (BRTA), Bizkaia Science and Technology Park, Astondo Bidea 700, Derio, Spain

ARTICLE INFO

Keywords:
Pattern recognition
Unsupervised clustering
Heating loads
Daily profiles

ABSTRACT

This paper presents a multistep methodology combining unsupervised and supervised learning techniques for the identification of the daily heating energy consumption patterns in buildings. The relevant number of typical profiles is obtained through unsupervised clustering processes. Then Classification and Regression Trees are used to predict the profile type corresponding to external variables, including calendar and climatic variables, from any given day. The methodology is tested with a variety of datasets for three different buildings with different uses connected to the district heating network in Tartu (Estonia). The three buildings under analysis present different energy behaviors (residential, kindergarten and commercial buildings). The paper shows that unsupervised clustering is effective for pattern recognition since the results from the classification and regression trees match the results from the unsupervised clustering. Three main patterns have been identified in each building, seasonality and daily mean temperature being the variables that have the greatest effect. The results concluded that the best classification accuracy is obtained with a small number of clusters with a classification accuracy from 0.7 to 0.85, approximately.

Nomenclature

Acronyms

CART	Classification & Regression trees
CVI	Cluster Validation Index
DBSCAN	Density Based Clustering
DH	District-Heating
DHW	Domestic Hot Water
DS	Dataset
EU	European Union
RES	Renewable Energy Source

* Corresponding author.
E-mail address: mikel.lumbreras@ehu.eus (M. Lumbreras).

https://doi.org/10.1016/j.jobe.2022.105732
Received 11 August 2022; Received in revised form 5 December 2022; Accepted 10 December 2022
Available online 16 December 2022

SH	Space-Heating
Symbols	
c_p	Cost Complexity parameter
Eps	Radius in dbscan
J	Cost function
minsplit	Minimum number of observations in a node
k	cluster
$q\ (t)$	Actual hourly heat load
$q_{norm}\ (t)$	Normalized hourly load
$q_{min}\ (t)$	Daily minimum heat load
$q_{max}\ (t)$	Daily maximum heat load
$\overline{q}$	Daily mean heat load
sd	Standard deviation
w	Relative weight
μ	Cluster center in K-means
x	Heat load observation

1. Introduction

The building sector is responsible for approximately 40% of the global primary energy consumption [1] and more than 30% of CO_2 emissions production [2]. Accordingly, the European Commission (EC) is focusing on increasing energy efficiency in buildings by means of directives [3,4].

Among other measures, it is well known that heat load characterization and prediction is by itself a very powerful tool to improve energy efficiency. Accurate heat load characterization methods are at the core of Measurement and Verification protocols, as they allow for a more accurate baseline definition and better-informed investments in energy efficiency. Load forecasting allows for the optimization of energy systems through load flexibility and demand response approaches [5]. Cao et al. [6] performed an extensive review of machine learning applications for buildings. This analysis identified that machine-learning was a very active field of research with multiple proposals for applications in the aforementioned areas. However, it attributed the lack of applicability of these methods to the lack of datasets for model training & validation; a lack of model transferability; insufficient justification of costs and benefits; and insufficient reliability and robustness of model performance.

District-heating (DH) networks gather a large number of buildings in the same energy grid [7]. In these settings, monitoring and actuation is standardized with substations at building level, and data are gathered in a systematic way. It provides an environment which facilitates the applicability of data driven models through the availability of large-scale datasets, while model transferability is facilitated through common data structures.

DH networks currently cover around 13% of the total thermal energy demand in the EU [8]. With a reduced baseload due to energy efficiency measures and variations in the heat production-mix due to the integration of renewables in the so-called 4th Generation District-Heating (4GDH) networks [9,10], there are expectations for an increased volatility in the operation of DH networks [11,12]. This results in an increasing need for accurate heat load modeling techniques.

Heating appliances in buildings require energy depending on a variety of factors, such as external climate, building usage, settings and scheduling of Heating Ventilation and Air Conditioning (HVAC) systems, among others. The heat load for space heating (SH) is highly correlated with external climate, but relevant transitory effects are generated with building usage and scheduling of HVAC systems. The Domestic Hot Water (DHW) heating load is principally correlated with building usage (i.e., scheduling of showers). Works such as [13] are focused on developing heat load characterization methods where the impact of climate data can be taken into account. Lumbreras et al. [14] showed that loads vary, not only depending on climate data, but also on the time of the day and calendar variables. Considering these variations and properly segmented data fed into the models resulted in an improved load prediction accuracy. However, this work did not optimize the process of identifying specific daily profiles. The proper identification of typical patterns arising from usage factors, such as calendar, type of building and scheduling of human activities, and the thermal variations in buildings may result in optimized data segmentation criteria and improved heat load prediction accuracy. This requires specific pattern identification processes based on historical data, as well as classification methods to define the most probable profile for each day in the future.

The relevance of building usage is also acknowledged by Cholewa et al. [15], which looks to indoor temperature as a further explanatory variable for a short-term prediction model. This variable, although highly informative, is discarded in the present work as it is not typically available for DH operators and would limit the applicability of the proposed clustering approach.

Heat meters gather accurate and frequent data on energy consumption and deliver these to the DH utility in a continuous process. The assessment of this data allows for the development of improved modeling and control of the network. Modern machine-learning techniques, and specifically unsupervised algorithms, are able to discover hidden patterns in unlabeled data [16–18].

Energy consumption patterns are daily loads or a fraction of the daily consumption profile that are repeated over time [19]. These energy consumption patterns may be caused by a repetitive consumption action by the users of the buildings or by energy management

strategies by the DH operator, and they may be repeated over different days within a heating season. A correct understanding of the energy consumption patterns and the causes will help in the characterization process of the heating demand in the building [20]. Unsupervised learning algorithms have been successfully applied to identifying usage patterns commonly used in electricity load analysis [21–24]; however, their use in heat-related applications has been limited so far. Amongst the existing references for electricity loads, Liu et al. [25] studied the daily electricity usage pattern of three office buildings with a combination of unsupervised and supervised clustering techniques and they also developed an application for anomaly detection. Carmo et al. [26] clustered the electricity profile of the distributed heat pumps' consumption located in more than one hundred buildings in Denmark. Two clusters representing weekend and weekdays were identified. A Demand-Response program is proposed by Ref. [27] based on electricity consumption patterns identified in the electricity consumption of a residential building, while Haben et al. [28] presented a feature-based clustering method in which the computational costs of the algorithm were reduced by using representative variables of the raw dataset.

Although the existing literature developed with data from electric loads [29] can be partially applicable to heat loads, it presents specificities through several effects [30], such as outdoor temperatures and activities taking place in the building. Furthermore, the literature on heat load patterns is more limited. A statistical approach to heating energy consumption patterns of buildings connected to a DH was presented by Ma et al. [31], based on such simplified variables as time and building types. A Gaussian mixture model was presented for heat load prediction with a relative error of 4–8%. A fault detection algorithm was proposed by Gadd et al. [32] based on the identification of two load patterns corresponding to Domestic Hot Water (DHW) and Space-Heating (SH) consumption. Tureczek et al. and Calikus et al. included clustering and classification methods to study the energy consumption patterns ([33,34], respectively). Tureczek et al. [33] demonstrated that unsupervised clustering can be applied to heat load data by analyzing data from 49 district heating substations and showing the autocorrelation existence between the clusters identified. Moreover, decision-trees were used by Calikus et al. [34] for mining the different consumption patterns in a unique office building located in New York.

Gianniou et al. [35] performed a clustering work over district heating data. It successfully identified a set of daily building heat load profiles, with specific patterns for weekdays and weekends. The likeliness of pattern changes in a building based on calendar due to changes in the space heating baseload magnitude was set with a monthly resolution. Johra et al. [36] also performed clustering over district heating data, resulting in the profiling of 1665 households to 4 profiles. This work was performed independently for all 4 seasons in a year, and the correlation between the clusters assigned for each of the 4 seasons was studied. In both cases, the data presented a quite stable baseload, mainly with one clear peak in the morning, which somehow limited the handling of more varied building usage. In addition, the clustering process was performed jointly for all the daily profiles in all the buildings, which hindered the possibility of adapting the cluster identification processes to the specificity of each building. What is more, the way to use the identified patterns in forecasting applications was not defined, which would anyhow be limited to the lack of any explicit relation to climate and calendar.

Even though some clustering works are applied over thermal energy, they are mostly focused on the methods to identify electric energy consumption patterns. This is mainly caused by the fact that smart meters for electricity consumption have been installed longer than smart meters for thermal loads. Furthermore, the identified energy profiles and their approach to the real causes have hardly been discussed to date. However, the impact of external variables such as climatic variables or seasonal patterns is even more important in thermal loads than in electricity. On the other hand, the effects of removing the outliers of the heat-load from smart meters in unsupervised clustering processes have scarcely been studied. A wrong pre-processing of the original data could lead to inaccurate results, even though the methodology and algorithms used are the optimal ones.

This work presents a novel method to approach the characterization of heat load profiles in a structured way. It applies existing knowledge on unsupervised clustering and classification techniques to this field of work. It presents the following technology innovations for the improvement of building heat load characterization:

- The recognition of daily (24 h) heat-load patterns using unsupervised learning algorithms, as well as cluster validation techniques to allow for the automated definition of the relevant heat load patterns in any given building.
- A classification process for days allowing for the attribution of specific days to each of the identified load patterns. The formulation of the classification process is based only on data available from external sources (calendar and climate), allowing its application over existing heat load data, as well as forecasts.

All this is performed with a dataset containing real heat loads from 3 buildings connected to a DH network in Tartu, Estonia. A standardized methodology is developed allowing the same pipeline to be applied to all the buildings in the network. Even though results are shown for only 3 buildings, this methodology would allow DH operators to gain information about the heat consumption patterns in the buildings connected to the network. Thus, reducing excess heat production and enabling flexible production strategies to be created. Several outlier detection and data normalization methods are tested, and clustering processes are tested against more than 30 clustering validation indexes (CVIs).

The main novelty of the paper resides on the combination of unsupervised and supervised machine learning methods to identify and understand the energy demand patterns in real buildings. Therefore, this work contributes to the current literature by the presentation of a widely applicable method for the identification of main heat load patterns in buildings, while also allowing the analysis of the causes for these patterns, building a classification model for the prediction of the pattern. For this purpose, advanced machine-learning models are proposed, combining unsupervised and supervised methods, as shown in Section 2.

2. Methodology

The used multistep methodology is a combination of different unsupervised learning techniques and a supervised classification

model. This framework includes the application of a commonly used K-mean algorithm [37,38] and a DBSCAN algorithm [39] for the detection of outliers.

- Step 1: Match the two data sources used in the study: Data from the substations of the DH network and climatic data source.
- Step 2: Outlier identification.
- Step 3: Heat load data normalization for each daily profile.
- Step 4: Identification of heat consumption patterns using a clustering algorithm.
- Step 5: Evaluation of clustering processes using cluster validation indexes (CVIs)
- Step 6: Classification and Regression Trees (CART) for cluster classification.

This multistep methodology is illustrated in Fig. 1.

Each step of the methodology is detailed in the following subsections, as follows: Section 2.1 introduces the data sources used; Sections 2.2 and 2.3 introduce the pre-processing activities; Section 2.5 explains the different clusters obtained after the application of the unsupervised learning and provides a methodology for their evaluation based on the study of 32 CVIs. Finally, Section 2.5 associates the clusters and energy consumption patterns with such external variables as climatic and calendar variables using CARTs.

2.1. Data sources

Heating load profiles consists of real data from different substations of the DH network in Tartu (Estonia). Climatic data is obtained from a weather station located and managed by the Physics Institute of the University of Tartu [40]. This weather station collects data with a 15-min frequency. For this study, outdoor temperature (T_{OUT} in °C) and global solar radiation (G_T in W/m^2) are used. Different studies, such as [41] or [42] have shown that these variables are those with the highest correlation to daily heat energy consumption. According to the Köppen-Geiger classification [43], Tartu's climate is classified as D_{fb}, so the most determining climatic variable is the very low outdoor temperature. Minimal ambient temperature in winter can reach −20 °C, and all monthly averages fall below 20 °C. Relative humidity has not been considered in the model, due to its little relevance in the absence of cooling demand.

Heat load information has been provided by the DH operator GREN [44]. All these substations are located in the sub-network of Tarkon and each substation contains a smart meter that measures different variables in the system with an hourly frequency. The energy meter installed in the buildings is the Multical® 603 from Karsmtrup [45]. The accuracy of this device is better than that specified in the European directive for this purpose (EN-1434–1:2015 [46]), so the measuring error remains below 5% in all the variables. Heat energy consumption is saved as a cumulated variable and read hourly. Each substation corresponds to one building. Among the substations under study, different building types are included. The smart heat meter measures other variables in the system, such as the supply and return temperatures of the primary and secondary sides of the substation. However, for this work, no information other than the heat load is used. Fig. 2 presents the monitoring scheme of the heat meters used.

2.2. Outlier detection using DBSCAN

First, the identification and removal of the outliers from the original data is carried out by means of the DBSCAN algorithm [39]. Due to the type of data-sources obtained, the identification of outliers is limited to heat-meter data, assuming that climatic data is not so critical in the process. The objective of the DBSCAN algorithm is to identify high density observations that are close together and the

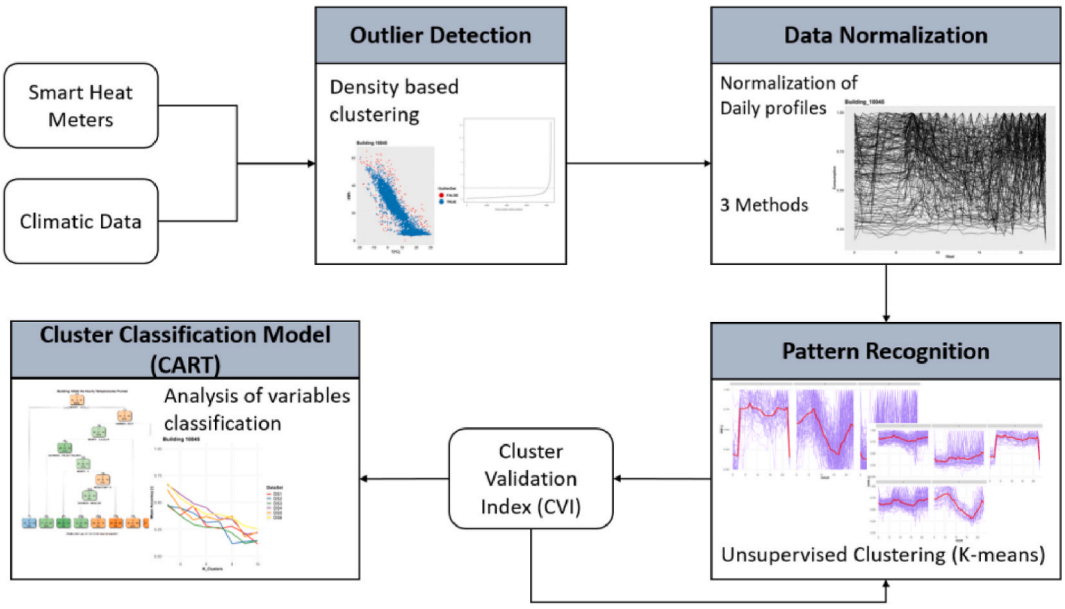

Fig. 1. General methodology.

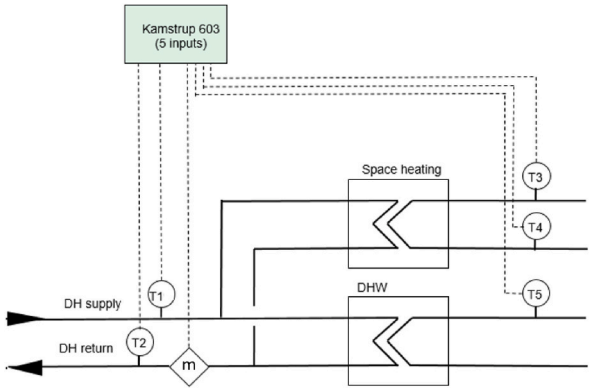

Fig. 2. Location and lay out of the smart energy meters in the DH in Tartu. Source [44].

points that are identified in low-density areas are considered outliers. This density-based clustering algorithm does not require the number of clusters to be pre-assigned, and outliers are usually coincident with isolated points in the low-density areas identified.

The outliers identified with this algorithm could have been caused by reading errors or by an unusual heating load. Regardless of the cause, these outliers are identified and removed from the original data in some of the datasets (DS1, DS2 & DS3 in Table 1).

The library "dbscan" [47] in R [48] is used to implement this algorithm. The observation radius (Eps) and the minimum amount of data points in each cluster (MinPts) are key parameters in the process. According to Hashler et al. (2021), the variable MinPts is initialized as the dimensionality of the dataset [47]. The calculation of the K-nearest neighbors' (k-NN) distances is carried out and the elbow of the ascending ordered distances correspond to the optimal Eps value. The elbow of that curve is calculated by means of the second derivative of the k-NN distances.

2.3. Definition of heat load profiles

Heat loads are known to vary in time. Lumbreras et al. [14] showed there are two types of heat-load variations:

- Intra-daily variations, where different load levels occur for each moment in time within the day. These variations might be caused by such variables and factors as climate, occupancy schedules, activation of thermostats and building management systems, as well as the transient response of buildings to the aforementioned issues.
- Inter-daily variations, where the variations are mainly associated to changes in how the building is used (i.e., bank holidays).

Within this work, heat load profiles are considered as 1-day long datasets. Each profile contains the variation of the heat load along the day. Considering the 1-h resolution in the data, arrays of 24 values are generated. In each building, profiles for all individual days are generated. Days with data gaps and/or outliers are discarded.

2.4. Normalization

The normalization of the daily heating energy consumption profiles is carried out for two reasons: shaping the energy consumption profiles for a better recognition of the patterns and the optimization of computational costs. When dealing with pattern recognition, the absolute value of the load is not considered as relevant as its variation throughout the day. All the values of the energy demand are ranged between 0 and 1 (except the normalization process using Eq. (3))). For this normalization process, three different equations are proposed in order to identify the best pre-processing conditions for each type of data:

$$q_{norm1}(t) = \frac{q(t) - q_{min}(t)}{q_{max}(t) - q_{min}(t)} \tag{1}$$

$$q_{norm2}(t) = \frac{q(t)}{q_{max}(t)} \tag{2}$$

Table 1
Generation of the 6 datasets (DS) and their pre-processing actions.

	Nomenclature	Outlier Removal	Norm. Eq. 1	Norm. Eq. 2	Norm. Eq. 3
DATA SET 1	DS1	YES	X		
DATA SET 2	DS2	YES		X	
DATA SET 3	DS3	YES			X
DATA SET 4	DS4	NO	X		
DATA SET 5	DS5	NO		X	
DATA SET 6	DS6	NO			X

$$q_{norm3}(t) = \frac{q(t) - \overline{q}}{sd(q)} \tag{3}$$

The generation of the datasets (DS in Table 1) for the next steps is a combination of the different pre-processing activities proposed. Therefore, these datasets are compared in terms of efficiency levels, to determine what the optimal preprocessing method is for this process. The characteristics of each data set are shown in Table 1.

2.5. Pattern recognition & cluster validation indexes (CVIs)

Energy demand patterns have been previously defined as daily loads or fractions of the daily energy profile that are repeated over different days in the same heating season. This repetitive character of the energy consumption patterns can be identified using unsupervised clustering algorithms. This paper explores the use of a commonly used algorithm, the so-called K-means [49] for the unsupervised clustering of the energy demand profiles. Hourly heat load data are re-structured so that each item in the dataset corresponds with one day and each day consists of 24 measurements.

The K-means algorithm is used to partition the dataset into K pre-defined groups or clusters, and each observation (i in Eq. (4)) belongs only to one group. This partition starts with a random selection of K centroids. The objective function (J) of this algorithm is the following (Eq. (4)):

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x_i - \mu_k\|^2 \tag{4}$$

where w corresponds to a relative weight, x is the measurement value (in this case, heat load), and μ is the cluster center.

After the random mapping of the initial centroids, the Euclidean distance between each point and the centroid is calculated to assign the point to its closest cluster center. Then, the centroid is updated with new values and this process is repeated until the centers do not change. Thus, the initially chosen K centroids may vary the clustering results and, consequently, for choosing optimal clustering, the algorithm is applied 50 times with different initial conditions for every K.

There are no initial indications to determine the optimal number of clusters to identify the different energy consumption patterns in the building. So the algorithm is applied for K = {3, 4 … 10}. K = 2 is skipped in order to avoid only weekday/weekend identification. Thus, for a specific building, eight different clustering processes are carried out.

The clustering effectiveness is evaluated using CVIs. For this study, more than 30 different CVIs are considered (concretely 32 indexes), including the most common indexes, such as the Silhouette index [50], the Dunn index [51], the C-index [52] and the Davies-Bouldin index [53], among others. These indexes evaluate the inter-cluster distances (between points in the same cluster) and intra-cluster distances (between points from different clusters). A low intra-cluster distance and high inter-cluster distance is indicative of separate, compact clusters. These CVIs are useful for the evaluation and comparison of the clustering processes from the same dataset, because the magnitude of these indexes is a function of the data used. Thus, the magnitudes of these indexes for different datasets are not comparable. It should be noted that some of the CVIs are optimized with the maximum value, others with the minimum value and yet more with the elbow method. Consequently, the resulting clusters from the process with more CVIs represent, in the best way, the different heat-load profile-types in each of the buildings.

2.6. Cluster classification model using CARTs

The CART algorithm is a supervised machine-learning technique that enables a predictive multiclass classification to be carried out [54]. It is used to infer the cluster classification with regards to external variables. Being a predictive tool, explanatory variables for the CART were selected based on such widely available information as climate and calendar information. The variables that are considered to be potentially relevant are the following: day of week (categorical), holiday/no holiday (categorical), month of the year (categorical), season (categorical), daily mean outdoor temperature (numerical) and daily total solar irradiance (numeric). The use of more variables in the classification model, such as hourly temperatures, was discarded due to the objective of finding easy logic to explain the identified clusters. The generation of classification trees is made via binary recursive partitioning. This algorithm is implemented using the rpart library [55] in R [48]. First, all the observations are partitioned at the root node, and then each of the two groups are divided into smaller subsets. The model in Ref. [55] is tuned with two parameters: *minsplit* (minimum number of samples in a node) and *cp* (cost complexity parameter). A low *cp* is first proposed to obtain an accurate model, and then different values for *minsplit* are tested. The one with the minimum error is used as the optimal value.

For the evaluation of the cluster prediction by the CARTs, the classification accuracy defined in Eq. (5) is used. This accuracy metric evaluates the number of correct classifications against the total number of predictions. Thus, this metric will vary from 0 to 1, where 1 corresponds to the perfect classification.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{5}$$

3. Results

3.1. Description of the buildings

A dataset with hourly heat load values for the year 2019 from a group of three buildings connected to the District Heating of Tartu (Estonia) is used. The methodology has been applied to the buildings independently, and most of the buildings show very different

energy consumption patterns. These three buildings are part of a larger dataset covering 42 buildings in a sub-network of the DH in Tartu. Thus, this selection has been made so that three buildings with completely different heat load patterns are shown caused by the different uses and energy consumption profiles. Each building is identified by a numeric code (ID number). The selected buildings are defined below, and their hourly demand observations are shown against the outside temperature in Fig. 3.

a) Building A. Residential building with space heating and domestic hot water demand throughout the year (Fig. 3a).
b) Building B. Educational building used as kindergarten showing heat load throughout the year, including space-heating and domestic hot water. The heat load in this building is not as linear as Building A and more than one trend is observed with low outdoor temperatures (Fig. 3b).
c) Building C. Commercial building with no, or at least very low, heat load in summer (Fig. 3c). Two main trends can be observed in winter (low outdoor temperature).

3.2. Outlier detection using DBSCAN

Outdoor temperature and solar irradiance are found to be the climatic variables with the highest correlation to heat consumption [41]. Density based clustering is applied to the hourly heat consumption against these two variables or dimensions and, consequently, MinPts is initialized as three.

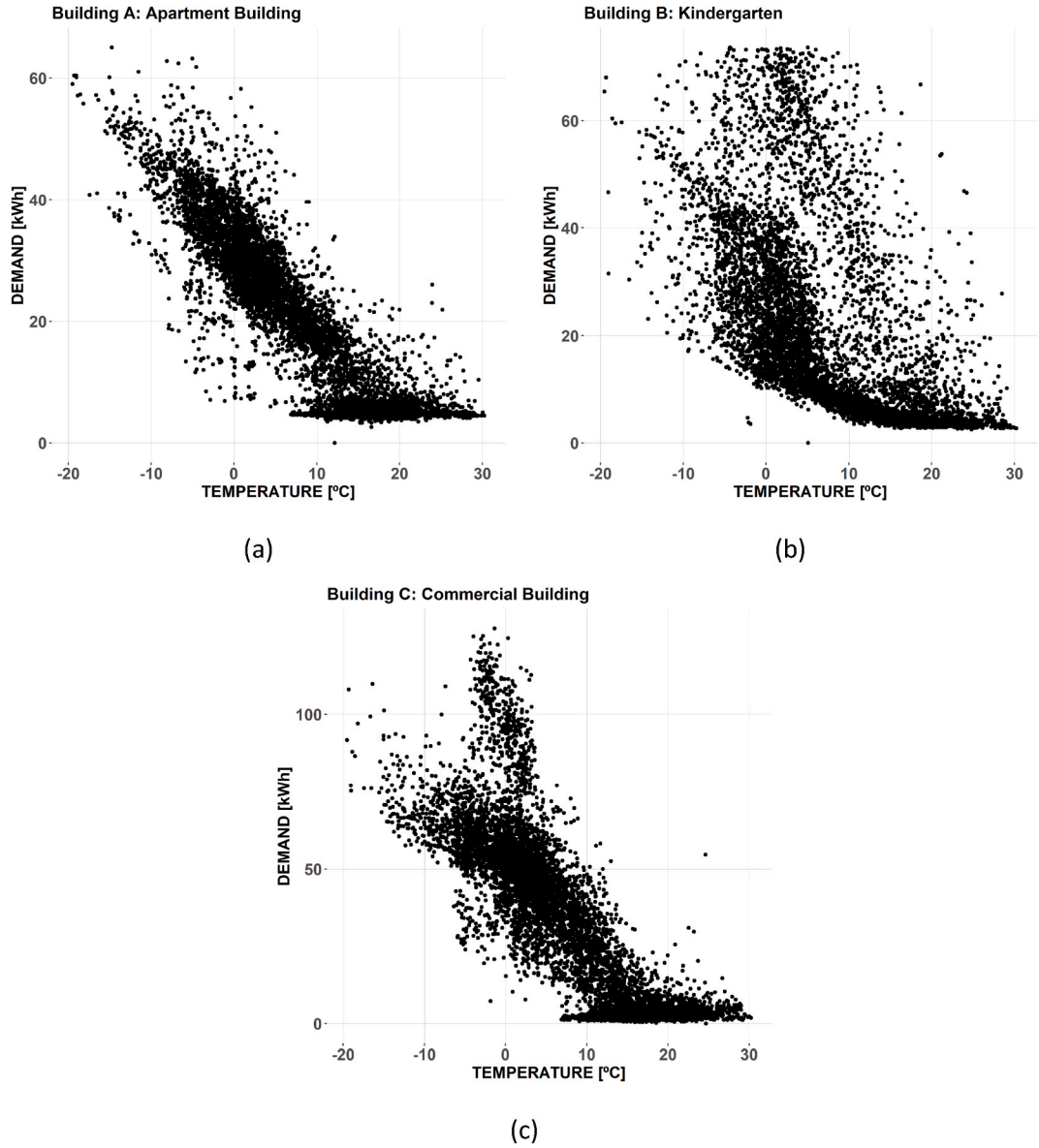

Fig. 3. Hourly energy consumption vs outdoor temperature for Building A (a), Building B (b) and Building C (c).

Following the methodology explained in Section 2.2, the 3-NN distance is calculated, and the *Eps* variable is calculated in the elbow of the sorted curve. Table 2 shows the number of observations (number of hours and complete days) in raw data and following the application of this algorithm. Moreover, the eps value that optimizes the process is also added.

The upcoming pattern recognition process requires complete observations of hourly load patterns along each day. To be consistent with this, full days are removed from the dataset if one or more hours within the day are considered as outliers. So, if one (or more) of the hourly energy demand points within one day is considered an outlier, the whole energy demand profile of that day is removed from the original dataset and is not used for pattern recognition.

3.3. Pattern recognition & cluster validation indexes (CVIs)

The K-means algorithm was used to identify daily pattern clusters. The significance of the identified clusters is assessed against CVIs. This method considered that the use of a large number of CVIs for cluster evaluation will contribute with a better pattern identification, since more different standards are included. Thus, Table 3, Table 4 and Table 5 summarize the number of CVIs that asserts which of the clustering processes developed is optimal for Building A, Building B and Building C, respectively. The quality of the clustering is assessed against datasets DS1 to DS6, considering the impact of outlier removal and normalization equations.

For Building A, the optimal number of clustering seems to be K = 3, for cases with normalized data, as per Eq. (2) (DS2 and DS4). There is also a relatively high number of CVIs that conclude that K = 5 is the optimal clustering process, with 7 CVIs with DS1. Consequently, Fig. 4 presents the clustering results for these two cases.

If the energy consumption patterns shown in Fig. 4 are visually analyzed, it can be inferred that both clustering approaches are not very different:

- Cluster 1 from DS4/k = 3 (Fig. 4a) corresponds to Cluster 3 in DS1/k = 5 (Fig. 4b).
- Cluster 2 from DS4/k = 3 and Cluster 5 & Cluster 1 from DS1/k = 5 correspond to the same pattern.
- Cluster 3 from DS4/k = 3 corresponds to Cluster 2 and Cluster 4 from DS1/k = 5.

Therefore, the following patterns were identified, based on the clusters from DS4, K = 3:

- In cluster 1, the heat load is heavily increased between 3am and 5am. It remains relatively constant and at very high values from 5am to 11pm. At 11pm, another strong demand variation is identified and the levels of the demand before 3am are maintained. The high demands along the day are caused by the very cold temperatures that Tartu (Estonia) usually presents in winter and requires a constant demand for SH. The strong variations are caused by a night setback induced by the DH operator, in which the set-point temperature is reduced. It is expected that the users of this residential building will be sleeping and there will be no need to maintain the same comfort conditions as at other times. This night setback means that the energy consumption differs from its dependency with climatic variables.
- The second cluster shows the most stable profile, grouping days with relatively constant energy consumption along the day in the same cluster. A relative peak demand is identified at 7–8am, coinciding with the same peak demand of the other clusters.
- In Cluster 3, the energy consumption gradually increases until approximately7-8am, when the peak demand is reached due to the DHW consumption in these hours. From 8am onwards, energy consumption decreases until 5pm, coinciding with the hours when the users of the building are supposed to be out of the building. After this hour, the demand starts to increase, up to the levels of the first hours of the day.

For building B, the results from Table 4 show a greater concurrence that the optimal clustering process is obtained with K = 3, especially in datasets DS3 and DS6. Fig. 5 presents the daily energy profiles obtained from this process with DS3. Similar to the analysis in Building A, these results were compared to results from K = 4 and DS4, which obtained 9 CVIs. As expected, different energy consumption profile types from those in Building A were found, since this building is used as a kindergarten.

As occurred in Building A, slight differences between profiles could be found in Building B:

- Cluster 1 from DS3/k = 3 (Fig. 4a) corresponds to Cluster 1 and Cluster 2 in DS4/k = 4 (b).
- Cluster 2 from DS3/k = 3 and Cluster 3 from DS4/k = 4 correspond to the same pattern.
- Cluster 3 from DS3/k = 3 corresponds to Cluster 4 from DS4/k = 4.

Therefore, based on the energy consumption profiles from K = 3, the patterns recognized from Fig. 5 are the following:

- Cluster 1 presents a quite stable energy consumption profile. This cluster groups the days in summer with no demand for SH and the days with very stable profiles with SH demand. A relative minimum demand is found at 12am. The energy profiles in this cluster show a very low correlation with climatic variables.

Table 2
Comparison of Raw data and clean data after the application of DBSCAN algorithm.

	Raw Data		After DBSCAN Data		Eps
	Hours	Complete Days	Hours	Complete Days	[−]
Building A	8408	320	8194	208	0.970
Building B	7973	198	7907	185	1.523
Building C	8403	314	8204	227	1.220

Table 3
Number of CVIs for optimal clustering process in Building A.

Nº of Clusters	Building A					
	DS1	DS2	DS3	DS4	DS5	DS6
K = 3	15	24	15	28	20	25
K = 4	0	1	0	1	3	2
K = 5	7	1	0	0	0	0
K = 6	2	1	1	0	3	1
K = 7	0	0	1	0	0	0
K = 8	4	1	3	0	3	0
K = 9	1	3	11	0	0	0
K = 10	5	3	2	5	4	6

Table 4
Number of CVIs for optimal clustering process in Building B.

Nº of Clusters	Building B					
	DS1	DS2	DS3	DS4	DS5	DS6
K = 3	15	17	28	19	22	26
K = 4	1	1	0	9	5	2
K = 5	0	3	0	0	1	0
K = 6	0	4	0	2	0	0
K = 7	0	3	0	0	0	0
K = 8	1	1	2	0	0	0
K = 9	0	0	0	0	1	1
K = 10	16	4	3	3	5	4

Table 5
Number of CVIs for optimal clustering process in Building C.

Nº of Clusters	Building C					
	DS1	DS2	DS3	DS4	DS5	DS6
K = 3	27	27	11	19	2	23
K = 4	0	0	15	5	3	5
K = 5	0	2	1	0	23	0
K = 6	0	0	0	0	1	0
K = 7	2	2	0	0	0	0
K = 8	0	0	1	0	1	0
K = 9	0	1	1	0	0	1
K = 10	5	2	4	10	4	4

- Cluster 2 is similar to Cluster 1, but a greater load reduction is observed at approximately 12am, increasing dependency on the climatic variables. At noon and coinciding with the hours with the highest ambient temperature and highest solar irradiance levels, the demand is reduced.
- Cluster 3 shows the profile with greater variability. The energy consumption profiles in this cluster remain relatively constant until approximately 7am. This time coincides with a common opening hour of kindergartens, or shortly before, so it is possible to condition the building before the arrival of the occupants. At this time, a steep increase in the demand is observed, reaching the first peak at around 10am. In the next hour, the demand slightly decreases, probably taking advantage of the thermal inertia of the building. Then from 12noon to 1pm (more or less), another increase in the demand is observed, and from 1pm to 3pm the demand increases again. A third maximum demand, in this case a relative maximum, is observed at around 4pm, before the demand starts to decrease to the levels of the first hours of the day. Finally, at around 6pm, the demand reaches a relatively constant value for the rest of the day.

Finally, for Building C, the most predominant optimal clustering results are obtained with K = 3 in all the datasets, but with K = 4 (DS3) and K = 5 (DS5) also showing good results, as can be observed in Table 5. Additionally, Fig. 6 presents the obtained clusters for K = 3 with DS2 and K = 5 with DS5, corresponding to the two processes with the largest amount of CVIs.

The following patterns are identified, based on the clusters from DS2, K = 3:

- In Cluster 1, days with relatively constant profiles are grouped, including days with very low demand and other days with intermediate loads.
- The daily energy profiles in Cluster 2 correspond to intermediate load days. As in Cluster 1, the maximum demand is identified at around 7am. However, from 7am onwards, the demand decreases, probably due to the more favorable climatic conditions outside.

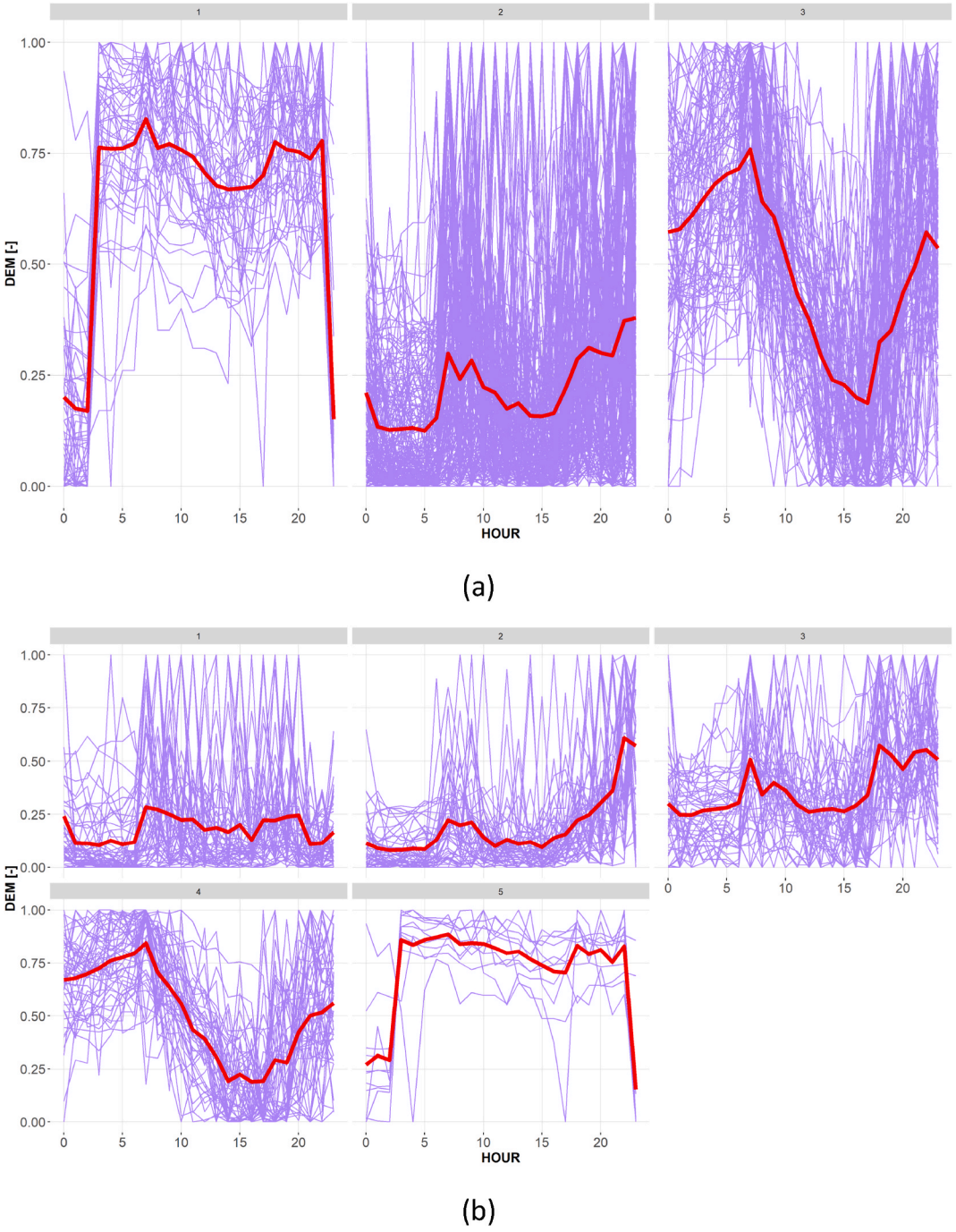

Fig. 4. Daily energy consumption clusters for normalized data in Building A (apartments building): a) K = 3 with DS4 and b) K = 5 with DS1.

The lowest demand period is found at around 1pm, and then the demand gradually increases until 9pm. At this time, the night setback decreases the set point of the SH and the demand returns to values of the first hours of the day.

- In Cluster 3, a relatively constant demand is observed until 7am, when the demand drastically increases. At this time, the shopping building may open, coinciding with when the potential customers start to use this building. A high demand is maintained from approximately 7am to 9pm, probably coinciding with the opening hours of this building. Finally, at 9pm, the heating demand returns to the same values as the first hours of the day.

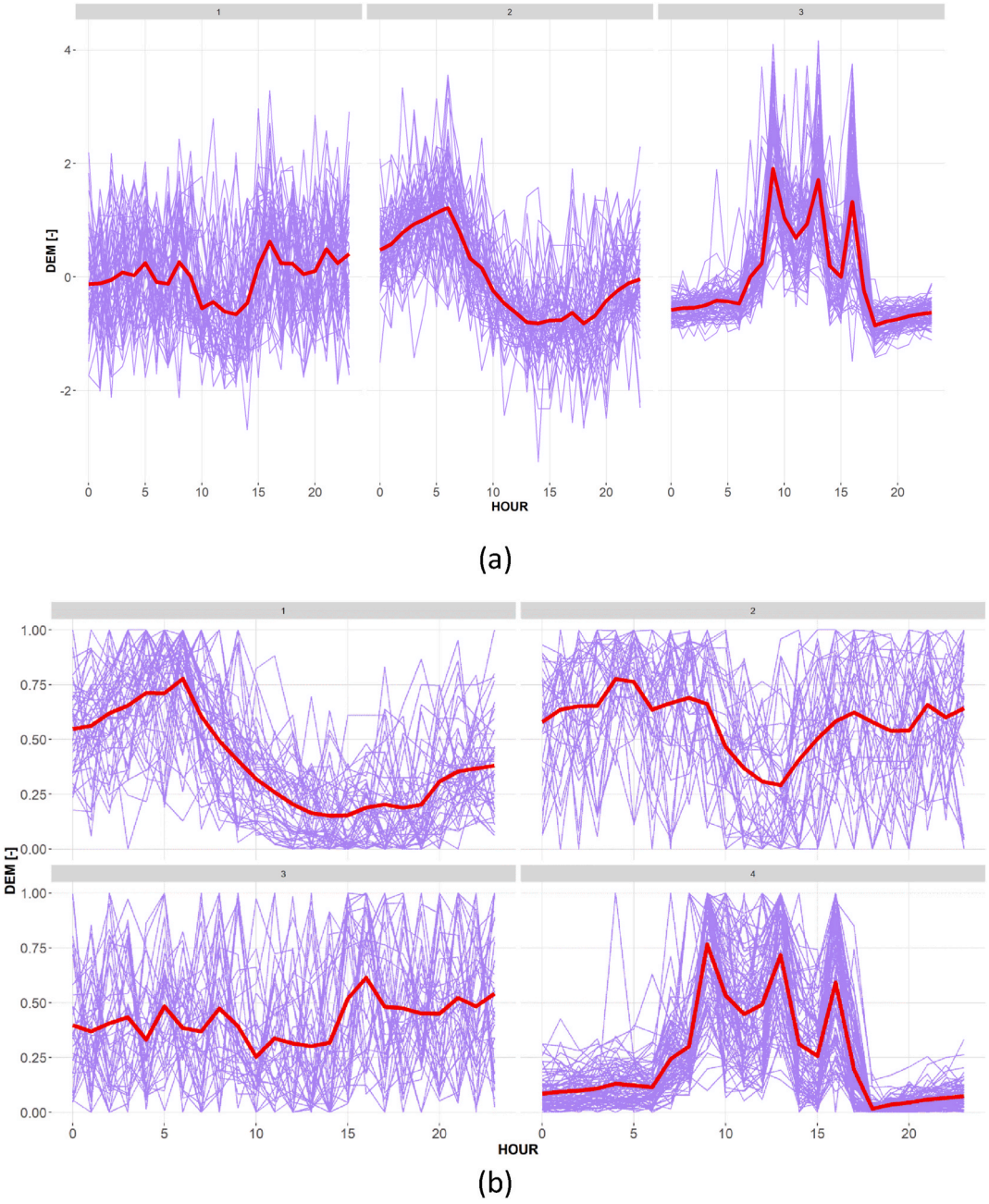

Fig. 5. Daily energy consumption clusters for normalized data in Building B (kindergarten): (a) K = 3 with DS3 and (b) K = 4 with DS4.

3.4. Cluster classification model using CARTs

Classification models using the CART algorithm are proposed to identify factors that lead to the different patterns identified in each building. This algorithm, based on a binary logistic classification, enables us to obtain a model in which the clusters are classified according to the variables introduced as predictors.

It is known that simpler CART models are more robust when using for prediction purposes. As well as more conditions/predictors are introduced an input to the CART model, the possibility to go wrong in each classification step increases the whole classification accuracy. For that reason, post-pruning processes were conducted to keep the generated binary tree as simple as possible. As a first approach, a very low complexity factor ($cp = 0.001$) was used for determining the optimal MinSplit and then, cp value was determined by one-standard error rule. Fig. 7 shows the obtained accuracy for each of the buildings, divided by datasets and number of clusters.

All the buildings show the maximum accuracy for clustering processes with 3 clusters (K = 3). This matches the conclusion from CVI analysis, in which K = 3 was the optimal clustering process for pattern recognition in the three buildings. In the following lines, the

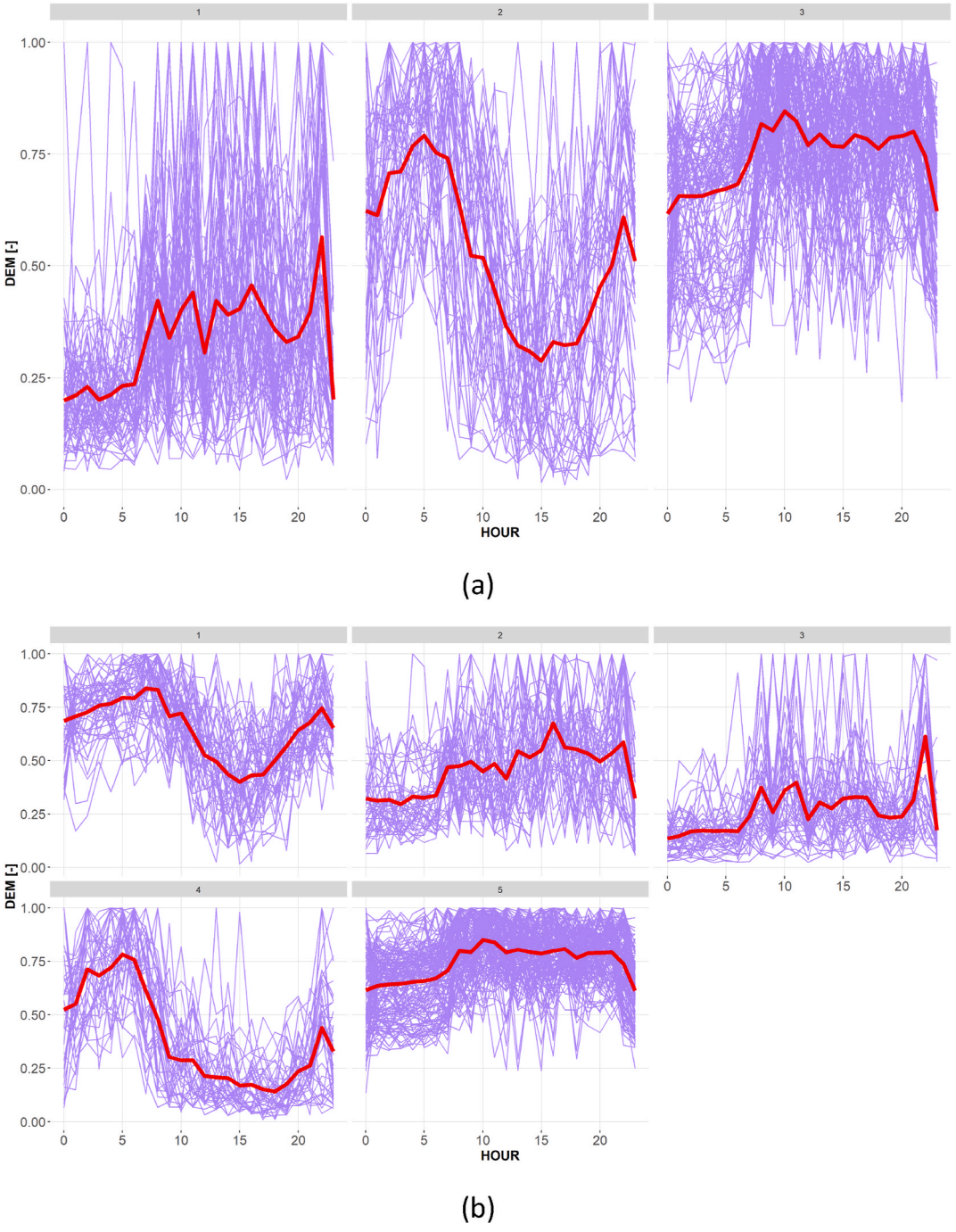

Fig. 6. Daily energy consumption clusters for normalized in Building C (Commercial building) (a) K = 3 with DS2 and (b) K = 5 with DS5.

dataset with the best accuracy is selected for each building, and the resulting CART models are analyzed to understand the relevance of predictor variables in setting the heat load pattern of the buildings. For this purpose, Fig. 8 enables the visualization of the optimal pruned CART and the variables that determine the decision making in these classification models. In each of the boxes that make up the models, the first number defines the number of the predominant cluster in that step of the model. In the second row, the distribution percentage of the existing clusters is presented and, finally, the number in the third row indicates the fraction of the data remaining after the previous classification step. Thus, in the first box of all the CARTs, the number in the third row is 100%.

For Building A (apartment building), Fig. 7a presents the CART model based on DS6, closely followed by DS2, which is the optimal clustering process determined by CVIs. In general terms, Cluster 3 incorporates days from summer and mid-season periods with low

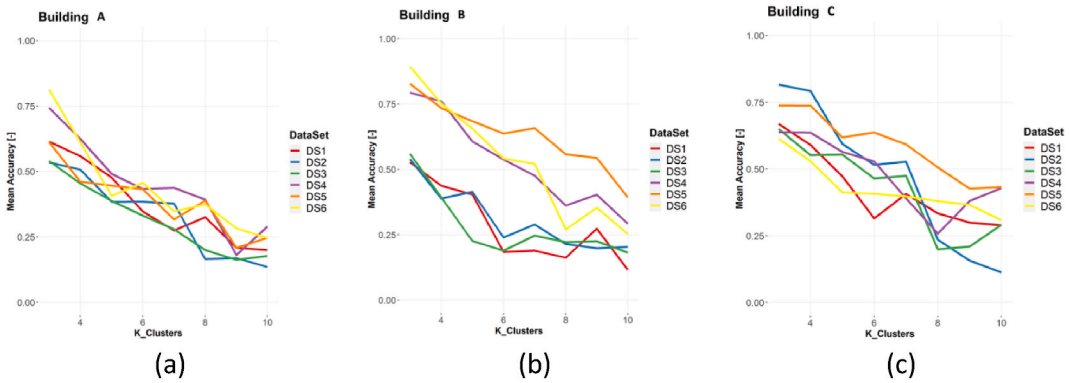

Fig. 7. CARTs accuracy evolution against number of clusters in (a) Building A, (b) Building B and (c) Building C.

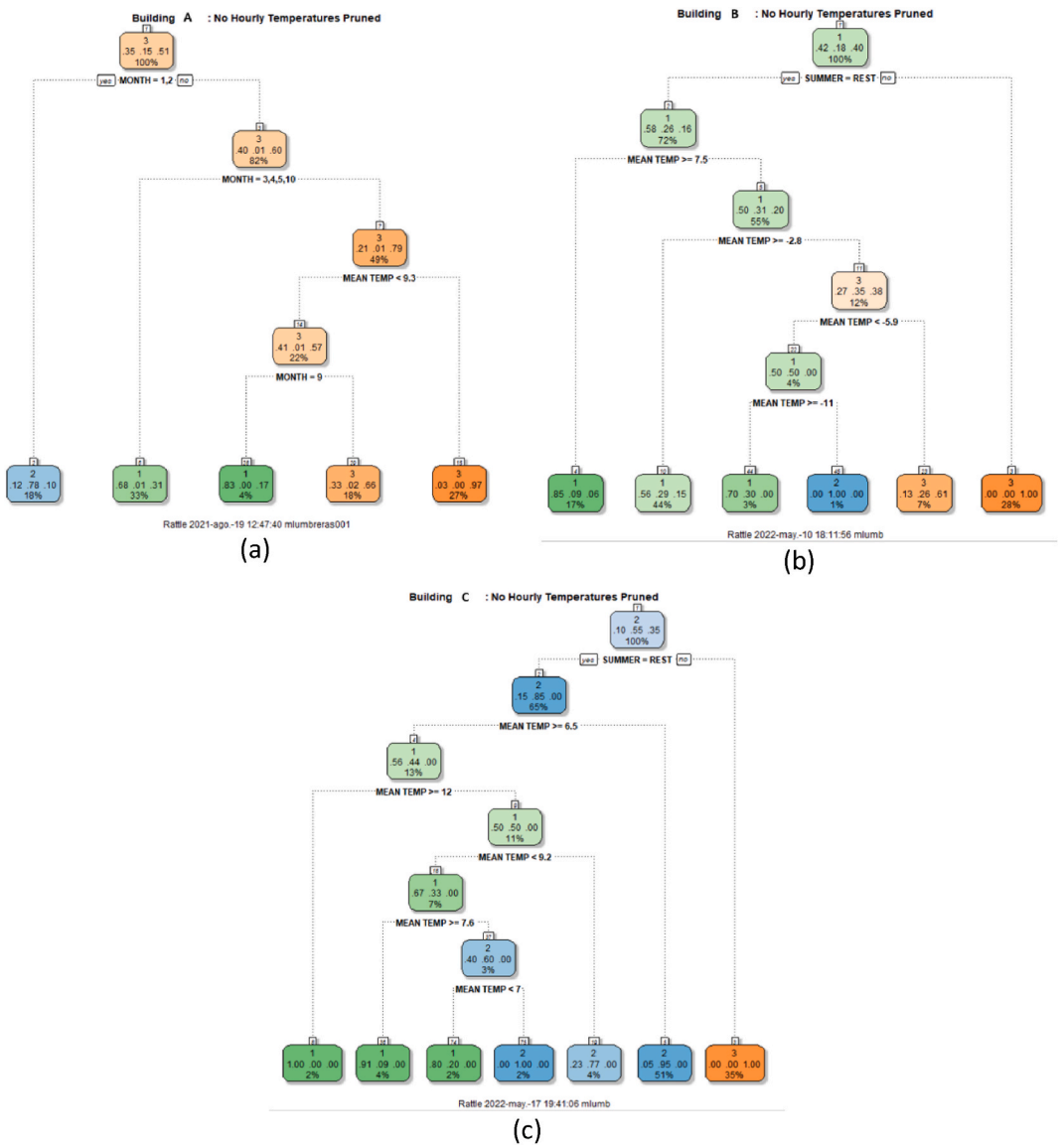

Fig. 8. CART pruned models in (a) Building A, (b) Building B and (c) Building C.

and stable loads. Cluster 2 groups days in January and February, coinciding with cold days (high heat load), when there is a night setback in the demand. Finally, Cluster 1 groups the rest of the days, when the demand is highly reduced at mid-day. Thus, the main variables affecting this model are the Month of the year (with particular focus on months 1 and 2), the summer period, the day of the week, and the daily mean temperature.

For Building B (kindergarten), Fig. 7b presents the CART model based on DS6. The CVIs also indicated that this process is optimal for pattern identification. Regarding the CART model, this classification model shows that the most determining variable for classification is the seasonal period (95% of the days were properly classified, considering only "summer" and "month" variables). Thus, Cluster 3 is composed by heating days in the summer period and some days in the heating season, when the daily mean temperature is above −5.9 °C. Cluster 2 gathers daily heating profiles of the days in the heating season, when the daily mean temperature is below −11 °C in the first and last months of the year. The days grouped in Cluster 2 are the rest of the days in the mid-season.

Finally, for Building C (shopping), Fig. 7c presents the CART model based on DS2 and three clusters. The clusters of this classification have been shown in Fig. 5a. This classification model also shows that the most determining classification variable is the season, followed by the daily mean temperature (91% of the days were properly classified, considering only the "summer" and "Mean outdoor temperature" variables). Thus, Cluster 2 is formed by days in the summer period and the mid-season with daily mean temperatures above (or equal to) 16 °C. Moreover, Cluster 1 groups demand profiles in the first and last months of the year (heating season) and the mid-season with low outdoor temperatures. Finally, Cluster 3 gathers the days that are not classified in Cluster 1 or 2.

4. Discussion

This paper proposes a framework combining different data-driven techniques to obtain deep insights into the energy consumption patterns inside a building. Unsupervised clustering is used for outlier detection and pattern recognition; whereas supervised learning techniques, by means of CART, are used to understand the potential factors influencing the unsupervised identification of energy consumption patterns.

In each of the presented buildings, the statistical analysis of the CVIs demonstrates that, in most cases, the case clustering process is optimal with K = 3. Table 6 presents the energy consumption patterns identified in each building, describing the most distinctive characteristic of the profile and the potential reasons for this demand based on the results from the classification model.

Thus, unsupervised learning enables us to identify and analyze different numbers of clusters in each of the buildings. The difference in the number of clusters identified is caused by the different heat consumption patterns in each building, which in turn is caused by the different occupants' behavior and the DH energy management strategies in the network.

Regarding Building A (residential building), three main patterns have been identified, and Table 6 describes the energy consumption profiles. The first pattern reveals the night setback that rules the energy demand in that building from 11pm to 3am. However, the energy demand (including DHW + SH) increases and stays relatively constant throughout the day. This consumption pattern matches the very cold months in winter, when the SH consumption is much higher than the DHW and, therefore, there are no relevant energy consumption peaks during these days. The second pattern reveals a typical energy consumption for mid-season, when the energy consumption for DHW is similar to that for SH and, consequently, the DHW consumption peaks are not very relevant. Finally, the third consumption pattern identified enables us to visualize the energy consumption for summer days, when there is no demand for SH. Thus, the energy consumption profile of this cluster roughly matches the DHW consumption profile in this building.

Regarding Building B, three main patterns were also identified, and the energy profiles are described in Table 6. The first pattern of

Table 6
Summary of patterns identified and their main characteristics.

Building	Pattern	Clusters	Description of the profile
A	1	Cl. 1 (K = 3 & DS4)	Strong increase between 3am and 5am.
		Cl. 5 (K = 5 & DS1)	Relatively constant and at very high values from 5am to 11pm. At 11pm, another strong demand variation is identified and the levels of the demand before 3am are maintained.
	2	Cl. 2 (K = 3 & DS4)	Relatively constant energy consumption along the day.
		Cl. 1,2 & 3 (K = 5 & DS1)	A relative peak demand is identified at 7–8am.
	3	Cl. 3 (K = 3 & DS4)	Energy consumption gradually increases until approximately 7–8am.
		Cl. 4 (K = 5 & DS1)	From 7am onwards, energy consumption decreases until 5pm.
B	1	Cl. 1 (K = 3 & DS5)	Stable energy consumption profile.
		Cl. 1 & 2 (K = 4 & DS4)	A relative minimum demand is found at noon.
	2	Cl. 2 (K = 3 & DS5)	Great load reduction is observed at approximately noon.
		Cl. 3 (K = 4 & DS4)	At noon, the demand is reduced.
	3	Cl. 3 (K = 3 & DS5) Cl. 4 (K = 4 & DS4)	A steep increase in the demand is observed, reaching the first peak at around 10am. From noon to 1pm (more or less), another increase in the demand is observed and from 1pm to 3pm the demand increases again. A third maximum demand at around 16pm. At around 6pm, the demand reaches a relatively constant value for the rest of the day.
C	1	Cl. 1 (K = 3 & DS2) Cl. 3 (K = 4 & DS3)	Relatively constant heating consumption profiles
	2	Cl. 2 (K = 3 & DS2)	The maximum demand is identified at around 7am. From 7am onwards, the demand decreases.
		Cl. 1 (K = 4 & DS3)	The lowest demand period is found at around 1pm. The demand gradually increases until 9pm.
	3	Cl. 3 (K = 3 & DS2)	Relatively constant demand is observed until 7am, when the demand drastically increases.
		Cl. 2 & 4 (K = 4 & DS3)	A high demand is maintained from approximately 7am to 9pm.

this building shows the most stable consumption profile and matches the mid-season consumption, when the demand for SH and DHW are similar. Thus, there are no relevant demand peaks throughout the day. The second pattern in Fig. 5 shows a similar profile to the third pattern in Building A and, similarly, this consumption profile matches the typical demand in summer. In these days, the unique demand is the one independent from climatic conditions; even though the kindergarten is supposed to be empty of children, there might be activity inside the building. Finally, the third pattern matches the heat consumption in the cold days in winter. The three consumption peaks are caused by the SH demand required over these days and probably matches the occupational pattern of the building.

Similar to the other two buildings, Building C also presents three main patterns, and the profiles of these patterns are described in Table 6. The first pattern corresponds to the typical profile in summer days, when the heat demand in the building is very low. There is residual heat consumption when the commercial building is supposed to be open. The second pattern in Building C is very similar to the second pattern in Building B and the third of Building A. Thus, the conclusions drawn are the same for this building. Finally, the third pattern identified in Building C corresponds to the heat consumption profile for winter days, when there is a high and relatively constant SH demand along the day due to the low outdoor temperature and the continuous comfort requirement in that building.

In the vast majority of cases, the optimal clustering process is obtained with three clusters, even though in some cases the CVI analysis suggests that a higher number of clusters is more effective. However, when observing the heat consumption profiles, the identified patterns show slight differences between the energy consumption profiles. The results may vary depending on the CVI chosen for the analysis. Therefore, the unsupervised learning of energy consumption patterns requires a post-processing to interpret the results. In addition, a high value for K in this unsupervised clustering process could result in small differences between the daily energy profiles in different clusters. A higher computational cost is required for sub-setting the original dataset into different clusters. In prediction models in general, sub-setting the data into a large number of clusters usually provides better results, since the optimal clustering process is the one with as many clusters as observations, in this case, energy demand profiles.

CARTs are modeled so that the potential factors influencing the hourly heating energy consumption are discovered. A higher accuracy when using CART means that there is a higher correlation between the external variables introduced to the models and the identified patterns. Thus, a higher correlation between the external variables and the consumption profiles causes more realistic patterns, since these consumption patterns may always be caused by occupational behavior or climatic factors. Therefore, the DS6 dataset obtains the highest accuracy levels in Building A and Building B, whereas DS2 is the one with the highest accuracy. The maximum classification accuracy obtained in Building A reached 0.71, that of Building B reached 0.89, and in Building C the model reached an accuracy of 0.82. A clear and high correlation between the results obtained from CVIs and the accuracy of the CART was observed. Therefore, lots of references between section 3.3 and section 3.4 have been used along this paper. The high correlation between these two sections illustrates the correct use of this methodology for the unsupervised pattern recognition.

5. Conclusions

A general framework for the identification of heating energy consumption patterns in buildings connected to a DH network has been presented. The methodology is based on a combination of unsupervised clustering learning methods and supervised classification models. The results for three representative buildings with different energy consumption profiles and uses are studied in detail. From the results, the following conclusions are drawn:

- The final use of the buildings and, consequently, the users' behaviors and energetic requirements determine the energy consumption differences. On the whole, buildings for residential purposes present a night setback in the heating season (no setback in summer); whereas other types of building patterns depend on the particular use and occupation of the said buildings.
- Seasonal classification turns out to be the most determining variable affecting the energy consumption patterns, followed by the outdoor daily mean temperature, which greatly affects the energy consumption. The daily mean solar radiation has little or no effect on the energy consumption pattern recognition.
- A high classification accuracy from the CART means that the developed model is close to the real behavior of the demand in the building. The maximum classification accuracy is obtained with a low number of clusters ($K = 3$), showing the same results as the clustering accuracy study using CVIs. Thus, the $K = 3$ clustering process and the CART model followed are optimal for the characterization of the buildings' demand.
- Regarding the analyzed datasets (Table 1), the statistical composition of each of the buildings' demand determines the optimal normalization process in each building. There is no unique optimal dataset for all the buildings under study. Moreover, even though the use of density-based clustering for outlier identification has proved to be effective, in general, better accuracy results have been obtained with DS4, DS5 & DS6 (datasets with outliers). This is due to the lower number of variables in datasets with no outliers and the lower number of variables to train the model.

CRediT author statement

Mikel LUMBRERAS: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing Original Draft, Writing Review & Editing and Visualization. Roberto GARAY-MARTINEZ: Conceptualization, Funding acquisition, Formal analysis and Writing - review &editing. Beñat ARREGI: Formal analysis and Writing - review &editing. Gonzalo DIARCE: Conceptualization, Methodology, Writing Review & Editing, Supervision, Project administration. Koldobika MARTIN-ESCUDERO: Conceptualization, Methodology, Writing Review & Editing, Supervision.

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.

Data availability

The data that has been used is confidential.

Acknowledgements

The authors would like to thank GREN Eesti [44] 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] A. Peltokorpi, M. Talmar, K. Castrén, J. Holmström, Designing an organizational system for economically sustainable demand-side management in district heating and cooling, J. Clean. Prod. 219 (2019) 433–442, https://doi.org/10.1016/J.JCLEPRO.2019.02.106.
[2] I.E.A, IEA, Energy Technology Perspectives Scenarios and Strategies to 2050, 2012. Paris.
[3] European Commision, Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on Energy Efficiency, Amending Directives 2009/125/EC and 2010/30/EU and Repealing Directives 2004/8/EC and 2006/32/EC Text with EEA Relevance OJ L, vol. 315, 2012.
[4] European Comission, Directive (EU) 2018/844 of the European Parliament and of the Council of 30 May 2018 Amending Directive 2010/31/EU on the Energy Performance of Buildings and Directive 2012/27/EU on Energy Efficienc, 2018.
[5] Y. Cao, J. Du, E. Soleymanzadeh, Model predictive control of commercial buildings in demand response programs in the presence of thermal storage, J. Clean. Prod. 218 (2019) 315–327, https://doi.org/10.1016/J.JCLEPRO.2019.01.266.
[6] T. Hong, Z. Wang, X. Luo, W. Zhang, State-of-the-art on research and applications of machine learning in the building life cycle, Energy Build. 212 (2020), 109831, https://doi.org/10.1016/J.ENBUILD.2020.109831.
[7] A. Kathirgamanathan, M. de Rosa, E. Mangina, D.P. Finn, Data-driven predictive control for unlocking building energy flexibility: a review, Renew. Sustain. Energy Rev. 135 (2021), 110120, https://doi.org/10.1016/J.RSER.2020.110120.
[8] H. Lund, Renewable energy strategies for sustainable development, Energy 32 (2007) 912–919, https://doi.org/10.1016/J.ENERGY.2006.10.017.
[9] H. Lund, S. Werner, R. Wiltshire, S. Svendsen, J.E. Thorsen, F. Hvelplund, B.V. Mathiesen, 4th Generation District Heating (4GDH): integrating smart thermal grids into future sustainable energy systems, Energy 68 (2014) 1–11, https://doi.org/10.1016/J.ENERGY.2014.02.089.
[10] H. Li, N. Nord, Transition to the 4th generation district heating - possibilities, bottlenecks, and challenges, Energy Proc. 149 (2018) 483–498, https://doi.org/10.1016/J.EGYPRO.2018.08.213.
[11] J. von Rhein, G.P. Henze, N. Long, Y. Fu, Development of a topology analysis tool for fifth-generation district heating and cooling networks, Energy Convers. Manag. 196 (2019) 705–716, https://doi.org/10.1016/J.ENCONMAN.2019.05.066.
[12] A. Sleptsov, E. Crisostomi, A. Bischi, Control schemes for district heating substations considering user-defined building's indoor temperature, Build. Environ. 191 (2021), 107598, https://doi.org/10.1016/J.BUILDENV.2021.107598.
[13] T. Cholewa, A. Siuta-Olcha, A. Smolarz, P. Muryjas, P. Wolszczak, R. Anasiewicz, C.A. Balaras, A simple building energy model in form of an equivalent outdoor temperature, Energy Build. 236 (2021), 110766, https://doi.org/10.1016/J.ENBUILD.2021.110766.
[14] M. Lumbreras, R. Garay-Martinez, B. Arregi, K. Martin-Escudero, G. Diarce, M. Raud, I. Hagu, Data driven model for heat load prediction in buildings connected to District Heating by using smart heat meters, Energy 239 (2022), 122318, https://doi.org/10.1016/J.ENERGY.2021.122318.
[15] T. Cholewa, A. Siuta-Olcha, A. Smolarz, P. Muryjas, P. Wolszczak, Ł. Guz, C.A. Balaras, On the short term forecasting of heat power for heating of building, J. Clean. Prod. 307 (2021), 127232, https://doi.org/10.1016/J.JCLEPRO.2021.127232.
[16] K. el Bouchefry, R.S. de Souza, Learning in Big Data: Introduction to Machine Learning, Knowledge Discovery in Big Data from Astronomy and Earth Observation, 2020, pp. 225–249, https://doi.org/10.1016/B978-0-12-819154-5.00023-0.
[17] M. Emre Celebi, Partitional Clustering Algorithms, Springer International Publishing, Cham, 2015, https://doi.org/10.1007/978-3-319-09259-1.
[18] M. Emre Celebi, A. Kemal, Unsupervised Learning Algorithms, Springer International Publishing, Cham, 2016, https://doi.org/10.1007/978-3-319-24211-8.
[19] Z. Dong, J. Liu, B. Liu, K. Li, X. Li, Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification, Energy Build. 241 (2021), https://doi.org/10.1016/j.enbuild.2021.110929.
[20] C. Fan, F. Xiao, Z. Li, J. Wang, Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: a review, Energy Build. 159 (2018), https://doi.org/10.1016/j.enbuild.2017.11.008.
[21] K. Zhou, S. Yang, Z. Shao, Household monthly electricity consumption pattern mining: a fuzzy clustering-based model and a case study, J. Clean. Prod. 141 (2017), https://doi.org/10.1016/j.jclepro.2016.09.165.
[22] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, J. Li, A review of data mining technologies in building energy systems: load prediction, pattern identification, fault detection and diagnosis, Energy and Built Environment 1 (2020), https://doi.org/10.1016/j.enbenv.2019.11.003.
[23] J.Y. Park, X. Yang, C. Miller, P. Arjunan, Z. Nagy, Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset, Appl. Energy 236 (2019), https://doi.org/10.1016/j.apenergy.2018.12.025.
[24] L. Wen, K. Zhou, S. Yang, A shape-based clustering method for pattern recognition of residential electricity consumption, J. Clean. Prod. 212 (2019), https://doi.org/10.1016/j.jclepro.2018.12.067.
[25] X. Liu, Y. Ding, H. Tang, F. Xiao, A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data, Energy Build. 231 (2021), https://doi.org/10.1016/j.enbuild.2020.110601.
[26] C.M.R. do Carmo, T.H. Christensen, Cluster analysis of residential heat load profiles and the role of technical and household characteristics, Energy Build. 125 (2016), https://doi.org/10.1016/j.enbuild.2016.04.079.
[27] A. Rajabi, M. Eskandari, M. Jabbari Ghadi, S. Ghavidel, L. Li, J. Zhang, P. Siano, A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications, Energy Build. 203 (2019), https://doi.org/10.1016/j.enbuild.2019.109455.
[28] S. Haben, C. Singleton, P. Grindrod, Analysis and clustering of residential customers energy behavioral demand using smart meter data, IEEE Trans. Smart Grid 7 (2016), https://doi.org/10.1109/TSG.2015.2409786.
[29] C.H. Jin, G. Pok, Y. Lee, H.W. Park, K.D. Kim, U. Yun, K.H. Ryu, A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting, Energy Convers. Manag. 90 (2015) 84–92, https://doi.org/10.1016/J.ENCONMAN.2014.11.010.
[30] F. Wernstedt, P. Davidsson, C. Johansson, Demand side management in district heating systems, in: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, 2015, p. 272.
[31] Z. Ma, H. Li, Q. Sun, C. Wang, A. Yan, F. Starfelt, Statistical analysis of energy consumption patterns on the heat demand of buildings in district heating systems, Energy Build. 85 (2014), https://doi.org/10.1016/j.enbuild.2014.09.048.

[32] H. Gadd, S. Werner, Fault detection in district heating substations, Appl. Energy 157 (2015), https://doi.org/10.1016/j.apenergy.2015.07.061.
[33] A.M. Tureczek, P.S. Nielsen, H. Madsen, A. Brun, Clustering district heat exchange stations using smart meter consumption data, Energy Build. 182 (2019), https://doi.org/10.1016/j.enbuild.2018.10.009.
[34] E. Calikus, S. Nowaczyk, A. Sant'Anna, H. Gadd, S. Werner, A data-driven approach for discovering heat load patterns in district heating, Appl. Energy 252 (2019), https://doi.org/10.1016/j.apenergy.2019.113409.
[35] P. Gianniou, X. Liu, A. Heller, P.S. Nielsen, C. Rode, Clustering-based analysis for residential district heating data, Energy Convers. Manag. 165 (2018) 840–850, https://doi.org/10.1016/J.ENCONMAN.2018.03.015.
[36] H. Johra, D. Leiria, P. Heiselberg, A. Marszal-Pomianowska, T. Tvedebrink, Treatment and analysis of smart energy meter data from a cluster of buildings connected to district heating: a Danish case, in: E3S Web of Conferences, vol. 172, 2020, pp. 2–9, https://doi.org/10.1051/e3sconf/202017212004.
[37] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297. California.
[38] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recogn. Lett. 31 (2010), https://doi.org/10.1016/j.patrec.2009.09.011.
[39] 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, 1996, pp. 226–231. Portland, Oregon.
[40] University of Tartu, Institute of Physics, Laboratory of Environmental Physics, 2021. http://meteo.physic.ut.ee/?lang=en.
[41] F. Margaret, PRISM: an introduction, Energy Build. 9 (1986) 5–18.
[42] L. Ferbar Tratar, E. Strmčnik, The comparison of Holt–Winters method and Multiple regression method: a case study, Energy 109 (2016) 266–276, https://doi.org/10.1016/J.ENERGY.2016.04.115.
[43] M. Kottek, J. Grieser, C. Beck, B. Rudolf, F. Rubel, World Map of the Köppen-Geiger climate classification updated, Meteorol. Z. 15 (2006) 259–263, https://doi.org/10.1127/0941-2948/2006/0130.
[44] G.R.E.N. Eesti, GREN Eesti, 2021. https://Gren.Com/Ee/.
[45] Karmstrup, 2021. https://www.kamstrup.com/en-us/heat-solutions/heat-meters/multical-603.
[46] En, EN 1434-1:2015, Heat Meters. Part 1: General Requirements, 2015.
[47] M. Hashler, M. Piekenbrock, S. Arya, D. Mount, R. Package, 'dbscan' 2020, 2021.
[48] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2013.
[49] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297, 14.
[50] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[51] J. Dunn, Well separated clusters and optimal fuzzy partitions, J. Cybern. 4 (1974) 95–104.
[52] L. Hubert, J. Schultz, Quadratic assignment as a general data-analysis strategy, Journal of Mathematical and Statistical Psychologie (1976) 190–241.
[53] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1979) 224–227.
[54] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth Inc., 1984.
[55] T. Therneau, B. Atkinson, B. Ripley, M.B. Ripley, R Package 'rpart', 2020.