University College London

Dept. of Electrical and Electronic Engineering

Analysis and optimisation of building management systems through Data Analytics and Machine Learning

MSc Thesis in Integrated Machine Learning Systems

Author

Polyxeni Kalliga

Student number

SN20159465

Supervisor

Dr. Ryan C Grammenos

University College London

Dept. of Electrical and Electronic Engineering

2020/2021

Abstract

The increasing demand for building services and user thermal comfort, along with the time spent indoors, have surged buildings' energy consumption, being an issue of high environmental and economic concern. Along with the energy consumption increase, there has been a mounting realization that superfluous power consumption is quite prevalent due to either malfunctioning equipment or incorrectly configured control systems. All the same, with the increased development of sensor technology along with communications and data analytics, energy systems of smart buildings have been moving well beyond their original function to intelligent systems, making use of technology that enables them to become efficient, reduce costs and emissions, detect faults and become more transparent in terms of operation. Ensuring the resilience of energy data and understanding the dynamics, or else the patterns, of building energy consumption are the key enablers for achieving such gains and has become a major concern of researchers. However, previous works have focused on identifying the energy usage patterns of buildings using either baseline limited statistics or black-box Data Mining methods, rather that interpretable analysis. To this end, this work proposes a general Data Wrangling-based framework for discovering and interpreting insightful knowledge hidden in building's energy patterns in a more time-efficient and automated way by using sophisticated statistical analysis, multiple data mining techniques and innovative visualizations. The proposed framework comprises three main phases. The first phase involves the transformation of unruly energy data into well-shaped and meaningful data through a process of robust and almost fully automated Data Wrangling steps; the second one identifies the energy consumption patterns in order to reveal energy demand behaviors of the underlying buildings, in conjunction with some exogenous potentially influential factors; while the last one performs an anomaly detection on daily energy profiles. For the detection of abnormal energy consumption, in order to determine whether a daily profile is anomalous or not, we define the degree of deviation from the normal behavior by using robust estimates of the cluster's interquartile range to which the profile under investigation belongs. The proposed framework is studied on a two-year energy consumption data from two building blocks of Hursley site, the industrial site of IBM, while its effectiveness is proved by a two-fold contribution; first, the efficiency on automatically handling highly inconsistent and unruly energy data, without losing important information; and second, the identification of robust and meaningful energy load patterns which optimize the detection of daily abnormal profiles by comparing the potentially anomalous profile with the mean profiles that are characterized by an overall similar behavior. The results indicate that it is possible to reduce the time and effort needed in analyzing and optimizing the building management systems, while concurrently, increasing their effectiveness and interpretability, comparing to manual interferences based on building operators' judgement.

Acknowledgements

I would like to express my deepest gratitude and thanks to my supervisor, Dr. Ryan Grammenos for his continuous support and valuable assistance throughout the completion of my dissertation project, from the early beginning to the very end. His uncommitted dedication and keen interest to help his students, his timely advice and meticulous scrutiny had been mainly responsible for completing my work. Dr. Ryan has been an ideal lecturer, mentor and supervisor, offering advice and encouragement with a perfect blend of insight and intimacy. He was always willing and enthusiastic to assist in any way he could throughout the research project, regardless the circumstances which have been quite challenging due to the COVID-19 situation. The weekly meetings and conversations were vital in inspiring me to think outside the box from multiple perspectives and stimulating questions that I would not have thought by myself, in order to form a comprehensive and scientifically-oriented critique. I would be amiss if I did not mention how extremely grateful, I am for our friendly chats at the end of our meetings and his valuable advice in my academic endeavours.

I would also like to express my sincere gratitude to Mr. Peter Ferguson, the industrial partner of this project, for providing me with valuable domain knowledge and timely suggestions, without which I could not incorporate meaningful explanations to my project.

To conclude, I cannot forget to thank my friends for all the unconditional support in this very intense academic year, who were always there for me, providing pleasant distractions to rest my mind. Of course, I could not skip expressing some more special thanks to my friends and classmates who were always more than willing to brainstorm and share our ideas any time, any day.

Contents

1. Introduction and Problem Statement

1.1 Background and Motivation

Over the past few years, the abrupt development of sensor technology along with communication and cloud computing technologies have aroused the penetration of more and more embedded systems in almost every aspect of our lives, allowing the connection between everything and everyone in an efficient and inconspicuous way [1]. The massive amounts of data that are generated by such systems in urban and industrial environments has set the scene for the development of sophisticated "Big Data Analytics" techniques, giving rise to intelligent and optimized Building Management Systems (BMS) [2]. Energy systems, having moved well beyond their original function to intelligent systems, in conjunction with the increasing energy consumption in buildings being an issue of high concern, has emerged the need of leveraging energy big data to derive meaningful insights and support applications for energy-efficient buildings. Therefore, in this work we focus on analyzing energy building data, provided by an industrial partner, with the view to interpret the performance of the underlying energy management system and identify anomalies in the energy pattern usage.

Since time immemorial, energy has been the most precious asset of human life, while its rapidly growing demand has raised more concerns over supply barriers, energy resources exhaustion and adverse environmental repercussions, being global warming and climate changes. Therefore, the necessity of the energy production infrastructure to adopt more efficient and innovative approaches has been highlighted [3]. According to [4], buildings, both residential and commercial, contribute to more than 40% of global energy consumption, producing more than 30% of carbon dioxide emissions [22], while in the U.K., non-domestic buildings are responsible for more than the half of country's total energy consumption, with Heating, Ventilation and Air Conditioning (HVAC) systems accounting for about 50% of total buildings' energy consumption [5, 29]. As Pérez-Lombard et al. [3] highlight, building energy demands follow a surging uptrend as population is growing and comfort requirements are getting even more demanding, hence dictating for reasonable solutions regarding the efficient use of energy in order to minimize the overall energy consumption along with the operational cost, without however compromising the user-preferred environment. Even though an alternative solution for alleviating the environmental impact of buildings energy would be the production of energy from substitute resources, it is quite expensive and time-consuming, besides questionable for its efficiency [6]. Therefore, a more efficient and sustainable approach for reducing the amount of buildings' energy consumed, which has been of increasing interest for researchers and energy companies, is the development of sophisticated energy management systems [22].

Energy management in buildings has been recognized as a crucial aspect for optimizing their energy consumption, comfort, equipment life and operational cost, with some of the most recent applications being the forecasting of energy consumption demands, the potential energy waste detection, and the supply of more financially and environmentally, economical pricing business models adapted to customer's habits and needs [8, 9]. With the increasing concerns in achieving global sustainability, energy management systems have played a key role in effectively enhancing energy consumption by identifying and restoring in real-time energy anomalies and failures usually related to incorrect energy-wasting occupant behavior, erroneous equipment, dysfunctional and poorly maintained control systems configured by humans, environmental factors, like temperature, besides other malicious activities that cause power outages [8, 9, 17, 25]. According to [8], the use of smart meters in a residential house

could save between $40 and $70 in annually expenses of energy consumption for each customer. It is quite worrisome that buildings can consume around 20% more energy than necessary due to the aforementioned reasons [17,18] and hence their efficient detection and confrontation should be a priority task. The detection of these anomalies can be defined as "the problem of finding patterns in data that do not conform to expected normal behavior" [16]. This procedure can be very overwhelming and hence unreliable by solely depending on human judgement, considering both the extreme mass of data and the assortment of anomalies caused by different malfunctioning, some of them being a high energy usage on bank holidays, spikes in the system during the night, unexpected changes in energy load due to weather, building occupancy, major events and so forth [19]. Therefore, for saving both energy and time on massive energy data analysis, it is of great potential to develop automatic and real-time anomaly detection systems that ensure the optimization of energy consumption, outperforming the manual anomaly detection methods, not only in terms of accuracy but also time.

With the recent surging integration of advanced metering infrastructures, building energy systems have been digitized, incurring the transition from simple one-way communication electrical grids to two-way communication smart grids, between infrastructures and operators [7]. Massive amounts of heterogenous building data are collected and stored in real-time and subject to extensive data analytics procedures in order to extract valuable information regarding the buildings' energy behavior and therefore enhance their operational performance [23]. Hence, it is very promising to develop data-driven approaches to learn a building's load behavior and perform a reliable energy anomaly detection. Particularly, Data Mining (DM) has recently gained increasing attention by scientists in handling massive energy datasets, since it can effectively discover energy consumption patterns, without the need for previously known domain knowledge, which in most cases is not easily provided. On top of that, DM also can efficiently describe the load patterns along with the exogenous conditions, like weather, time, day of the week, season, customer's occupancy and so forth, which have the most influence in building's energy consumption [23,24]. For example, these patterns are profiles that represent electricity load profiles for weekdays or weekends when an office building is mostly not occupied, hence electricity usage levels will be low; profiles based on the time and duration that peak electricity loads occur, like peak loads taking place between 9:00 to 17:00 representing a high-level electricity usage during office hours; profiles revealing small variations of consumption during the day; profiles with patterns presenting peaks at late evening hours, indicating that overtime working is likely to occur and so forth. According to an MIT Review, DM is considered as one of the top 10 emerging technologies that will change the world [29].

Contrary to DM techniques, current traditional energy analysis applications are mostly based on simple statistical analysis, limited to calculating monthly or annual total usage density for providing some overall consumption standards, which cannot provide any insightful feedback for optimizing the BMS. Therefore, to realize the full potential of the aforementioned challenges, in this work we are going to apply data analytics and machine learning techniques, in order to develop robust methods that add granularity to the BMS under consideration, by discovering meaningful information behind the energy usage. Particularly, the identification and interpretation of energy load patterns, as well as the detection of energy usage anomalies will be examined, which in turn will be able to be integrated into data-driven decisions and enable smart buildings to reduce operating cost and become more transparent in terms of operation.

1.2 Problem Statement and Aims

The increasing energy consumption of both residential and commercial buildings, currently reaching the 40% of all carbon emissions in developed countries [4], along with their superfluous power consumption due to malfunctioning control systems, currently accounting for a 20% surplus [18], have highlighted the importance of enhancing BMS for optimizing their energy performance. Most building systems fail to meet the performance expectations due to various faults, often related to incorrect occupant behavior, improper control system functions or poor maintenance [18]. One of the most promising solutions to tackle energy wastes in buildings is anomaly detection through building management techniques. Energy management systems can collect and store massive quantities of energy data from such buildings and through extensive data-driven methods, extract actionable insights that could confront any inconsistency of the energy system. Despite already existing building optimization systems, most of them fail to perform accurately. Current conventional BMSs underperform alert operations when anomalous energy events take place, since operators need to manually set the thresholds for alarms [26]. This cannot be effective when the building has more than one energy consumption profiles depending on exogenous variables, such as weather, day, occupancy and many others and therefore a tight threshold can trigger many false alarms, while a loose threshold can neglect serious system's failures [26]. It has been proven that analyzing building energy behaviors into clusters, each of them representing a specific energy pattern, could be a robust approach for effective management and building operation [28]. Therefore, powerful and efficient tools coming from Data analytics and Machine Learning realm is the key enabling factor to understand the performance of BMS and develop automatic and reliable anomaly detection systems to ensure the optimization of energy consumption. Although some recent intelligent building optimization systems have been developed, most of them still fail to perform accurately. The reason is that they do not consider important factors that lead to unexpected anomaly behaviors [18], and overlook the feature engineering approaches to understand the building's energy consumption dynamics, besides neglecting to handle the corrupted measurements by the sensors affected by external factors [16], focusing solely on model development [12]. If inconsistent data, like missing values and outliers are used in data analysis process, then the results will be hardly reliable. Meanwhile, advanced DM techniques are constantly emerging, overwhelming building professionals to keep up with the most recent technologies. Knowledge discovered by DM can be quite enormous and difficult to interpret without thorough domain knowledge, making difficult to select practically valuable knowledge [30]. Therefore, more focus needs to be placed on developing a reasonable and reliable framework for automatically converting massive poor quality building data into actionable and interpretable knowledge, without requiring the human interference, that will retrofit and accelerate the optimization of energy systems.

More specifically, this work focuses on analysing building data from electricity, heating and cooling meters, provided by the industrial partner, IBM, for its research and development laboratory, located in Hursley Site, UK, consisting of two different building blocks located nearby, the Hursley House (HH) and the D East block (DE). The general aim of this project is to develop a general Data Wrangling-based framework with a two-fold purpose; first, there is a need to understand more about the performance of the underlying BMS, by converting unruly energy building data into actionable granular information and following, model the trends of the energy load profiles for electricity, heating and cooling separately, in order to provide us with interpretable knowledge regarding the patterns and some dynamic influencing factors (weather, day, season); second, to automatically detect unusual energy consumption or else anomalies, within each occurring energy pattern group. For example, a potential abnormal

event would be to cluster a profile captured during unoccupied building night hours, in a group with low electricity consumption during the non-working night hours, but in reality, this profile would follow a high late-night electricity consumption pattern and it should be reported as a potential abnormal event. The outer contribution of this work is to provide Hursley House's building operation staff with valuable knowledge regarding the characteristics of building energy usage patterns and its anomalies and therefore help them to develop actionable measures for energy saving. A potential application of the proposed framework, in conjunction with some highly accurate predictive modelling on a fault-free dataset for the expected energy daily profile, would be to process incoming data on-the-fly and alert the users to potential issues in the system by comparing the actual with the expected energy data using the proposed anomaly detection. Particularly, the underlying objectives that this research is dealing with, are the following:

1. Convert the raw unprocessed BMS sensor dataset into a clean and prepared final dataset, by implementing a sophisticated data wrangling Python algorithm, that outperforms the conventional data mining techniques, and identifies and handles data inconsistencies, like missing values and outliers that correspond to either independent non-plausible datapoints or they are part of hidden anomalous trends, caused by system's glitches and follow specific patterns. An example is the issue of consecutive missing values followed by the accumulated electricity load which should be equally apportioned across the missing data points in order to provide a clean and consistent dataset.
2. To develop an algorithm for imputing the first 5 months of missing DE Electricity data and particularly, build a Machine Learning model for predicting the missing data, using the other energy data and the extracted time-scaled features as candidate predictors and evaluating the performance of the best fitted model by means of Root Mean Square Error (RMSE).
3. To implement some EDA for inspecting the overall yearly, monthly, hourly, weekly trends of each attribute for each building by calculating the mean and the IQR range of the attribute across each of the time-scaled features.
4. To cluster similar energy load profiles and therefore identify energy patterns for each individual building for the Electricity, Heating and Cooling attributes, separately, with the view to provide further knowledge discovery. By using multiple unsupervised learning approaches, we aim to identify the best clustering algorithm that groups daily load profiles into meaningful and compact clusters based on their similarities and afterwards to interpret each cluster's behavior in conjunction with exogenous related factors, like the weather, the day and the season. The metrics used for choosing the best algorithm will be geometrical, as well as statistical (variance), both indicating the compactness of clusters, while the metrics of Silhouette score and Dunn Index will be used for choosing the best hyperparameters of each clustering method.
5. To perform an anomaly detection method, where possible unusual energy consumption profiles within each cluster are identified by determining the amount of variation from normal using robust statistical estimates. For example, a potential abnormal event would deviate from the mean of the cluster it belongs to, more than the Maximum value of the cluster's interquartile (IQR) range.

1.3 Organization of Report

The remainder of the report is organized as follows.

Section 2 presents a thorough and comprehensive review of the previous research relevant to this work along with a theoretical description of the background terms and methods used to allow the reader to follow the flow of the analysis presented throughout the report.

Section 3 describes the research methodology adopted, explains the experiments performed in this research, and finally analyzes the results and compares their performance. Particularly, this section is divided into 4 subsections. The first subsection presents the reader with an overview of the framework adopted in this work, while the following three subsections analyze its underlying steps. The second subsection describes the process of cleaning the data and bringing them into a coherent shape for being used for further Data Mining analysis; the third one covers the process of knowledge discovery, along with the various experiments that were implemented for adopting the best method, including the clustering analysis and the interpretation of results; and the last subsection discusses the anomaly detection process as a post-mining step to the aforementioned processing.

Section 4 summarizes the findings and conclusions of the underlying work, while future work proposals and improvements are drawn in Section 5.

2. Literature Review and Background Theory

In this section we introduce a comprehensive summary of the related works carried out by other researchers, besides some fundamental background theory, aiming to provide a concise understanding of the concepts being studied in that research domain and demonstrate the methods that set the scene and inspired this project's approach.

2.1 Literature Review

Building energy consumption accounts for a considerable portion of the overall energy consumption, following an increasing trend over time [4]. Therefore, building energy analysis has become an emerging area of interest, with energy optimization becoming a milestone. In order to understand more about Building Management Systems (BMS), real-time tremendous amount of building energy data is collected and stored, ready to be utilized within data analysis methods. However, most recent BMS perform baseline data analysis, like historical data tracking, averages and total consumption data and neglect important knowledge extraction [30]. Even thought, over the last decade, more sophisticated data analysis techniques and tools have been leveraged for extracting more information from energy data, still feature engineering methods for converting the raw time-series data into clean and actionable information have remained unreliable. Hence, in order to retrofit the energy systems and extract as much valuable knowledge as possible, BMS industry needs more robust and powerful tools to analyze the massive and heterogenous building energy data [30]. A new process, called Data Wrangling has been introduced in order to tackle the issue of inefficient data cleaning and transformation.

Data Wrangling or Data Munging, comparing to conventional data mining techniques, handles massive amount of data, not simply by finding patterns between them, but by transforming big data through removing information that do not benefit the overall set, structuring and enriching

them in order to ensure meaningful insights with metadata statistics, in less time [11]. With data insights having the most decisive role for every business decision-making, data wrangling has become a ubiquitous process when complex and massive data require interpretation to derive highly accurate results. Indeed, data wrangling has been a recent breakthrough, that has been following a surging uptrend for use in industry during the last few decades, with the most of previous works being limited on data mining techniques for their models' development and more focused on sophisticated Machine Learning models rather than on understanding and extracting meaningful insights from the data [12]. As Zhang et al. [12] highlight, Li et al. [13] showed great results on hourly building electricity load prediction, using an SVM model with fine-tuned hyperparameters, proving that SVM's quick and robust learning performance can outperform the conventional complex back-propagation neural networks. Following the same logic as [13], Rodger [14] neglects any knowledge extraction method from raw data and focuses on predictive model development, proposing a fuzzy K-nearest neighbour statistical neural network model. Each of these research works hinges upon robust black-box prediction models, efficient on extracting patterns from raw data, rather than using transparent data wrangling and machine learning methods to extract meaningful insights. However, as many later works have proved, an additional preprocess of data, comprising data cleaning, preprocessing and feature engineering, before being fed to any Machine Learning algorithm, could yield unprecedented results on any data analytics task, since features with the higher impact on models' efficiency, along with knowledge insights are extracted and used as input to further processes.

Zhang et al. [12], trying to fill the gap between conventional data mining and contemporary data wrangling techniques, introduce three consecutive feature engineering techniques, namely feature visualization, feature selection and feature extraction, to dive further into an efficient energy management system. For feature visualization, Exploratory Data Analysis (EDA) is adopted, a group of various techniques aiming to capture the correlations of data set's features, discover patterns or trends of data, besides spotting abnormal behaviors by applying statistics and visual representations. EDA could be deemed as an initial filtering stage where features, highly related to outputs, are chosen to be fed into further ML models. Following, Zhang et al. [12] applied a feature selection technique, called Random Forest in order to perform dimensionality reduction given that their data used were high dimensional, hence an ensemble of tree predictors was used to discard features with the least influence on outputs, downsizing the dimensions of their data to a desired feature subspace. Last, a feature extraction method, called Principal Component Analysis (PCA) was applied to further reduce the feature space by creating a new one in which the most relevant features were projected. Vidal et al. [1] follow similar approach for preparing the data to be fed into various models for further energy consumption predictive analysis. However, these methods of dimensionality reduction provide promising results only when energy data consist of high dimensional features, by narrowing down the most important ones, in order to implement as more insightful and interpretable energy management systems as possible.

Another work that focuses on the concept of energy data wrangling in order to derive interpretation for the underlying building management system is that of Silipo and Winters [9]. The analysis applied in this project starts from structuring, cleaning and transforming massive energy data into qualitative data. After concatenating 6 CSV file into one, structuring columns by renaming, formatting datetime columns and resampling samples using interpolation techniques, they extract new time scaled features, being hour, month, day of week and year, which are leveraged to describe the energy usage patterns, besides a few average and percentage statistics. An enrichment step of using these extracted measures in order to group data into clusters is following, while at the end, time-series predictive analytics are applied

using an auto-regressive model. An important contribution of this paper, has been the feature engineering technique of clustering energy data based on some behavioral characteristics, which enriched the knowledge information derived from smart meter measurements. Given the huge amount of 6000-meter IDs for this analysis, clustering has been used as a dimensionality reduction technique to downsize the individual meters to only 30 groups, each comprising the meters with similar behaviors based on daily and weekly distributing values, while the model used was the unsupervised k-Means algorithm. Energy-related time series data are characterized by high dimensionality, defined by the time interval of measurements, therefore, this could distort the intrinsic characteristics that are captured by DM techniques, the so-called problem of "Curse of Dimensionality" [31]. Therefore, many other dimensionality reduction strategies have been proposed to overcome this issue, with feature definition being extensively used to describe energy load daily profiles by a limited number of statistical features [24]. Haben et al. [33] divided each date sample into four time periods (non-working hours, breakfast, daytime and evening) and for each one period they extracted the average consumption, hence converting a 24-Dimensioned dataset into 4-Dimensioned dataset, whereas the authors of [32] defined seven statistical features (mean, standard deviation, skewness, kurtosis, chaos, energy and periodicity) to represent the whole 24-hour day. Liu et al. [24], inspired by the previous two works, divided daily profiles into four segmentations according to the working schedule (off time, rise time, daytime, evening) and defined the mean value for each period, besides the peak-to-valley difference rate, proving that clustering DM techniques on featured-based data outperform these on raw time series.

Data mining (DM) techniques have been widely used in the realm of building energy profiling, where sequences of energy data are grouped in order to identify typical energy consumption patterns, while they have been applied in various applications, such as forecasting, energy management, anomaly detection and so forth. Particularly, unsupervised DM learning approaches have been rather powerful at identifying building energy load patterns from sensor data, with clustering techniques being mostly used [24]. An assortment of clustering algorithms has been used for grouping similar building energy load profiles, mostly in daily basis, from partition and hierarchical methods, to density-based and model-based methods; however, classic partitioning methods, with k-means algorithm being the most famous, have been the most used in literature [24, 39]. For example, Carmo et al. [40] by using the k-means algorithm on 139 buildings, identified two main clusters of daily heating electricity load, one for weekdays and one for weekends. However, when dealing with massive large-scale time-series, the classic partioned-based algorithms missperform, both computationally and qualitatively, the so-called problem of "Curse of Dimensionality". Therefore, researchers have worked on various alternatives. Liu et al. [24], as mentioned previously performed some statistical feature engineering in order to reduce daily electricity load's dimension before using the k-means, while k-means with a different distance metric, called dynamic time warping (DTW), which was designed for finding the similarity between time series has, been proved effective for clustering problems without the need of dimensionality reduction, but with a computational cost [41]. Some other papers that had to deal with profiles from several (more than one thousand) different buildings, used fuzzy C-means methods to extract different profiles associated with consumer's categories [25, 42], when others like Yang et al. [28] in order to capture as many as possible invariances, proposed k-shape clustering for identifying similar patterns.

Nevertheless, even though Zhang et al. [12], Vidal et al. [1], Silipo and Winters [9], Liu et al. [24] and many others have discovered meaningful patterns of their energy, most of them have neglected to handle appropriately the issue of inconsistent values. Inconsistent values, caused by various malicious activities, instrumental errors, environmental factors or human errors, can

highly distort the extraction of accurate and useful information; therefore, they should be either handled appropriately, like replaced or removed before any knowledge extraction approach, or be left as it is for further knowledge discovery [19]. By simple replacing them with constant values without elaborating on what events caused them or whether they should be considered as inconsistent or not, we lose important information for the upcoming decision-making. Particularly, inconsistent values can be either missing values when no data is recorded by sensors or outliers, where, in time-series there are two main types: the point anomalies and the contextual anomalies [34]. The point anomalies, or else the statistical outliers, are the individual datapoints that deviate more than a pre-defined threshold from the rest of the data and usually are caused by measurement errors affected by external factors, whereas an anomaly is characterized as contextual when a data instance is anomalous in a specific context but not otherwise, like in specific weather conditions, holidays, and so forth [34]. The understanding of inconsistent values can be quite overwhelming if depending on human judgement, due to the different types of malfunctioning of systems, therefore robust methods based on data analytics should be developed for automatically identifying them [19].

As far as statistical outliers are concerned, most works, like [1, 9, 12], replaced any outlier and missing value as well, with fixed values, either with zero or with the mean of the whole data, without elaborating on what should be considered as an outlier, hence losing important information regarding the distribution and the hourly energy patterns. However, some other researchers like [15] and [19] tried to automate and improve the procedure of outliers' identification and replacement based on statistics. Particularly, researchers in works [15, 19, 20] developed an automated detection of outliers based on the concept of Hampel filter [15], where a sliding window of configurable width goes over the data and any value within a window that deviates more than x standard deviations from the median of that window is considered an outlier, the so-called method cycled-based Z-score. This concept has been proven quite beneficial and informative, since if we would consider the whole data distribution for detecting outliers, then actual outliers that are not much far away from the normal behavior, but they are within their surrounding window, they would have been missed [19]. Given that these are time-series problems, it's not efficient to remove the inconsistent values since periodicity is a key value in their analysis, unless the entire load profile of the day is removed [15], therefore replacement techniques like linear interpolation or filling with the mean value of the surrounding area, that retain a smooth transition between time-steps, have gained a lot of interest [15, 19].

The aforementioned step of identifying and replacing the outliers using sophisticated statistics, gave rise to efficiently distinguishing the measurement outliers that should be filtered out, from the plausible outliers (contextual anomalies). Contrary to point anomalies, where parametric approaches are mostly applied for their detection, contextual anomalies should not be considered as inconsistent values and then replaced, but detected with the help of unsupervised non-parametric or machine learning techniques in order to provide users with further knowledge, since most of the times they are plausible measurements that don't behave as statistical outliers. Compared with supervised analytics, unsupervised anomaly detection is more practical in analyzing real-world building operational data, given that anomaly labels are typically not available [35]. As Hurst et al. [37] demonstrate, clustering methods can seamlessly detect unusual energy consumption within naturally occurring groups with similar characteristics. Khan et al. [18], exploited a non-parametric density-based clustering algorithm, called DBSCAN (density-based spatial clustering of applications with noise) to automatically detect records with abnormal energy consumption, or else dates that do not follow the most common behaviors and group them together.

It is obvious that unsupervised learning has been successfully applied in many building-related management problems [18, 24, 30], however their potential in discovering insightful knowledge from continuously increasing amount of time-based data has not been fully exploited. Particularly, weaknesses like the difficulty to interpret raw forms of time-series data, not considering the seasonality of data and neglecting information derived from the energy consumption profiles, are the most common in literature review [34, 38] and hinder the interpretation of abnormal profiles. Hence, some other researchers, with the view to provide more robustness and accuracy to their anomaly detection methods, combined unsupervised learning algorithms to identify the dominant energy daily profiles, with supervised algorithms to predict the expected daily profile, given some exogenous parameters (weather, day, season) and finally compare it with the actual profile. In fact, clustering analysis does not provide knowledge discovery of energy profiles, therefore the use of supervised techniques on extracted profiles could fill the gap between clustering results and external dynamic factors, such as weather, day, occupancy and many others. Particularly, Liu et al. [24], taking advantage of DBSCAN's effectiveness to remove any abnormal daily profile, grouped similar daily energy profiles by means of k-means algorithm and trained a CART model on fault-free training set with clusters being the response value, in order to compare the actual daily energy sequence with the expected energy profile and deduce if a profile is anomalous or not. Hence, they developed a mining-based framework that extracts insightful knowledge hidden in the energy load patterns. Similarly, Piscitelli et al. [15], developed a clustering analysis for identifying the infrequent load profiles by using a partitive algorithm based on the "Follow the Leader" approach. Afterwards after removing the anomalous dates from their dataset, they used Neural networks along with a Regression Tree to train a fault-free predictive model for predicting the daily hourly energy consumption and by comparing the predicted with the observed energy consumption, they were estimating the presence of anomalous events. This model achieved a 93.7%5 accuracy on detecting anomalous profiles, while only the 5% of fault-free days were wrongly predicted as anomalous. Chou et al. [36], instead of grouping the daily patterns into clusters, developed an ARIMA model to predict the daily consumption and following identified anomalies by differences between real and predicted consumption by applying the two-signa rule; however, with interpretability totally be missing.

Nevertheless, there are some main limitations of existing unsupervised anomaly detection methods [35]. First of all, when statistical methods, that are based on mathematic assumptions, are applied to big data, model's performance and computational efficiency can be highly degraded, while the post-mining can be very overwhelming. Second, the conventional unsupervised methods are applied into features extracted from raw time-series which are constructed based on domain expertise and statistical properties, hence it is impossible to generalize and automate the anomaly detection process, especially when external factors highly affect the presence of abnormal behaviors. Therefore, recently, many scholars have investigated a very promising approach to the problem of anomaly detection which is the use of autoencoder architectures, where by using neural networks (NN), high-level data representations are extracted, instead of using the aforementioned weak statistical feature engineering methods. Fan et al. [35] leveraged various autoencoder architectures with different aspects (whether denoising training and conditional information like weather, day and so forth are used or not), aiming to reconstruct each daily energy consumption sequence and detect anomalous profiles by minimizing the means of root mean squared error (RMSE) between actual and reconstructed signals. Nevertheless, this method works efficiently in a high-dimensional and large-scale data scenario, but not in smaller-scale datasets where the samples are not representative of all the profiles under investigation and therefore the selection of

training data requires considerable effort, being also susceptible to overfitting [34]. On top of that, NN work as black-boxes, hindering the researchers from interpreting their results.

Overall, numerous techniques and methods have been used in order to understand and optimize BMS; however, many different weaknesses have been detected in each work. The most common ones that we identified in the literature review from which we were inspired, are the following:

a. The structured of time-series data was not taken into account.
b. Raw data not properly prepared for Data Mining.
c. Inconsistent data not detected and handled inappropriately.
d. The seasonality of data was neglected.
e. Generalization problem.

2.2 Background Theory

2.2.1 Data Pre-processing

Data pre-processing is a mandatory task for converting raw time-series data into clean and organized dataset before applying any data mining technique. As a first step, the raw data is analyzed through diverse statistical methods in order to identify any inconsistent value (missing or outliers) and either remove them or replace them [30]. Following, when dealing with time-series energy data it is quite common to perform some feature engineering steps for chunking all observations in sub-sequences, usually daily as it is presented in literature review, either in the raw time series format or in dimensionally reduced extracted features.

2.2.1.1 EDA

In statistics, exploratory data analysis (EDA) is an approach of examining data sets in order to extract some key properties and this is usually implemented by observing their behavior through statistical visualization techniques. EDA was initially introduced as a preliminary step before any statistical analysis in order to explore data and formulate hypothesis based on observations and not randomly, while also EDA techniques have been widely used to discover unexpected patterns in data that could determine the subsequent trajectory of the underlying research [60].

2.2.1.2 Data Cleaning

A. Inconsistent data in Energy domain

Inconsistent data in energy domain can be missing data or data corrupted from its original value. Understanding the reason of this values' occurrence in energy time series data is of pivotal importance for identifying the real inconsistent values and not the false positives. Energy data, most of the times, comes from smart meters attached to building's infrastructures. Smart meters are devices that measure, capture and transform data related to the electricity, heating, gas or water usage in real time, consisting a wider smart grid network. The smart grid massive data, after being captured, are sent upwards to the cloud, that the company provider hosts, where they are subject to data analytics techniques, comprising power optimization in real time, forecasting of electricity demand, consumption patterns and anomalies detection, besides dynamic pricing business models [8]. The challenge however that smart grids

encounter, is that each smart meter device can produce a huge amount of data per day ranging from thousands to millions, depending on the defined frequency of measurements, hence resulting in a deluge of high-resolution data that are difficult to manage, much less in real-time [7]. Smart grid data in their raw format can be widely assorted, characterized by high volume (on the scale of TB), ranging velocity or sampling frequency (real-time, minutes/hours resolution) and variability because of different data formats, as well as it can be highly corrupted and hence inconsistent, affected by external factors [16]. Smart meter data inconsistencies, which appear in the form of missing values, negative values or outliers, can be caused by diverse reasons, some of them being instrumental errors, environmental factors, like temperature, malicious activities causing power outages besides human errors [7, 10]. Missing data occur when there are no data values for specific observations and primarily results from errors in data collection or data transmission. Negative energy consumption values are usually caused as a result of a system misconfiguration, since energy cannot be negative. Last, outliers are data points that deviate a lot from the rest of the data or else do not follow the general trend and can be caused either by system's misconfigurations or external factors, depending on each system's nature. In order to implement further decision-making on energy data, it is a prerequisite to acquire a clean and reliable data set by identifying any inconsistency and either discard it or impute it, otherwise the distribution of data can be highly skewed and subsequently distort any other data mining process.

B. Identifying Inconsistent values

While missing values are easily identified as NULL values, outliers need a statistical approach to de detected, considering that they are outlying observations that deviate from the remainder of the other observations [16]. Even though graphical approaches, like Boxplots, Histograms and Scatterplots can highlight outliers, when it comes to large scale datasets, these approaches can be very tedious and overwhelming, considering the human interference is necessary to point out these values. Therefore, more automatic and probabilistic techniques are usually preferred. Probabilistic approaches use probabilistic distribution functions (pdf) to estimate the distribution of the data and, given a chosen threshold, they identify outliers as data points that are less probable than that threshold, or else that deviate considerably from the rest population [44]. There are two types of probabilistic approaches: parametric and non-parametric. Parametric methods assume a predefined distribution function for the data under investigation, whereas nonparametric methods estimate the density function. For the detection of statistical outliers in this work, parametric approaches will be used.

When a variable follows the normal or almost normal distribution, Z-scores can quantify the unusualness of an observation. To calculate the z-score for an observation, we subtract the mean from that observation and divide by the standard deviation, as follows:

$$z = \frac{x - \bar{x}}{s},$$

where $\bar{x}$ is the mean value and S the standard deviation. The further away a z-scaled observation is from zero, usually more than 3 standard deviations away, the more unusual its value is and therefore it could be considered an outlier. However, when considering the normal z-score, data is normalized using the mean and standard deviation of the whole data, hence some outliers that are not far away from the normal behavior may not be identified [19].

In order to confront the problem of Z-score method, where outliers affect the computation of the mean value and therefore the skewing they bring neglects to find all the actual outliers, the modified z-score has been proposed [61]. Modified z-score method does not suffer from that

limitation since it replaces the mean with the median, averting the outliers from affecting the computation of the variables that formulate the overall dataset's distribution. The median is more robust for measuring the tendency of data and hence the median absolute deviation used, instead of the standard deviation, is respectively more robust in measuring the dispersion of data.

Another statistical approach for identifying outliers, is the use of Interquartile Range (IQR). The interquartile range is defined as the difference between the values ranking in 25% and 75% in a data set, denoted as Q1 and Q3, respectively, while the outlier fences are defined as follows:

$$Minimum = Q1 - c * IQR$$

$$Maximum = Q3 + c * IQR$$

$$IQR = Q3 - Q1,$$

where c is decided by the user and usually set to 1.5.

Usually, the box plots or else whisker plots are used to visualize the quartiles of the IQR method, which are also often used in EDA. Box plots visually illustrate the distribution of the data and their skewness is presented through displaying the data quartiles and median. In other words, box plots show the five-number summary of data, including the Minimum score, the first quartile (Q1), the median, the third quartile (Q3) and the Maximum score. Any value that lies outside the Minimum or Maximum value, is considered an outlier. Nevertheless, IQR can also be inefficient when Q3 is located within the outliers [46].

However, the main drawback of the parametric methods is that the distribution of the observations needs to be known in advance and in real data sets, the underlying distribution of the data is not known [45]. On top of that, there is not an optimal rule for choosing the threshold over which values are considered as outliers, therefore manual configuration is required.

C. Imputing Inconsistent values

After identifying the inconsistent data points (both missing values and outliers), they should be imputed. Usually there are three main approaches to handle these values. They can be discarded, replaced by known values or replaced by estimated values either using the variable's univariate distribution or using regression models between one or more dataset's variables, depending on the dataset's characteristics. When dealing with time-series data, it is quite rare to discard the irregular values, especially when dataset is small scaled, because in that case, given that time datapoints have temporal relationships between them, all date's records should be removed and not only the faulty ones, hence losing important information, usually when missing values constitute more than 5% of the whole data [1]. Replacing inconsistent values with fixed values is mostly used in literature review and usually that value is chosen as an estimate of the variable's distribution, like the mean. Mean imputation is a good approach when the inconsistent observations are very few since it preserves that variable's distribution; however, when the problem is large scale, then this method can distort the estimates of variance, besides the temporal trends [47]. Therefore, regression models have been widely used in imputing inconsistent time-series data, where the replacement value is based upon the relationship (usually linear or polynomial) to the other observed values, which are used as the independent variables of the regression model. These regression models can be either univariate or multivariate when the dataset consists of more than one variable, which can capture substantial correlations between the inconsistent data and the other variables.

Another technique for handling multivariate data, proposed by Stekhoven et al. [49], which has been proved to outperform established imputation methods by reducing the imputation error by more than 50%, is the MissForest algorithm. MissForest is a random forest nonparametric imputation algorithm that can be used to impute continuous and/or categorical data, capturing complex and nonlinear relations, being widely used when there is large amount of consecutive anomaly values.

Nevertheless, data imputation creates artificial data that are treated as real data, without providing any degree of precision. Many statistical methods used to study time series require samples with actual and accurate values, otherwise in case of massive inconsistent data they yield biased insights [48].

2.2.1.3 Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of features, or dimensions under consideration, while preserving the integrity of the dataset as much as possible. While higher data dimensionality values generally yield more accurate results, they can also degrade the computational efficiency, along with the clustering quality when the dataset does not consist of an adequate number of samples, while it is getting even more challenging to visualize high-dimensional data simultaneously. Therefore, obtaining a reduced representation of the dataset should be more efficient and produce more robust results [43]. Some dimensionality reduction methods, used in this work are the following:

A. Principal Component Analysis (PCA)

Principal Component Analysis is a type of dimensionality reduction algorithm that reconstructs or else projects the original feature space into a new low-dimensional feature space [12]. This method uses a linear transformation to create the new data projection, whose new features are called “principal components”. These principal components represent different directions of the new feature space and they are derived such that the covariance between those components and outputs are the largest [12]. In this work, PCA will be used in order to facilitate the visualization of clustered samples in a smaller feature space, understandable by human eye.

B. Autoencoders

Autoencoders can be considered as different neural networks designed for unsupervised learning [50]. An autoencoder consists of an encoder and a decoder, while the input and the output are set identical. The encoder transforms the input signal into a high-level representation, while the decoder tries to reconstruct that intermediate representation into the original signal. Therefore, by projecting the initial data into a lower-dimensional representation, we can use this latent space as the new features extracted from the initial ones and perform any following clustering on the extracted low-dimensional data. The autoencoder is trained trying to minimize the loss (usually the mean squared error or the cross-entropy loss) between the initial signal and the reconstructed and hence the lower the total loss is, the more accurate and representative is the latent space. The network architecture can be anything, such as Convolutional Neural Networks (CNN), Long short-term memory (LSTM) and many others. 1D Convolutional architecture has been proved to better preserve the information embedded in temporal building energy data into the latent space [51].

2.2.2 Machine Learning techniques

2.2.2.1 Supervised Learning

A. Classification and Regression Tree (CART)

The CART algorithm is a supervised machine learning algorithm that is based on classification and regression trees and can perform predictive modeling of multiclass classification. Particularly, a CART tree is a binary decision tree, where each node contains the samples that satisfy the condition of its parent node, while the root node contains the whole sample [62]. Therefore, the rules for splitting data at a node are based on the value of one variable selected while the stopping rules decide when a branch is terminal and cannot be split anymore. Several methods can be used for identifying the best split in a CART tree; however, the most used is the *Gini Impurity* which is defines as:

$$Gini\ Impurity = 1 - \sum_{i=1}^{n} p_i^2,$$

where p_i is the fraction of items belonging to class i and n is the total number of classes or else target variables.

B. MissForest algorithm

MissForest is a random forest imputation algorithm for missing data introduced by Stekhoven and Buhlmann, who proved that it outperforms all other imputing algorithms, with missing values ranging from 10% to 30% of the total sample [63]. MissForest initially uses the mean of data to impute all missing data, while afterwards it fits a random forest on the clean data and predicts the missing values as being the unseen set. This process of training and predicting, is repeated in an iterative way until a stopping criterion is satisfied, while the reason for the iterative process is that the training is conducted on better and better quality data each time than the data itself has predicted on the previous iteration.

C. Model Evaluation Metrics

In order to evaluate a machine learning supervised model's performance, we need to define some universal metrics that both accurately quantifying the performance of a model and also can be used as a comparing measure between different models.

When dealing with Classification problems, when the response value is a discrete value to be predicted, a simple and the most common used evaluation metric is the Accuracy. Accuracy is the ratio of the predictions classified as Positive to all the predictions made, and it is defined as:

$$\text{Accuracy} = \frac{TP+TN}{Total},$$

where TP = True Positives, TN = True Negatives and Total is the total number of predicted samples. Accuracy always ranges from 0 to 1 and the higher it is, the more accurate is the model. However, Accuracy is not always the most appropriate metric to evaluate the performance of a model, especially when the classes of our samples are imbalanced.

Regarding the Regression problems, when continuous values, rather than discrete, are predicted, the most popular metric used is the Root Mean Squared Error (RMSE). It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

and intuitively it represents the Euclidean distance between the vector of the actual values and the predicted ones by the underlying model, averaged by N which is the number of the predicted values. The lowest the RMSE, the most accurate the model, since it means that the predicted values do not deviate a lot from the actual ones.

2.2.2.2 Unsupervised Learning

Unsupervised learning is the process where machine learning techniques are used to analyze and cluster unlabeled datasets by discovering hidden patterns. Clustering can be considered as the most important unsupervised learning problem, where unlabeled data is partioned into meaningful groups based on their similarities and dissimilarities. The selected algorithms used in this work can be classified into two categories: (a) Partitioning methods and (b) Density-based methods. Partitioning methods subdivide the data sets into a set of k groups, where k is the number of groups pre-defined by the analyst and can identify only linear-separable clusters, with k-means being the most popular [52]. Density-based methods identify separate clusters in the data, based on the idea that similar data in a data space is characterized by a high density, while dissimilar data is separated by contiguous regions of low density, with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) being a famous algorithm on this category [53]. Density-based methods, contrary to the partitioning methods can deal with non-spherical shaped clusters and identify outliers with a high accuracy given that outliers mostly lie in low-density regions.

A. Partitioning Model (k-means)

In K-means method, data is divided into k partitions, where k is set as an input parameter by the user. Each cluster is represented by the centre or means of the data points belonging to that cluster, called centroid and it is defined through an iterative process, where k samples of the dataset are randomly chosen as the initial centroids [52]. During every iteration step, each sample is assigned to the closest centroid, by means of a predefined distance metric, while afterwards centroids are relocated by computing the new mean of the samples assigned to each cluster. The main idea of k-means is to partition a dataset into k clusters with the view to minimize the within-cluster similarity and increase the inter-cluster dissimilarity. K-means performs well when clusters are intuitively well separated. However, k-means algorithm requires a good initial estimate of the number of clusters to prevent converging to local minima, while it is quite sensitive to outliers since every observation becomes part of a cluster even if it is scattered far away in the vector space, affecting the compactness of clusters as well [52].

B. Density Based (DBSCAN)

DBSCAN is a density-based clustering algorithm that is effective with non-linearly separable distributions, when clusters follow arbitrary shapes and it can identify outliers as a low-density area. It does not require to pre-define the number of clusters, but only two input parameters in order to define the density threshold in the data space. These parameters are the *minPts* which defines the minimum number of points that can be grouped together in order to form a cluster and the *eps(ε)* which is a distance measure between data points in order to form a neighbourhood [53]. If there are more than *minPts* objects with a distance less than *eps* from the considered object, the object and its surroundings form a new cluster. However, if the dataset is too sparse, the method may fail to detect clusters, while the setting of its parameters can highly affect the effectiveness of that method.

C. Similarity measures

All clustering methods require the definition of a metric in order to compute distances between data points in a dataset. The choice of the metric is pivotal for the efficiency of the clustering algorithm, since it defines the similarity between two objects [54]. Two of the most often used distance metrics for time-series clustering are the Euclidean and the Dynamic Time Warping (DTW). Euclidean distance calculates the point-to-point distance between two time-series, which means that the i-th point on one time series is aligned with the i-th point on the other, therefore it cannot capture temporal trends [54]. This produces a poor similarity score, hence another metric that preserves the temporal relationships should be used in time-series data. DTW, by minimizing the distance between two sequences which may vary in time or speed, measures the similarity between them and allows a non-linear mapping of them. In that way, it produces a more intuitive similarity measure that allows shapes with similar trends but out of phase to match, since DTW looks dynamically for the best alignment between the two-time series [54]. The metric's major flaw is the required effort to calculate the path of least cost.

Ultimately, Euclidean metric is preferably used in feature-based models, while DTW in raw-time-series data.

D. Cluster validation

Many clustering validation indexes (CVIs) have been published in the literature for finding the optimal number of clusters in a dataset, before k-means is implemented. Every CVI has their limitations and none can always perform robustly. However, in this study, the Dunn index was selected to evaluate the clustering results, based on a comparison of different experiments implemented by Liu et al. [24] who proved that the Dunn index had the highest performance on choosing the optimal number of clusters. The Dunn index defines the ratio of the shortest inter-cluster distance to the longest intra-cluster distance and the largest the Dunn index the more compact and well separated are the clusters [24]. Another CVI used in this work in order to provide a more accurate and generalized evaluation of clusters, is the Silhouette score which calculates the average distance between clusters and assesses how effectively an observation is grouped [55]. Similarly with the Dunn Index, the higher the Silhouette score, the better the clustering estimation.

E. Clustering for Pattern Discovery in Time Series

The goal of clustering time series is to discover patterns in order to find a building's profiles that represent its operations over the time. The highlight, though, of time-series data, compared to non-time dependant data, is that each time sequence entails a temporal dependence between its dimensions that provides important information and should be taken into account in any subsequent data mining process. Therefore, time-series clustering can be approached in two ways: (a) developing a *feature-based* model where raw time series data is transformed into statistical features using some feature extraction methods or parametric models, like autoencoders, or (b) *raw-data-based* where clustering is applied directly to the raw time-series data [54]. Both approaches have their positives and negatives, while their performance can deviate from problem to problem. Within a general context, *feature-based* techniques can neglect important temporal information, while the *raw-data-based* ones are susceptible to the problem of the curse of dimensionality [56] given that time-series are high dimensional data, besides that they add important computational effort. Nevertheless, *raw-data-based* techniques have been proved to be easily generalized in larger-scale problems, while they can easily

capture temporal correlations between time sequences when the appropriate distance metric is selected.

Nomenclature

Abbreviations

ML	Machine Learning
NN	Neural Networks
EDA	Exploratory Data Analysis
DM	Data mining
BMS	Building Management System
HH	Hursley House
DE	D East
CART	Classification and regression tree
PCA	Principal Component Analysis
kNN	K-nearest neighbour
CVI	Clustering validation indexes
IQR	Interquartile

Symbols

Eps	Radius
MinPts	Minimum number of observations in the *eps* region

3. Methodology, Analysis and Results

3.1 Description of the framework

Although DM techniques can extract valuable information from the data they are applied to, when it comes to massive building energy data, great challenges arise, as described in literature review. Advanced DM techniques are continuously emerging, while the knowledge extracted can be quite overwhelming and onerous to interpret into valuable information for further decision making, especially when domain knowledge is missing. Therefore, in order to provide a pipeline of suitable DM techniques and automate the process of knowledge discovery from massive building energy data, we introduce the following framework, illustrated in Figure 1.

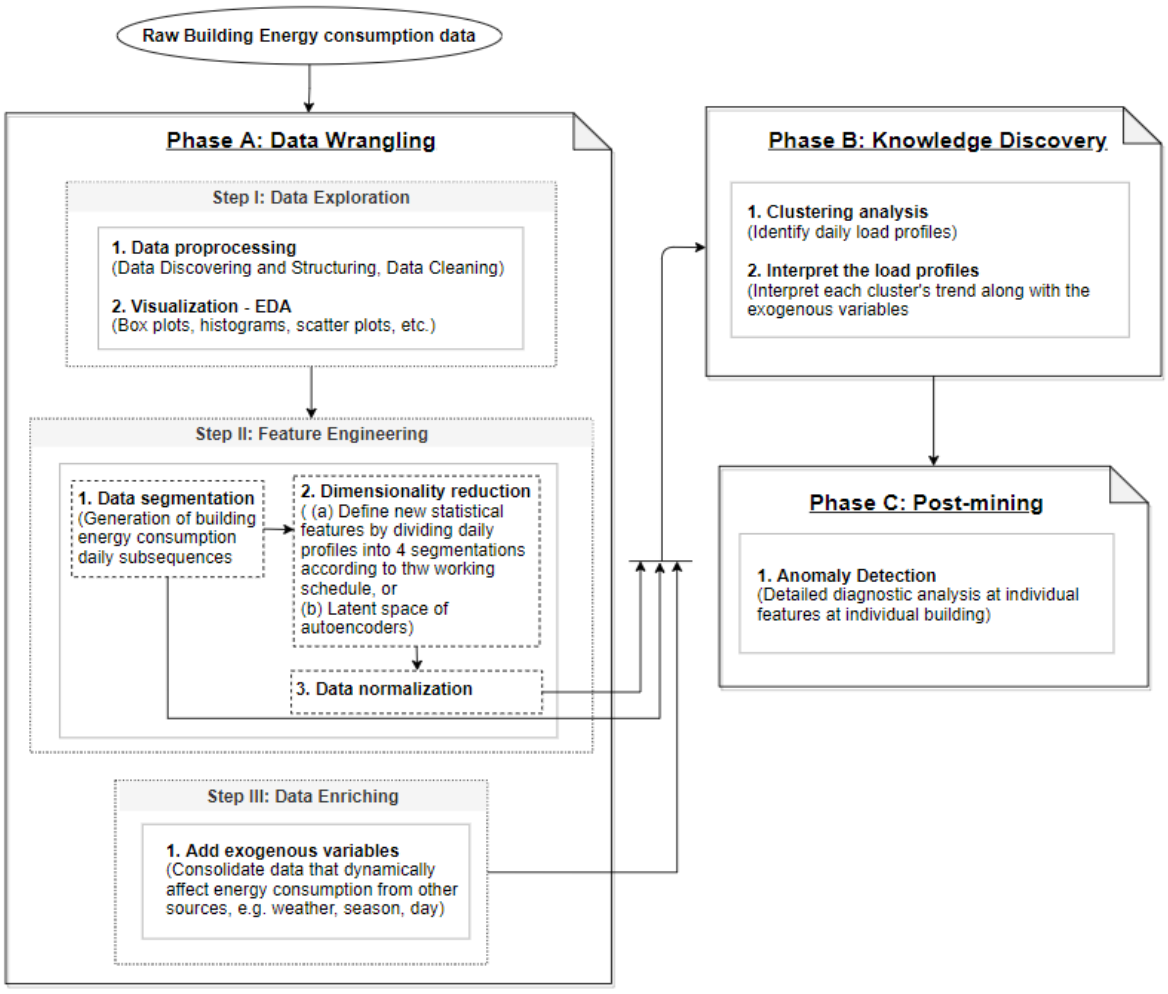


Fig. 1. The proposed framework outline

This framework, designed to address the aforementioned challenges, consists of 3 main phases, the Data Wrangling, the Knowledge Discovery and the Post-mining, which have been all evaluated in the case study of the Hursley site's BMS. Initially, the phase of Data Wrangling aims to transform raw data into a suitable format before applying DM techniques, taking into consideration the poor quality of raw building energy data. Particularly, during that phase, we first develop a data exploration process for organizing and cleaning the unruly data, before transforming it into appropriate dimensions while afterwards, we gather data from multiple sources in order to reveal a deeper intelligence. Subsequently, during the second stage, we

develop a number of DM techniques, mostly focused on clustering analysis in order to discover the hidden knowledge, whereas the third phase of post-mining deals with the issue of anomaly detection, by making use of the knowledge discovered in the previous steps. Each phase is thoroughly explained in the following subsections. All code files related to the proposed framework and any visualization result in this work, are provided in the GitHub project link.

3.2 Data Wrangling

Understanding the underlying dynamics of our data is the very first step towards achieving intelligent decision-making for the Building Management System under consideration. Previous studies have neglected uncovering valuable information from raw data and focused on robust predictive models [14]. Others have been following manual and tedious brute-force processing steps for cleaning and preparing raw data by either removing inconsistent values, or by imputing fixed values to them, without taking into account the attribute's distribution nor the reason that caused this inconsistency, which could be pivotal for the overall model [1, 9]. Nevertheless, as Bhattarai et al. [10] discuss, suitable data analytics combined with visualizations, before any predictive modelling, can ensure meaningful results, that can lead to better awareness and efficient, besides interpretable decision-making. Therefore, in this section, we elaborate on the whole pipeline of preparing raw massive data into actionable and organized data, in order to discover deeper insights about the performance of the underlying BMS and prepare them to be usable for further analytics.

The first phase of our proposed framework is the data wrangling, a fundamental step to prepare time series data for any subsequent DM analysis, comprising three main tasks: the data exploration, the feature engineering and the data enriching. As Zhang et al. [57] highlight, data exploration or else data pre-processing is a crucial step in knowledge discovery, accounting for up to 80% of the overall effort. Data exploration involves data structuring, data cleaning and data transformation, where the collected raw data are analyzed through statistical analysis and EDA in order to identify potential inconsistent values, including missing values and outliers that must be replaced or removed, before being transformed into a different type or scale so that all variables are equal regarding the quantity. Feature engineering, which is the next step after data cleaning, aims to reshape time series data either by chunking original time series in fixed length windows representing constant temporal subsequences, with the daily scale being the most popular, or by defining new features that represent these subsequences in order to yield a dimensionality reduction that will improve the computational efficiency. In case of defining new features, it is important to perform a data scaling afterwards, in order to normalize data, with max-min and z-score normalization being the most-used methods. The last important step in data wrangling, is the data enriching, where influential exogenous variables to building energy consumption are added to the already existing data. These exogenous factors usually are weather variables, like temperature and humidity and time variables like month, time, day type, season, which all have great influence on buildings' operations [23]. Given that an observation might be classified as an anomaly in one context but not in another, incorporating information from relevant external factors into the anomaly identification process or just the interpretation process, might be beneficial.

3.2.1 Data Discovering and Structuring

The data considered in this work was provided by the industrial partner, IBM, regarding the Building Management System of its research and development laboratory in Hursley Site. Particularly, the data comprises tens of thousands of observations from individual electricity,

heating, cooling and water meters deployed in two different building blocks, located nearby, the Hursley House and the D East block. The data collection had taken place over 2 years between January 1, 2018 and December 31, 2019, while data was captured at 30-minutes intervals. Each building block consists of its own smart meters, while a separate excel file is provided for each meter, for each building, for each year. Along with the meter values, the timestamp during which each value was captured is provided in timestamp format of Pandas library as "Month/Day/Year Hour:Minute:Seconds". Following, we describe the sensor network across the Hursley Site, while Table 1 describes the attributes, 7 in total, that were provided in the given dataset.

Sensor Network

Both blocks consist of four floors each with different operations and hence different demandings, regarding the energy consumption.

- Hursley House consists of some housing support areas including kitchen, LAN room, IBM Museum, storage and plant in the Basement floor, an auditorium in the ground floor and some office rooms in first and second floors, while generally it is considered as "low tech" in its operation, given that is a more than 200 years old building. It is provided with a CHW cooling system for cooling and dehumidifying building's air, along with a LTHW heating system, whose energy consumption is measured by a CHW meter and a LTHW meter, respectively, for the whole building. The chilled water from CHW system is only used for the ground floor auditorium Air Handling Unit and its use is periodic rather than constant. Ventilation is predominantly natural other than for the ground floor auditorium and a couple of technology rooms. An electrical meter and a CWS water meter are provided for measuring electricity consumption and chilled water capacity.
- D East block consists of Plant areas and IT labs on Ground floor, office open space areas on first and second floor and Plant rooms equipped with electrical equipment (Chillers and AHU's) besides an office space on the third floor. Electricity, heating LTHW and chilled water CWS meters are provided. A CHW cooling system is not provided since HVAC plant is installed instead. Contrary to Hursley House, D East block's ventilation is mechanical with HVAC plant, implying much higher power consumption measurements of the Electrical meter than for the Hursley house.

It is also important to highlight that for both buildings, there are days, even periods of days when little or no recorded data is available due to communication line failures or some other unknown nature faults causing missing gaps in the data.

Building	Feature	Description	No of Instances
Hursley House	Electricity	Electricity consumption meter (kWh)	34891
	Cooling (CHW)	Cooling meter (kWh)	35035
	Heating (LTHW)	Heating meter (MWh)	35036
	Water (CWS)	Chilled Water capacity meter (m^3)	34892
D East block	Electricity	Electricity consumption meter (kWh)	28711
	Heating (LTHW)	Heating meter (kWh)	35040
	Water (CWS)	Chilled Water capacity meter (m^3)	34934

***Table 1**: Meter Attributes Description of given Dataset*

As mentioned above, the raw data are in the form of separate spreadsheets for each individual attribute, for each building, for each year, hence in total 14 spreadsheets. Therefore, the first step towards the preparation of data is to combine these individual data sheets into a single

structured table. Figure 2 displays the data structuring process for generating the final dataset before being subject to the cleaning pipeline.

First, we filter out any NaN column of each individual datasheet that issued when importing the datasheets using the Pandas library. Following, we observe that some sensor data are accompanied by an extra column, for which we perform an "eye-balling" given that they consist only from few values and assuming that data on these columns was added later manually, we decide whether to keep them or not if they appear to be within the distribution of the 1st column or not. If they do and the corresponding value of the first column is missing, then we replace these missing values with the 2nd column's values. Another challenge we face, is the checking of time mismatches, which in our case don't exist. Due to sensor failures, there are many data losses, hence when merging horizontally the individual datasets, the timestamps with no values for specific attributes are filled with NaN values in order to have a final 2D table of sensor data, with a total of 35040 observations, which is the size of the biggest individual dataset, as Table 1 reveals. Ultimately this provides the evidence of no mismatches, since, considering that the index is timestamped, any timestamp that is not involved in a column, it is filled with NaN value for that column and added to the whole final dataset. Hence, if final dataset's observations were more than the longest individual dataset, then we would have had to deal with time mismatches.

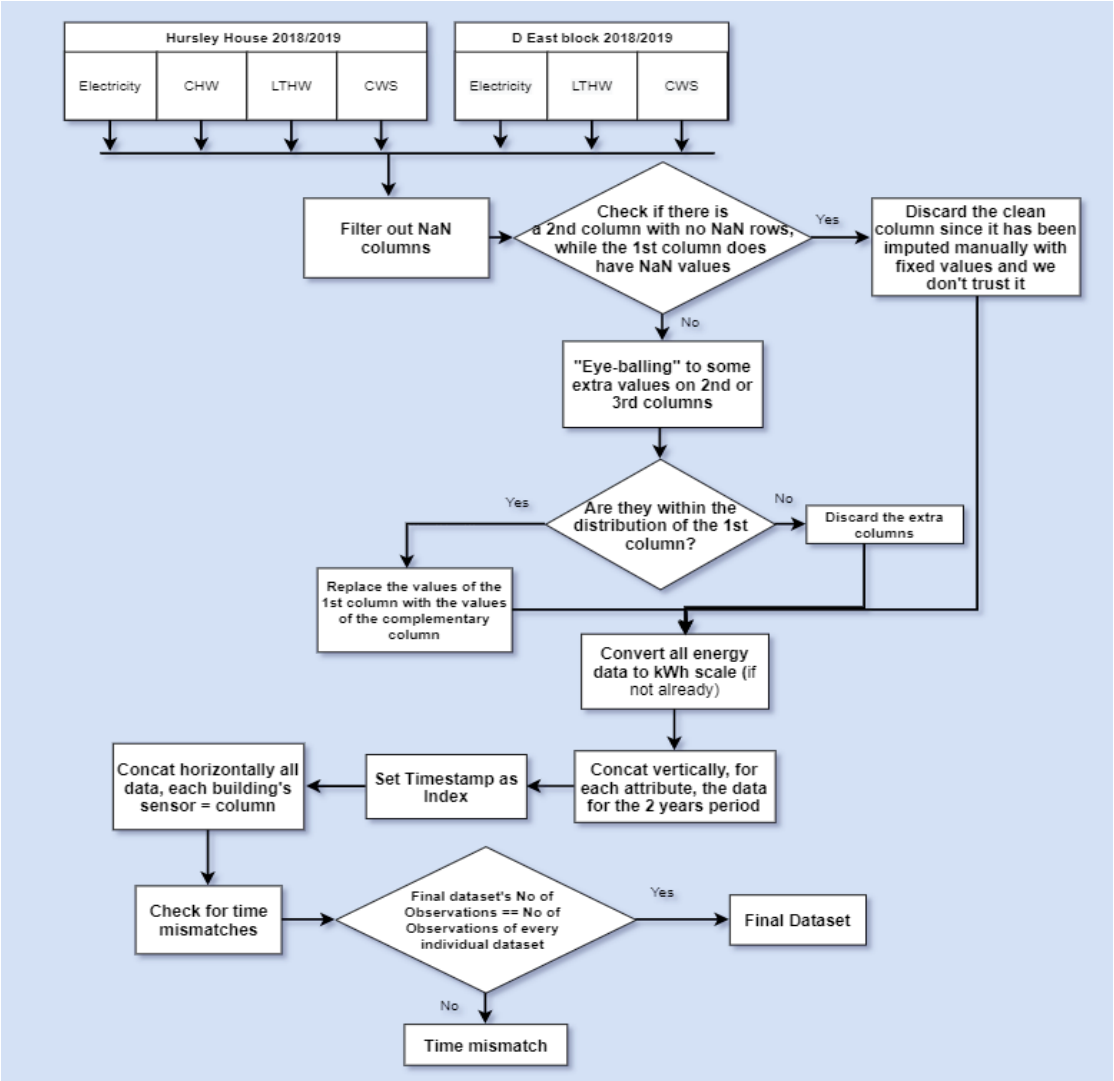


Fig. 2: Data Structuring process for generating the merged dataset

3.2.2 Data Cleaning and Preparation

After structuring the final merged dataset, we need to ensure resilience of the data by cleaning it from any noise, including missing data and outliers, otherwise data is skewed and any further decision making is distorted and less interpretable. Hence, a preprocessing task is to be preliminarily accomplished for detecting and replacing punctual inconsistencies. The first step performed, is to identify any outlier and missing value.

3.2.2.1 Identify missing values

It is clear from Table 1, that each attribute has a different number of missing values. By using the function *isnull()* of Pandas library, and filtering the final dataframe in order to print all the timestamps of each missing value for every feature, we observe some patterns of consecutive missing values over specific periods across some attributes, implying that during these periods some power outages or special events took place and affected many smart meters' measurements, while for the Electricity data of D East block the first 5 months are totally missing. An array with the missing values is included in Appendix A.

3.2.2.2 Identify outliers

3.2.2.2.1 Approach without domain knowledge

Outlier values are irregular data points that do not follow the expected normal behavior, as Chandola et al [16] state. However, outliers can be distinguished into two main categories, the point anomalies and the contextual anomalies. The contextual anomaly datapoints make sense under specific circumstances, such as high energy usage on bank holidays, spikes in the system during the night, and generally any unexpected change in energy load due to weather, building occupancy or other major event, whereas the point anomalies are unrealistic measurements caused by instrumental or human errors that do not offer any knowledge for further decision making [19]. Therefore, before elaborating on the abnormal energy usage patterns detection, in this step, we need to validate our data by identifying and cleaning any non-plausible point measurement error [16].

Boxplot is a common statistical technique to identify outlier data values according to how much they are spread out around the median value, through displaying the data quartiles (or percentiles) and averages. Boxplots show the five-number summary of a set of data, including the minimum score, first quartile, median, third quartile, and maximum score and are based on the approach of IQR. However, this approach is not ideal for our problem since the only outliers we wish to identify in this step are the non-plausible ones and box-plots capture all the outliers, even those that are part of realistic anomalous load profiles and should not be replaced. This assumption is supported by plotting the scatterplots of each attribute (Appendix C), where we observe that values that lie close to the overall pattern of observations, that could be part of a plausible profile, appear as outliers, as they do in the boxplots (Appendix C). Nevertheless, boxplots can be useful as they show the dispersion of a dataset, from which we can visually deduce whether data follow a normal distribution or it is skewed (negatively or positively).

In that case, we examine the efficiency of z-score technique for outliers' detection. As a first step we plot each variable's histogram in order to infer about their distribution and have an "eye-balling" on the values that fall outside the distribution. However, none of the variables follow a normal distribution, but either a highly skewed or multimodal distribution, as we can see in Appendix D. Contrary to boxplots, z-score approach is more susceptible to extreme outliers since the amount of variation from normal is determined using robust estimates of the

mean and the standard deviation, where mean itself is sensitive to extreme values. For that reason, z-score approach identifies as outliers only the extreme values, while boxplot that identifies the outliers based on a specific distance from the median, which is independent to the extreme outliers, considers as outliers, mild high measurements that correspond to consistent high consumption values, given that data does not follow a normal distribution. Therefore, even though z-score cannot stand alone as a method for outliers' detection, we leverage its property to detecting the most salient statistical outliers, which can be also identified by the human eye on the scatterplots in Appendix C, even when the distribution is not normal. Particularly, after normalizing data for each attribute separately, we set the threshold = 3 and hence any value surpassing the threshold is considered an outlier. Appendix B depicts an array of these identified values. However, by solely applying the z-score method for identifying outliers, two important issues arise that undermine the efficiency of this method:

I. Some identified outliers are not actual outliers, but actual values that represent high consumption.

Given that our data is not promptly normally distributed, there was an imply that the z-scaling along with the threshold would identify as outliers, datapoints that might not be, but instead belonging to some high energy consumption profile. Therefore, in order to prove this assumption, we developed a function for visualizing the dates that appeared to have outlier measurements based on the z-score. Particularly, after identifying all outliers for each meter for each building, we defined the dates that appeared to have at least one outlier value based on the z-score method, as the outlier dates and then we plotted the distribution of each date separately based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). In Figure 3 we present an example of some of the outlier dates' box plots for Electricity meter's measurements of HH building, only to see that for some dates, like 31/01/18,30/1/19, 23/05/19 and many others, the values, previously identified as outliers, given the date's distribution, are not actually outliers. Hence, we assume that for the dates where boxplot does not appear to have outlier values, there were probably no outliers, but an increased electricity consumption that falsely was handled as an outlier.

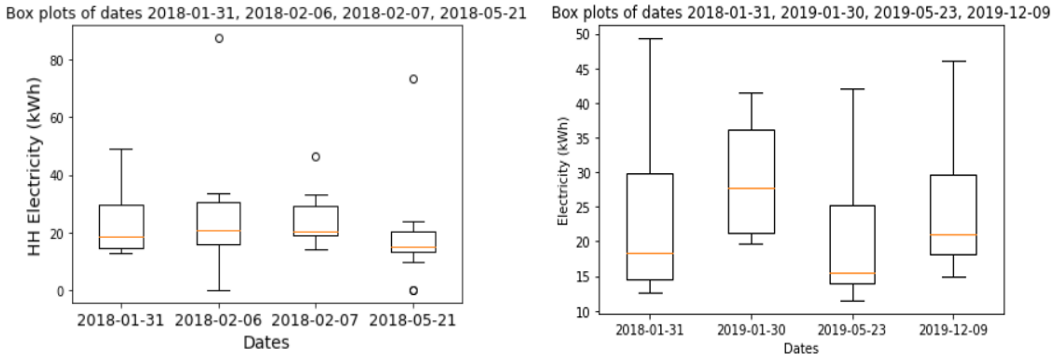


Fig. 3: Box plots of specific dates with identified outliers on HH's electricity data

II. Topical non-plausible outlier values are not detected.

The second problem coming out of the z-score method is that it considers the mean and the standard deviation of the whole data for normalization, which might impede the detection of outliers that are not far away from the normal behavior. For example, given the measurements from the heating meter, a value might be considered as an outlier if

it highly deviates from its surroundings during the summer where the energy consumption is supposed to be low, but not in the context of the whole distribution, which takes into account the winter season, as well, when energy consumption is high [15]. Therefore, a more sophisticated approach should be developed that addresses small neighbors separately, like the one that Habib et al. [19] applied, who performed a cycle-based z-score between On and Off periods of the chillers. Particularly, this algorithm computes the median of each window that includes the k surrounding data points and if a data point inside that window differs from the median by more than a standard deviation, then it is considered as a statistical outlier and should be replaced through a linear interpolation. This method ensures that all outliers are detected, even if they would not be considered as outliers regarding the distribution of all data, but for the specified window they appear to be abnormal.

Overall, using the z-score for automated outliers' detection is not an efficient approach for energy data, since there is a great chance that we neglect or consider the correct data points as erroneous data points. A more sophisticated and robust method should be developed, that takes into account some general domain knowledge about energy data.

3.2.2.2.2 Approach with domain knowledge

Before applying any imputation method, we need to have a proper understanding of what is considered as a measurement outlier, because if an actual outlier is not detected, this could distort the data distribution by underestimating or overestimating the inconsistent data point, while if a plausible data point is considered as outlier and get replaced then valuable information gets lost. In order to face the two aforementioned problems, it is required to integrate some domain knowledge in our approach, regarding the building energy data. According to the HH industrial site's BMS maintainer's comments, a lot of meter reading glitches have been caused due to a number of factors, with the most common being connectivity failures, BMS issue picking up data and software errors. These issues have been 'common' with this type of meter installation, however the reasons that caused them are mainly theories rather than proven issues given that most of the meters are 'black boxes' that cannot be easily interrogated and they were initially installed for indication of load rather than recording absolute values. After identifying the missing values along with some extreme outliers, which are displayed in the arrays of Appendix A and B, we observed that outliers and missing values appear to be present one after the other, implying that certain events took place that period of time for which we should acquire a better understanding. Two main patterns of inconsistencies have been dominantly observed in these smart meter infrastructure; first, a peak and a sag value are often identified, a system's common issue, that if these two values are added and then equivalently apportioned across the intermediate time slots, then an estimation of the load during that time is derived; and second, the pattern of missing values followed by a high value representing the accumulated load during the missing time slots period is commonly detected and should be apportioned across the missing time periods. Particularly, each meter experiences some specific patterns, displayed in Table 2, which were identified by the algorithm illustrated in Figure 4. In Appendix E, we present an example of these patterns identified in our dataset.

The procedure of identifying these inconsistent patterns using that domain knowledge can be very overwhelming and hence unreliable if depending on human judgement, considering both the extreme mass of data and the assortment of outliers caused by different malfunctioning. Therefore, by integrating that domain information, we are going to develop an algorithm which automatically detects the aforementioned patterns of inconsistencies in Hursley site's energy

data by using statistics, without any manual interference. A flowchart outlining the process of identifying outliers, performed by the data cleaning and preparation algorithm is shown in Figure 4.

In order to identify the patterns presented in Table 2, we need to define some statistical thresholds that automate the process of patterns recognition. Particularly, for each feature for each building, we create a dataframe of differences between consecutive values and by dividing each dataframe into three quartiles, we measure the data dispersion through the IQR range. Before doing so, we set temporarily all missing values to 0 in order to avoid having NaN values in the new dataframes. Therefore, for each feature we use the Maximum value of their distribution as the threshold for capturing the differences that deviate from the normal pattern. The Maximum value equals Q3+1.5*IQR, where Q1 and Q3 are the first and third quartiles, respectively, and IQR is the interquartile range between Q3 and Q1. Based on this definition and the outliers that were identified using the z-score approach we set the following values:

- *"**Low value**"*: a value whose difference with the previous observation is more than the Maximum threshold of the differences dataframe and is less than the previous value.
- *"**High value**"*: a value whose difference with the next observation is more than the Maximum threshold of the differences dataframe and is higher than the next value.
- *"**Peak**", "**Sag**"*: high positive outlier and low negative outlier respectively.
- *"**Outlier**"*: outlier value based on the z-score approach.

Normally, the values defined as "High values" should have been identified as outliers through the z-score approach we applied for identifying the outliers without having any domain knowledge. However, there are some extreme outliers due to system's glitches that highly skewed the distribution and therefore the so called "High values" were considered as normal values within a normal range, rather than local outliers. Hence, the approach of defining the distribution of consecutive values' differences can effectively detect local outliers since the difference between them and the previous or following observations will deviate from the mean difference.

Features	Outlier Patterns
Electricity	• "Low value-Missing value(s)-Outlier" • "Missing value(s)-Outlier" • "Low value-Outlier" • "Low value-High value"
Heating	• "Peak-Missing value(s)-Sag" • "Missing value(s)-High value"
Cooling	• "Peak-Sag"
Water	• "Missing values-Outlier".

Table 2: Smart meter's patterns of inconsistent values due to system's glitches

3.2.2.3 Handle inconsistent data

While sophisticated techniques exist for filling in missing values and replacing outliers in buildings energy data, there is a challenge involved when outliers and missing values appear to follow specific outlier patterns. Particularly, we observed that for each feature, missing values and outliers follow similar patterns. Based on that, we automate the procedure of identifying and cleaning them by apportioning the sum of the outlier values around the outliers' period, across the whole inconsistent period, with equivalent values. However, some missing

gaps or outliers have been irrelevant to these outlier patterns, hence, inspired by [19, 21], we filled these inconsistent values using linear interpolation. Linear interpolation produces data between two consistent time stamps in order to provide a smooth transition from the start point of the missing values period to the end point. Even though linear interpolation provides a smooth linear trend between two data points, this imputation method would not be appropriate to be used for replacing the outlier patterns, since linear interpolation fills the gaps with constructed and not real values, while the method of apportioning that we applied, distributes the real accumulated load during that period. Besides that, linear interpolation is suitable for short periods since it cannot capture other dependencies regarding the time or other variables, but only follows a specific trend between the two edge values, assuming a linear relationship between the data points.

Regarding the water consumption data, according to the industrial partner's comments, it is quite inconsistent and unreliable, while their contribution to the overall aim of this project is almost negligible; therefore, we are going to ignore this feature for this work.

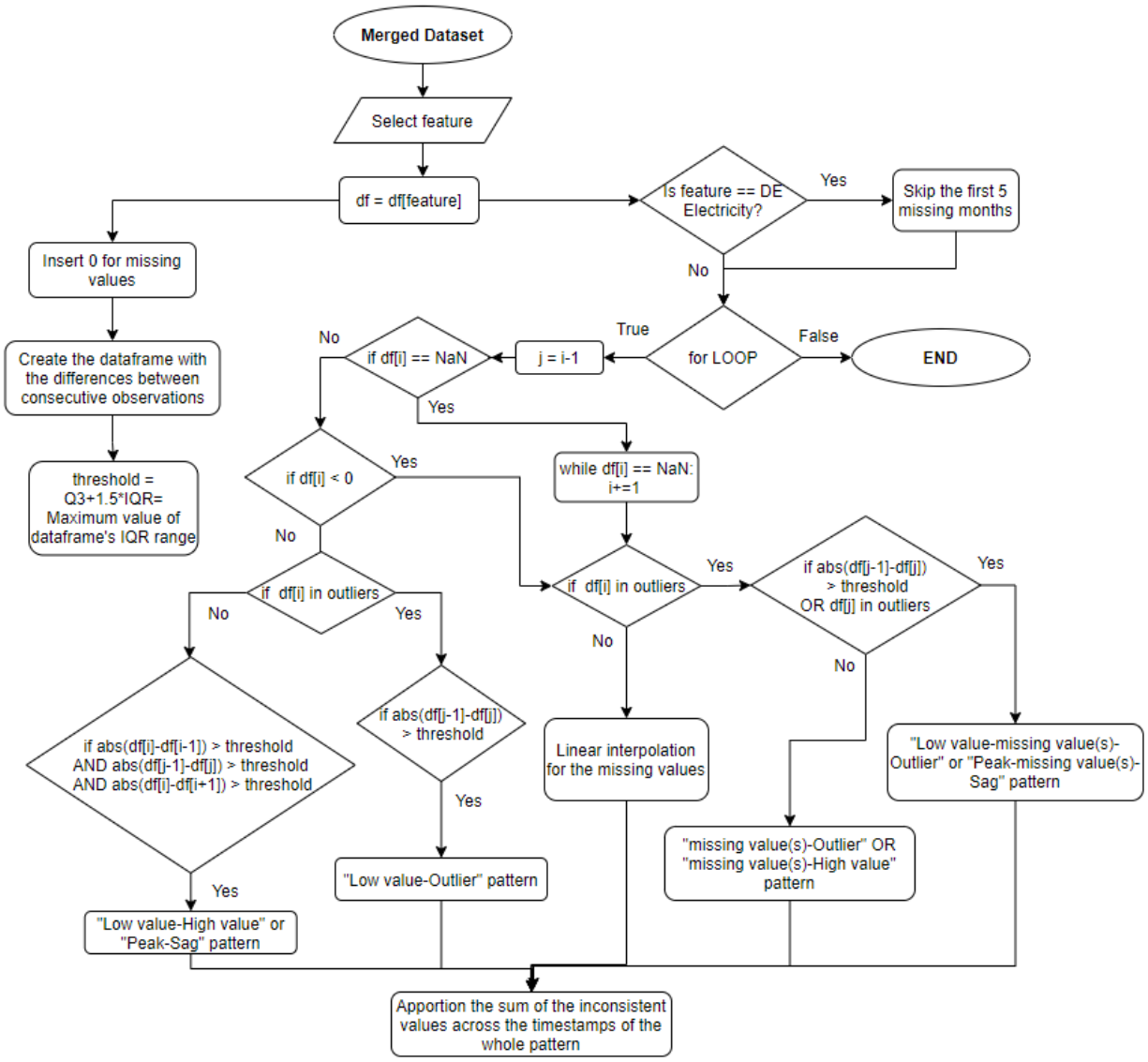


Fig. 4: Flow chart of outlier patterns cleaning function

Ultimately, by developing this cleaning algorithm, using statistical analysis, displayed in Figure 4, we managed to automate a very tedious procedure for dealing with some common

outlier patterns that appear in building energy data. As Table 3 presents, all missing values were replaced, except for the 5 months of missing information for the DE Electricity feature, which will be handled through a different sophisticated ML approach, elaborated in the following section. Besides the missing values, outliers were also resolved, which can be validated by comparing the electricity consumption plots across time in Figure 5, before and after the cleaning process, where we can see that the abrupt peaks have been smoothed over. In Appendix F, the comparing plots of the other two features for each building are include as well.

Feature	*Before cleaning*	*After cleaning*
HH Electricity	149	0
HH Cooling	5	0
HH Heating	4	0
DE Electricity	6329	6211
DE Heating	40	0

Table 3: Missing values before and after the automate cleaning

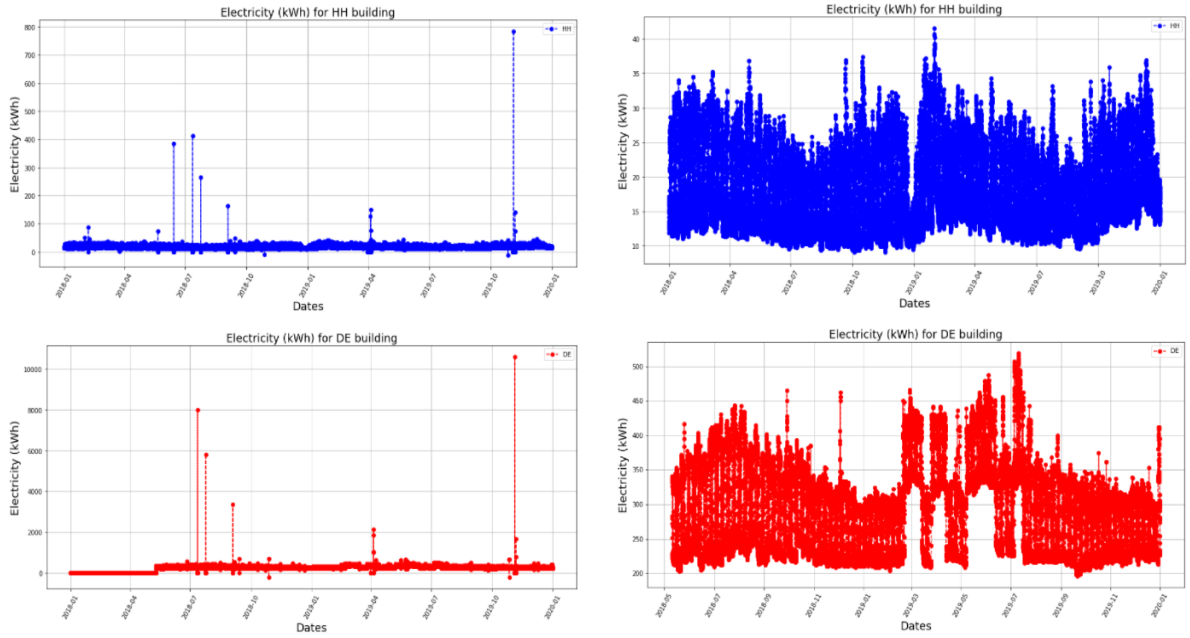


Fig. 5: Electricity consumption plots before and after the automate cleaning process

3.2.2.4 Imputation method

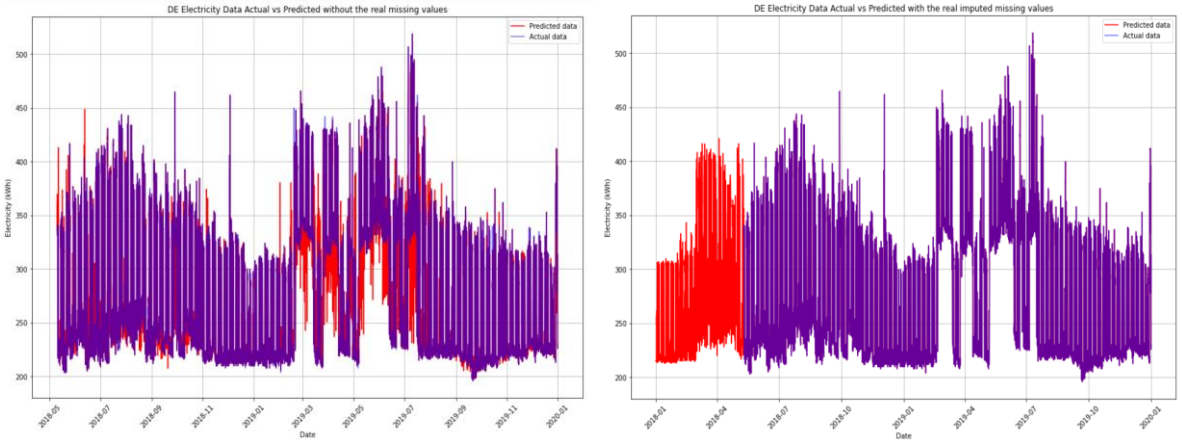
As mentioned, the DE electricity feature is missing values for the first 5 months of 2018; therefore, in order to align all the energy features given to us in time-domain, we tried to impute that period of missing values. In total 18% of the total sample, this missing period cannot be imputed by classic methods, since we desire to be as more realistic as possible. Hence, in order to do so, we need to capture the time dependencies, as well, since it is obvious from the Fig. 5 that they affect the DE building's electricity consumption. After extracting the variables of *hour*, *month* and *day of the week*, and proving that there are no linear relations between these time variables and the energy features, as illustrated in Appendix I, inspired by Stekhoven and Buhlmann [49], we are going to employ the MissForest imputation algorithm.

First of all, in order to evaluate MissForest's performance, we experiment through three different models with different predictors, as displayed in Table 4. Particularly, for the

evaluation process, we split our dataset into two subsets, the train and the test set, where the test set consists of all the samples with a missing value on DE feature. Following, in train set, we produce randomly artificial missing values, so that they account for the 18% of the whole train set, before fitting our train data on the MissForest model, in order to fit a model that satisfies the same conditions as the test set. The new artificial missing values account for the evaluation, or else dev set, on which the performance of the MissForest model will be evaluated, in terms of RMSE. As illustrated in Table 4, the model trained on all features, both building data and time-scaled data, yields the best performance, since it gives the lowest RMSE. Taking into account the overall data's mean and standard deviation which account for 277,931 kWh and 63.903 kWh respectively, the RMSE value is relatively small and within a single standard deviation from the original data's mean. Fig. 6 depicts the best MissForest model's performance on training (left) and test data (red color in right figure), proving that the model captured the time dependencies from the train set and hence, it follows the overall trend of DE electricity consumption on test set, as well.

Features - Predictors	***RMSE***
All building data	38.269
All building data + month	26.800
All building data + All time-scaled data	23.183

Table 4: MissForest evaluation on artificial missing data of training set (dev set)



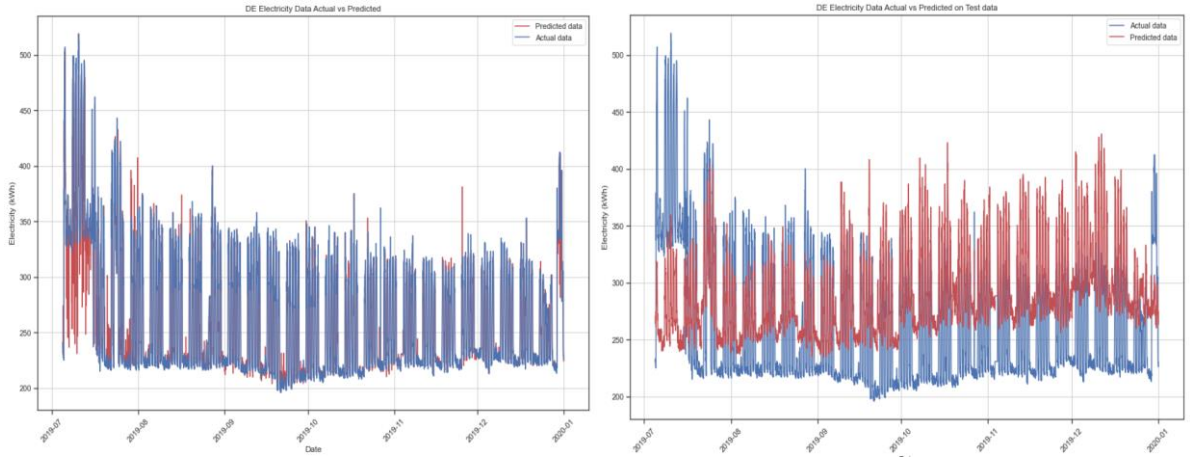
(a) (b)

Fig. 6: Best MissForest model's prediction on a) train set and b) test set

In Fig. 7, we compare the performance of the best MissForest model and a baseline Linear Regression model for the period starting at the beginning of July 2019 until the end of the year, only to show the difference of the performance between the proposed model and a simple Linear Regression. It is obvious, that even though the Linear Regression captures the magnitude difference between working days and weekends, cannot correlate the other time-scaled variables with the DE electricity feature. Besides that, the Linear Regression models gives an RMSE of 54.027 on the dev set, which is worse than the RMSE of any MissForest employed, as displayed in Table 4.

MissForest is excellent, since, first of all, it does not require extensive data preparation, given that Random Forest algorithm can determine which features are important, and therefore we do not need to predefine which features to make use of as predictors, and second it does not

require any hyper-parameter tuning as most ML models do. However, the imputed DE electricity data won't be used within the scope of this work, since they are artificial and not actual measurements; hence they cannot contribute to the following DM analysis we are to perform, considering that the aim is to discover hidden knowledge and not to perform predictive analytics.



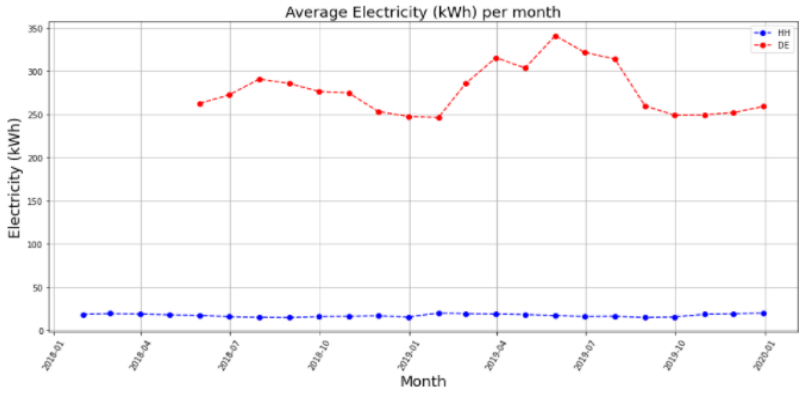
(a) MissForest (b) Linear Regression

Fig. 7: (a) MissForest vs (b) Linear Regression performance on a specific period measurements of DE electricity consumption

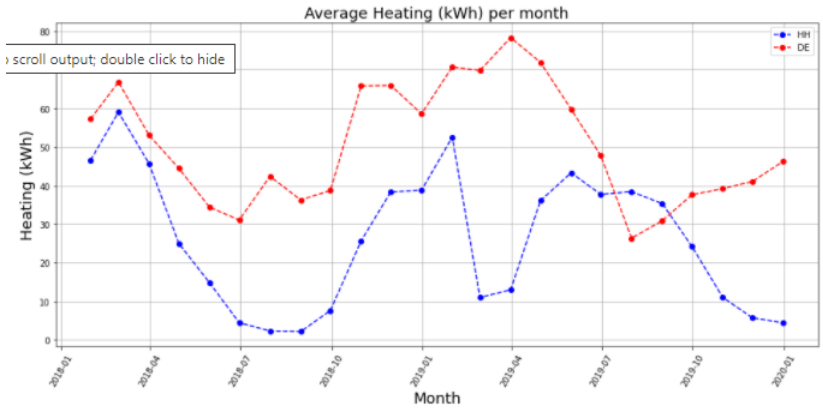
3.2.3 EDA

Following the data preparation and cleaning process, it is of paramount importance to have a good understanding of the building energy data patterns before applying DM techniques to such massive data. Therefore, it would be wise to apply some Exploratory Data Analysis (EDA) for getting some further usage insights that will inform our next choices. As a preliminary step, given that time is a decisive factor for building occupancy and therefore energy consumption, we extract some new time-scaled features, being the day of the week, the hour, the month and the year in order to acquire extra knowledge regarding the energy patterns over different time features.

First of all, by plotting the average electricity per month, per year, illustrated in Figure 8a, we observe that the DE block has an overall higher electricity consumption comparing to the HH block during the whole year, which is supported by the fact that DE's ventilation is mechanical with HVAC plant, while HH's ventilation is predominantly natural, besides the fact that the HH block, given that is a 200-year-old building, is 'low tech' in its operation. Heating consumption seems to follow similar patterns for both blocks during the first year, with DE consuming slightly more energy, whereas during the year 2019, we observe inconsistent patterns that do not follow the general expected energy behavior, which is low heating consumption during summer months and higher during the cold months. On top of that, it is worth observing that during the period March-July 2019, DE building's average electricity consumption has been higher than the other months, implying the presence of a specific event. Particularly, according to the industrial partner's comments, during that period, a replacement work took place where the chillers of Hursley Site's Data Centers were temporarily replaced by DE's chillers, causing an increased electricity consumption.



(a)



(b)

Fig. 8: Average electricity (a) and heating (b) consumption over the examined period

After inspecting the overall energy consumption of both buildings, we present some seasonal trends, by plotting the median and the interquartile range (range between 25% and 75% quartiles) of all data across the two years that we examine, averaged across all days for each datapoint of x-axis. From Fig. 9, it can be inferred that the two buildings follow different monthly trends regarding the electricity consumption, which could be corroborated by the fact that HH's ventilation is predominantly natural, while DE's ventilation is mechanical with HVAC plant. Therefore, HH appears to have the least electricity consumption during summer season, where ventilation is natural, besides the fact that during the summer months more working people go on holidays rather than during the winter.

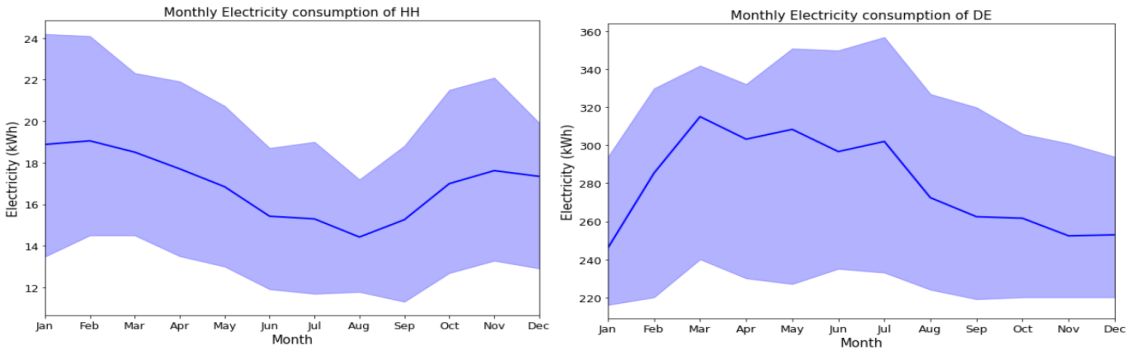


Fig. 9: Monthly trends of electricity consumption

Contrary to HH, DE presents higher electricity consumption during the spring and early summer season, which could be explained by the fact that ventilation needs are higher, and given that ventilation is mechanical, higher electricity loads are expected. The same observations can be supported by the heatmaps of Fig. G.1 and G.2 in Appendix G.

Regarding the heating consumption, the monthly trends for both buildings reveal an overall downtrend during the summer months; however, we observe some inconsistencies that seems to have been caused by other external parameters, rather than weather factors, given that they do not follow the expected pattern that the first year's consumption does. Particularly, for the HH block, during the period of March-April 2019, heating consumption is almost negligible, while it was expected to be high, whereas during the summer months of 2019, it is so high as during the cold months, implying that perhaps some pumping testing was implemented. This behavior corroborates also the trends of Figure 10b, where the mean consumption surpasses the IQR range during the summer months, implying that during that period an unusual extremely high consumption effect of boilers took place. Additionally, during the period of November 2019-January 2020, there is an indication of low HH's heating consumption equal to that during the summer months, implying that probably some replacing operations of radiators pumps took place and the building's heating was provided by temporal electrical heaters.

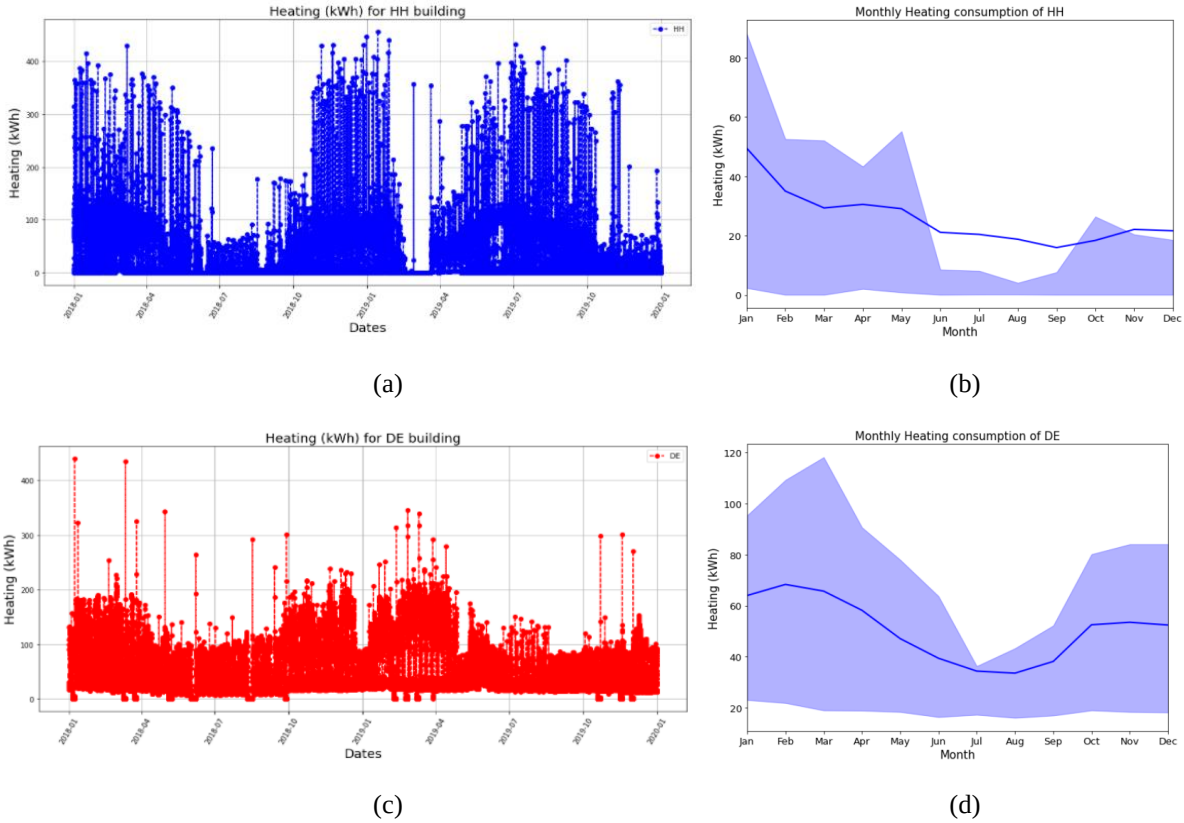
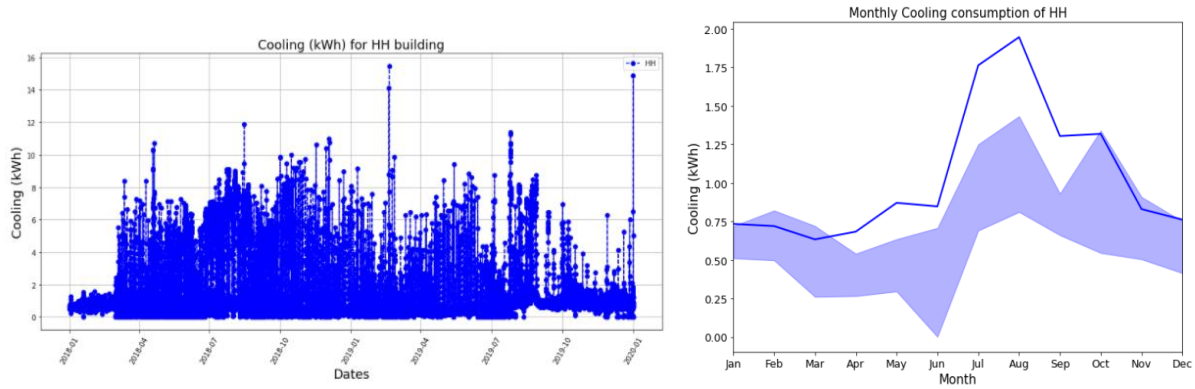


Fig. 10: Monthly trends of heating consumption

As far as the HH's cooling consumption is concerned, Figure 11b reveals that there is a higher cooling consumption during the summer months, however, we can see at Figure 11a that there are many fluctuations, that are due to the fact that cooling is mostly dependent on the human occupancy inside the Auditorium room, and therefore every time someone enters the room, it bursts until sensors detect no occupancy.

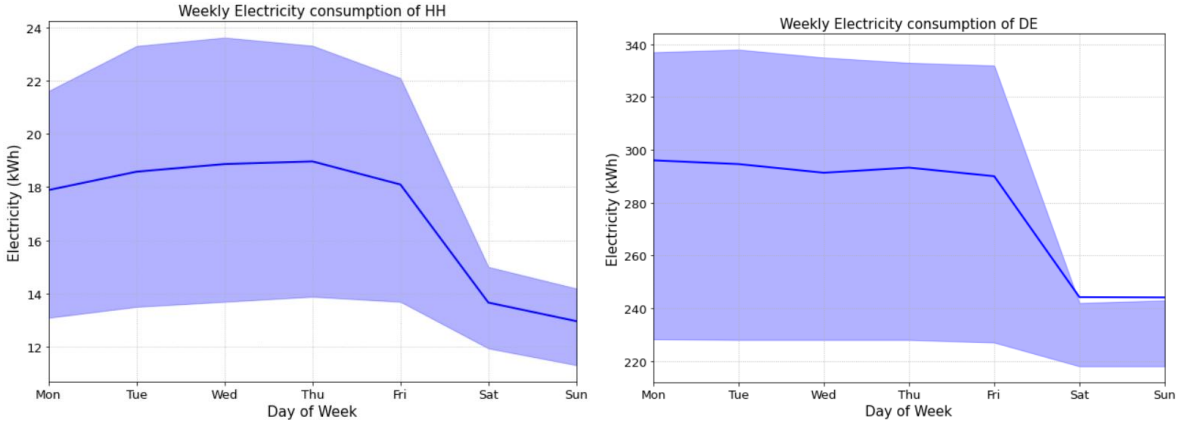


(a) (b)

Fig. 11: Monthly trends of HH's cooling consumption

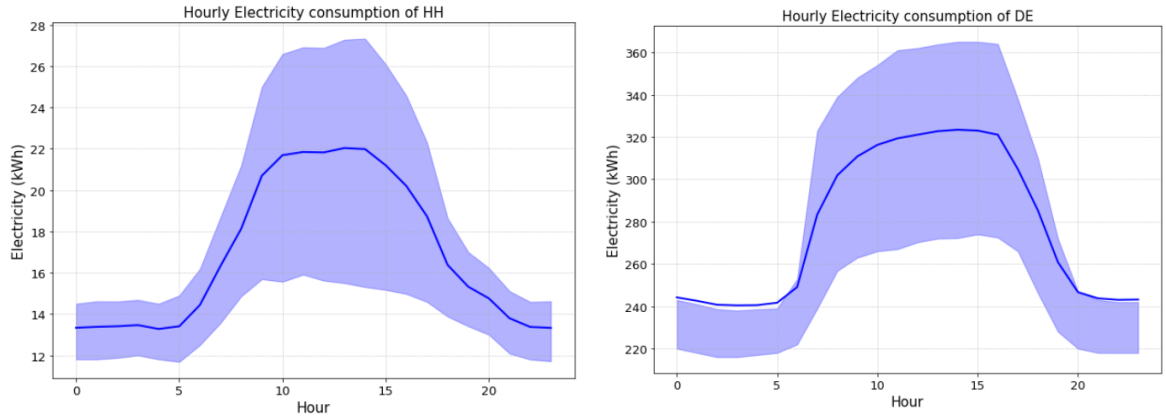
Following, Figure 12 demonstrates the weekly trend of the electrical load, one for each building, while Figure 13 depicts their hourly trends. As expected, the averaged consumption and interquartile ranges referred to unoccupied periods, particularly the night hours, the weekends and holidays are smaller than the values of working hours and days. The quite wide range of the IQR suggests the presence of different building systems and patterns during the year. Particularly, as Figure 13 reveals, the working schedule of both buildings consists of 4 main time periods; 22:00-05:00, 05:00-10:00, 10:00-15:00 and 15:00-22:00, representing the off time, rise time, daytime and evening, respectively.

Last, Figure 14 show the hourly consumption of heating and cooling consumptions during the period between January 1, 2018 and December 31, 2019 where it is obvious that although the overall trends approximately follow the aforementioned working schedule, there are many fluctuations, implying that other factors besides the working schedule, affect the heating/cooling consumption. It is also important to notice that heating consumption trends present morning peaks, which account for the initialization of the boilers every morning.



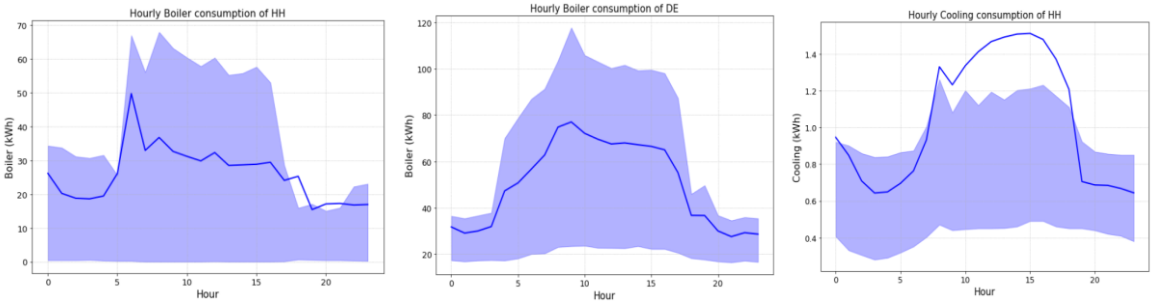
(a) (b)

Fig. 12: Weekly trends of electricity meter



(a) (b)

Fig. 13: Hourly trends of electricity consumption



(a) (b) (c)

Fig. 14: Hourly trends of heating and cooling consumption

3.2.4 Feature Engineering

Feature engineering, which is the next step after data cleaning and EDA, aims to reshape time series data, either by chunking original time series in fixed length windows representing constant temporal subsequences, with the daily scale being the most popular, or by defining new features that represent these subsequences in order to yield a dimensionality reduction that will improve the computational efficiency. In case of defining new features, it is important to perform a data scaling afterwards, in order to normalize data, with max-min and z-score normalization being the most used methods.

3.2.4.1 Time series data reshaping

Time-series data is characterized by many individual timestamped data values, recorded between a specific time interval, which in this case is a 30-minutes interval. However, in order to extract knowledge for our buildings' behavior, we need to define particular profiles that preserve the temporal information. The daily scale is the most common time scale adopted in this field of investigation; therefore, we need to transform the massive time-series into daily energy profiles. Even though this method could have potential drawbacks, since patterns in some cases could be better characterized by different length of windows in the segmentation procedure, in this work, the framework developed aims to extract knowledge by defining atypical daily subsequences. The subsequences, representing the daily load profiles, are organized into a MxN matrix where M is the number of days under consideration while N represents the total number of daily measurements and depends on the granularity of the

collected measurements. Therefore, we convert the cleaned dataset of 35040 rows x5 columns, where 35040 is the total timestamped data points across the two-year period and 5 the features under consideration for both buildings (HH electricity, HH heating, HH cooling, DE electricity, DE heating) into two separate datasets, one for each building, as illustrated in Figure 13. The first dataset concerning the HH building consists of 730 rows, which is the number of days during the two-year period and 144 columns, comprising 48 times the three energy features, while the second dataset of DE building has a shape of 730x96, where 96 equals 48 times the two energy features for which we have data for the DE building. It is worth mentioning that for the DE electricity feature, only 601 days are ultimately used, since the rest 129 have been imputed as mentioning in section 3.2.2.4 and do not represent actual measurements.

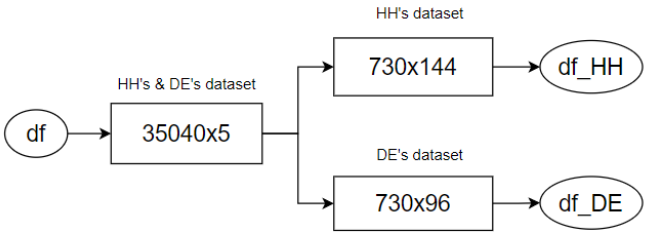


Fig. 13: Dataset's transformation

3.2.4.2 Feature definition of daily electricity load profiles

In order to perform a dimensionality reduction to our datasets, but still leverage as much as possible information that the timestamped features can provide, it would be wise to perform a feature definition analysis based on the energy usage insights we extracted from the EDA. Particularly, we aim to identify some typical daily load profiles that represent the energy behavior by defining statistical daily features. However, as Figures 13 and 14 reveal, while hourly electricity consumption patterns follow well defined trends, heating and cooling hourly consumptions do not follow compact trends, hence the segmentation based on the working schedule would be noisy and unreliable, rather than representative for these two features. Therefore, daily load profiles will be created only for the electricity consumption.

We first divide each date into four segmentations according to the working schedule that we identified from Figure 13, defined from the following four time periods: 22:00-05:00, 05:00-10:00, 10:00-15:00 and 15:00-22:00, representing the off time, rise time, daytime and evening, respectively. Then for each period, we calculate some statistical features, which will be used as the new extracted daily features for electricity consumption. Particularly, the mean value of each segmentation was calculated and used for the first four features, along with the min, max and peak-to-valley values of the whole daily profile, inspired by [24], therefore, accounting in total for 7 new statistical features. By defining these 7 features, we aim to better capture the shape characteristics of each daily electricity consumption profile, while the peak-to-valley feature, representing the ratio of the difference between the daily maximum and minimum load to the daily maximum consumption, was introduced as an additional statistical feature to capture the relation between the peak and the minimum value. Therefore, the original 48-dimansional dataset was reduced to a 7-dimansional dataset, as depicted in Figure 16. It is worth mentioning that for the DE electricity feature, only the 601 actual daily profiles have been used, rather than the first 129 days that were replaced with imputing and hence they are artificial. As a final step, we apply a Standardization technique using scikit-learn's *StandardScaler* function as it is usually performed prior to machine learning model fitting. *StandardScaler* function removes the mean and scales all variables to unit variance. The reason

we need to normalize our data as a final step is because usually different variables are measures using different units and scales. For instance, a variable ranging from 0 to 100 will outweigh the effect of a variable that has values between 0 and 1, which may result in a bias. Therefore, it is of high importance to transform data, so that they range between the same scale.

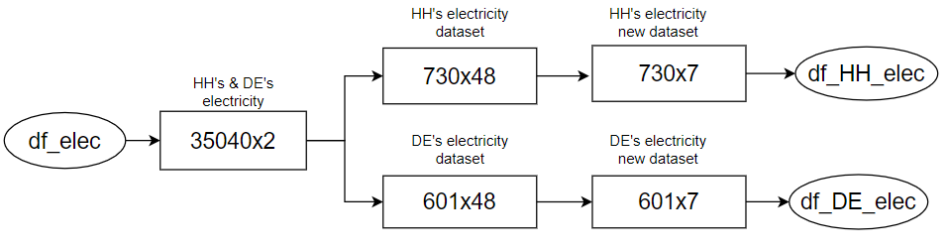


Fig. 16: Electricity Dataset's segmentation

3.2.5 Data Enriching

Another important subsequent step is to identify the most influential exogenous variables to building energy consumption. Consolidating data from other exogenous resources, could benefit any decision making, by integrating extra knowledge that is tightly correlated with energy consumption. Such variables could be time variables (Month, Day Type, Hour), that may greatly change the behavior of building operations or weather-related variables that highly affect cooling and heating systems behavior. Similar works that deal with energy building data have reported improved results when weather data were added to their models [7,12,15]. In this work, given that the industrial site under consideration is located in the English county of Hampshire, we acquire weather-related data, particularly the temperature and humidity, from the closest weather station which is located in Southampton, through the rp5.ru online resource. Additionally, another parameter that could help discover deeper insights regarding the energy usage would be the bank holidays. Given that we are provided a time-series dataset, using the Pandas calendar we can extract a new feature that defines the bank holidays. Therefore, a new dataset of shape 730x6 is introduced, where 730 stands for the dataset's samples and 6 for the following exogenous features:

1. Weekday
2. Month
3. Season
4. Holiday
5. Mean Temperature (°C)
6. Mean Humidity (%)

Such type of information could be used in a decision tree approach, due to its ease of model visualization and interpretability, in order to find the most influential exogenous factors that affect building's energy consumption. Within the context of this work, considering that an observation may be classified as an anomaly in one context but not in another, it is useful to incorporate the information of influential exogenous factors into the proposed framework for knowledge discovery and anomaly detection.

3.3 Knowledge Discovery

With the view to reveal the most representative profiles of energy behavior, the extraction of daily energy patterns according to profiles' similarity generated under the same load

conditions, is the next step towards knowledge discovery. Although load profiles can be analyzed at different scales and different building levels, in this work, we consider daily profiles at individual building level. The process of daily load profiling is mainly implemented by grouping similar load profiles using domain expert-based procedures, statistical methods and data mining algorithms. However, given that domain knowledge is usually missing, like in this case study, unsupervised DM approaches and statistical methods are going to be used. The shape of similar load profiles is usually representative of a specific building's operational pattern depending on the weather, the day of the week, the season or specific features of the building or its occupants. Clustering algorithms have been proved to be particularly effective in discovering robust energy patterns from massive time series data, according to literature review, since they can discover similar shapes based on similar trends or magnitude. Therefore, as a first step, a detailed diagnostic clustering analysis of energy time series is performed to discover energy usage patterns for every feature for each building, while secondly, an interpretation of the discovered patterns, along with the exogenous factors, is implemented.

In the first phase of clustering analysis, two different clustering techniques will be examined; first a clustering on the original clean time-scale based subsequences; and second on artificial features derived from feature engineering techniques and particularly on the extracted statistical features based on a segmentation phase, as explained in 3.2.4.2. By comparing the robustness and compactness of the patterns extracted from both approaches, we will be able to assess the efficiency of raw time series data versus the reduced statistical features.

3.3.1 Clustering Analysis

3.3.1.1 Baseline clustering

In section 3.2.4.2, based on the overall daily trend of electricity consumption, depicted in Fig. 13a, b, it is obvious that electricity is totally dependent on the hourly working schedule. Therefore, we segmented each day into 4 distinct parts, for which we calculated some statistics, converting the raw time-series electricity dataset from 48 dimensions into 7 dimensions. However, it is not wise to apply the same logic on the other two features, cooling and heating, since their hourly trends seem to be quite inconsistent and hence the segmentation into the same 4 parts would not be representative of the overall time series date sample.

Leveraging the definition of the new statistical, dimensionally reduced features and inspired by [24], we are going to apply a 2-stage clustering in order to identify the electricity load profiles in HH and DE buildings, separately. The first stage of clustering aims at identifying any abnormal profiles, by using the DBSCAN clustering technique and filtering them out in order to obtain a clean dataset of normal profiles before applying the k-means, which is the 2nd stage with the view to group similar daily electricity load profiles. DBSCAN aims to remove any outliers in order to prevent distorting the accuracy of k-means clustering, while given that the outlier profiles are mostly the irregular electricity load profiles that lie in the low-density areas, it could automatically implement the anomaly detection phase.

A. HH's Electricity consumption

i. 1st stage – DBSCAN

DBSCAN algorithm, based on the observations' density in the underlying space, clusters the observations that are closely packed together into similar groups, while its efficiency lies on the fact that it can create clusters with arbitrarily shapes without needing to predefine the

number of groups. The concept of DBSCAN is that each neighbourhood of a given radius has to consist of at least a minimum number of samples, which is determined by the two hyper-parameters, *eps* and *minPts*. In order to effectively choose the most appropriate value for the *eps* hyper-parameter, we perform a fast k-nearest neighbour (kNN) search which calculates the mean distances of each point to its k-th nearest neighbours, with k representing the minPts value, where minPts is usually set to the dimensionality of data plus 1, as suggested by [58]. A small *eps* can lead to considering a massive part of observations as outliers, while a large one can partition all observations into the same cluster. Therefore, after setting *minPts* equal to D+1, or else 8, we perform the kNN distance method and plot the distances in ascending order as depicted in Fig. 17. Based on [58], in order to define the optimal *eps*, we need to find the point where the plot makes a knee, which according to the red line is approximately to 0.1. After making some tries with values around the 0.1, we deduce that DBSCAN performs better when *eps* = 0.12, where 2 main clusters are created, representing the weekend (green) and working days (blue) consumption, besides the outlier profiles (red) as illustrated in Fig. 18.

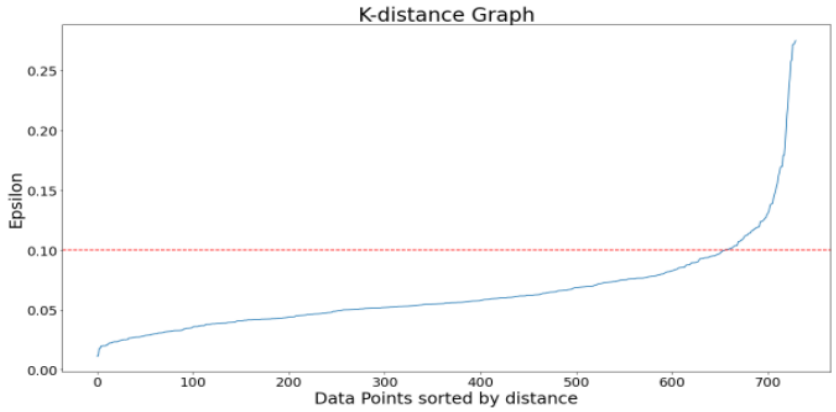


Fig. 17: kNN distance plot for determining the eps value of DBSCAN for HH' electricity

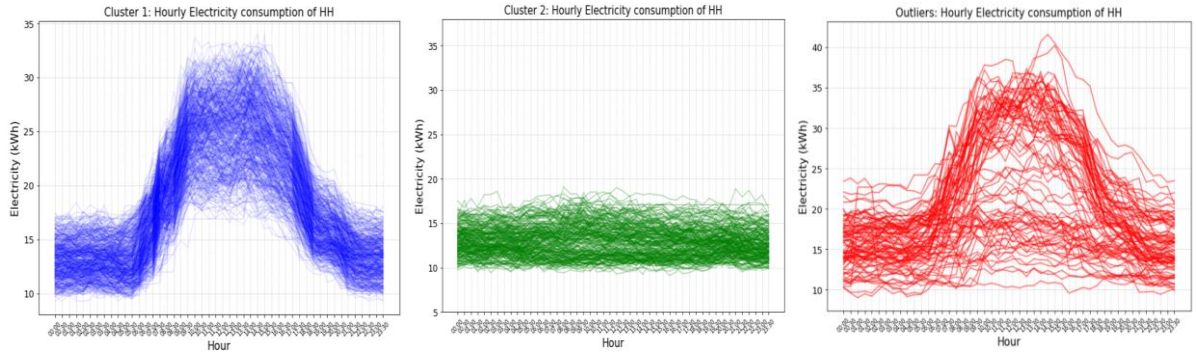


Fig. 18: Clusters identified from DBSCAN (blue and green), along with the abnormal dates (red) for HH electricity

With the view to visualize the performance of DBSCAN technique, Fig. 19 presents a convex hull plot with two dimensions, which are generated by PCA analysis and represent 76.9% and 20.5% of the dataset's variation, repsectively. Fig. 19 visualizes the relations between the two clusters and the outliers, where the blue dots represent the weekends, the orange dots the working days, and the green ones that are far away from the convex hull represent the days with abnormal consumptions. It is notable that only two electricity load patterns dominate, which was expected from the beginning; however, it is not sufficient since potentially there might be more than two patterns as subgroups to the two dominant clusters. Therefore, we will investigate this assumption by conducting a subsequent k-means clustering on the clean dataset,

which consists of 645 samples, given that 85 out of 730 initial samples were identified as outliers and were filtered out.

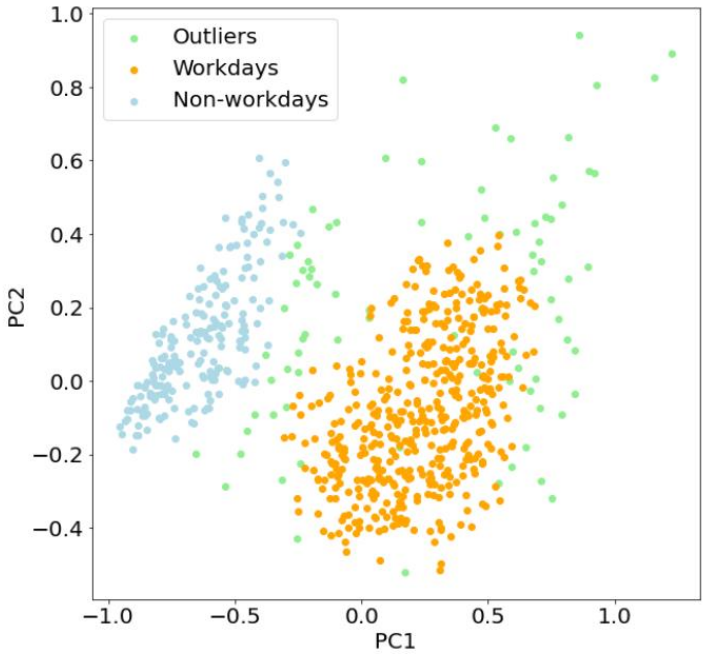


Fig. 19: Visualization of DBSCAN's results for HH electricity consumption

ii. 2nd stage - k-means

Having acquired a clean dataset, free of the outliers detected by DBSCAN, we use the k-means clustering to find more accurate electricity load profiles than the density-based clustering, based on the 7-features dataset. Having defined the 7 new statistical features, k-means with the Euclidean distance used for dissimilarity measure, is an efficient approach with the least time complexity, commonly used for load profiling clustering [24]. K-means requires only one hyper-parameter, the number of k clusters, which is predefined by the user. For choosing the optimal k, we perform a grid-search for k values ranging from 2 to 31, where we evaluate the performance of each value based on two clustering validation indexes (CVIs), the Dunn Index and the Silhouette score. The largest the both values are, the more compact and well separated are the clusters identified for the corresponding k value. Using too many clusters to discover the different load profiles could result in creating artificial boundaries within real clusters, while using too few could neglect important separation boundaries. Fig. 20 shows the average silhouette score, while Fig. 21 shows the Dunn index per number of clusters. Even though the Silhouette score is the highest for k = 2, we do not choose this value since it gives the same clusters that the DBSCAN identified, representing the weekends and the working days. By observing both Fig. 20 and 21, we can see that both have a peak on k = 6 and after experimenting with some other k values we infer that k = 6 gives the optimal number of clusters as displayed in Fig. 22.

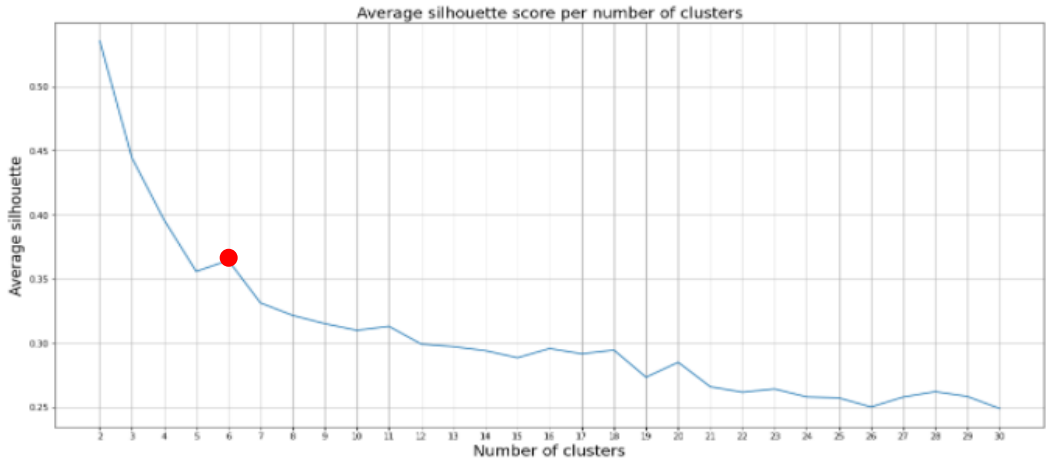


Fig. 20: Clustering performance based on Silhouette score for different number of clusters for the HH electricity feauture

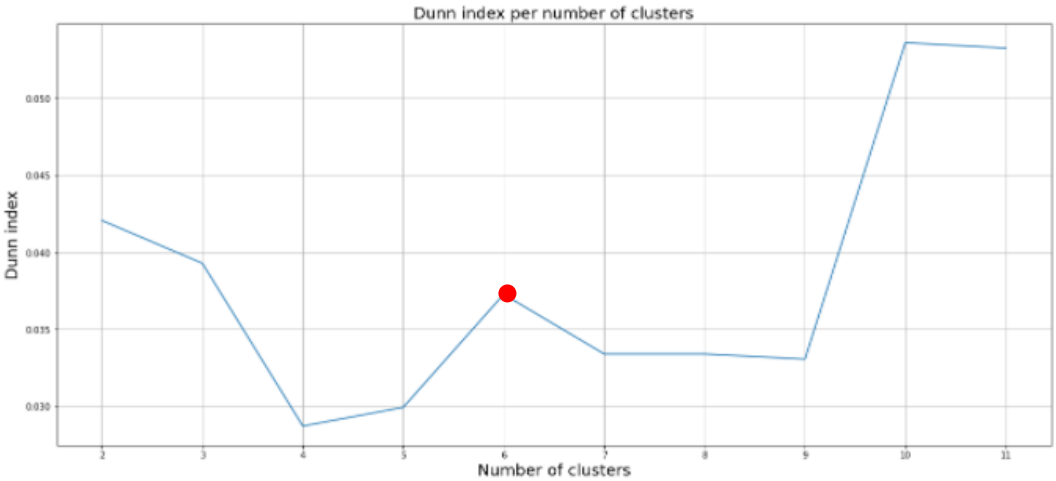
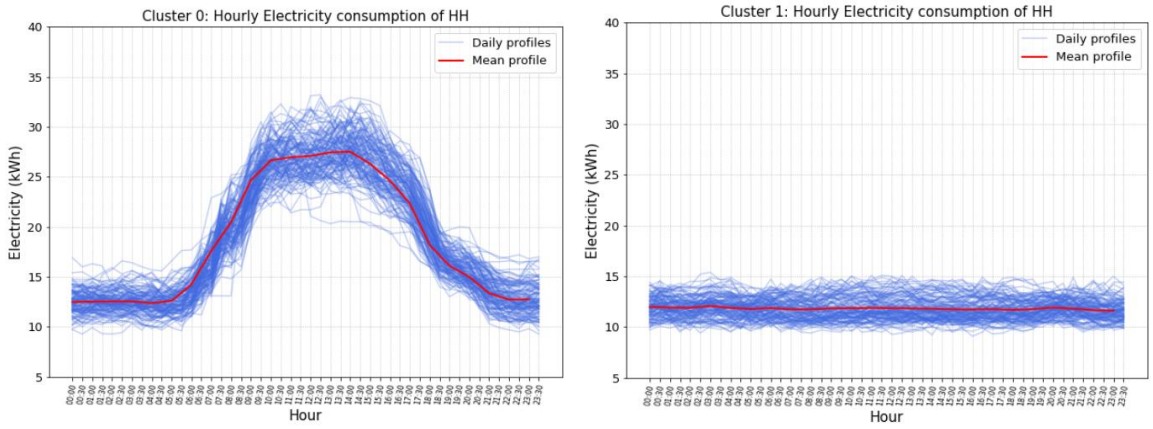


Fig. 21: Clustering performance based on Dunn index for different number of clusters for the HH electricity feauture



(a) (b)

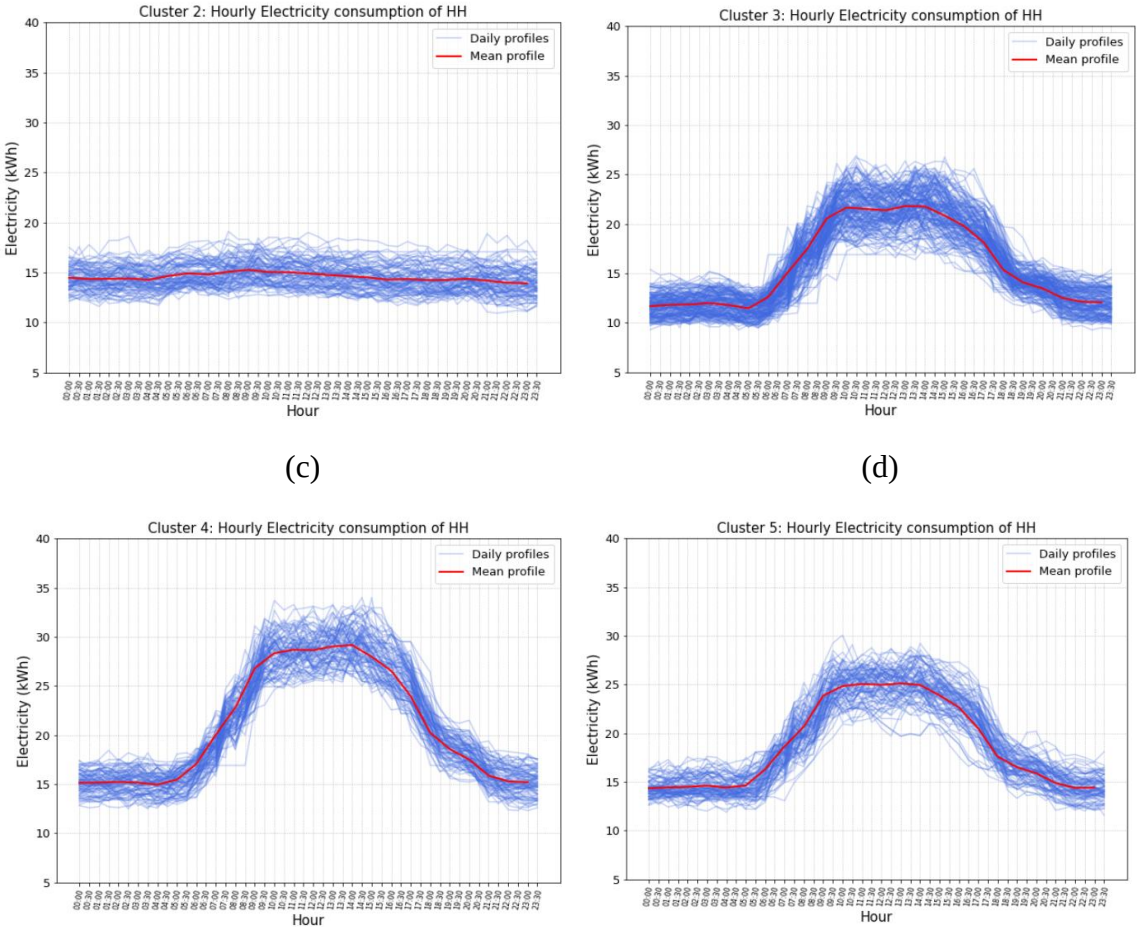


Fig. 22: Electricity load patterns identified for the HH electricity feature by the k-means clustering method

From Fig. 22, we observe that 2, out of 6 clusters, belong to weekends, while the other 4 to weekdays, since they follow the working schedule trend, where a rise, a stable and a descending period are identified. Particularly, clusters 0, 3, 4 and 5 have similar time durations but different peaks and high-level values of consumption during the daytime between 10:00 and 15:00, while clusters 1 and 2 have similar, almost steady, consumption during the day without a clear rising or falling trend, and a low electricity usage level, indicating the rare occupation of the building during these days. Besides the dissimilarities between the low-level and high-level consumption magnitude, we also observe that cluster 3 presents two peaks at around 10:00 am and 14:00 pm, which might be associated with the variable of occupancy. Additionally, cluster 4 has a small peak at around 20:00 pm, implying an overtime working or another operation that takes place at evening.

B. DE's Electricity consumption

i. 1st stage – DBSCAN

Similarly, we implement the same process for the DE's electricity feature. After defining the optimal values for *eps* and *minPts* through the K-distance graph, as 0.12 and 8 respectively, DBSCAN identifies the following 4 clusters illustrated in Fig. 23, along with the outliers, which account for 41 samples, out of the total 601, observations (~6%). Fig. 24 depicts a convex hull

plot with two dimensions, which explain the most of the variance from the total 4 dimensions of PCA dimensionality reduction method (68.5% and 28.9%). While it is obvious that four distinct high-density clusters are intuitively created, they are not very compact, and therefore, we further perform a k-means on the clean dataset to either support this conslusion or reject it.

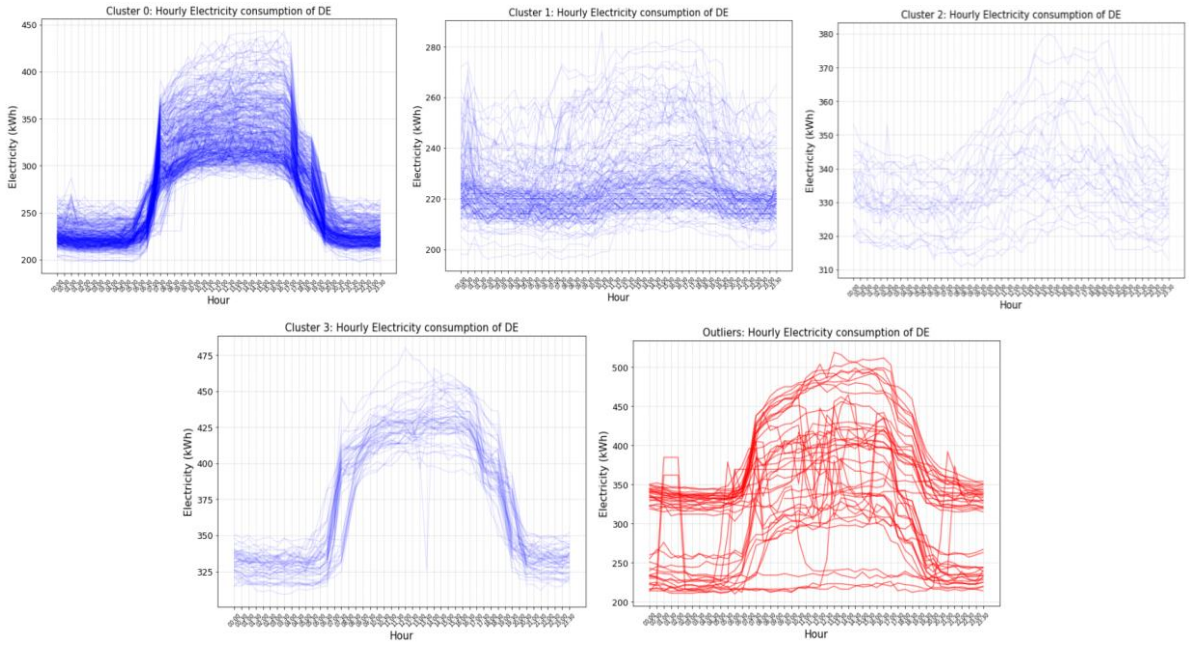


Fig. 23: Clusters identified from DBSCAN (blue and green), along with the abnormal dates (red) for DE electricity

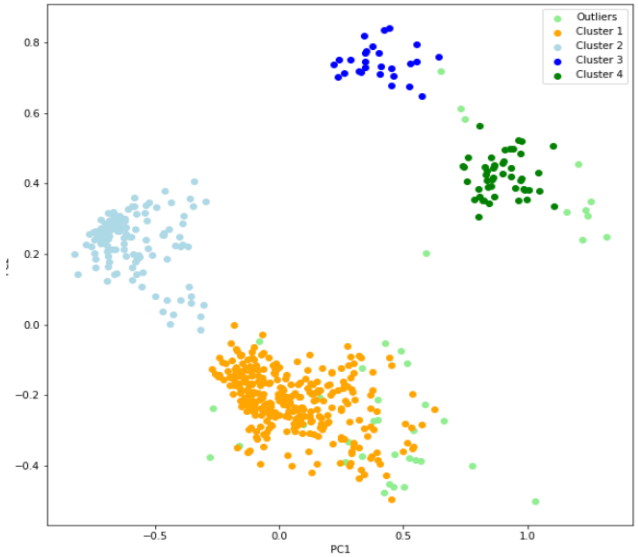


Fig. 24: Visualization of DBSCAN's results for DE electricity consumption

ii. 2nd stage - k-means

Similarly with the 2nd stage of clustering for HH electricity consumption, after defining the k = 5 as the optimal k, based on the two CVIs, the Silhouette score and the Dunn Index, the clusters depicted in Fig. 25 are derived. As previously, it is obvious that 2, out of 5 clusters, belong to weekends, while the other 3 to weekdays, since they follow the working schedule

trend. Overall, we observe that between the clusters of the same category (weekends or weekdays), the magnitude of the electricity consumption is the main dissimilarity criterion.

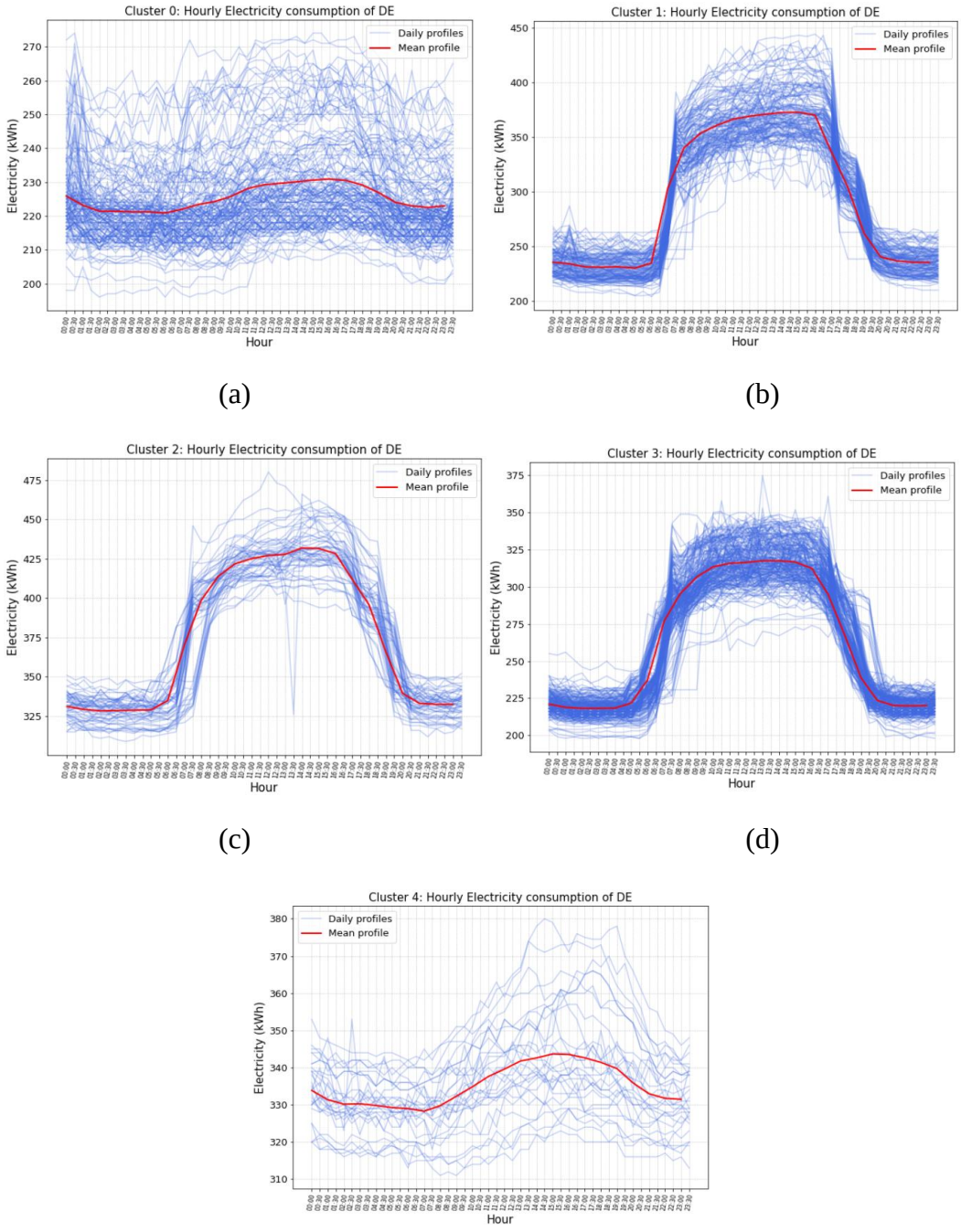


Fig. 25: Electricity load patterns identified for the DE electricity feature by the k-means clustering method

Ultimately, after performing a baseline clustering on the dataset that has been dimensionally reduced, we imply that k-means with the Euclidean distance metric uses the magnitude of energy consumption as the main dissimilarity criterion. Therefore, in order to discover more solid clusters, we are going to apply the k-means algorithm on the raw time series data, while the DTW will be used as distance metric, instead of the Euclidean distance. Besides that, this

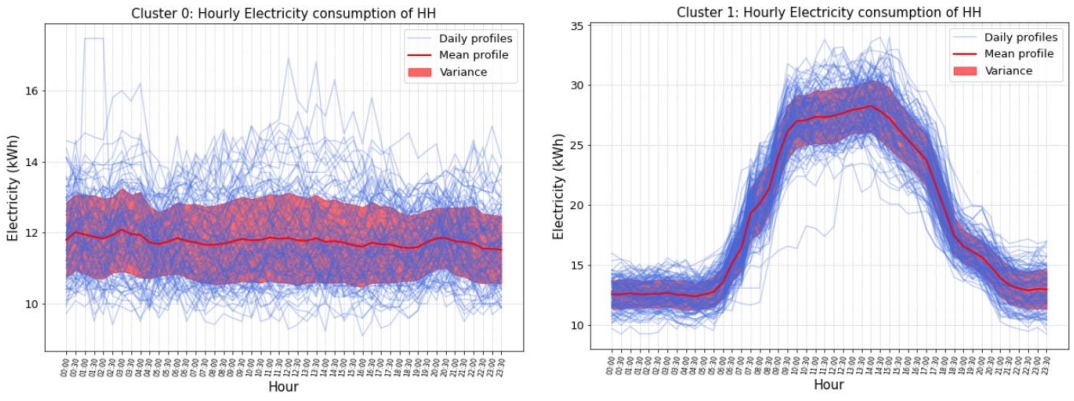
baseline method cannot be effectively applied to the other two features (cooling, heating), because they cannot be segmented into 4 houly periods, given that they do not follow the same accurate working schedule as electricity does.

3.3.1.2 k-means with DTW

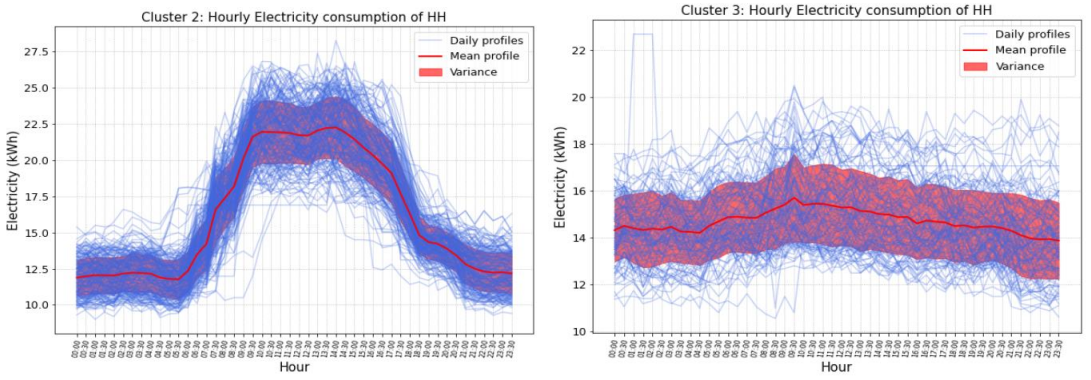
Expert segmentation is not always a robust preliminary step towards the clustering of time series. Contrary to feature extraction techniques, time series data are, most of the times, distorted, meaning that in order to compare time series, it is of importance to take into account the invariances between them [28]. Such invariances may vary accordingly, from different time series datasets to different scenarios; however, in this work, the invariances that we encounter, are shift invariances, where two time series days follow similar trends but with a slight phase difference. This phase difference might be only half or one hour, but even a slight difference could consider two time series totally dissimilar if an inappropriate distance metric is used, like the Euclidean one. Therefore, in order to automate and improve the process of clustering by capturing different group profiles based not only on the magnitude but also the trends that do not align exactly in time, speed or length, we are going to apply the k-means method with the Dynamic time Warping (DTW) distance metric to the whole raw time series data. DTW metric, being a shape-based distance metric can address the shift invariance in a way that robust and consistent profiles can be discovered, without the definition of a pre-determined feature space. In order to specify the number of clusters k, we follow the same evaluation process as for the baseline method, using the metrics of Silhouette score and Dunn Index. Nevertheless, k-means algorithm is highly dependent to the initialization of clusters' centroids which are randomly placed on feature space. Therefore, in order to find the ideal groups, we ran multiple times the algorithm until it discovers the most representative clusters. The metric used for evaluating the efficiency of each run was the total compactness of all clusters, which was measured by their variance; the smallest the variance of each cluster, the better the initialization of centroids, and ultimately the most distinct and compact the clusters.

Hence, for each of the 5 energy features of our dataset, we apply the k-means clustering with the DTW metric, imported from the *tslearn* Python's package and we receive the following clusters.

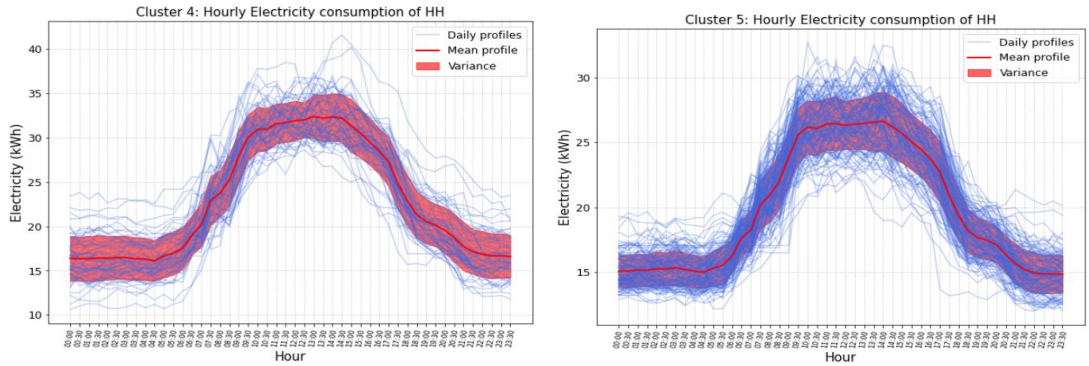
A. HH electricity



(a) (b)



(c) (d)

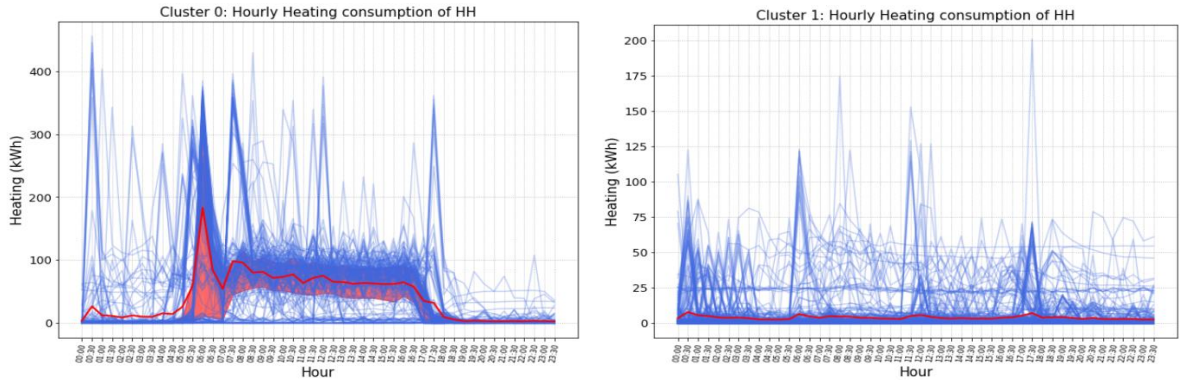


(e) (f)

Fig. 26: Electricity load patterns identified for the HH building by the k-means DTW

As expected, similarly to Fig. 22, we observe in Fig. 26, that 2, out of 6 clusters, belong to weekends, while the other 4 to weekdays. While we can see that both approaches seem to capture similar groups, k-means with DTW creates more distinct groups which do not have only magnitude dissimilarities but also different trends. Particularly, in cluster 1, the peak at 14:00 is more prominent than in cluster 0 at Fig. 22. The same applies to cluster 2 of the k-means with DTW whose two peaks are more concrete and compact. Last, we can observe that k-means with DTW identified a new cluster that the baseline method did not, which is the cluster 4 that consists of all the working days with the highest consumption during working hours. The baseline method did not capture that group of observations mainly because most of them were identified as outliers by the DBSCAN method, considering that they were diverted from the normal behavior, given that they account for a minority.

B. HH heating



(a) (b)

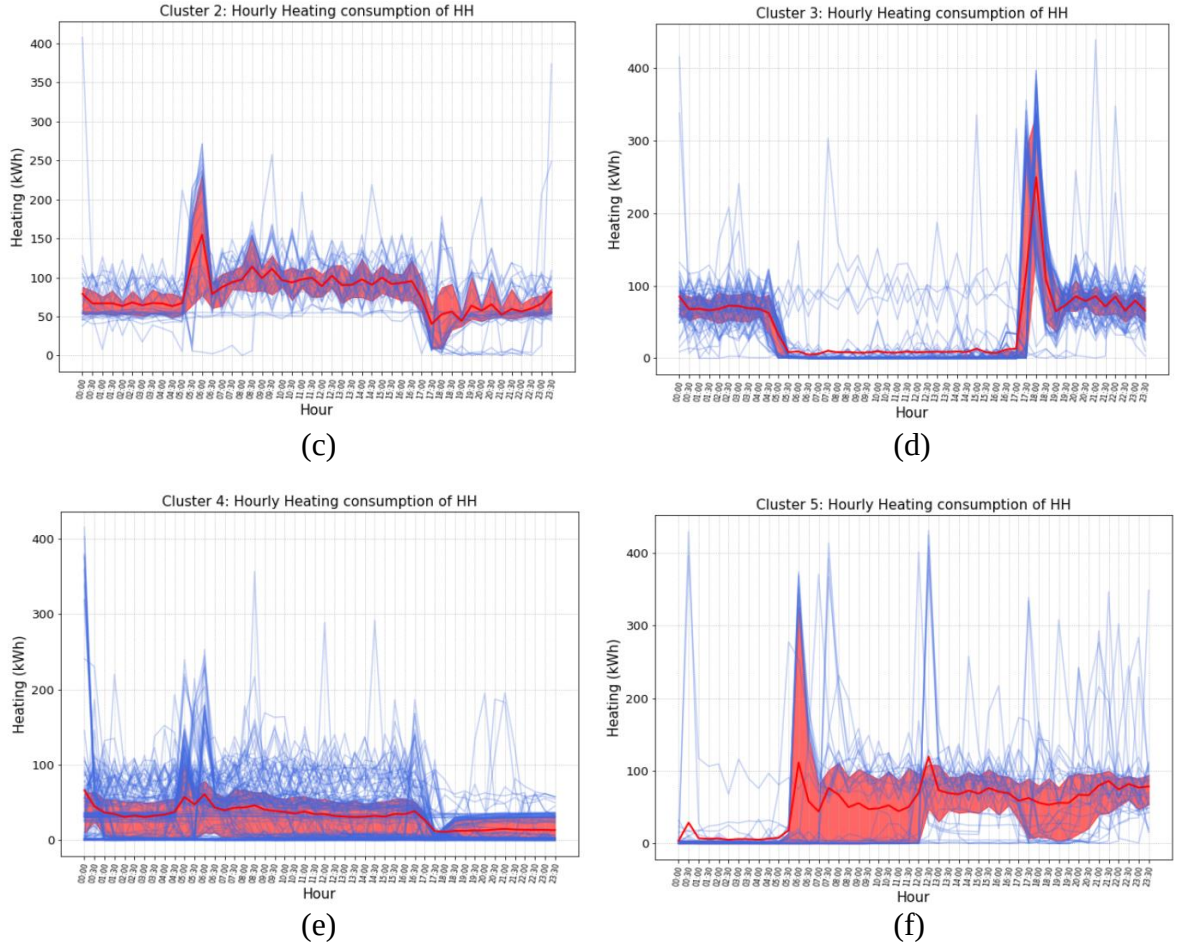
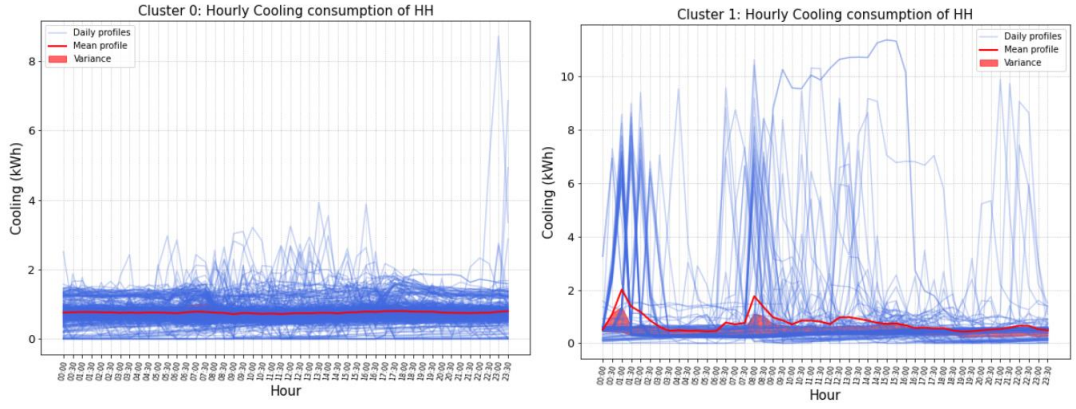
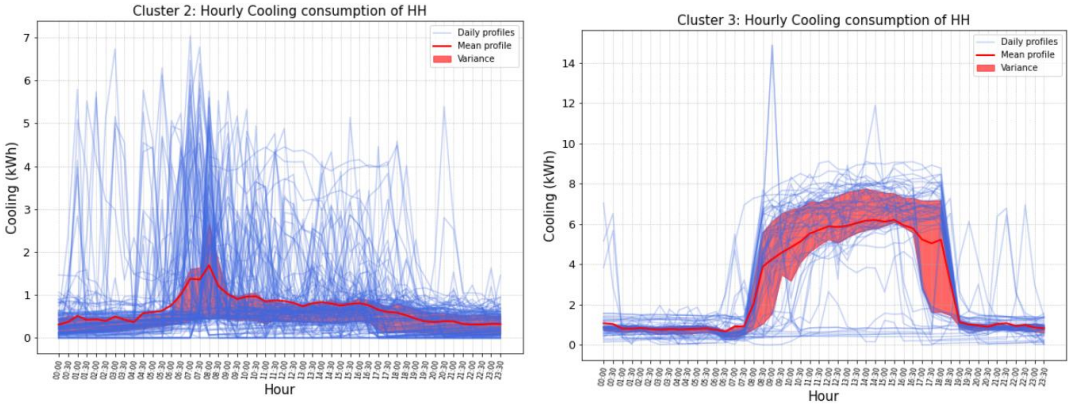


Fig. 27: Heating load patterns identified for the HH building by the k-means DTW

As we discovered during the EDA, when the working schedule starts every morning, the boilers of the HH building, most of the times, launch at a peak (Fig. 14a) and afterwards they are adjusted to building's and occupant's needs. Some sharp patterns are discovered, as illustrated in Fig. 27, however, we observe many random fluctuations of heating consumption that do not form a specific pattern, besides the peaks that are mostly identified and form the displayed groups.

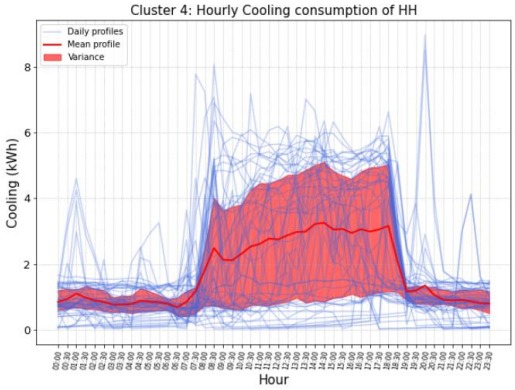
C. HH cooling



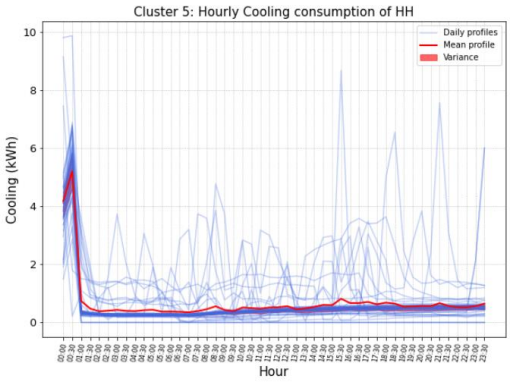


(c)

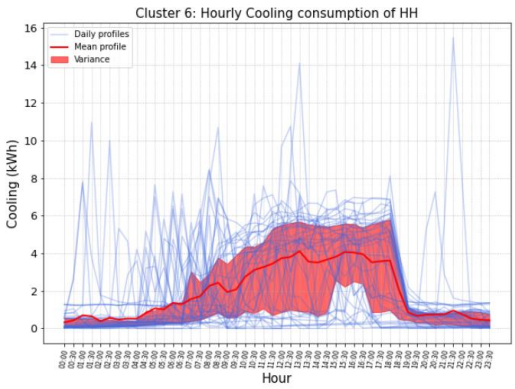
(d)



(e)



(f)



(g)

Fig. 28: Cooling load patterns identified for the HH building by the k-means DTW

Similarly with the heating consumption, cooling cannot form compact patterns considering the random fluctuations of chillers; however, some distinct groups are identified.

D. DE electricity

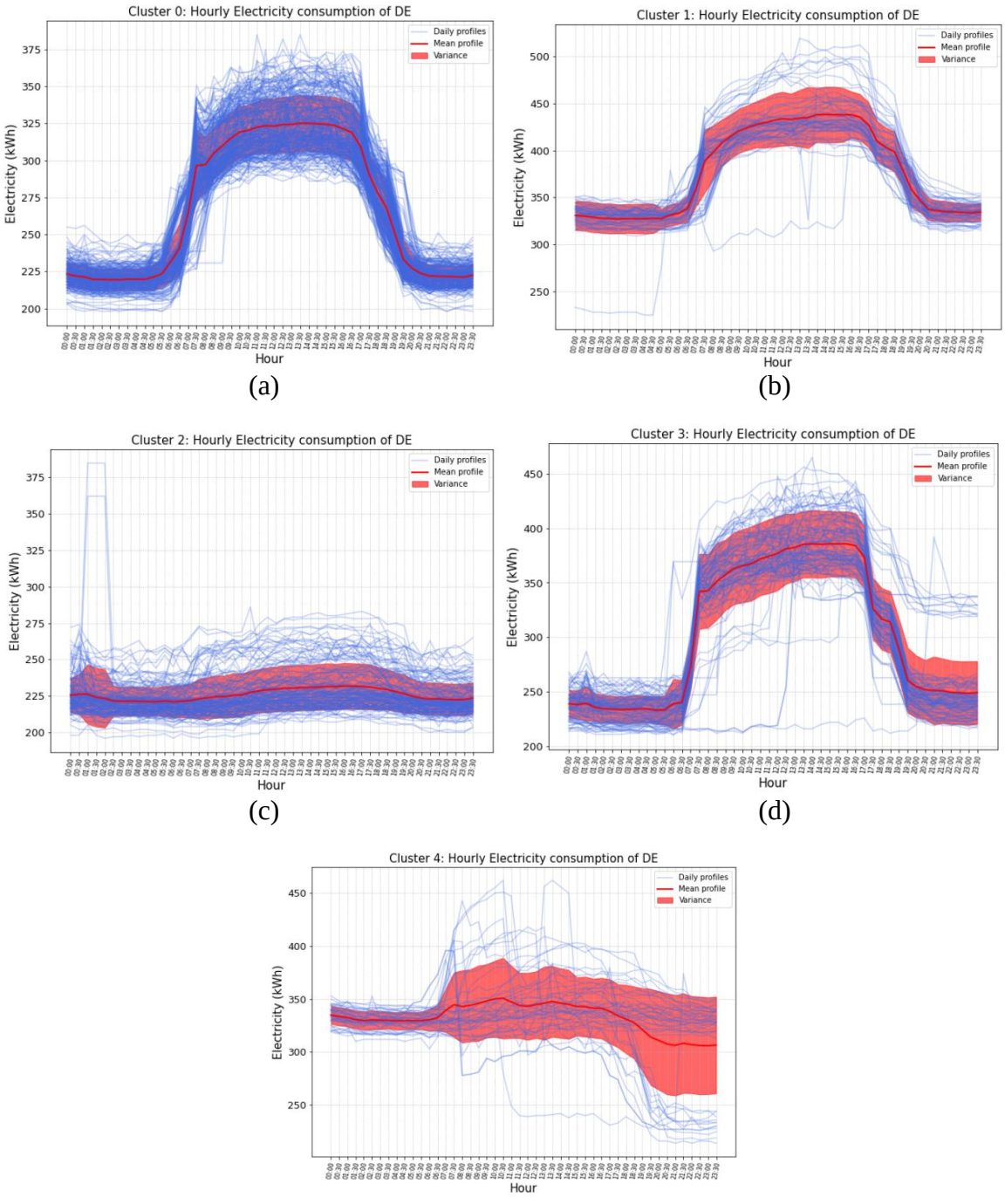
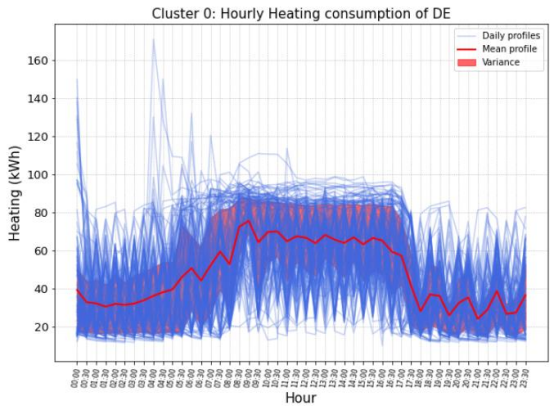


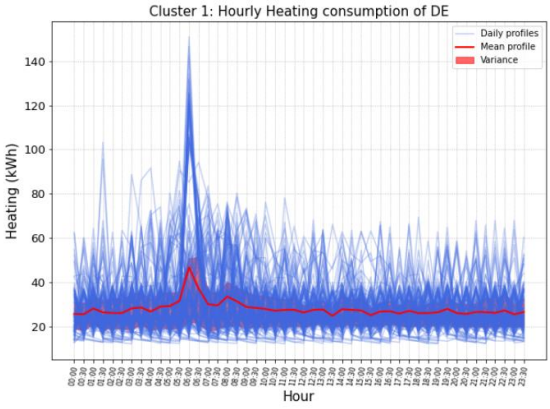
Fig. 29: Electrcity load patterns identified for the DE building by the k-means DTW

As discussed in HH's electricity clustering, k-means with DTW captures trends that the baseline clustering cannot, since it ignores the time shifts.

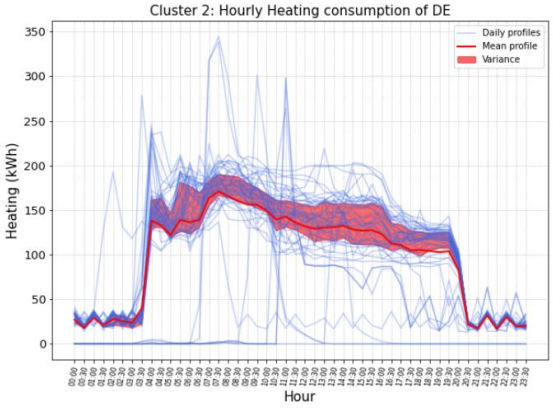
E. DE heating



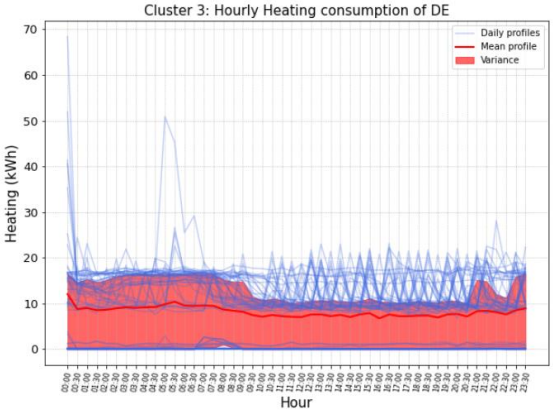
(a)



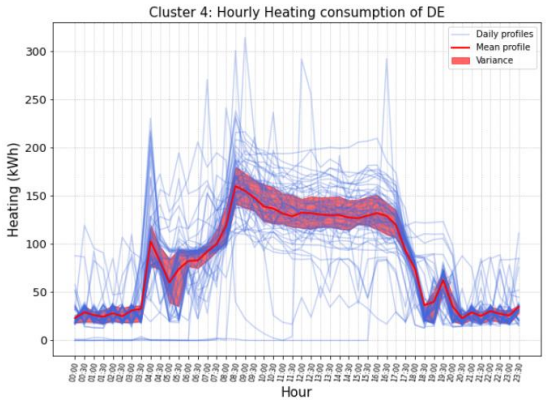
(b)



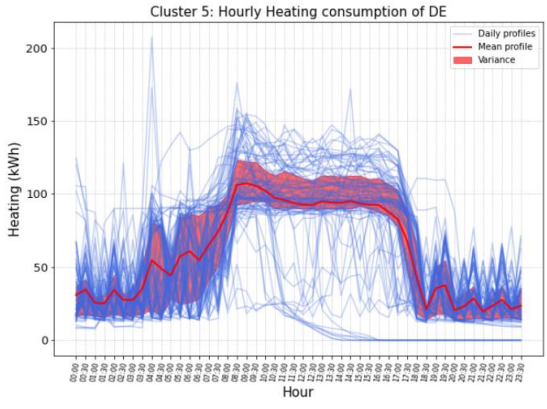
(c)



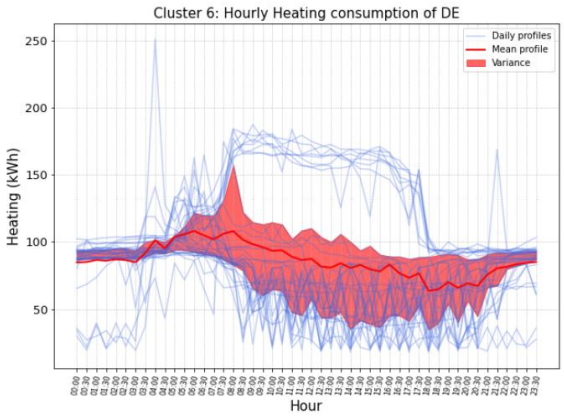
(d)



(e)



(f)



(g)

Fig. 30: Heating load patterns identified for the DE building by the k-means DTW

Even though DE's heating consumption performs random fluctuations, as well, the clusters identified reveal that the patterns of weekday profiles, approximately follow the working schedule, at least more accurately than HH's heating, which might be supported by the fact that DE building is modern and 'high-tech' and therefore the heating controllers are better scheduled, contrary to the HH. Overall, it is obvious that consumption's magnitude differs a lot from day to day; however, clustering analysis can mostly capture the uptrend or downtrend.

To sum up, the identified energy load patterns from the proposed clustering method seem almost reasonable, for both buildings, since each cluster has a unique energy load pattern, especially for the electricity feature whose clusters have clear variations compared with the other clusters; nevertheless, more investigation needs to be done in order to further understand the reasons that lead to those prominent identified profiles.

3.3.1.3 Compare clustering techniques

To further demonstrate the advantages and disadvantages of the two clustering methods under investigation, we compared their performance on HH's electricity load profiles. The first method has been the 2-stage clustering, applying DBSCAN for filtering out the outliers and following, the k-means on the new features extracted for discovering the patterns, while the second method has been the k-means with the DTW distance metric on the raw time series data. Both clustering techniques have been widely adopted in energy time series data in the literature; however, the nature of each dataset determines the efficiency of each method. Particularly, even though the first approach is more time and resources efficient, considering that instead of a 48-dimensinal dataset, a 7-dimensional dataset is used, from the experiments conducted, we imply that data needs to be quite neat in order this method to be effective. We observe that DBSCAN filters out the profiles with the high consumptions, while DTW clusters them into a separate cluster. The disadvantage of DBSCAN algorithm is that it does not provide interpretable knowledge about the outliers, while its hyper-parameters should be carefully defined, otherwise values that should be defined as outliers, they are not and the opposite. Even by bypassing the 1st step of the baseline clustering, k-means with the Euclidian distance metric applied on the extracted features, still underperforms the k-means with the DTW metric on raw time series data, as supported by the compactness of the discovered clusters for each of the two methods in Table 5. Normally, given that the clusters have different densities, as Fig. 19 and 24 depict, k-means clustering algorithm would underperform and output inadequate clustering results if the outliers were not removed since it is difficult to identify clusters with widely

different sizes or densities. However, within the context of our scenario this is not applied, since the observations that are considered as outliers by the DBSCAN, in reality, most of them are not outliers, but observations with less frequency than other patterns. Similarly, DBSCAN could not stand by its own because as illustrated in Figure 19, only two patterns are identified and reflect the characteristics of electricity usage for working and not working days.

Clusters	k-means with Euclidian's variances	k-means with DTW's variances
Cluster 0	1.86	1.40
Cluster 1	1.68	1.07
Cluster 2	1.97	1.71
Cluster 3	1.88	1.49
Cluster 4	1.74	1.51
Cluster 5	1.75	1.71

Table 5: Clusters' variances for k-means with different distance metric on HH electricity, indicating the compactness of clusters

Notably, the baseline clustering approach seems to have good performance only when building's consumption follows particular trends and hence data can be robustly represented by dimensionally reduced statistical features; however, still the proposed method slightly outperforms, considering that time series data should not be considered as independent features, but as sequences that experience many time shifts. The above comparison demonstrated the effectiveness of the proposed clustering method when energy data is not well defined due to building's or system's weaknesses, besides when domain knowledge is missing and the human interference to model's approach is not desirable. Unlike the expert segmentation required for the baseline method, the proposed cluster analysis allows load patterns, probable previously unknown, to be identified in a not pre-determined time domain and with no dependency on domain expertise. In this way, robust and consistent reference profiles can be discovered.

3.3.2 Interpretation of knowledge discovered

3.3.2.1 Investigate CART potential

To understand the reasons that lead to the identification of the aforementioned energy load profiles for each feature, based on the proposed clustering analysis, we need to explore the underlying relations between these patterns and the potential influencing exogenous factors. To do so, we are going to investigate the potential of the CART classification algorithm. Compared with other popular supervised learning algorithms such as random forest or extreme gradient boosting, CART has a higher interpretability, given that it forms a decision tree with the clusters we identified used as target variables, and the potential exogenous influencing factors used as predictors [24]. The exogenous factors that have been selected for exploration, have been commonly used in literature for interpreting building energy data and they are summarized in Table 6.

Variable	*Description*
Weekday	Day of the week
Month	UK Bank holidays
Holiday	Month of the year
Mean temperature (℃)	Daily average outdoor temp
Mean humidity (%)	Daily average outdoor hum

Table 6: Selected potential influencing exogenous factors as input predictors to CART

In order to examine the efficiency of CART algorithm, we apply the CART for the target variable of HH electricity's clusters which have been the most compact and neat clusters, comparing to the other features. The two hyper-parameters *mdepth* and *minleaf* of CART algorithm, were determined by manually evaluating the best pair that avoids over partitioning or early stop that could neglect further decision rules. Our dataset is relatively small for applying any evaluation estimator, such as the k-fold cross validation. Therefore, after manual tuning, the *mdepth* value was set to 6, meaning that the CART tree can make up to 6 splits until coming to a prediction and the *minleaf* value to 8, meaning that a node can be further split when the number of samples in each split is more than 8. The final accuracy of the trained CART model was calculated by the ratio between the predicted and the actual labels (or else clusters), being 65.4% for the training data and 54.8% for the test data. Our dataset, consisting of only the exogenous variables of Table 6 and the cluster that each observation belong to, was split into train and test data with an analogy of 80% and 20% respectively, randomly, using the *train_test_split()* function from the sci-kit Python library, considering that each observation is a whole date, hence we do not take into account the temporal relation between dates.

From Table 7, it is obvious that there are not clear, but vague decision rules that lead to specific patterns and hence the same rule can output more than one electricity patterns, implying that other exogenous factors may affect the daily electricity consumption profile, probably the occupancy factor, building's operations or the events taking place every day and so forth or simply implying that the samples are not enough for this kind of problem. The predicted patterns' numbers in Table 7, match to the clusters in Fig. 26. This assumption is also corroborated by the accuracies achieved on train and test data.

Decision rule	Boundary condition	Predicted cluster
Rule 1	If weekend & temp≤12 & hum≤74	4-0
Rule 2	If weekend & temp≤12 & Jan≤month≤Mar & hum≤76	1-3
Rule 3	If weekend & temp≤12 & 67≤hum & Mar≤month≤Sept	4-0
Rule 4	If weekend & 12≤temp	0
Rule 5	If working day & Jan≤month≤Feb & 4≤temp	0-1
Rule 6	If working day & month=Jan & temp≤4	5-4
Rule 7	If working day & Feb≤month & hum≤76	5
Rule 8	If working day & Feb≤month & 76≤hum≤83	1
Rule 9	If working day & Feb≤month & 83≤hum & temp≤9	5
Rule 10	If working day & Jun≤month≤Jul & hum≤60	5
Rule 11	If working day & Jun≤month≤Jul & 76≤hum	5
Rule 12	If working day & Jun≤month≤Jul & 60≤hum≤76	2
Rule 13	If working day & Jul≤month≤Sept	2
Rule 14	If working day & Sept≤month≤Nov	1
Rule 15	If working day & Dec≤month & hum≤79	1
Rule 16	If working day & Dec≤month & 79≤hum & temp≤7	0
Rule 17	If working day & Dec≤month & 79≤hum & 7≤temp	4

Table 7: Decision rules generated by CART for HH electricity variable

According to Table 7, the three main variables that contributed to the different load patterns were the weekday, the temperature and the month. Particularly, CART algorithm selects "weekday" as the first splitting variable at the root node and then for each of the two subgroups, the "weekends" and the "working days", it uses different splitting variables. For the "weekend", the variable "temperature" (temp ≤12) is mostly splitting the samples into the two clusters that belong to weekends (cluster 0 & 3), with some exceptions thar are defined by some deeper decision rules, according to Table 7. For the working days though, "Month" is the first decision rule, with "temperature" and "humidity" following, without however accurately distinguishing the different clusters referring to the working days, by the aforementioned exogenous variables. For example, when "month = [Dec, Jan]" and the "temperature" is very low (temp<4), daily profiles belong to cluster 4, which represents the dates with the highest electricity consumption, implying that electrical heating demands were high during that period due to the low temperature. Nevertheless, the decision rules for choosing between the clusters 1, 2 and 5, are not well defined, given that the consumption magnitude is not too different between the three clusters, but they differ on the periods when peaks are taking place, which most likely is dependent on other exogenous variables that we have not taken into account or we have no domain knowledge about them.

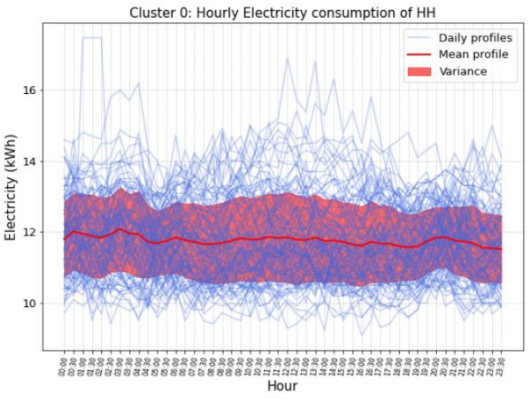
Even though CART is one of the most effective algorithms for knowledge discovery, since its output is a set of interpretable decision rules in the format of "if-then" rules and therefore it could be highly trusted when used for business decision-making, our use case scenario lacks of important variables that affect energy load, hence CART underperforms. On top of that, the labels of our dataset fed into CART, have been defined through an unsupervised learning, which cannot reassure us about their quality. It is generally accepted that any ML model is as good as its data is. Therefore, considering that the performance of CART was not highly accurate for the HH electricity feature that was represented by the most distinct and compact clusters, we won't further examine the CART algorithm for the rest of our scenario's features, but instead, we will follow a more manual approach, where the distributions of exogenous variables will be studied for each separate cluster.

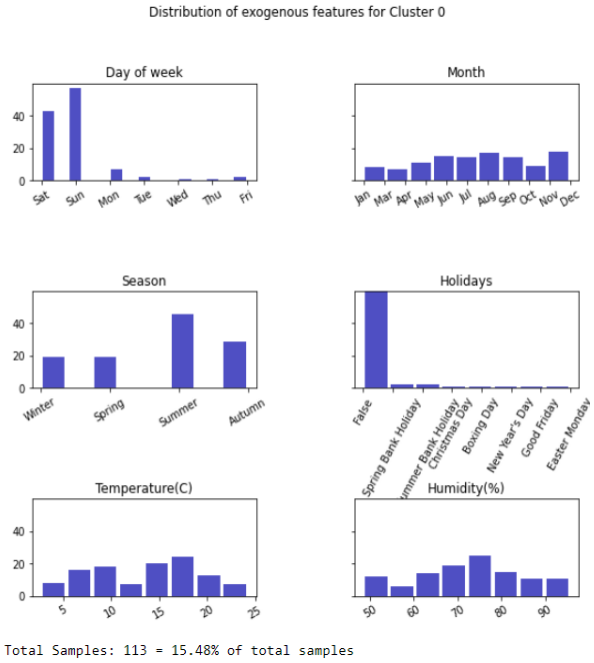
3.3.2.2 Exogenous variables distributions

Following the experimenting over the CART algorithm, we propose a visual knowledge discovery method where we interpret the identified clusters by plotting the distributions of the potential influencing exogenous variables for each and every cluster and explaining their relation. Therefore, we are going to perform a visual separate interpretation for each cluster for each of the 5 energy features of our dataset, while subsequently we provide a knowledge discovery for some prominent relations between clusters and exogenous variables.

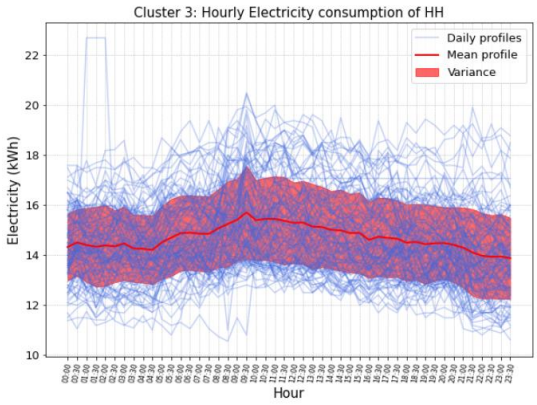
HH electricity

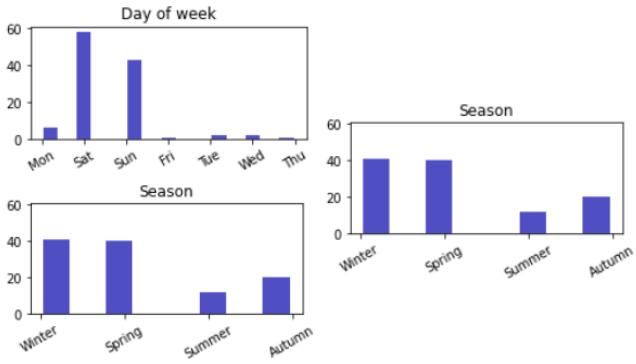
In Fig. 31a, b we observe that both profiles belong to weekends; however, the second profile dominates during the cold months, when likely some electrical heating would remain turned on in low levels. Cluster 2 which has the lowest magnitude comparing to the other two clusters of working days, is mostly observed during the hottest months of the year, while, on the contrary, cluster 4 which contains the profiles with the highest consumption, dominates over the coldest months of the year.


Cluster 0: Hourly Electricity consumption of HH
Daily profiles
Mean profile
Variance
Electricity (kWh)
Hour

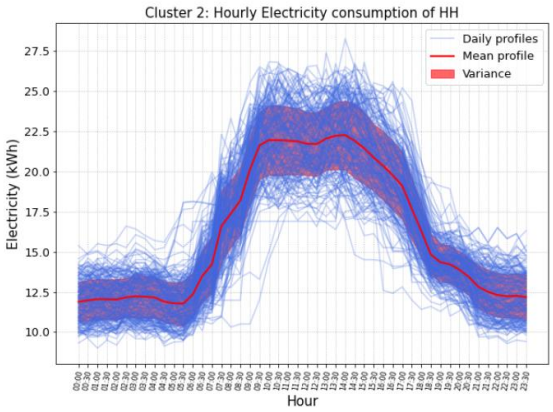

Distribution of exogenous features for Cluster 0
Day of week
Month
Season
Holidays
Temperature(C)
Humidity(%)
Total Samples: 113 = 15.48% of total samples

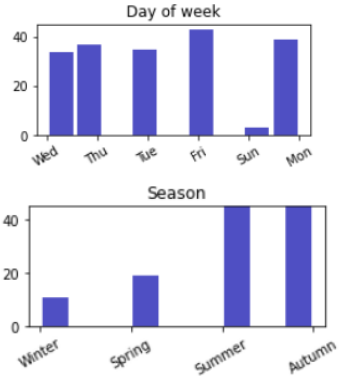
(a)


Cluster 3: Hourly Electricity consumption of HH
Daily profiles
Mean profile
Variance
Electricity (kWh)
Hour

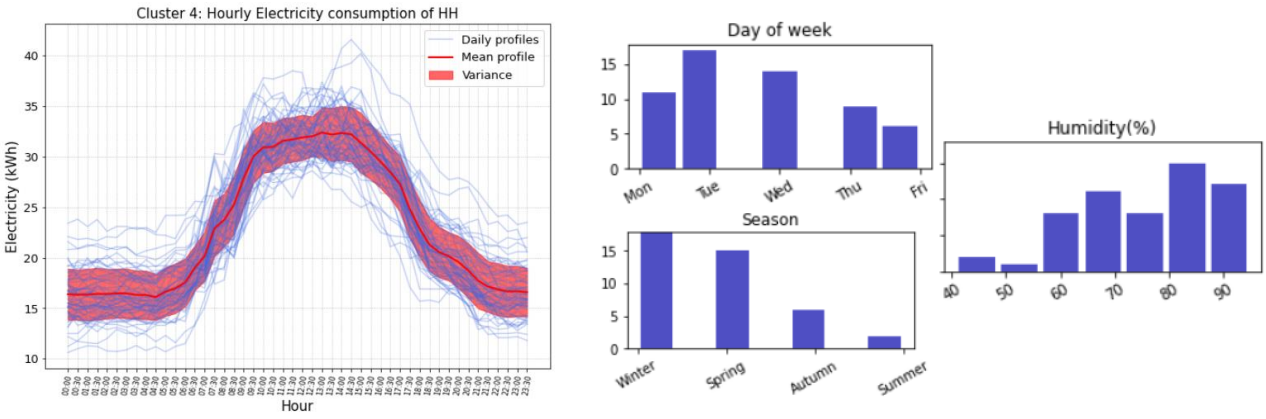

Day of week
Season
Season

(b)


Cluster 2: Hourly Electricity consumption of HH
Daily profiles
Mean profile
Variance
Electricity (kWh)
Hour


Day of week
Season

(c)



(d)

Fig. 31: Clusters along with selected exogenous variables distributions for HH electricity

HH heating

Although the cluster 3 seems to prevail during the summer season, as depicted in Fig. 32, the other clusters are mostly independent of the exogenous variables under investigation, since their distributions appear to have almost equal frequencies for every state.

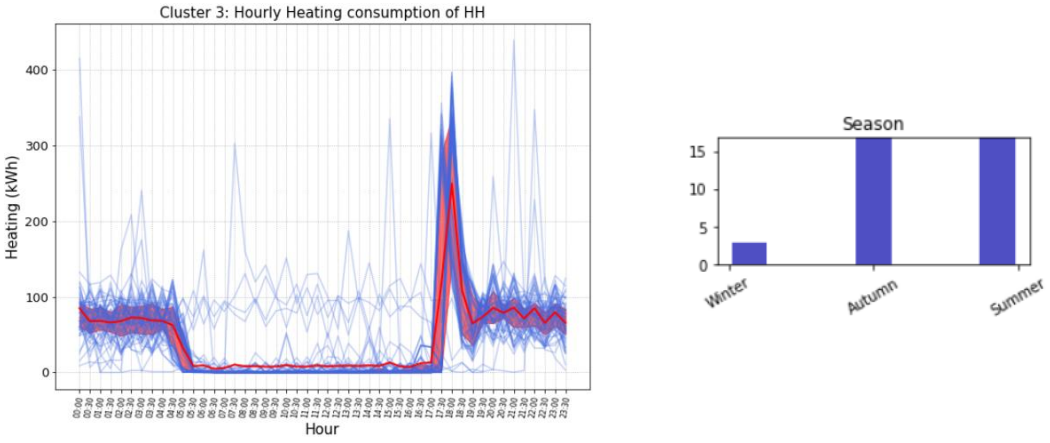
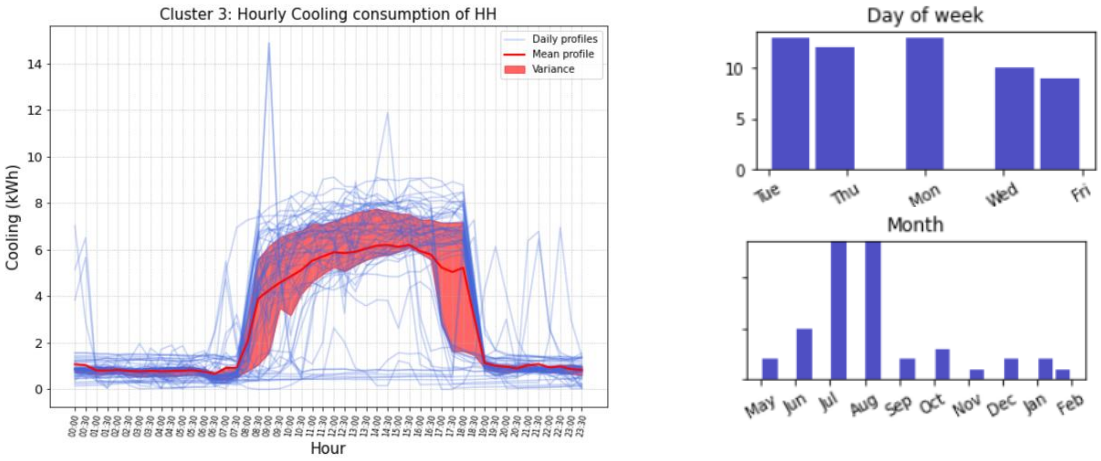


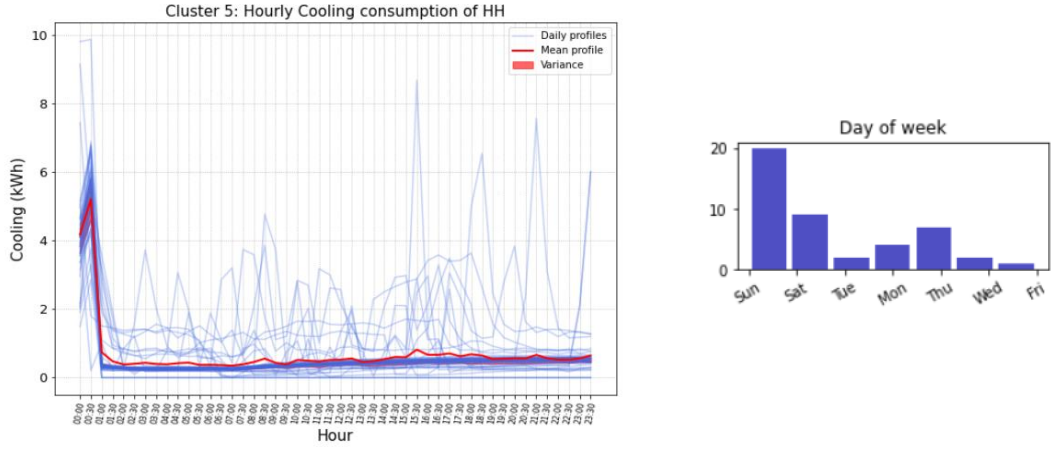
Fig. 32: Clusters along with selected exogenous variables distributions for HH heating

HH cooling

Regarding the cooling load consumption, as illustrated in Fig. 33, the samples of cluster 3 are mainly observed during summer months, which is totally corroborated by the shape of working hours when cooling demands are higher, while cluster 5 seems to dominate during the weekends when probably, chillers perform some kind of restart or another operation at midnight.



(a)

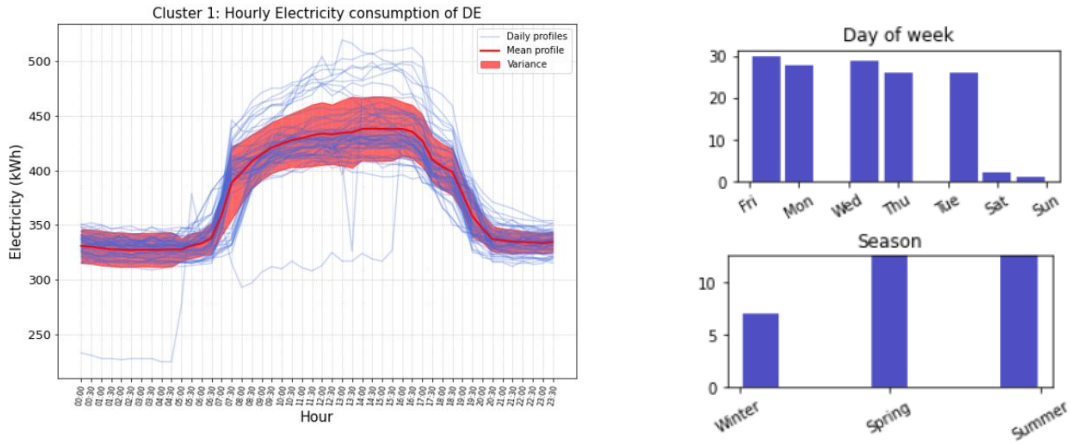


(b)

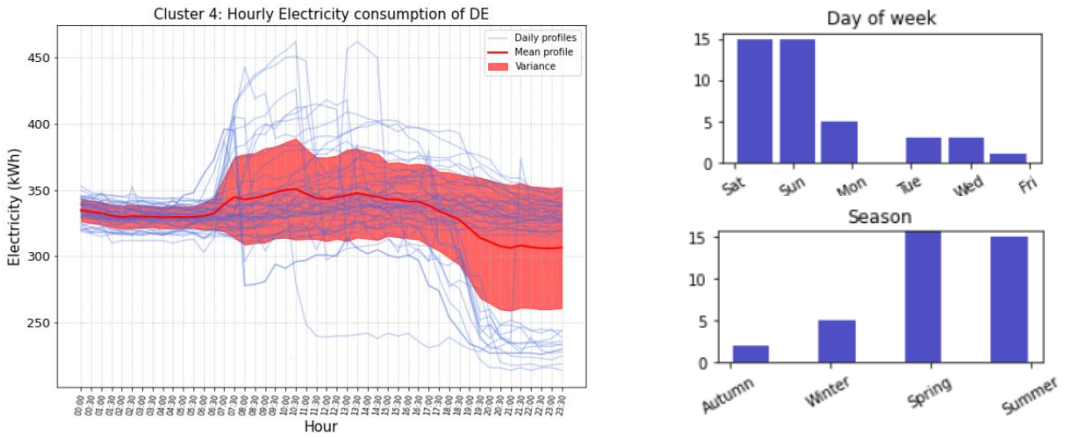
Fig. 33: Clusters along with selected exogenous variables distributions for HH cooling

DE electricity

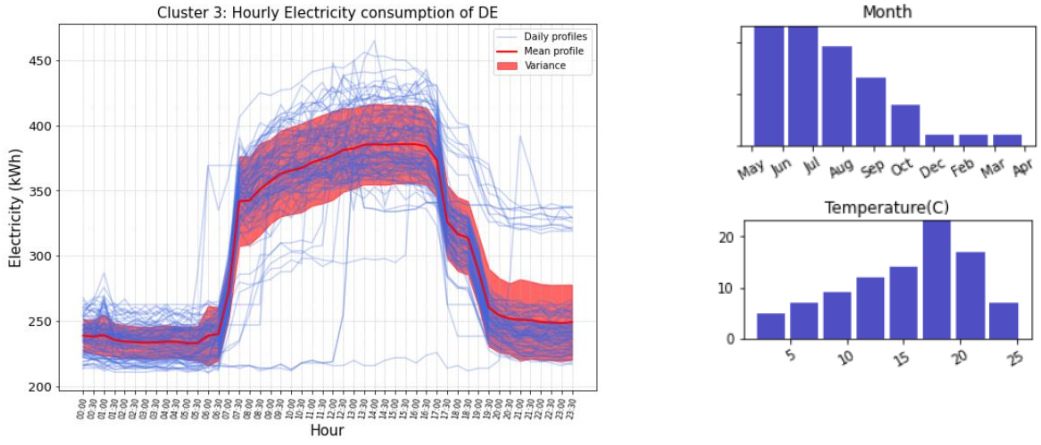
Contrary to HH's electricity patterns, in Fig. 34a we observe that for a small number of days during the hotter months, particularly during spring and summer, daily profiles reach the highest consumption during the whole day, even during the non-working hours, perhaps because of some cooling operations still running at night, while cluster 4 reveals that during that seasons, on weekends, there is also a higher electricity consumption than the rest of the weekends. This probably implies the running of additional operations during that period, considering that only 9% of total samples belong to Cluster 1 and 7% to Cluster 4, while normally, during hot periods, cluster 3 (Fig. 34 c) represents the working days' consumption (16% samples), where it is obvious that non-working hours have less consumption rather than the samples of Cluster 1.



(a)



(b)



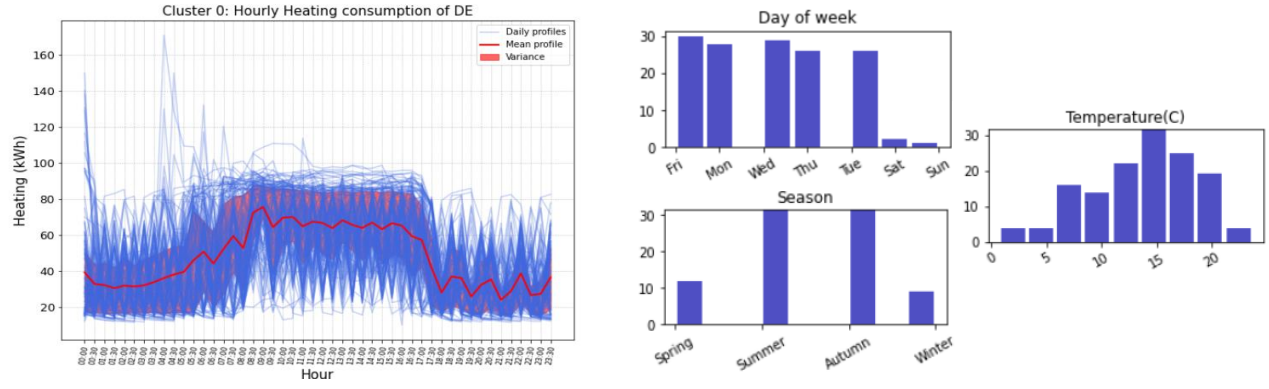
(c)

Fig. 34: Clusters along with selected exogenous variables distributions for DE electricity

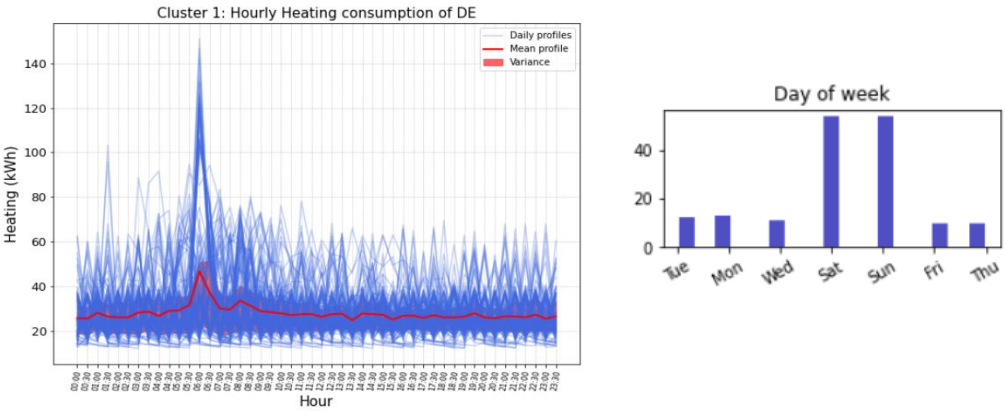
DE heating

Although cluster 2, as expected, follows a high consumption pattern, as Fig. 35 reveals, given that it mostly consists of the cold days, we observe an extended hourly schedule (03:30 – 18:30)

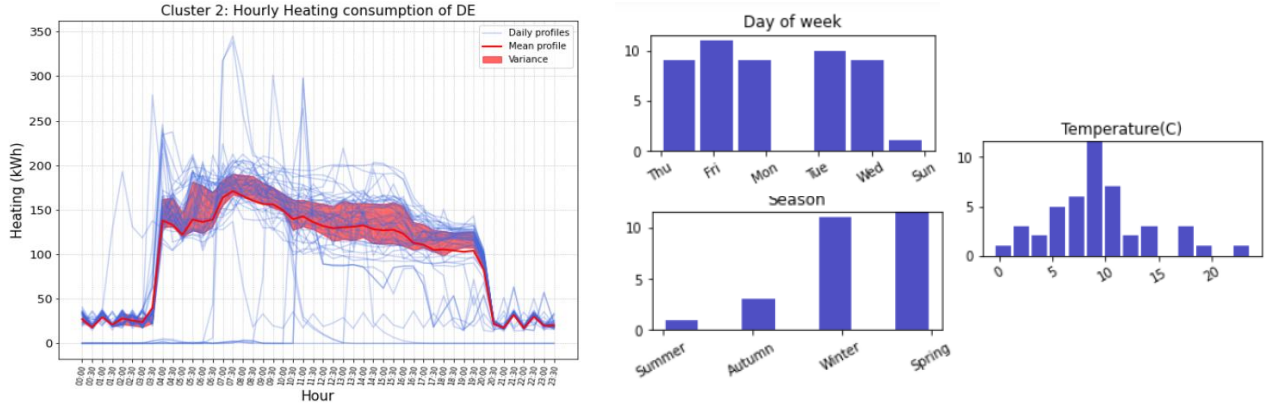
than the usual working schedule, during which heating is turned on. In cluster 1, which represents most of the weekends we observe a peak at early morning between 05:30 and 07:00, while Cluster 0 reveals that heaters are also turned on during hotter days, however with continuous fluctuations implying some inconsistency on system's operation either caused by the user or by poor configuration.



(a)



(b)



(c)

Fig. 35: Clusters along with selected exogenous variables distributions for DE heating

3.4 Post-mining

The knowledge discovered by DM techniques can be used for a variety of purposes, including energy consumption forecasting, diagnostic analysis, preventative maintenance and so forth. In fact, the mining of energy load patterns can be also viewed as a powerful pre-processing phase for the implementation of post-mining models regarding the building energy consumption [8]. In this work particularly, we investigate the abnormal energy behavior of Hursley site's buildings. After having defined both buildings' energy consumption profiles, we are to carry out a subsequent step of statistical analysis in order to detect any abnormal daily profile.

3.4.1 Anomaly Detection

Particularly, after having identified the potential energy consumption patterns for each feature, we detect any abnormal energy profile that is significantly different from the overall pattern of the cluster it has been grouped to, through the proposed clustering analysis. In order to define the condition under which a daily profile is considered anomalous, inspired by Seem [26] who leveraged the modified z-score to detect abnormal profiles, we are going to implement a similar anomaly detection method that is based on each cluster's outlier identification. Precisely, possible unusual energy consumption profiles are identified within each cluster by determining the amount of variation from normal, by using robust estimates of the median and the interquartile range (IQR). Therefore, a potential abnormal event would deviate from the median of the cluster it belongs to, more than the Maximum value of the cluster's interquartile (IQR) range, in terms of the DTW distance. IQR method was selected instead of the z-score method given that for all features, each cluster's distribution is mostly right skewed; therefore, it would be troublesome for accurately identifying all the outliers by using the z-score method considering that the mean and standard deviation are highly affected by outliers. Hence, similarly to [26] who leveraged the efficiency of median when data is skewed and proposed a modified z-score based on median, instead of mean measures, we proposed the use of IQR, considering that median and interquartile range are more robust measures of central tendency and dispersion, respectively, when the dataset does not follow a normal distribution. The right (upper values) skewedness of HH' electricity clusters is obvious in Fig. 36, while Appendix H illustrates the boxplots of all the other features' clusters of our dataset.

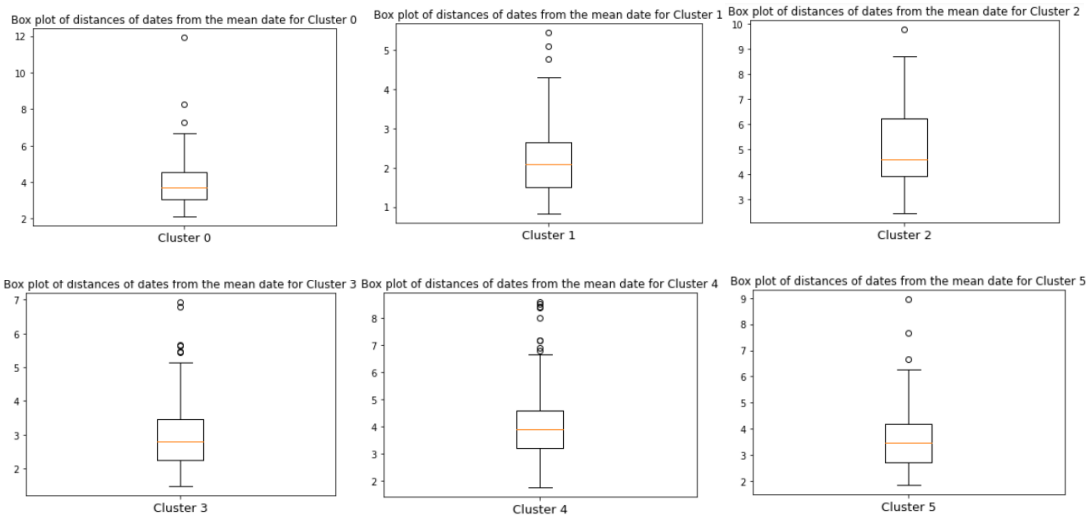


Fig. 36: Right skewed boxplots of HH electricity's clusters

In order to boost the performance of the proposed approach for identifying the anomalous days, we introduce an additional step, during which, for the outlier days detected we check if the underlying day should be grouped into a different cluster. Specifically, we achieve this by comparing the outlier's DTW distance from the mean of its cluster with the DTW distances from the mean of all the other clusters. If the shorter distance belongs to a different cluster, then this day is most probably not an outlier and has been mistakenly classified as an outlier, while it should belong to the cluster that the distance from its mean is the shortest one. The following flow chart in Fig. 37 explains the proposed robust method for the anomalous profiles detection and how the model determines if the detected outlier is an actual anomaly or a misclassified sample.

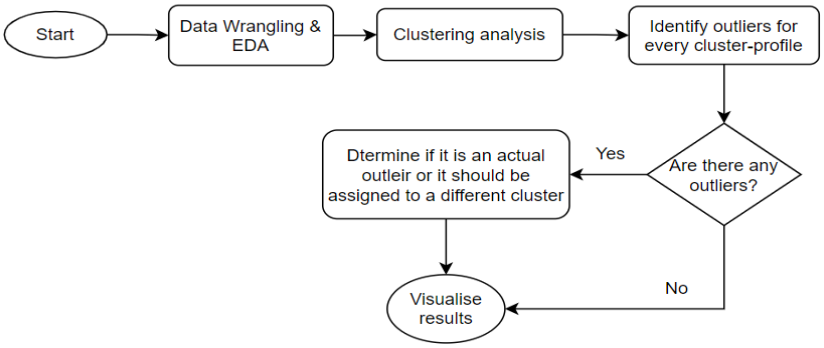


Fig. 37: Flow chart for detecting abnormal energy daily profiles

One potential application of this proposed method is to provide building's operators with a visual interpretation of the whole day's energy consumption, by revealing which should be the actual consumption profile, along with the influential exogenous factors of this day that probably caused the underlying anomaly, with an outer view to confront this behavior and prevent similar anomalies in the future.

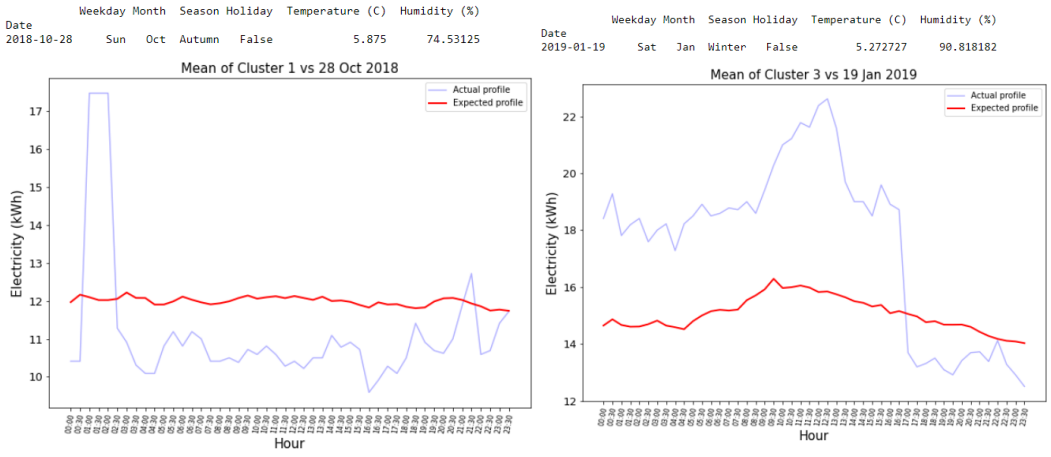
Following, we present some distinct anomalous daily profiles identified through the proposed method of Fig. 37, for each energy feature, while Table 8 shows the dates when more than one energy features have been detected as anomalous, implying that a general event took place during that day, affecting either the whole building ore the whole industrial site, such as a power cut. The process of finding the common anomalous loads was implemented by simply comparing the lists of anomalous profiles of each energy feature.

	HH			DE		Event
Dates	**Electricity**	**Heating**	**Cooling**	**Electricity**	**Heating**	**-**
2018-07-24	✓			✓		Missing values 05:30-14:30
2018-10-28	✓			✓		Peak during 00:30-02:30 (Sunday)
2019-11-04	✓			✓		Missing values 00:00-09:30
2019-10-27	✓			✓		Peak during 00:30-02:30 (Sunday)

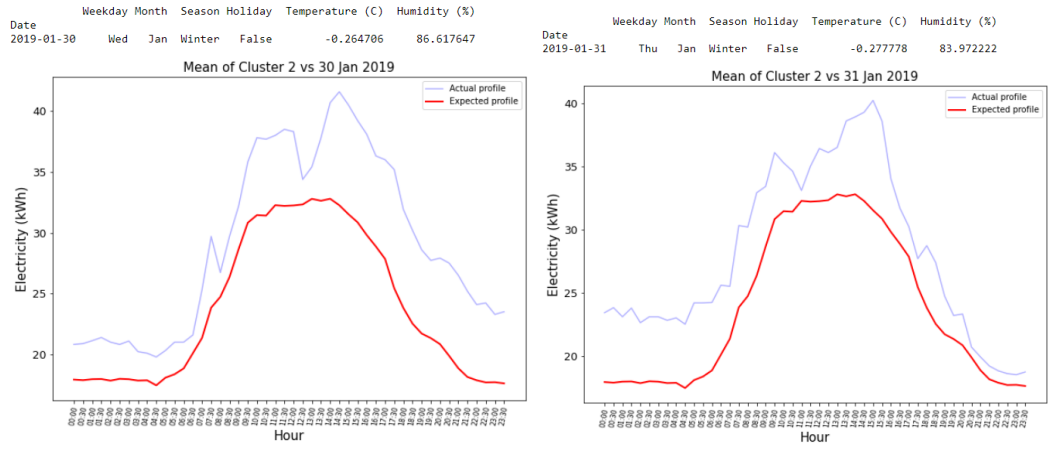
Table 8: Common anomalous daily profiles across different features

HH electricity

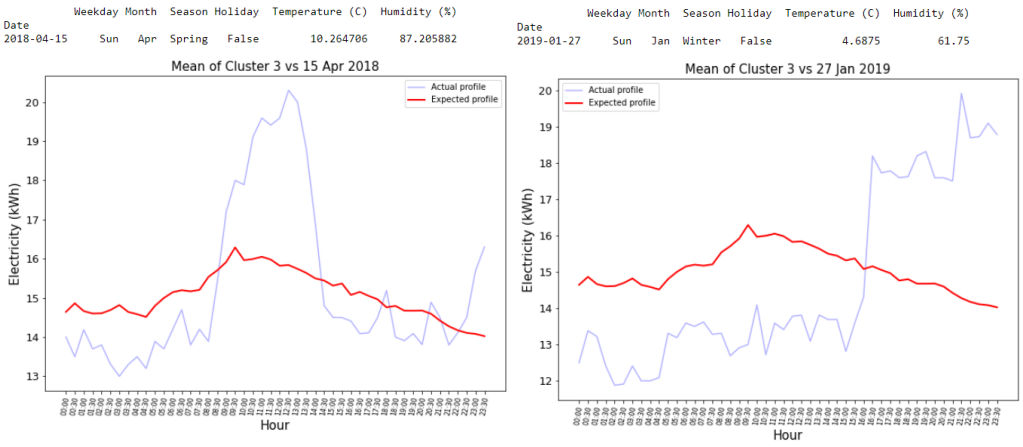
Fig. 38 depicts some indicative anomalous daily profiles of HH electricity, identified through the proposed anomaly detection method. Particularly, in Fig. 38a, we observe an unexpected peak between 12:30 and 02:30 on Sunday, which most probably would have been a system's malfunction or a kind of system's scheduled restart, while Fig. 38c, d reveal that, besides a higher electrical consumption that has taken place during the working hours due to extreme low temperatures (<0°C), an unexpected high consumption also was recorded during the night which indicates that some electrical systems in the building were still operating from 20:00 pm to 05:00 am next day, when the building should be unoccupied or partly occupied. Similarly, according to Fig. 38b, an overnight high electrical consumption is observed which could also be corroborated by early morning cleaning operations, besides an even higher consumption pattern between 08:30 and 16:00 which implies the presence of a Saturday event, such as a community event. A community event must have taken place on Sunday 15 April 2018, whose pattern is illustrated in Fig. 38e, while Fig. 38f implies a building's operational event which kicked off at 16:30 and lasted until late night.



(a) (b)



(c) (d)

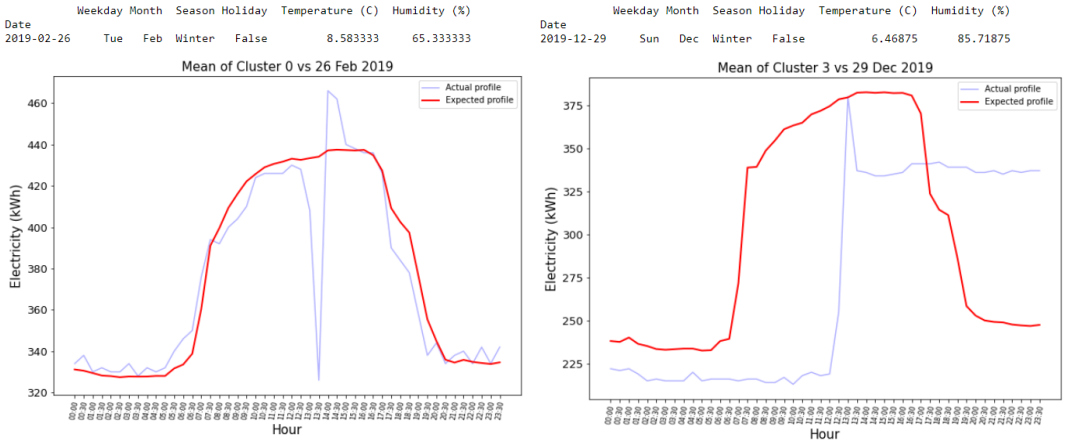


(e) (f)

Fig. 38: Anomalous profiles vs expected patterns for HH electricity consumption

DE electricity

Regarding the DE building's electricity consumption, an anomalous consumption that plunges, is perceived between 13:00 p.m. and 14:00 p.m. in Fig. 39a, implying a meter's failure, while in Fig. 39b, e, f, we can see there is an increased late-night consumption, most likely because some electrical systems were left turned on. Fig. 39c, d, which represent the electricity load consumption profiles of two Bank holidays, the Late summer Bank Holiday and the Good Friday respectively, seem to follow a low consumption working schedule, implying that mainainers forgot to reschedule the control system and therefore, some basic operations were running during the usual working hours. The same anomalous pattern is observed for all the Bank Holidays that are on week days, which are particularly the Late Summer Bank Holiday , the Good Friday, the Easter Monday, the Christmas day and the Boxing Day for the given 2-year time period.



(a) (b)

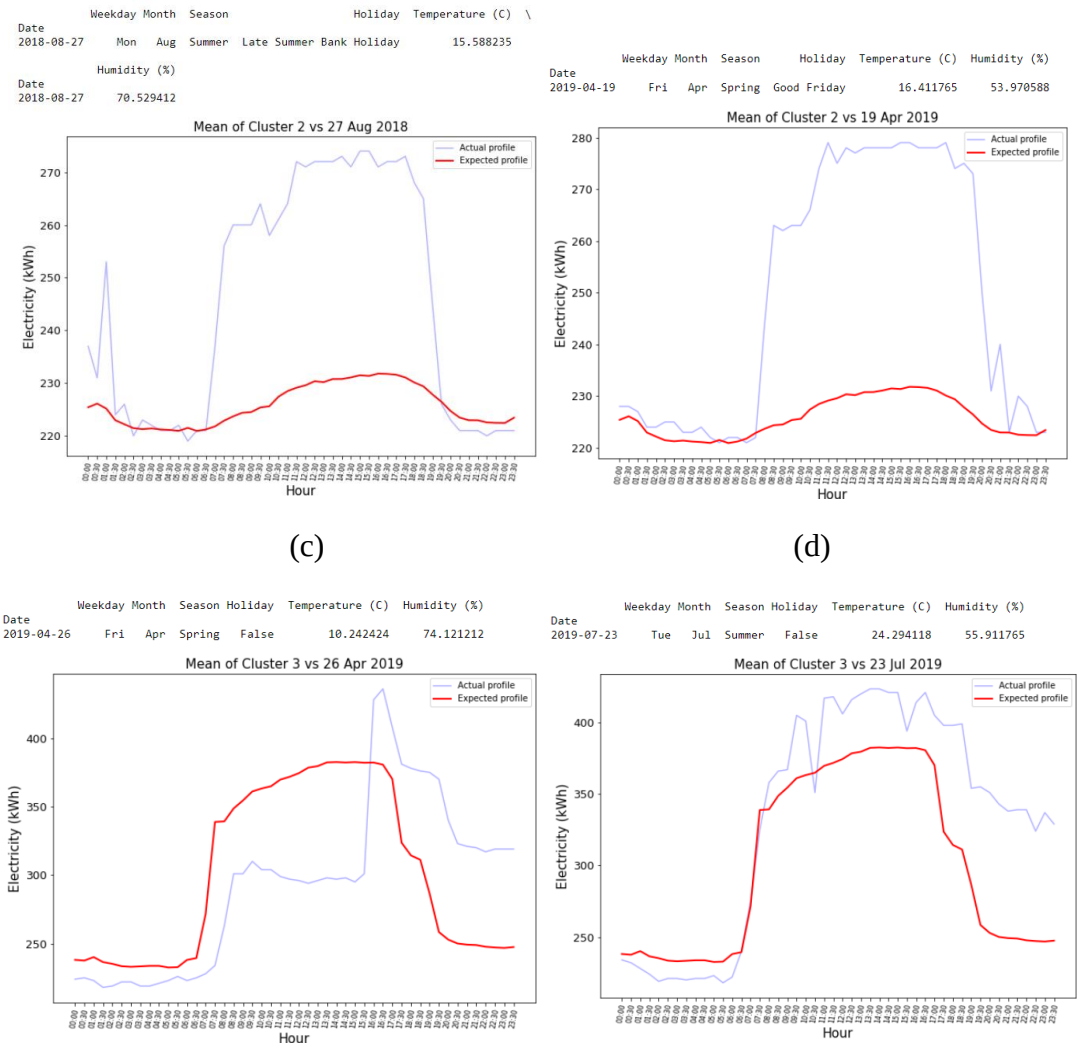


Fig. 39: Anomalous profiles vs expected patterns for DE electricity consumption

HH heating, HH cooling, DE heating

As previously discussed, heating and cooling consumption follow erratic patterns which, even though form some distinct clusters, they are not compact and perform many random fluctuations during the day, irrespective of the working schedule. Either these fluctuations are random or caused by external factors, it is not plausible to identify anomalies in heating and cooling load patterns given that the patterns they belong to, are not accurate. When there is not a particular and compact pattern that should be followed, then there is not a robust condition as well, to determine what is an outlier and what is a normal profile. Fig. 40 shows some distinct heating anomalous profiles where abrupt peaks take place during late hours, while in Fig. 41b a cooling anomalous Saturday profile is observed with 3 peaks, probably implying a human occupancy during those hours.

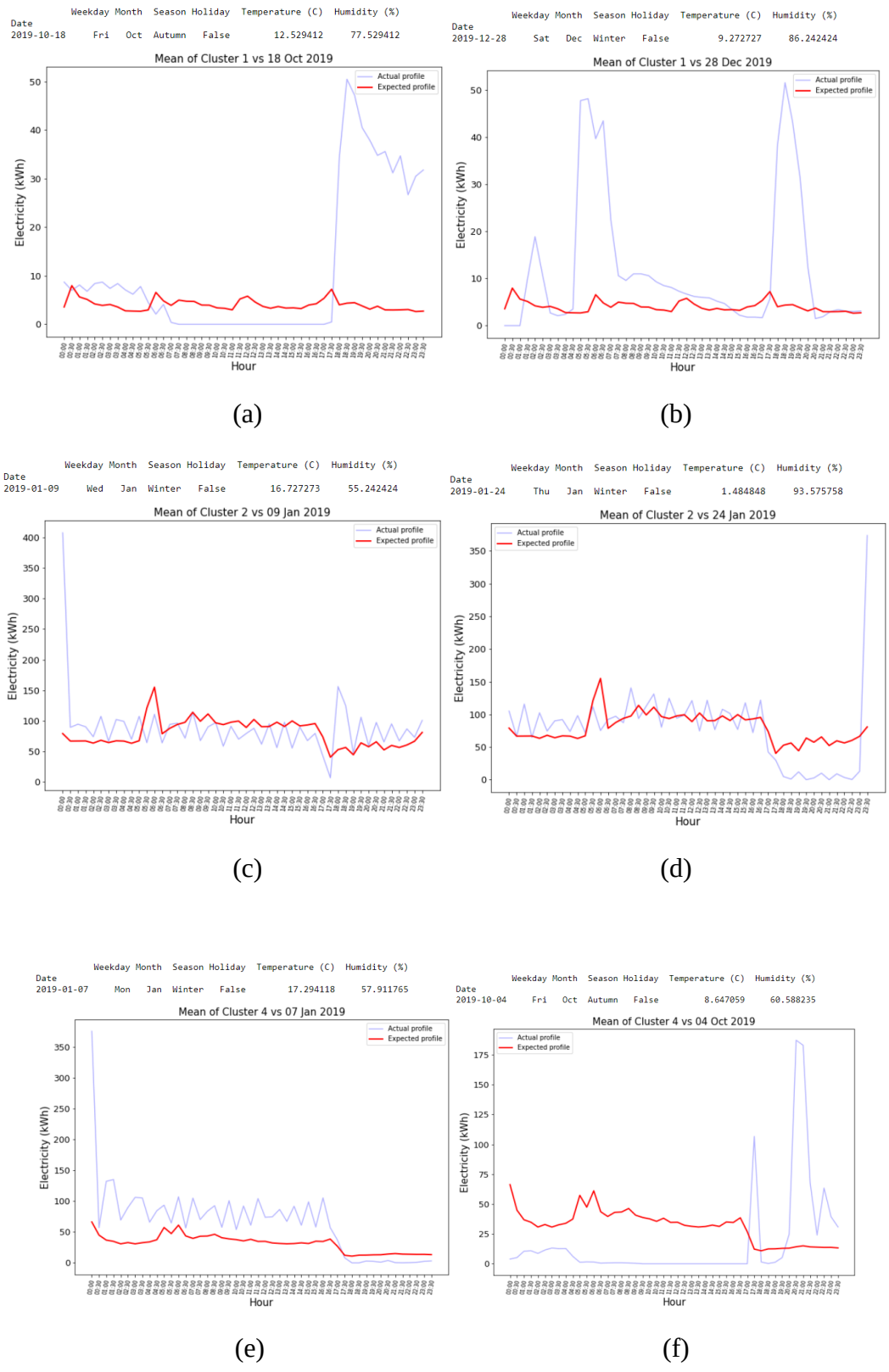
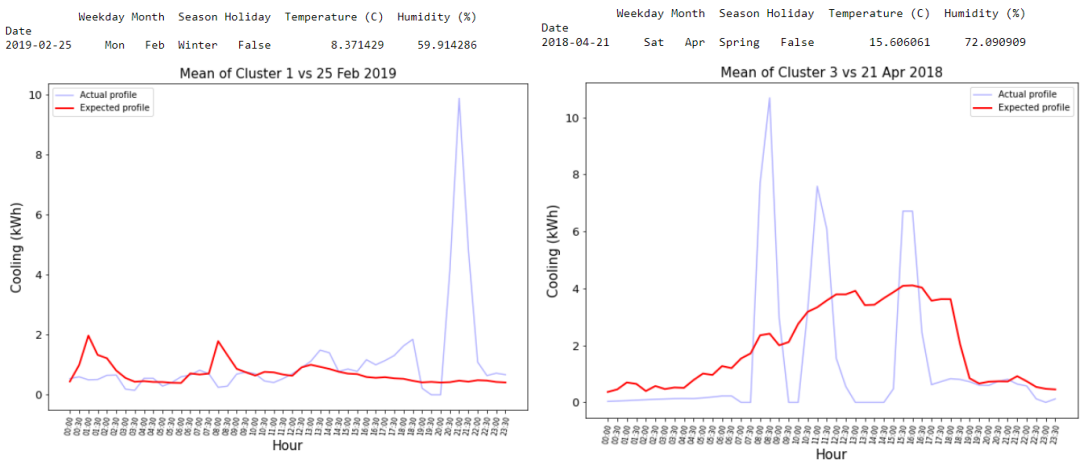


Fig. 40: Anomalous profiles vs expected patterns for HH heating consumption



(a) (b)

Fig. 41: Anomalous profiles vs expected patterns for HH cooling consumption

Overall, the proposed anomaly detection method efficiently identified the anomalous electricity consumption profiles for both buildings, considering that electricity consumption forms distinct and compact profile clusters, as it was previously proved, and therefore, the profiles that deviated from the normal behavior were seamlessly identified as abnormal. After idenitifying the anomalpous profiles, along with the information of the exogenous factors, an investigation into the abnormal consumption can reveal control problems that were caused either by system's failures or misconfigured schedules, besides the anomalies that were caused due to building operational changes, such as individuals working overtime or weather changes and subsequently inform the building's operators about taking appropriate and preventative measures for better management in the future. However, in order to achieve even more efficient measures, domain knowledge is a prerequisite since it can indicate the circumstances under which an anomalous profile was caused and reject the hypothesis of being anomalous. For example, the electricity consumption on a day might not actually be anomalous because the energy requirements were low during that period of the year.

Nevertheless, the heating and cooling consumption profiles are not easily handled for identifying anomalies, at least within the context of these buildings' measurements, given their vague behavior that cannot be framed into robust and compact clusters. Therefore, many daily profiles can be wrongly classified as anomalous. Further information regarding the influential factors that affect heating and cooling systems' behavior, along with further domain knowledge would probably improve both the identification of consumption patterns and the detection of anomalous profiles.

4. Conclusions

The need of reducing and optimizing the consumption of energy from building operations has given rise to the incorporation of sophisticated Building Management Systems into more and more buildings, enabling them to become more efficient, reduce costs and emissions and become more transparent in terms of operation. The availability and accurate analysis of the data generated by such buildings through powerful and efficient tools coming from Data analytics and Machine Learning realm, is the key enabling factor for achieving such gains. Data analytics-based BMS can provide a better knowledge of building's energy behavior, allowing a better management of the energy demand on one side, and make it possible to develop reliable forecasting and automatic anomaly detection systems to ensure the optimization of energy consumption, on the other side [15].

This work proposed a Data Wrangling-based framework for analyzing the building data provided by the industrial partner, IBM, for its research and development laboratory, located in Hursley Site, with an outer two-fold purpose; first, an automate method for handling unruly time-series components was proposed that significantly improved the ability to obtain energy load patterns along with interpretable knowledge regarding the energy behavior and some dynamic influencing exogenous factors; and second, a diagnostic phase was developed in order to detect abnormal daily energy consumption profiles, which could be results of inefficient operating strategies, equipment failures, or errors in sensing and transmission [27]. The outer contribution of this study has been the development of a highly interpretable and reasonable framework for automatically converting poor quality building energy data into actionable knowledge regarding the building's behavior and inconsistencies, without requiring the human interference. This work also proved that accurate and interpretable, rather than complicated and computationally consuming analysis of energy data generated by such buildings is the key enabler for achieving trustfully such profits.

Regarding the first fold of project's contribution, initially, we proposed a robust and automate Data Wrangling process for cleaning and preparing the raw unprocessed building energy data, by handling any inconsistency generated by malfunctions of the metering system, by means of statistical and DM analysis, without needing any human manual interference as most of the conventional energy data cleaning processes in literature require. The main contribution of this cleaning process is that it efficiently differentiated the point anomalies from the contextual ones and handled them without distorting any feature's distribution, therefore, eliminating any chance of acquiring a biased outcome from the upcoming knowledge extraction processes, given that they are fed with clean and neat data. Following, we performed a highly interpretable clustering analysis, after comparing two widespread methods for handling time-series data. The comparison proved that feature extraction techniques for reducing data dimensionality, do not always outperform the clustering on raw time-series, implying either that data was not enough to allow the learning of robust representations or that the feature extraction technique tested was not robust enough to capture important information from the whole time series. The proposed technique, however, which was the k-means method with the DTW distance metric, discovered meaningful compact clusters based not only on the magnitude of the energy load, but the trend, as well. In order to discover insightful knowledge on the clusters identified and improve the interpretability of clustering analysis in conjunction with some exogenous influential factors, such as weather, day, season and so forth, the CART algorithm was tested but achieved poor results, implying that probably, the different profiles identified are affected by other exogenous parameters that we are not aware of. Therefore, we provided a visual interpretation of the exogenous factors by plotting their distribution for every cluster's samples.

Subsequently, in the proposed framework, a diagnosis phase was conceived for identifying unusual or else abnormal energy consumption profiles. Particularly, according to the proposed method, potential abnormal daily events were compared in terms of DTW distance with the mean of the cluster they belong to, according to the previous stage, with the threshold for considering a profile as anomalous, being determined by the Maximum value of the cluster's interquartile range. In order to achieve robust estimates of anomalous events, we introduced an additional final step where for each potential abnormal profile, a comparison between the DTW distances from every cluster's mean was calculated and only if the shortest distance was that from its own cluster's mean, the profile was actually considered as an outlier. This additional step minimized the false negative abnormal events; however, only regarding the electricity consumption, since the heating and cooling loads have been quite inconsistent and difficult to frame compact clusters, which hindered the comparison process for identifying anomalies. The pattern recognition analysis showed various anomalous events, from high late-night or early-morning electricity consumption patterns and high energy consumption loads during unoccupied hours, to power cuts, bad scheduling and higher consumption due to extreme weather conditions. This proposed method is highly interpretable, and therefore trustful, providing that the whole time series is displayed along with the expected mean behavior and the exogenous factors that may influence that event, which is pivotal especially when domain knowledge is missing.

Ultimately, this framework proposed in this work provides a generic and automate solution for identifying the energy usage patterns of a building and additionally uncovering hidden knowledge regarding the various energy profiles, improving the interpretability of clustering analysis, besides detecting any abnormal load profile. Such analysis and anomaly detection tool can support building managers or homeowners, by providing them with interpretable, previously hidden, knowledge and giving back to them important indicators of excessive energy usage during unsuspected time periods in order to understand building's operation conditions, perform energy performance assessments and reduce the operating cost and time of their operational systems. Such a robust and interpretable framework, in conjunction with a highly accurate energy forecasting model, could identify in advance or in real-time anomalous patterns and hence lead to important energy savings.

5. Future work

Future investigations could fruitfully further explore this issue under investigation. First and foremost, as it is recommended, in almost every research work, future studies should aim to replicate the results discovered either in a larger dataset of the same case study or in other similar datasets, for benchmarking purposes. The proposed framework was tested in a relatively small dataset for the scope of the clustering analysis; however, in order to efficiently discover all the identified groups that define a building's energy consumption profile, a representative number of samples for each energy consumption group is required.

Regarding the clustering analysis, the proposed algorithm cannot be considered always as the best solution for the task addressed in this study, but considering the dataset given to us, k-means with DTW has been proved as the best performing between the clustering methods under examination. Another approach that could be examined for the given dataset, is the Gaussian Mixture Models (GMM). GMM is a probabilistic approach used for clustering analysis, which is based on the assumption that all the data points are generated from a mixture of some Gaussian distributions, where each of them represents a cluster, while contrary to k-means it provides us with the probabilities that a given point belongs to each of the possible clusters. Contrary to k-means which only considers the mean to update the centroid, GMM takes into account the mean as well as the variance of the data, therefore GMM is worth being leveraged for capturing non-spherical based clusters, as happens in our case study.

Additionally, some further efforts could be placed towards the developing of more effective feature extraction methods than the one presented in that work, in order to perform a clustering analysis on the new dimensionally reduced dataset. Particularly, autoencoders is a type of neural network (NN) that can be used to learn a compressed representation of a raw signal or else the latent space of the autoencoder. The latent space of autoencoders can be a very robust representation of raw time series, since the decoder attempts to reconstruct the input from the compressed signal provided by the encoder. Instead of manually defining new features for a raw time series based on knowledge extracted through EDA and jeopardizing neglecting important information, autoencoders can be used in order to learn a very robust representation of the signal through sophisticated neural networks. These neural networks, are trained by minimizing the loss (usually the mean squared error or the cross-entropy loss) between the initial signal and the reconstructed and hence the lower the total loss is, the more accurate and representative is the latent space. The network's architecture can be anything, such as Convolutional Neural Networks (CNN), Long short-term memory (LSTM) and many others. 1D Convolutional architecture has been proved to better preserve the information embedded in temporal building energy data into the latent space [51]. Therefore, by projecting the initial data into a lower-dimensional representation, we can use this latent space as the new features extracted from the initial ones and perform any following clustering on the extracted low-dimensional data. Even though the development of robust neural networks was not within the scope of this work, considering that NNs work as black boxes and therefore the interpretability would be missing if used in this study, we performed some experiments on autoencoders using CNNs only to show that they underperformed the other two techniques that we presented in Section 3. We could give a possible explanation that, in general, NNs require a quite large dataset to efficiently be trained, given that NNs consist of thousands of weighting parameters that need to be learnt and in order to avoid overfitting, the number of samples must be large. However future investigations are necessary to validate these assumptions.

Future works will also be aimed at enhancing the knowledge discovery procedure by employing a more efficient, than CART, algorithm to explain the relations between the

dynamic influencing exogenous factors to energy consumption. Nevertheless, CART is a classification and regression algorithm based on regression trees which can provide high interpretability on the relations between the predictors and the response value. Hence, providing that literature review has proved its efficiency and robustness on energy-related data, a proposing future work would be to acquire more influencing exogenous factors related to energy, such as building's occupancy or specific features of the building itself, besides having many more samples for each cluster identified.

Moreover, following the knowledge discovery and identification of anomalous daily energy patterns, future work should be devoted to the development of a model that will process incoming data and alert the users to potential issues in the system in real-time. This study's anomaly detection was mainly approached as a diagnostic phase; however, detecting, in advance, anomalous patterns leads to a significant reduction of energy waste during building operation. Therefore, more efforts should be made on accurately predicting the energy consumption, rather than the expected pattern, by leveraging sophisticated models that take into account not only influential exogenous factors but historical data as well, as literature review proposes [59]. After predicting the energy consumption for a given future period, then by comparing the expected energy consumption with the actual one, we could identify any anomalous values according to the degree of deviation between predicted and actual value. On top of that, given that domain knowledge is mostly missing, considering that a bigger dataset could be provided, autoencoders would be a quite robust approach for detecting anomalies, while the pre-processing step of cleaning could also be neglected since the inconsistent data points would be undermined by the network. Similarly, the reconstructed signal form the autoencoders can be compared with the actual values and determine if there is an inconsistency or not. Also, the subsequence's length does not need to be necessarily equal to the length of a day's duration. Another approach that could be adopted, especially when data is not enough, is to perform a data augmentation technique where the number of samples is increased by defining a rolling window of fixed length k and therefore, instead of having N daily samples, acquire Nxk samples in total. This approach could identify atypical subsequences that extend during two or more dates and therefore reveal long term anomalies that the previous approached could not do so.

Therefore, this work provides a good and highly interpretable initial approach to the problem under investigation; however, future work is certainly required to investigate the proposed and not only, future approaches.

6. References

[1] A. González-Vidal, V. Moreno-Cano, F. Terroso-Sáenz and A. Skarmeta, “Towards Energy Efficiency Smart Buildings Models Based on Intelligent Data Analytics”, *Procedia Computer Science*, vol. 83, pp. 994-999, Jan. 2016.

[2] C. L. Stimmel, “Big Data Analytics Strategies for the Smart Grid”, *Auerbach Publications*, Jan 2019.

[3] L. Pérez-Lombard, J. Ortiz and C. Pout, “A review on buildings energy consumption information”, *Energy and Buildings*, vol. 40, no. 3, pp. 394-398, Jan. 2008.

[4] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali and A. Amira, “Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives”, *Applied Energy*, vol. 287, p. 116601, Apr. 2021.

[5] J. Serra, D. Pubill, A. Antonopoulos and C. Verikoukis, “Smart HVAC Control in IoT: Energy Consumption Minimization with User Comfort Constraints”, *The Scientific World Journal*, vol. 2014, pp. 1-11, Jun. 2014.

[6] M. Fayaz and D. Kim, “Energy Consumption Optimization and User Comfort Management in Residential Buildings Using a Bat Algorithm and Fuzzy Logic”, *Energies*, vol. 11, no. 1, p. 161, Jan. 2018.

[7] S. Pawar, B. F. Momin, “Smart electricity meter data analytics: A brief review*”, In 2017 IEEE Region 10 Symposium (TENSYMP)*, IEEE 2017, pp. 1-5, Jul. 2017.

[8] K. Zhou, C. Fu and S. Yang, “Big data driven smart energy management: From big data to big insights”, *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 215-225, Apr. 2016.

[9] R. Silipo, P. Winters, “Big data, smart energy, and predictive analytics.”, *Time Series Prediction of Smart Energy Data*, vol. 2, pp. 37, 2013.

[10] B. Bhattarai et al., “Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions”, *IET Smart Grid*, vol. 2, no. 2, pp. 141-154, Feb. 2019.

[11] F. Endel and H. Piringer, "Data Wrangling: Making data useful again", *IFAC-PapersOnLine*, vol. 48, no. 1, pp. 111-112, May 2015.

[12] C. Zhang, L. Cao and A. Romagnoli, “On the feature engineering of building energy data mining”, *Sustainable Cities and Society*, vol. 39, pp. 508-518, May 2018.

[13] Q. Li, Q. Meng, J. Cai, H. Yoshino and A. Mochida, “Applying support vector machine to predict hourly cooling load in the building”, *Applied Energy*, vol. 86, no. 10, pp. 2249-2256, Oct. 2009.

[14] J. Rodger, “A fuzzy nearest neighbor neural network statistical model for predicting demand for natural gas and energy cost savings in public buildings”, *Expert Systems with Applications*, vol. 41, no. 4, pp. 1813-1829, Mar. 2014.

[15] M. Piscitelli, S. Brandi, A. Capozzoli and F. Xiao, “A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings”, *Building Simulation*, vol. 14, no. 1, pp. 131-147, May 2020.

[16] V. Chandola, A. Banerjee and V. Kumar, “Anomaly detection”, *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, Jul 2009.

[17] S. Wu and J. Sun, “Cross-level fault detection and diagnosis of building HVAC systems”, *Building and Environment*, vol. 46, no. 8, pp. 1558-1566, 2011.

[18] I. Khan, A. Capozzoli, S. Corgnati and T. Cerquitelli, “Fault Detection Analysis of Building Energy Consumption Using Data Mining Techniques”, *Energy Procedia*, vol. 42, pp. 557-566, Jan 2013.

[19] U. Habib, G. Zucker, M. Blochle, F. Judex and J. Haase, “Outliers detection method using clustering in buildings data”, In *IECON 2015-41st Annual Conference of the IEEE Industrial Electronics Society*, pp. 694-700, IEEE, Nov. 2015.

[20] U. Habib, G. Zucker, M. Blochle, A. Wendt, S. Schaat, L. C. Siafara, “Building energy management and data analytics”, In *2015 international symposium on smart electric distribution systems and technologies (EDST)*, pp. 462-467, IEEE, Sep 2015.

[21] A. Chong, K. P. Lam, W. Xu, O. T. Karaguzel, Y. Mo, “Imputation of missing values in building sensor data”, *ASHRAE and IBPSA-USA SimBuild*, 6, pp. 407-14, Aug 2016.

[22] A. Capozzoli, G. Serale, M. Piscitelli and D. Grassi, "Data mining for energy analysis of a large data set of flats", *Proceedings of the Institution of Civil Engineers - Engineering Sustainability*, vol. 170, no. 1, pp. 3-18, 2017.

[23] A. Capozzoli, M. Piscitelli and S. Brandi, "Mining typical load profiles in buildings to support energy management in the smart city context", *Energy Procedia*, vol. 134, pp. 865-874, Oct 2017.

[24] H. Tang and F. 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*, vol. 231, p. 110601, Jan 2021.

[25] A. Capozzoli, M. Piscitelli, S. Brandi, D. Grassi and G. Chicco, "Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings", *Energy*, vol. 157, pp. 336-352, Aug 2018.

[26] J. Seem, "Using intelligent data analysis to detect abnormal energy consumption in buildings", *Energy and Buildings*, vol. 39, no. 1, pp. 52-58, Jan 2007.

[27] C. Fan, F. Xiao, Y. Zhao and J. Wang, "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data", *Applied Energy*, vol. 211, pp. 1123-1135, Feb 2018.

[28] J. Yang et al., "k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement", *Energy and Buildings*, vol. 146, pp. 27-37, Jul 2017.

[29] 10 breakthrough technologies, *MIT Technology Review*, Massachusetts Institute of Technology, Cambridge, MA, USA, Jan/Feb 2001.

[30] C. Fan, F. Xiao and C. Yan, "A framework for knowledge discovery in massive building automation data and its application in building diagnostics", *Automation in Construction*, vol. 50, pp. 81-90, Feb 2015.

[31] M. Köppen, "The curse of dimensionality*", In5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, vol. 1, pp. 4-8, Sep 2000.

[32] T. Räsänen, M. Kolehmainen, "Feature-Based Clustering for Electricity Use Time Feature-Based Clustering for Electricity Use", *In: Proceedings of international conference on adaptive and natural computing algorithms (LNCS 5495)*, Berlin, Germany: Springer-Verlag, pp. 401–412. 2009.

[33] S. Haben, C. Singleton and P. Grindrod, "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data", *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 136-144, 2016.

[34] X. Serrano-Guerrero, G. Escrivá-Escrivá, S. Luna-Romero and J. Clairand, "A Time-Series Treatment Method to Obtain Electrical Consumption Patterns for Anomalies Detection Improvement in Electrical Consumption Profiles", *Energies*, vol. 13, no. 5, p. 1046, Jan 2020.

[35] C. Fan, F. Xiao, Y. Zhao and J. Wang, "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data", *Applied Energy*, vol. 211, pp. 1123-1135, 2018.

[36] J. Chou and A. Telaga, "Real-time detection of anomalous power consumption", *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 400-411, May 2014.

[37] W. Hurst, C. Montañez and N. Shone, "Time-Pattern Profiling from Smart Meter Data to Detect Outliers in Energy Consumption", *IoT*, vol. 1, no. 1, pp. 92-108, Sep 2020.

[38] H. Shareef, M. Ahmed, A. Mohamed and E. Al Hassan, "Review on Home Energy Management System Considering Demand Responses, Smart Technologies, and Intelligent Controllers", *IEEE Access*, vol. 6, pp. 24498-24509, 2018.

[39] S. Aghabozorgi, A. Seyed Shirkhorshidi and T. Ying Wah, "Time-series clustering – A decade review", *Information Systems*, vol. 53, pp. 16-38, 2015.

[40] C. do Carmo and T. Christensen, "Cluster analysis of residential heat load profiles and the role of technical and household characteristics", *Energy and Buildings*, vol. 125, pp. 171-180, 2016.

[41] A. Sardá-Espinosa, "Time-Series Clustering in R Using the dtwclust Package", *The R Journal*, vol. 11, no. 1, p. 22, 2019.

[42] M. Fernandes, J. Viegas, S. Vieira and J. Sousa, “Analysis of residential natural gas consumers using fuzzy C-means clustering”, *Proc 2016 IEEE Int Conf Fuzzy Syst* , 2016.

[43] J. Han, J. Pei, M. Kamber, “Data mining: concepts and techniques”, Elsevier, Jun 2011.

[44] G. Buzzi-Ferraris and F. Manenti, "Outlier detection in large data sets", *Computers & Chemical Engineering*, vol. 35, no. 2, pp. 388-390, 2011.

[45] M. Bouguessa, “A probabilistic combination approach to improve outlier detection”, *In 2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, vol 1, pp 666–673, 2012.

[46] J. Yang, S. Rahardja, P. Fränti, “Outlier detection: how to threshold outlier scores”, *InProceedings of the international conference on artificial intelligence, information processing and cloud computing*, pp. 1-6, Dec 2019.

[47] R. Little, "Regression with Missing X's: A Review", *Journal of the American Statistical Association*, vol. 87, no. 420, p. 1227, 1992.

[48] R. Little and D. Rubin, "The Analysis of Social Science Data with Missing Values", *Sociological Methods & Research*, vol. 18, no. 2-3, pp. 292-326, 1989.

[49] D. Stekhoven and P. Buhlmann, "MissForest--non-parametric missing value imputation for mixed-type data", *Bioinformatics*, vol. 28, no. 1, pp. 112-118, Jan 2011.

[50] P. Baldi, “Autoencoders, unsupervised learning, and deep architectures”, *JMLR Workshop Conf Proc*, pp. 37–50, 2012.

[51] C. Fan, F. Xiao, Y. Zhao and J. Wang, "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data", *Applied Energy*, vol. 211, pp. 1123-1135, Feb 2018.

[52] B. Juang and L. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 9, pp. 1639-1641, 1990.

[53] M. Ester, H.P. Kriegel, J. Sander, X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise”, *proceedings of the 2nd international conference on knowledge discovery and data mining*, 226-231.

[54] F. Iglesias and W. Kastner, "Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns", *Energies*, vol. 6, no. 2, pp. 579-597, Feb 2013.

[55] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.

[56] G. Zervas, S. Ruger, “The Curse of Dimensionality and Document Clustering*”, In Proceedings of 1999 IEE Colloquium on Microengineering in Optics and Optoelectronics*, London, UK, pp. 19:1–19:3, Nov 1999.

[57] S. Zhang, C. Zhang and Q. Yang, "Data preparation for data mining", *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 375-381, 2003.

[58] M. Hahsler, M. Piekenbrock and D. Doran, "dbscan: Fast Density-Based Clustering with R", *Journal of Statistical Software*, vol. 91, no. 1, 2019.

[59] C. Fan, Y. Sun, Y. Zhao, M. Song and J. Wang, "Deep learning-based feature engineering methods for improved building energy prediction", *Applied Energy*, vol. 240, pp. 35-45, Apr 2019.

[60] J. Behrens, "Principles and procedures of exploratory data analysis.", *Psychological Methods*, vol. 2, no. 2, pp. 131-160, Jun 1997.

[61] T. Crosby, "How to Detect and Handle Outliers", *Technometrics*, vol. 36, no. 3, pp. 315-316, 1994.

[62] T. Fearn, "Classification and Regression Trees (CART)", *NIR news*, vol. 17, no. 6, pp. 13-14, 2006.

Appendices

Appendix A: Missing values for all attributes

Timestamp (dd/mm/yy tt)	HH Electricity (kWh)	HH Cooling CHW (kWh)	HH Boiler LTHW (kWh)	HH Water CWS (m^3)	DE Electricity (kWh)	DE Boiler LTHW (kWh)	DE Water CWS (m^3)
01/01/18 00:00 - 10/05/18 09:00[1]					✓ (!)		
06/02/18 13:30	✓			✓			
25/03/18 01:00 - 01:30		✓	✓				
21/05/18 10:00 - 10:30	✓						
13/06/18 18:00 - 14/06/18 07:30	✓			✓			
11/07/18 18:00 - 12/07/18 08:00	✓			✓	✓		✓
24/07/18 06:30 - 13:00	✓			✓	✓		✓
15/08/18 14:00		✓					
03/09/18 02:30 - 08:00	✓			✓	✓		
13/09/18 12:00	✓			✓	✓		✓
31/03/19 01:00 - 01:30	✓	✓		✓	✓		✓
04/04/19 10:00 -11:30, 13:00, 14:00 - 15:30, 17:30	✓			✓	✓		✓
20/08/19 13:00-13:30			✓				
03/11/19 10:30 - 04/11/19 08:30	✓			✓	✓		✓
06/11/19 08:30 - 10:00, 13:30	✓			✓	✓		✓

Table A.1: Missing values of dataset

Appendix B: Outliers for all attributes and relationship with missing values

Date	Time	HH Electricity (kWh)	HH Boiler LTHW (kWh)	HH Water CWS (m^3)	DE Electricity (kWh)	DE Boiler LTHW (kWh)	DE Water CWS (m^3)
31/01/18	15:30	49.3	-	-	-	-	-
06/02/18[2]	14:00	87.5	-	-	-	-	-

[1] 5 missing months

[2] The measurement is captured right after a missing value

07/02/18	15:30	46.41	-	-	-	-	-
10/02/18	17:00	-	-	-	-	-628905.6	-
	17:30					628946.7	
12/02/18	13:00	-	-	-217.89	-	-	-
21/03/18	12:00	-	-	-	-	-	480
	12:30						480
	13:00						480
14/04/18	09:00	-	-	-100.48	-	-	-
21/05/18[3]	11:00	73.41	-	-	-	-	-
14/06/18[4]	08:00	383.7	-	-	-	-	-
25/06/18	11:30	-	-	263.25	-	-	-
12/07/18[5]	08:30	411.91	-	-	7984	-	-
15/07/18	19:30	-	-825771.5	-	-	-	-
17/07/18	03:00	-	825772.6	-	-	-	-
24/07/18[6]	13:30	264.72	-	-	5796	-	-
30/08/18	08:30	-	-	-2363.9	-	-	-
	09:00			1844.3			
	10:00			519.7			
03/09/18	08:30	162	-	-	3346	-	-
04/09/18	13:30	-	-	-	-	-	-6563.6
	14:00						4177.6
13/09/18	12:30	47.5	-	-	687	-	-
17/09/18	14:00	-	-	-2496.8	-	-	-
	14:30			2497.23			
23/10/18	10:00	-	-	-245.6	-	-	-
28/10/18	01:00	-	-	-	692	-	-
	01:30	-10.41			-230		
	02:00	-			692		
30/01/19[7]	14:00	40.69	-	-	-	-	-
	14:30	41.59					
	15:00	40.5					
04/04/19[8]	12:00	126.22	-	-	1863	-	-
	13:30	76.31			1047		
	16:00	149.78			2146		
	18:00	43.22			999		
05/04/19	12:30	-	-	-	-	-	-5615.1
23/05/19	13:00	42.12	-	-	674	-	-
15/07/19	19:30	-	-825771.5	-	-	-	-
16/07/19	03:00	-	825772.6	-	-	-	-
27/10/19	01:30	-13.5	-	-	-217	-	-

[3] The measurement is captured right after 2 consecutive missing values
[4] The measurement is captured right after consecutive missing values from 13/06 18:00 to 14/06 7:30
[5] The measurement is captured right after consecutive missing values from 11/07 18:00 to 12/07 8:00
[6] The measurement is captured right after consecutive missing values from 24/07 6:30 to 13:00
[7] Maybe should not be considered as outliers
[8] All these measurements are captured right after consecutive missing values

	02:00	41.31			-		
04/11/19[9]	09:00	785.38	-	-	10610	-	-
06/11/19[10]	10:30	139	-	-	1676	-	-
	14:00	71.94			775		
09/12/19	10:00	46.19	-	-	-	-	-

Table B.1: Outliers for Electricity, Heating and Water Meters

The CHW cooling meter data appear to have some spread out outlier values (probably measurement errors) from the timestamp 11/09/19 11:00 and so on with value >= 41536.9 or <=-41666.54.

Appendix C: Scatterplots of all meters' data

The following scatter plots visualize how outliers are scattered through time for each meter for each building.

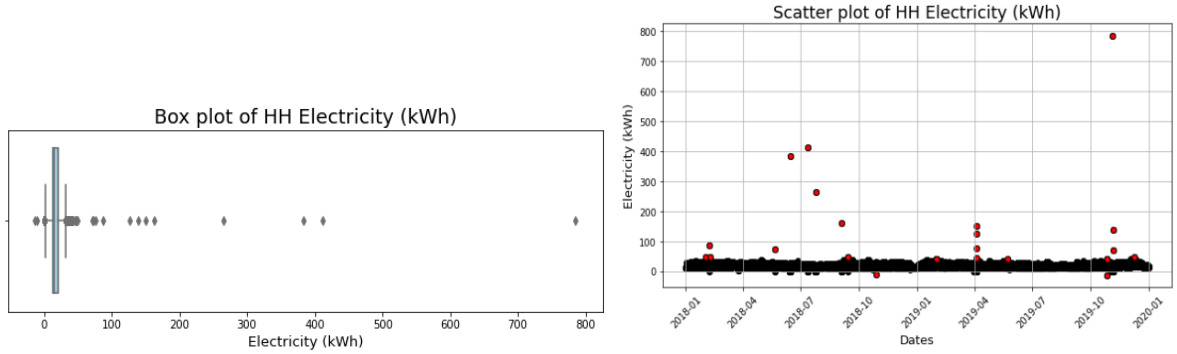


Fig. C.1: Box plot and Scatterplot for HH electricity

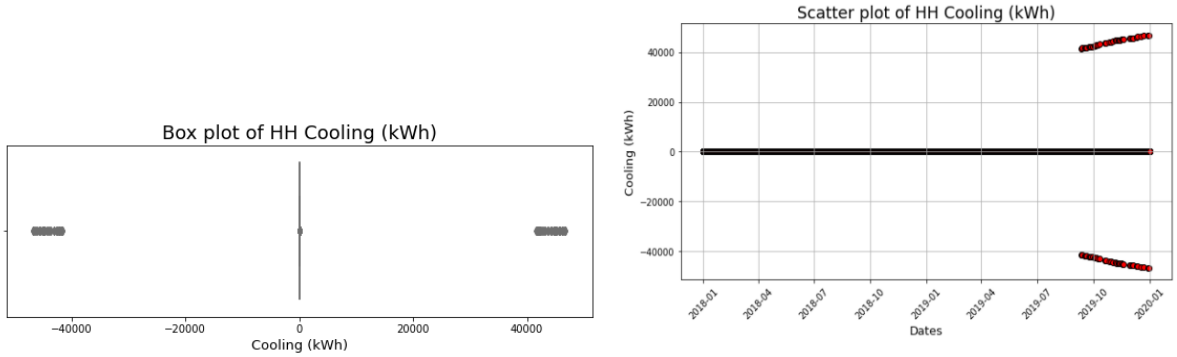


Fig. C.2: Box plot and Scatterplot for HH cooling

[9] The measurement is captured right after consecutive missing values from 03/11/19 10:30-04/11/19 08:30
[10] Both measurements captured after missing values

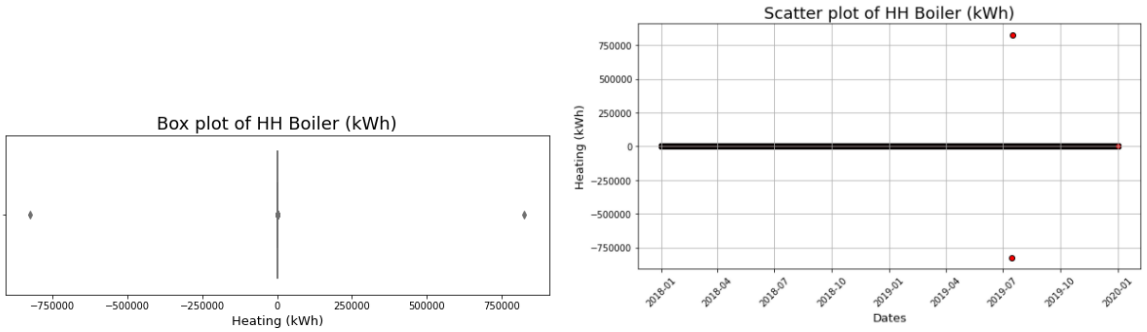


Fig. C.3: Box plot and Scatterplot for HH heating

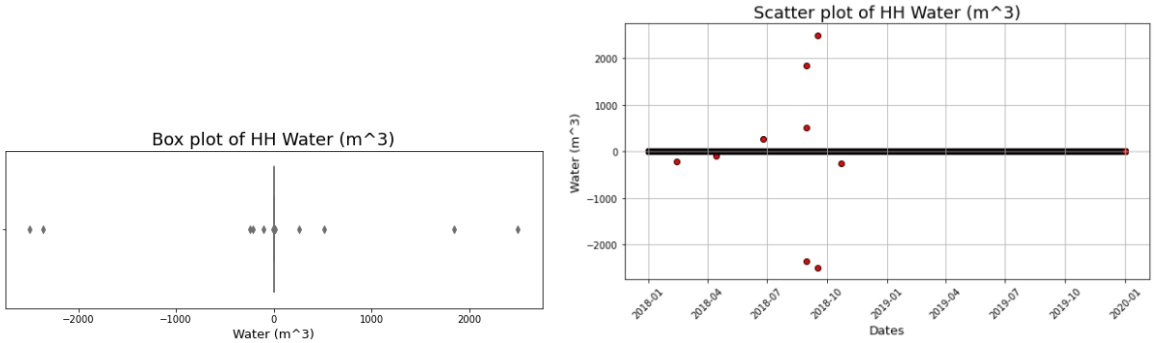


Fig. C.4: Box plot and Scatterplot for HH water

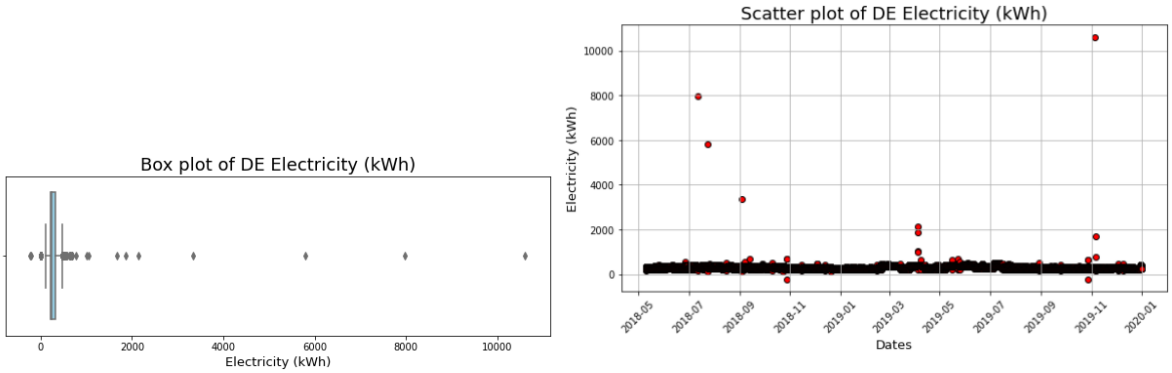


Fig. C.5: Box plot and Scatterplot for DE electricity

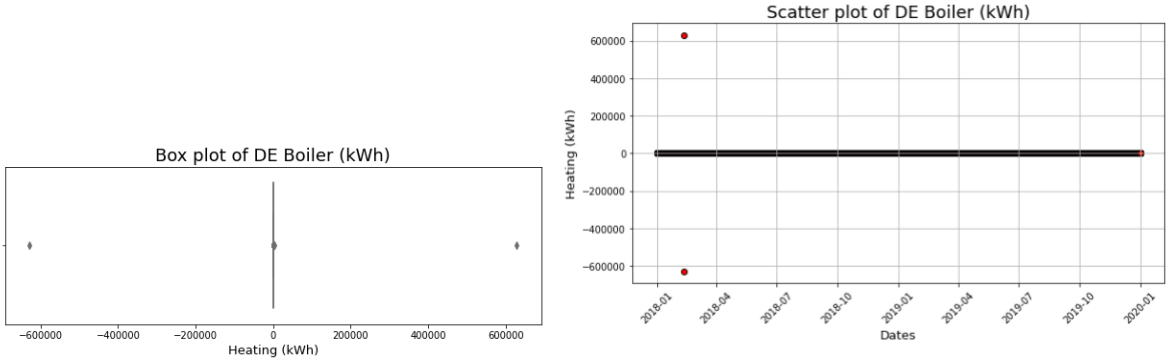


Fig. C.6: Box plot and Scatterplot for DE heating

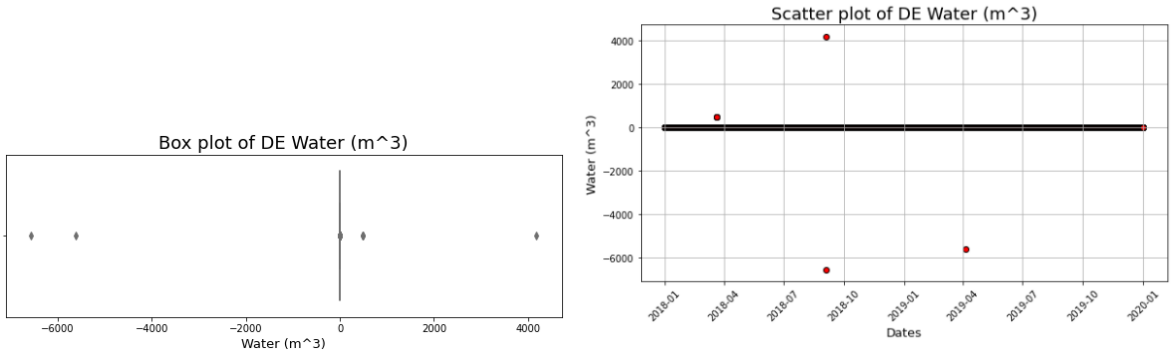


Fig. C.7: Box plot and Scatterplot for DE water

Appendix D: Distributions of un-processed energy features

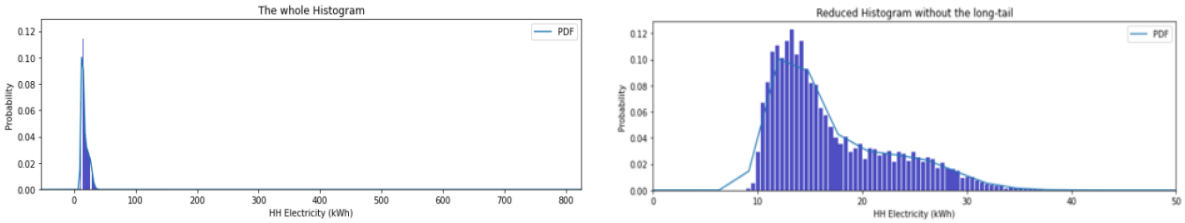


Fig. D.1: Distribution of HH electricity

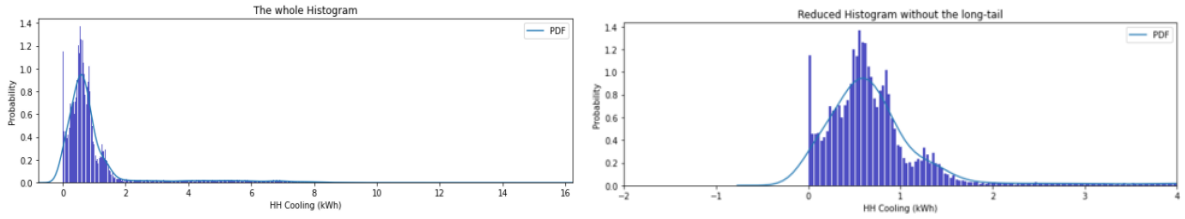


Fig. D.2: Distribution of HH cooling

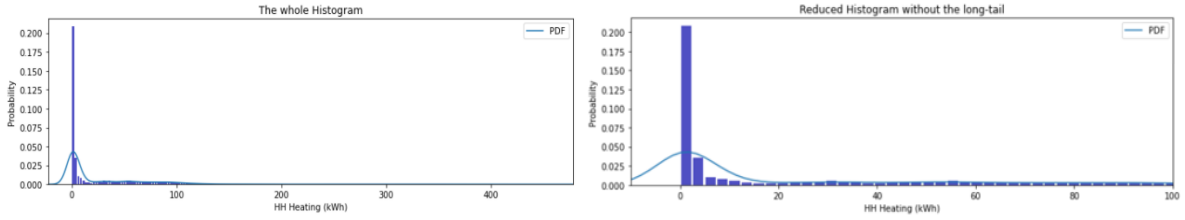


Fig. D.3: Distribution of HH heating

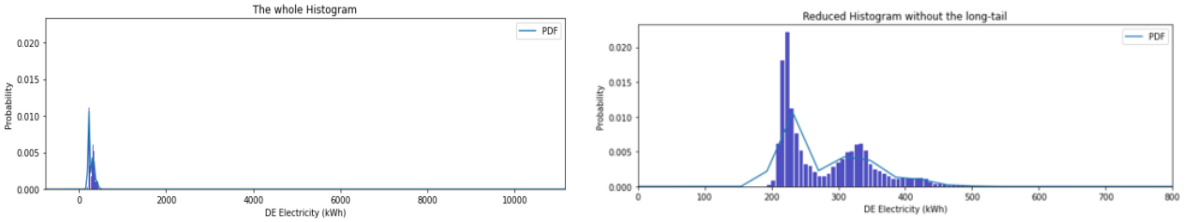


Fig. D.4: Distribution of DE electricity

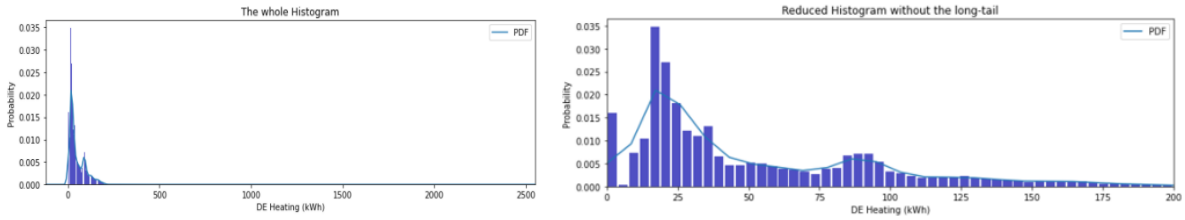


Fig. D.5: Distribution of DE heating

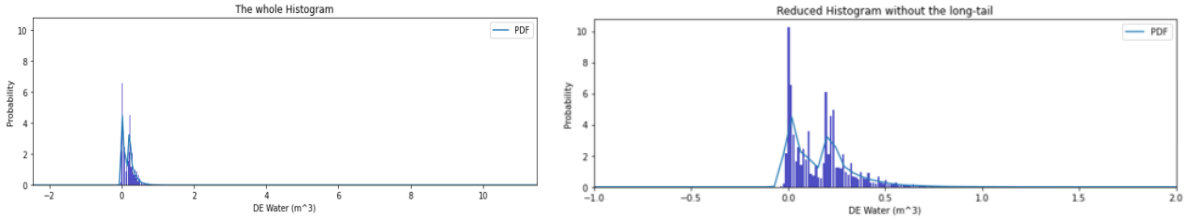


Fig. D.6: Distribution of DE water

Appendix E: Example of the patterns identified in our dataset

In the first screenshot, we can see a patter of "Low value-missing values-accumulated load", while in the second one, the pattern observed is "Low value-accumulated load".

5/21/2018 9:00	20.89	Consistent value
5/21/2018 9:30	9.5	Low value
5/21/2018 10:00		Missing values
5/21/2018 10:30		
5/21/2018 11:00	73.41	Outlier
5/21/2018 11:30	20.39	Consistent value

1/31/2018 14:00	30.8
1/31/2018 14:30	32.09
1/31/2018 15:00	16.61
1/31/2018 15:30	49.3
1/31/2018 16:00	30.2
1/31/2018 16:30	29.5

Appendix F: Plots of features before and after the cleaning process

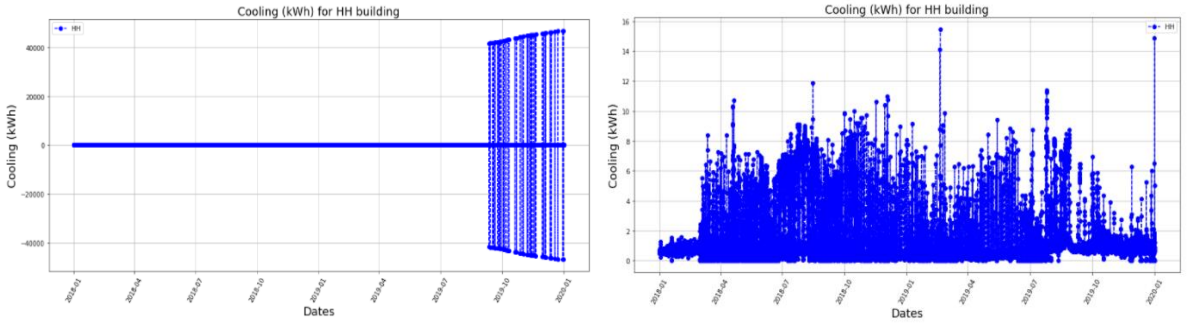


Fig. F.1: Cooling consumption plots for HH before and after the automate cleaning process

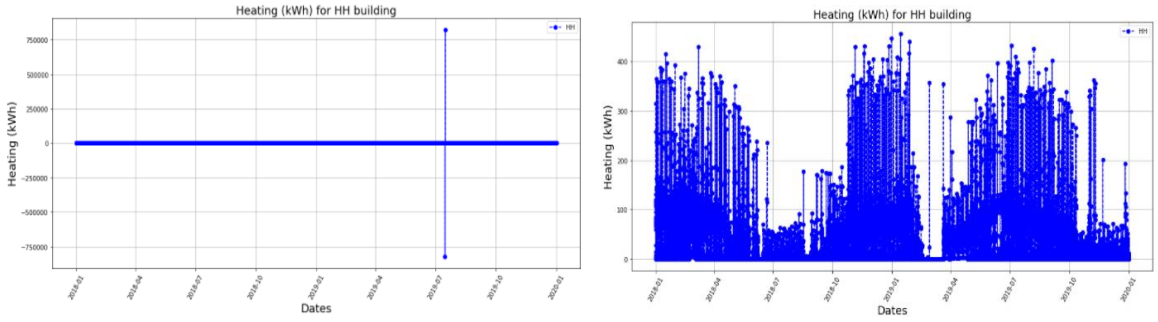


Fig. F.2: Heating consumption plots for HH before and after the automate cleaning process

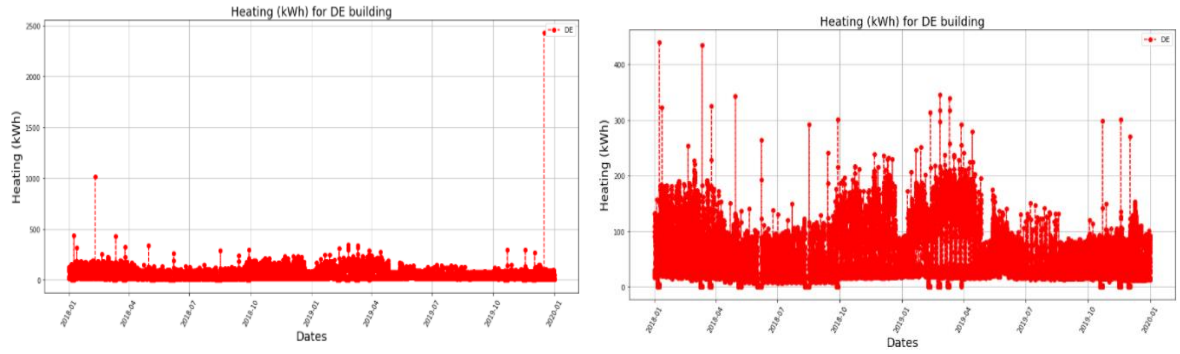


Fig. F.3: Heating consumption plots for DE before and after the automate cleaning process

Appendix G: Heatmaps of time-scaled variables for HH & DE electricity consumption

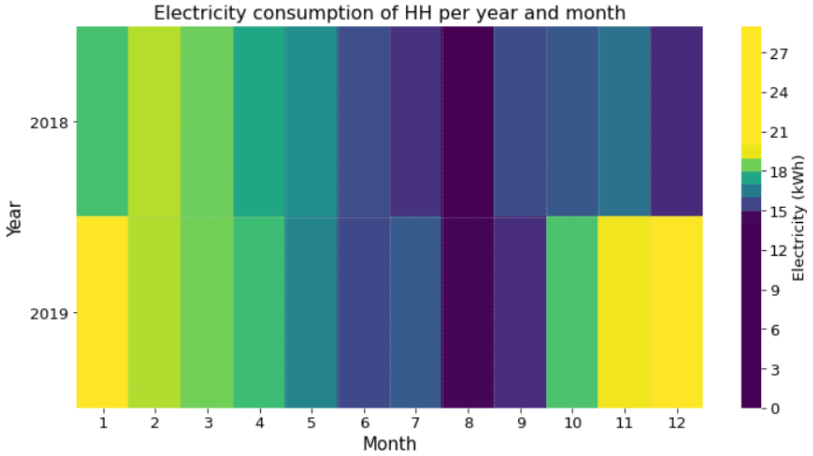


Fig. G.1: Heatmap on year and month of HH's electricity meter

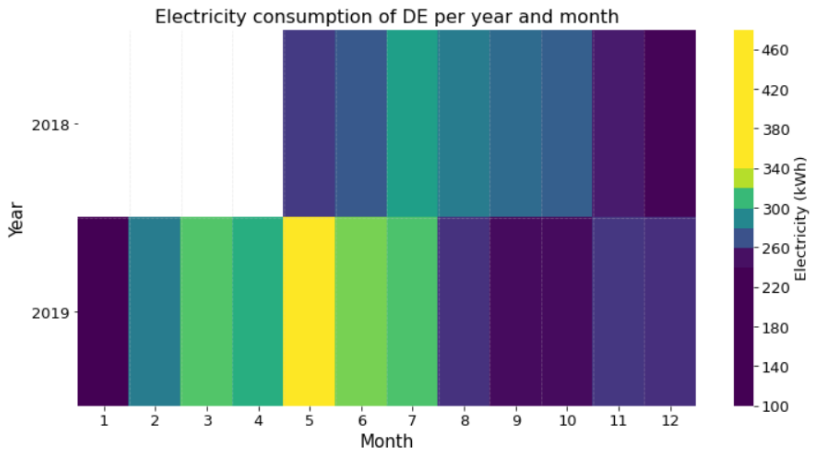


Fig. G.2: Heatmap on year and month of DE's electricity meter

Appendix H: Boxplots of every feature's clusters

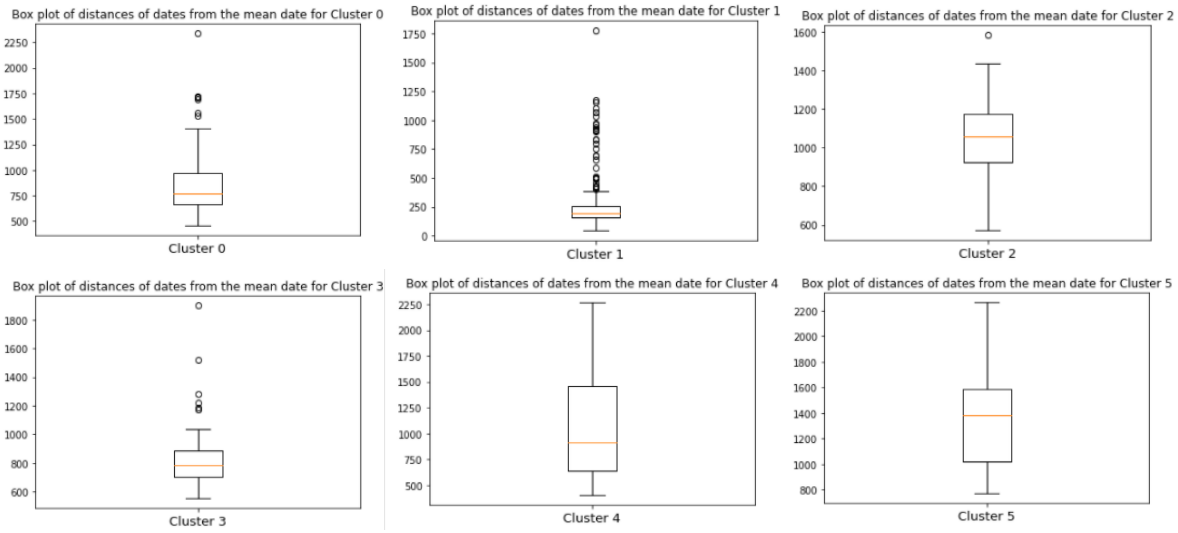
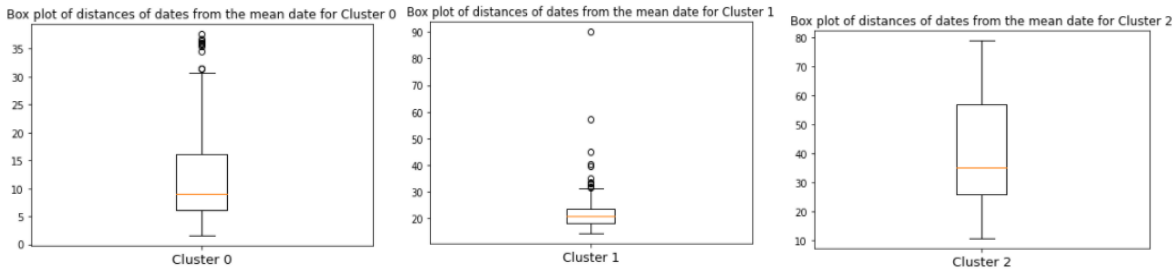


Fig. H.1: Right skewed boxplots of HH heating's clusters



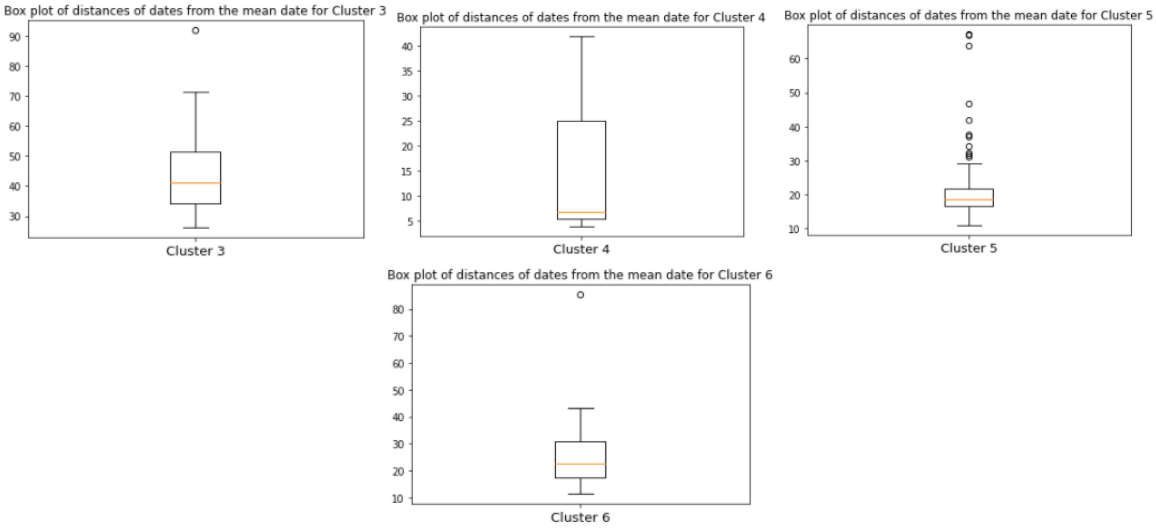


Fig. H.2: Right skewed boxplots of HH cooling's clusters

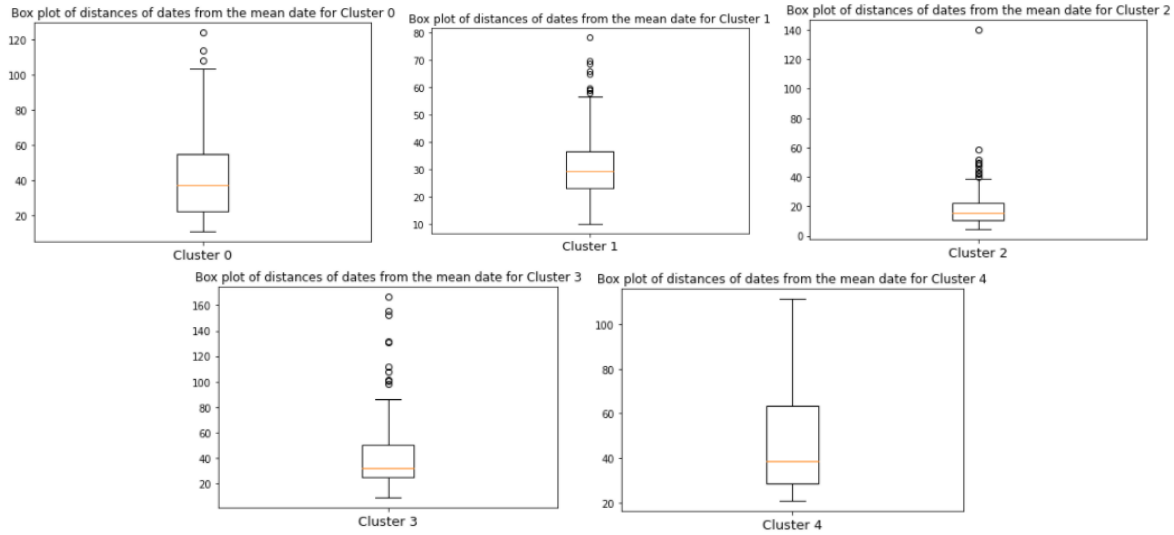


Fig. H.3: Right skewed boxplots of DE electricity's clusters

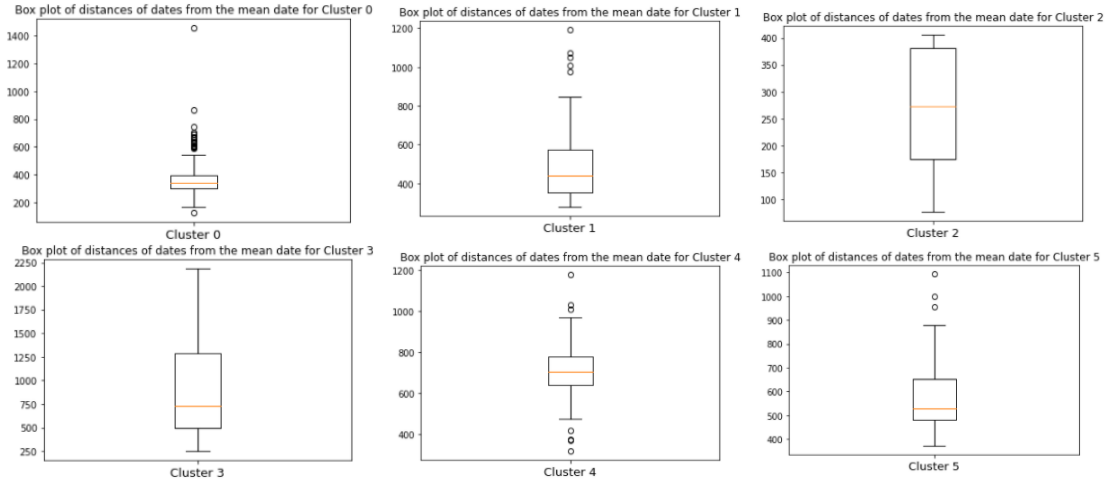


Fig. H.4: Right skewed boxplots of DE heating's clusters

Appendix I: Feature Correlations

The Pearson correlation function *corr()* in the Pandas provides a linear measure of features relationship. The pair-wise correlation values between variables in the dataframe indicate if a predictor (feature) value increases as another feature value increases (the Pearson correlation value is positive) or decreases when the input variable (the Pearson correlation value is negative) in a linear way.

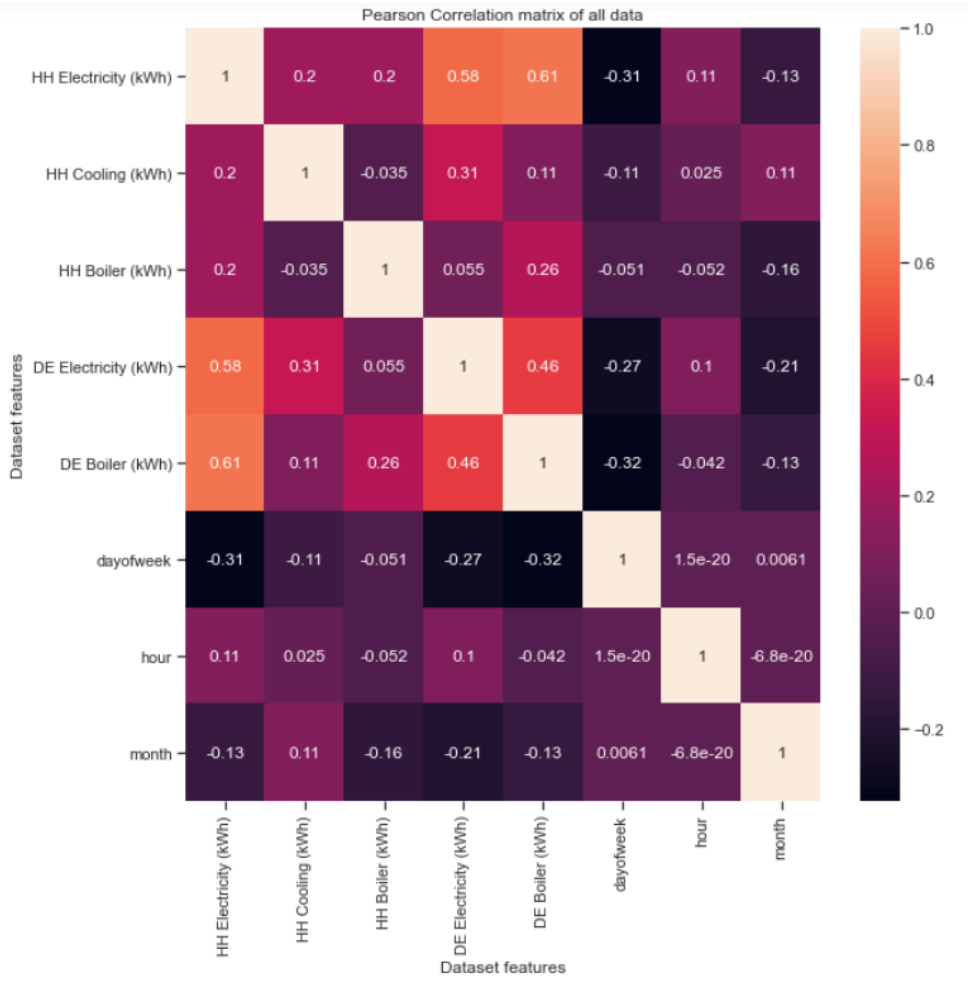


Fig. I.1: Pearson correlation matrix for all features along with time-scaled variables