Meter-level Electrical Load Anomaly Detection in buildings using Contextual Matrix Profile

Roberto Chiosaa, Marco Savino Piscitellia, Cheng Fanb , Alfonso Capozzoli a[[1]](#footnote-2)\*

a Department of Energy “Galileo Ferraris”, TEBE Research Group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy

b affiliation

Abstract

The ……

*Keywords:* Anomaly Detection, Matrix Profile, Energy Information System

1. Introduction

In the last few years, the increasing widespread use of IoT sensors in buildings for the pervasive monitoring of energy-related data, led to an unprecedented acquisition of reliable and accessible knowledge of the building performance during its operation. Considering that in Europe the building sector accounts for 40% of final energy use [1] and that almost 90% of the total energy consumed during the life cycle of a building depends on its operation [2], supporting building owners and energy managers to extract valuable information from collected data to reduce energy consumption, increase appliances efficiency, prevent energy wastes and operate their buildings more efficiently is of paramount importance.

Although a great deal of research has been done, the increasing volume of building-related data still overwhelms end-users [3], making it hard to spot energy reduction opportunities, find the root cause of anomalies or simply be aware of energy usage in buildings and systems. Building data are heterogeneous and reflects the complex interaction that occurs between occupants, energy systems, the building envelope, and external conditions. Managing those data is not trivial, however if properly managed, ingested and analysed they provide the opportunity to gain insight on the building operational behaviour discovering opportunities for savings [4].

A robust coupling of IoT sensors data, artificial intelligence approaches and energy domain knowledge proved to be effective in achieving high energy saving potential by leveraging a variety of energy management applications. The tools that provide such capabilities are the so called energy management and information systems (EMIS) which are employed to monitor, analyse and control energy systems leveraging on advanced data analytics technologies and are intended to help facility staff to enhance building systems performance and efficiency [5]. Depending on the level of detail of the measured data, EMIS can be classified as meter-level or system-level: the first includes high level measurements (e.g., data related to the whole-building electrical load or of the main sub-loads) while the second involves more detailed data related to the operation of specific systems or components.

A subset of EMIS conceived for the collection and analysis of meter-level aggregated data are the so-called Energy Information Systems (EISs). EISs typically focus on data not usually collected through building automation systems (BAS) providing graphical and analytical capabilities enabling predictive energy modelling analysis such as energy consumption forecasting, anomaly detection and diagnosis, advanced energy benchmarking, load profiling, and schedule optimization of building energy systems [6], [7].

﻿Among these, anomaly detection and diagnosis (ADD) as a tool for immediate and automatic recognition of abnormal and non-optimal performance of energy systems is essential to provide quantification of energy waste and prioritize corrective intervention. While fault detection and diagnosis (FDD) tools analyse system/component-level data to detect faults and anomalies, ADD tools rely on aggregated meter-level data to automatically detect anomalous energy trends at whole-building scale. Although performing a meter-level analysis poses several challenges regarding the complex interaction between buildings, occupants, and energy systems, it is of considerable value in real world case studies where the number of measured variables is not so high. In this context, the main objectives for an ADD process are: (i) the recognition of typical patterns in the whole-building energy consumption timeseries (ii) detection of anomalous load patterns and (iii) diagnosis of the detected anomalies performing an inference on the main sub-loads.

In accordance with these goals, this paper conceptualizes and develops an EIS tool capable of performing meter-level ADD on buildings electrical load. The proposed approach employs Matrix Profile algorithm for the automatic detection of energy consumption anomalies at the whole-building level and their diagnosis at the sub-load level, revealing which sub-loads impacted the most. The following paragraphs review the main works related to ADD in energy field with a particular focus on the capabilities of Matrix Profile algorithm as an anomaly detection method and how the authors addressed the issues related to its application in the buildings energy field.

* 1. Anomaly detection and diagnosis in buildings: related work

An anomaly is a region of data with significantly different behaviour from other data and that do not conform to expected values [8]. It can be referred as discord, outlier, deviation or exception and its definition is significantly different depending on the field of application and the analysis performed. In the energy and buildings field, the definition of anomaly is very domain-specific and may include abnormal behaviour of occupants, faulty operations of appliances, incorrect management of energy systems, abnormal sub-load consumption and technical and non-technical energy losses [9]. Thus, the very nature of buildings timeseries data requires to carefully address the definition of anomalies, which can be classified as point, collective, or contextual. A *point anomaly* is one individual instance or observation that can be considered anomalous when compared to the remaining data. On the other hand, a *collective anomaly* is a collection of instances anomalous with respect to the entire dataset. Finally, *context anomalies* are anomalies only if considered in a certain context (i.e., boundary conditions) and may not be considered an anomaly if it happens in a different context [10].

From the perspective of the analysis ADD has been traditionally tackled under statistical point of view, however the gradual introduction of advanced machine learning and data analytics techniques coupled with the large volume of available data opened new possibility to develop more sophisticated data-driven anomaly detection and diagnosis models [3].

The detection phase is usually accomplished by estimating a reference baseline representing normal behaviour, according to specific boundary conditions, and label every observation that does not belong to this region as anomalous [11]. The discrimination between normal and abnormal behaviour is essential, so both the robust development of a reference model and the proper selection of the feature used to define anomalies are crucial.

Supervised methods permit to train a machine learning algorithm using labelled dataset (i.e., ground truth) to create a reference model though which it is possible to classify anomalous consumption or not. Support vector machines (SVM) and multi-layer perceptron are largely used to perform model based anomaly detection [12]. Zhao et al. [13] developed an SVM-based method that was proven to be effective in detecting chiller operation anomalies of different severities. Autoencoders are employed in ﻿[14] to model the normal ﻿solar photovoltaic energy generation and anomaly detection is performed using reconstruction metrics. Regression and decision tree classifiers are other widely used supervised techniques used to develop the energy consumption reference model used to compare abnormal behaviours [15]. An evolutionary tree model was employed as a detection method in [16] and was effective in discovering frequent and infrequent patterns in meter-level electrical loads. An hybrid neural net ARIMA model is employed in [17] to predict the energy consumption and then identified anomalies comparing the actual and predicted energy consumption using the two-sigma rule. In [18] a regression model for detecting the anomalous trends in electrical energy consumption was developed by coupling artificial neural network (ANN) and regression tree (RT). Although supervised approach can achieve very accurate results, its adoption in real-world is still limited compared to unsupervised methods, mainly due to the difficulty in obtaining high quality datasets [19] and the absence of a reliable power consumption annotated datasets [9], [20].

On the other hand, *unsupervised* anomaly detection is more promising for practical applications since they permit to detect rare and unknown anomalous patterns without any a priori knowledge. Beside statistical unsupervised methods like the Principal Component Analysis (PCA) or Generalized Extreme Studentized Deviate (GESD) the data mining methods like association rule mining (ARM) and clustering gained popularity thanks to the capability to automatically extract significant relations between complex and massive data [4], [21], [22]. Li et al. [24] employed GESD to identify anomalous observations in electricity usage timeseries data collected from 40 buildings. Fan et al. [23] identified typical daily load profiles through ﻿entropy-weighted k-means (EWKM) clustering and then abnormal daily energy consumption profiles were identified through GESD. An anomaly detection framework that exploits ARM to identify collective anomalies and categorical clustering to identify potential anomalies is presented [25] where it is tested in smart grid context. Other unsupervised frameworks propose a distance-based approach for the evaluation of distance between a potentially anomalous observation and its neighbour samples [26], [27]. In this context, one of the most promising data analytics techniques for unsupervised distance-based anomaly detection in timeseries is Matrix Profile (MP) that will be extensively reviewed in the next paragraph.

* 1. Matrix profile method for anomaly detection

Introduced by [28], Matrix Profile (MP) is a novel algorithm that performs *all-similarity-join-search* among two timeseries, i.e. finds the nearest neighbour for each object of a data collection. Trivialimplementations of *all-similarity-join-search* algorithms result in excessive computation al time even for modest datasets. Common variants of this problem involve the search of k-nearest neighbour by setting a threshold parameters, which is both critical and difficult to set [29]. Others perform similarity search by reducing the dimensionality of dataset through PAA ﻿[30], [31] to speed up computation, however, this method causes loss of valuable information. Conversely, MP proposes an ultra-fast similarity search under the z-normalized Euclidean distance that does not reduce dimensionality, but calculates the full join, eliminating the need of setting a threshold making the method almost parameter free and exact. The exact and scalable algorithm allows the method to be incrementally maintainable, deterministic in time and so parallelizable on multicore processor to speed up even further the computations.

Given two timeseries and a given subsequence length, the MP algorithm produces two new series: the MP and Matrix Profile Index (MPI). MP is a one-dimensional timeseries that stores the z-normalized Euclidean distance between each subsequence of the first series and the closest matching subsequence (i.e., nearest neighbour) of the second timeseries. MPI is a one-dimensional timeseries that contains the index of where the nearest neighbour is in the second timeseries. By joining information of MP and MPI many insights could be extracted. Finding the minimum value of the MP is possible to find the best matching subsequence in a series (i.e., motif discovery) on the other hand by finding the maximum value of the MP it is possible to find the subsequence with the largest distance to its nearest match (i.e., discord discovery). In this sense, discord discovery may be interpreted as an anomaly detection method that identifies the most unique subsequences in a dataset.

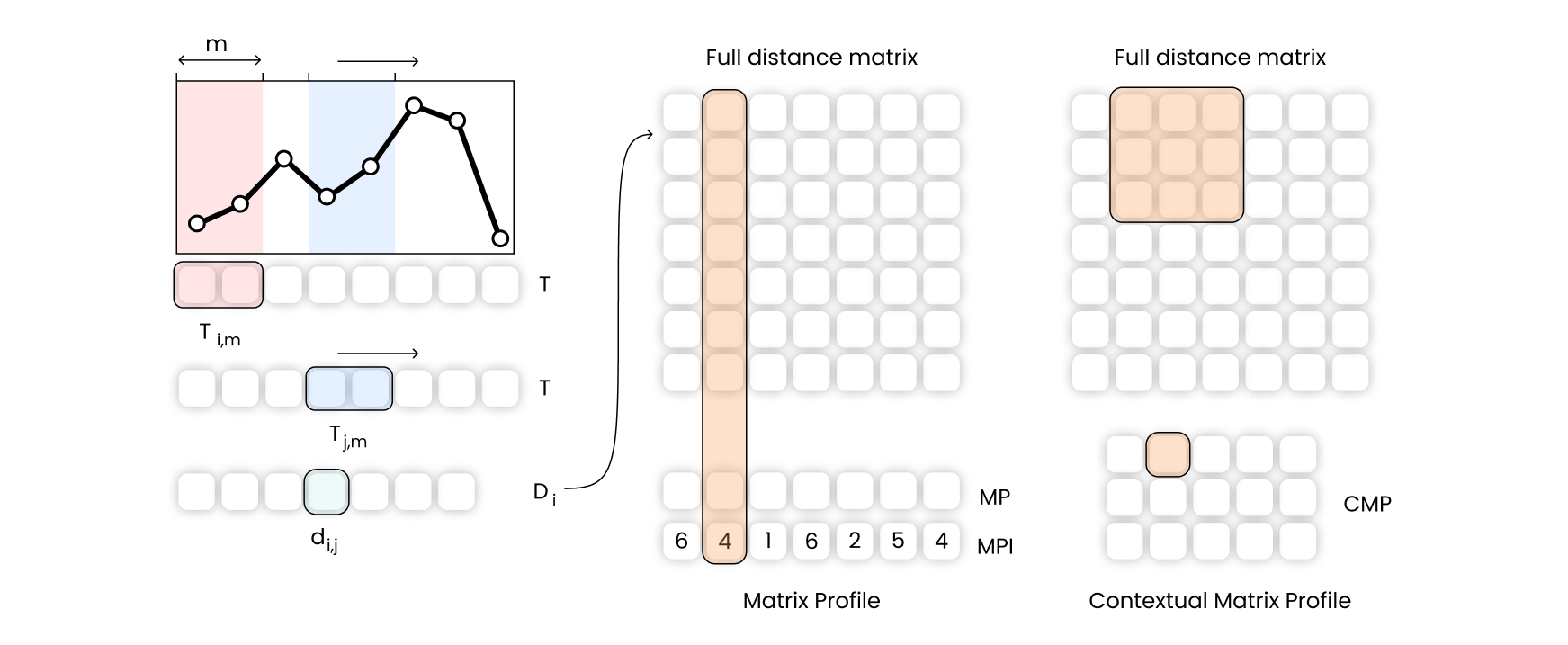
With reference to Figure 1 some fundamental concepts and definitions need to be introduced before going deeper into the topic. First, a *timeseries* is a sequence of real-valued numbers with where is the length of . Since the focus is on local properties of timeseries (i.e., portion of timeseries) a *subsequence* is defined as a continuous subset of values from of length starting from position ; formally defined as with .

An ordered set of all possible subsequences of obtained by sliding a window of length across is called *all-subsequences-set*  of a timeseries and is formally defined as follows: where is a user-defined subsequence length.

By computing the distance between a given query (i.e., subsequence ) and each subsequence in an all-subsequences set it is possible to define is a vector of distances called *distance profile* of a timeseries . Formally, where ) for all where and is the distance metric applied. It is possible to adopt different kind of distances to compute the distance profile [32], [33], [34] but the original method employs the Euclidean distance between the z-normalized subsequences.

If the distance profile is calculated between a query in and the all-subsequences set of (i.e., self-join), by definition the location of the distance profile is zero since the distance is calculated between the query and itself, . Moreover, the distance is close to zero just before and after this location. Those matches are called *trivial matches* and are usually avoided during similarity search by imposing an *exclusion zone* (as function of , usually set to ) before and after this location.

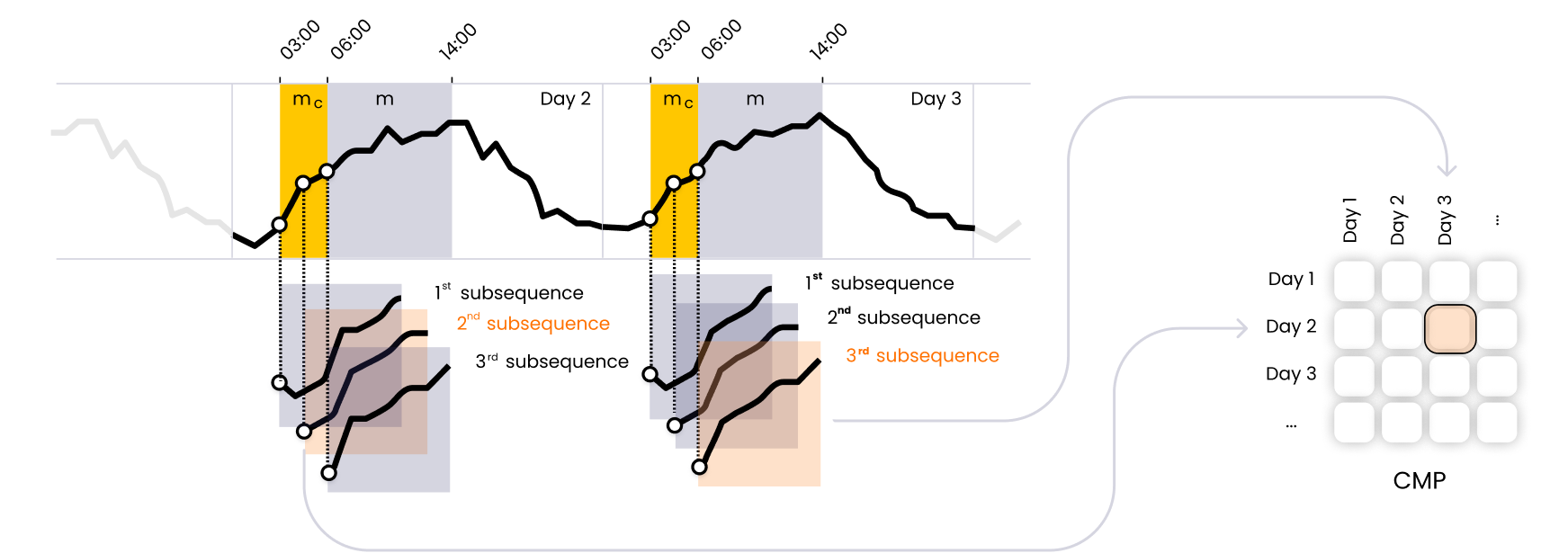
It is possible to finally define Matrix Profile (MP) as the vector that stores the z-normalized Euclidean distance between each subsequence and its nearest neighbour. Formally, where is the distance profile corresponding to query and timeseries . In other words, it can be generated by extracting the smallest value in each row/column of the full distance matrix. With reference to Figure 1, the MP is the column wise minimum over the entire full distance matrix, meaning that if finds the best matching subsequence (i.e., minimum distance) for any subsequence in . Of course, the construction of the full distance matrix is the most straightforward method but even the less computational efficient, this is the reason why many algorithms has been proposed for the MP calculation to reduce time and dimensionality complexity such as STAMP, STAMPI and STOMP based on MASS algorithm [35], approximated AMPSA and AMP [36] and multidimensional mSTAMP [37].



**Figure 1.** Description of Matrix Profile and Contextual Matrix Profile calculation steps in case of self-join of a timeseries . From left to right is explained the calculation of the element of the distance vector given the query . By calculating the distance vector for the all-subsequences set of , and storing those values in a matrix, the full distance matrix is obtained. MP is the row wise minimum while the CMP is the minimum over rectangular regions.

The original MP method has been successfully applied in different fields of anomaly detection. In medical field was proposed an unsupervised real time anomaly detection method based on continuous learning of timeseries shaplets extracted though MP algorithm [38]. Those shaplets are extracted and stored in an anomaly library and then used for sliding-window based anomaly detection in an electro-cardiogram (ECG) timeseries (﻿MIT-BIH database [32]). An industrial application is presented in [39] where the classical approach of MP is combined with the hamming distance to automatically detect intrusions in the network of a water processing facility. A generalization of MP algorithm called PanMP is proposed in [40] to find different length anomalies in ﻿automated pedestrian counting system developed in Taipei. MP has been largely employed to identify anomalies in IT field. Herath et al. [41] introduces ﻿a real time anomaly detection framework based on MP called Real-Time Aggregated Matrix Profile (RAMP), that can identify anomalies in scientific workflows. De Paepe et al.[42] applies a noise elimination technique based on MP on real Yahoo! internet traffic metrics to detect anomalous behaviours; [43] demonstrate how the elimination of noise can help in anomaly detection of noisy date by testing the algorithm on Numenta Benchmark [44].

In the energy field there are few implementations of MP algorithm. The possibility to use the MP discord detection capabilities to provide insights about the dominant energy usage pattern in large academic buildings was explored by [45]. The authors applied the classical MP approach with daily, weekly, and monthly time window lengths as a feature extraction method to identify unusual behaviours in energy consumption timeseries of a large commercial building dataset on one year period. The process was tested on 422 buildings [46] which end uses were classrooms, offices, laboratories and dormitory. Zhu et al. [34] demonstrated how MP can be useful in detecting rare anomalous electricity consumption occasionally produced by a meter swapping event. The algorithm was tested on a synthetic meter swapping event built on top of two timeseries of household electrical power demand and was proven to be effective to discover the suspicious similarity between the two timeseries. Park et al. [47] applied MP as a part of an automated load profile discord identification (ALDI) based on statistic comparison between normal and anomalous patterns in a large portfolio of buildings. The MP method was used to quantify the similarities of daily subsequences in timeseries meter data under z-normalized distance. The computed MP values were then compared with typical-day MP distribution and was proven to be effective to identify unique load shapes patterns and discords.



**Figure 2.** Focus on the definition of subsequence length and contexts in an electrical load timeseries and their importance in the CMP calculation. The timeseries has one hour frequency and for each day a context of (yellow) and a subsequence length of (grey) are defined. When calculating the CMP, the nearest neighbour search is performed only on the subsequences that start within the defined context. The nearest neighbour subsequences between Day 2 and Day 3 are highlighted in orange. The Euclidean distance between the nearest neighbour subsequences is calculated and stored in the CMP shown on the right side.

Despite it proved to be effective in several application fields; the original MP method compares regions of timeseries that belongs to different context, operating conditions or different boundary conditions and may result into misleading results. To address this issue, in [48] was introduced the Contextual Matrix Profile (CMP), defined as the minimum over rectangular regions of the full distance matrix (see Figure 1), allowing to find the best matching subsequence in ranges over and . This allows to group data in custom way comparing only portions of with portions of . The CMP calculation is led by the definition of contexts which are a lapse of time in which a subsequence of length may start. Figure 2 shows an example definition of context subsequence length and groups. Suppose that in an electrical load timeseries with frequency an analyst wants to find for every day the nearest neighbour between the 3 subsequences of length 8h that start between the 03:00 and 06:00 AM. To this aim the subsequence length must be set to , (grey in Figure 2) and the context length (yellow in Figure 2) and the resulting self-join CMP will have a rows/columns﻿ for each day. Each point of the CMP would display the distance between the best matching long subsequence of the two days: lower the distance better the match and vice versa. While context is suitable to create a priori grouping of timeseries, once the MP is calculated it is even possible to further divide the MP into groups that reflect a broader comparison among contexts. For instance, it is possible to subset the CMP by keeping only weekdays or weekends.

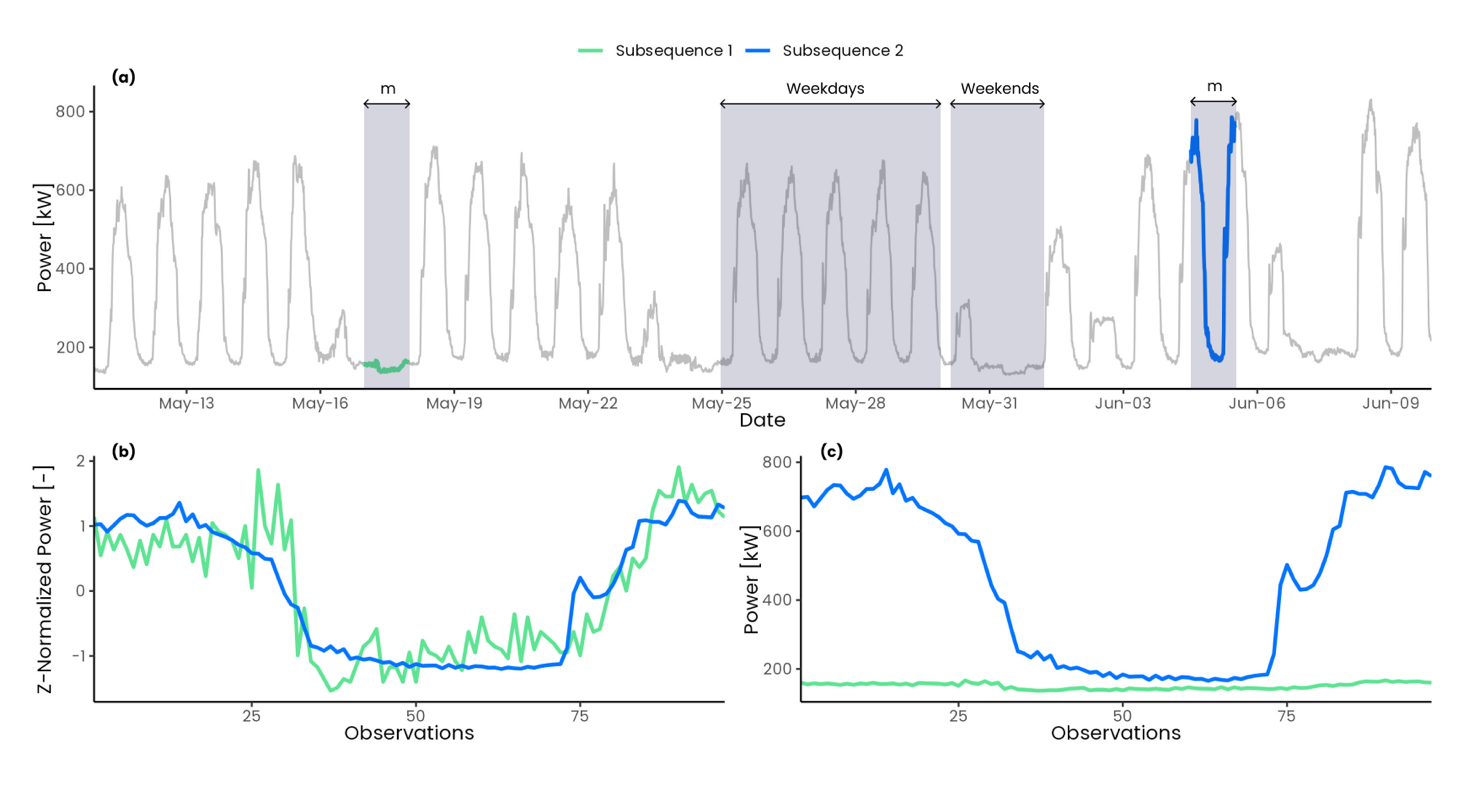
An application of the CMP method for anomaly detection in buildings is presented in [48] where the method was applied to a dataset containing different indoor air quality related variables (i.e., temperature, humidity, CO2 etc.) measured in residential built environments. CMP was applied on the CO2 concentration timeseries, defining a 2-hour context and subsequence length of 3 hours and considering only weekends. In this way it was possible to compare only weekend morning subsequences and identify 6 anomalies. Moreover, the authors were able to identify periodic behavioural patterns even if there was not a time alignment among them.

This case study demonstrated the flexibility of the anomaly detection method and its effectiveness when coupled with domain knowledge. Therefore, it emerges a great potential in the application in the building energy field, in which the definition of anomaly is strongly related to the expert definition of contexts and the boundary conditions, leading to the recognition of patterns hardly detectable otherwise neglected.

* 1. Research gap and contribution of the paper

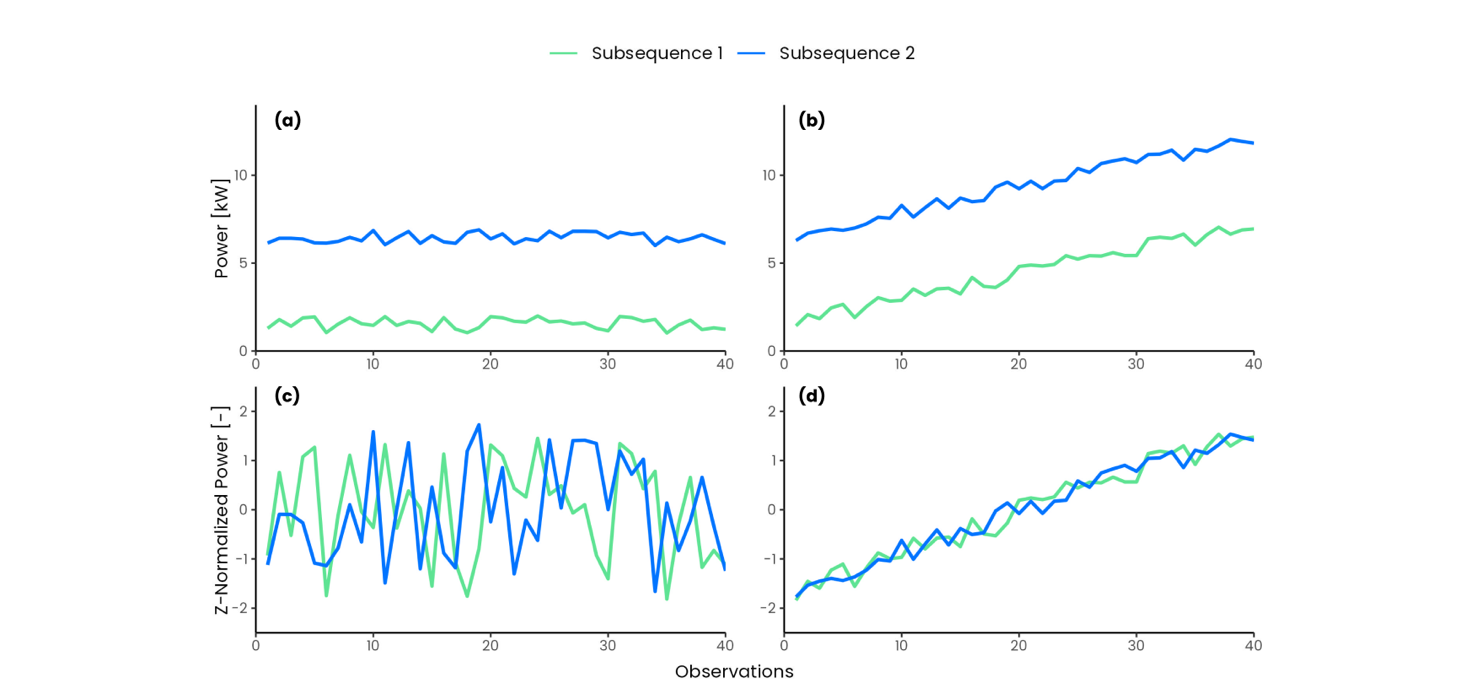
From the previous literature review it emerges that the MP method has been employed successfully in different fields of anomaly detection and authors have proposed different implementations according to the field of interest. In fact, even if MP is an unsupervised method useful for discord discovery, every field has constraints and peculiar boundary conditions that cannot be overlooked. In the case of our interest, related to the identification of anomalies in building energy consumption timeseries, it is essential to approach the problem also considering different variables such as occupation, weather conditions, energy system configuration and so on. A completely unsupervised method may fail to consider the relations with those variables and extract ineffective or trivial results, not useful for anomaly detection. In the following paragraphs some examples are presented.

In buildings, anomalous subsequences in the energy consumption timeseries are defined as unexpected behaviours that result in atypical shape and/or magnitude. The classic MP through z-score normalization of the subsequences searches the nearest neighbour based only on shape similarity losing crucial information related to the magnitude. Figure 3(a) shows a real electrical load timeseries for a non-residential building (university campus) in May and June. It is possible to observe how the electrical load changes significantly from weekdays to weekend when the load profile is almost flat. Applying the classic MP method with a subsequence length of one day, the two subsequences highlighted respectively in blue and green are identified as nearest neighbour. As shown in Figure 3(b) under z-score normalization they are almost overlapping. However, from Figure 3(c) it is possible to see that the not normalized subsequences have very different amplitudes and are referred to distinct energy consumption patterns typical of weekdays and weekend respectively. This is a clear example of how the subsequence normalization led to misleading results.



**Figure 3.** Effect of z-score normalization on two electrical load timeseries subsequences (blue, green) of length 24h (96 observations): (a) full electrical load timeseries; (b) comparison between z-score normalized subsequences; (c) comparison between not normalized subsequences.

Z-score normalization not only minimizes magnitude effects in the research of motifs and discords but also tends to emphasize any fluctuation and noise in the timeseries. By comparing two relatively flat subsequences under z-score normalization the resulting Euclidean distance is higher compared to non-flat subsequences, leading into higher values of MP in flat regions of the timeseries. In Figure 4 a comparison between two noisy synthetics timeseries is shown. In Figure 4(a) the two timeseries are relatively flat while in Figure 4(b) the two timeseries present a positive slope. In the first case, Figure 4(c), the normalization led to an amplification of the noise of the original subsequences resulting into a high Euclidean distance (d = 9.25) while in the second case, in Figure 4(d), the Euclidean distance is much lower (d = 1.5). This issue have been largely analysed in [42] where a smoothing process is proposed as possible solution, beside the trivial solutions of discarding flat regions or change the subsequence length. A clear consequence of this issue is that, referring to Figure 3(a), the MP method would identify the weekends as discords since they present almost flat profiles compared to weekdays subsequences and this is a critical aspect when dealing with energy consumption timeseries of buildings that by their nature often present a strong pattern periodicity among working and not working days.



**Figure 4.** Effect of z-score normalization on noisy relatively flat subsequences (a-c) and on noisy subsequences with positive slope (b-d).

Comparing two subsequences belonging to different energy patterns could be not so useful, therefore, introducing domain knowledge to find motifs or discords only in some portions of the timeseries became in specific cases extremely important. The concept of Annotation Vector (AV) is used to introduce domain knowledge in the process of motif and discord discovery [49], which allows to find results that respect user defined constraints and produce robust results, closer to expectations of the analyst. Annotation vector is a meta timeseries used to correct a posteriori the values of the original MP manipulating the motif/discord search. This method overcomes some weakness of the traditional definition of motifs by solving issues related to stop-word bias and simplicity bias [50]. However, this method does not modify the MP calculation: *all-pairs-similarity-search* is always performed and then a downstream processing is conducted. In some applications it can be useful to exclude some region or to split subsequences into different groups and then perform the similarity search to discover anomalies by comparing only the interesting regions and excluding others. A solution to this problem have been proposed by [48] introducing the Contextual Matrix Profile (CMP) algorithm. CMP allows search intervals to be defined along two timeseries and look for the best matching subsequence among them, allowing to compare homogeneous subsets of the original timeseries providing novel insights.

The application of unsupervised methods coupled with expert knowledge is crucial for the accurate discovery of anomalies in building electrical load and it is the key to reduce energy wastes and enhance energy management in buildings. To this aim the objective of this work is the introduction of an unsupervised anomaly detection procedure based on MP algorithm to detect anomalous electrical load at building level in quasi real-time and identify the responsible sub-loads though a diagnosis phase. According to the previous literature review and excursus on implication of MP as anomaly detection method, this paper intends to address the following issues by contributing as follows:

1. Develop a contextual anomaly detection and diagnosis methodology by introducing a MP-based method that employs automatic and unsupervised processes (clustering and decision trees) for the identification of parameters such as subsequence length, groups, and contexts.
2. Explore and propose a solution to the challenges that arise from the application of MP as anomaly detection method in the energy field. To overcome the issues of z-score normalization and its implication by using the Euclidean distance between not normalized subsequences and introducing, thanks to the CMP method, domain specific boundary conditions upstream the MP calculation allowing a knowledge-based comparison between subsequences.
3. Introduction of a robust anomaly score definition based on different statistical methods and domain knowledge that allows potential anomalies to be identified and ranked, within a given group and context, by considering only those that negatively impact the energy use (i.e., high energy consumption in absolute and relative terms).

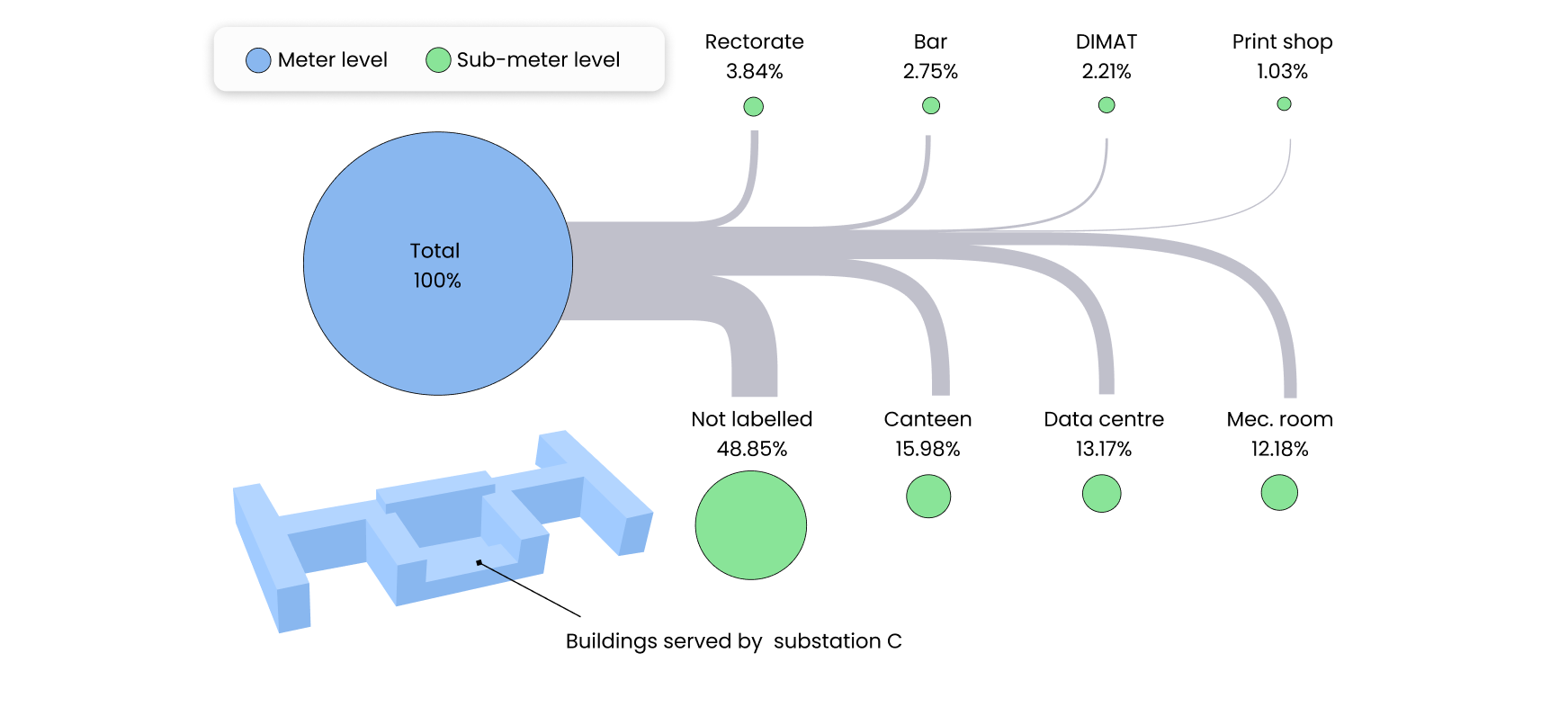
The rest of the paper is organized as follows. Section 2 provides the description of the case study. Section 3 presents the methodology adopted. Finally, Section 4 presents the anomaly detection and diagnosis results and Section 5 critically discusses the outcomes and contains the concluding remarks.

1. Case study

The case study analysed refers to the energy consumption of a MV/LV transformer cabin identified as “substation C”, that serves a part of the main campus of Politecnico di Torino (PoliTo), an Italian university located in Turin. Data related to the total electrical load and to some sub-loads are available with 15 min timesteps from 1 January 2015 to 31 December 2019. The hierarchical structure of the available data is shown in Figure 5: The first level (blue) refers to the total electrical load of substation C, while the second level (green) shows the available sub-loads. In addition, the load breakdown in terms of average annual energy consumption was provided.

In particular, a bar and a staff canteen were at the disposal of students and campus staff and on average, respectively represent the2.75% and 16.03% of the yearly electrical energy consumption of substation C. The university data centre accounted for 13.16% of the total energy consumption. The administration offices (Rectorate) corresponded to 3.83% of energy consumption and the mathematics department (DIMAT) for 2.21%. A large share of energy consumption (12.22%) was related to the mechanical room. The equipment located in this room included hot and chilled water circuits and auxiliaries such as recirculation pumps. The chilled water was provided by two chillers of nominal electrical power of 220 kW and a rated cooling capacity of 1120 kW, and a reversible water-water heat pump, with nominal a power and cooling capacity of 165 kW and 590 kW, respectively.

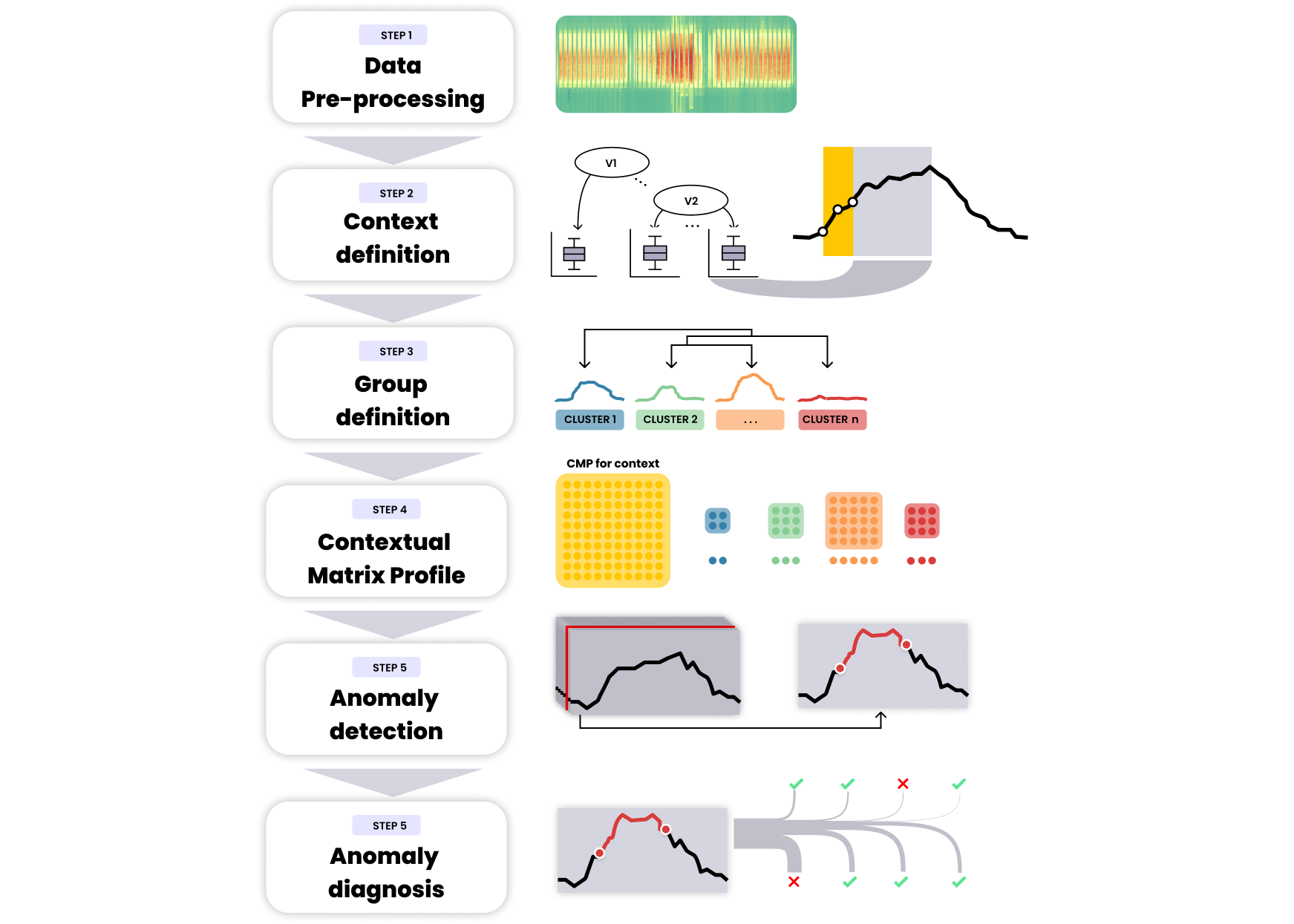
The remaining energy consumption was aggregated under a unique synthetic sub-load tagged as “Not labelled” as showed in Figure 5. It accounted for 48.76% of the total energy consumption, and since it was not directly measured, cannot be characterized, as the other sub-loads, respect to the energy service or building served.



**Figure 5.** Hierarchical structure of the electrical load database under study.

1. Methodology

In this section the methodological framework is presented. The method is based on the application of the CMP coupled with unsupervised and supervised data analytics techniques such as cluster analysis and Classification And Regression Trees (CART) to perform a parameter-free and automatic anomaly detection and diagnosis in electrical load timeseries of a University campus. The multi-step procedure, reported in Figure 6 consists in four steps, described in detail in the following paragraphs.



**Figure 6.** Graphical description of the methodological framework.

* 1. Pre-processing.

The first step consists in data pre-processing and it is a crucial task for the data analysis workflow. The proposed methodology does not focus on advanced pre-processing techniques since the considered dataset had a good quality with a missing value and inconsistence ratio less than 5% on the overall observations [51]. Thus, the pre-processing is performed through univariate statistical approaches, aimed at removing statistical inconsistencies and replace them together with missing values through linear interpolation.

* 1. Contextual matrix profile.

The second step of the analysis deals with the application of the contextual matrix profile method and goes through the following phases: (a) context definition (b) group definition (c) CMP calculation and its further splitting according to previously defined groups.

Within the daily electrical load timeseries it is possible to identify different regions and different behaviour (e.g., base load, peak load, ramp-up ramp down) whose relative time window length can be defined statistically or inferred from the typical building operational and occupational schedule [52], [53]. By identifying daily electrical load subsequences it is possible to extract information of particular interest for building energy management. The methodology proposed in this paper identifies sub-daily time windows () through the recursive partitioning Classification and Regression Tree (CART) [54], [55]. Starting from the root ( that contains all the available instances) this method proceeds through a binary decision fashion to split the instances in purer subsets (nodes) in a froward stepwise fashion maximizing at each step the purity of each node [8], [55], [56], yielding local optimum [57] once a stopping condition is satisfied. The identification of these region in an unsupervised way has a twofold meaning: (i) automatically identify time windows based on historical operational data, (ii) define the two CMP parameters, subsequence length () and context length () that usually are set a priori based on domain knowledge. The regression tree is developed using the electrical load as numeric target attribute and the hour of the day as explanatory attribute. This allows to identify, through a cost complexity process, a set of non-overlapping time windows and consequently contexts and subsequence length. Thus, the subsequence length for the context is set equal to the relative time window length (). Moreover, since the CMP provides the flexibility to investigate similarity of shifted subsequences, context is defined as the half of the smallest time window length (). If the smallest time window is two hours long from 6:00 to 8:00 the context is defined as one hour long from 5:00 to 6:00.

As a second step a group definition based on daily load profiles was performed. A supervised expert approach was first applied to group holidays and non-working days (i.e., flat daily profiles) and half working days (e.g., Saturdays). Then hierarchical clustering with Euclidean distance was performed on the remaining daily profiles. The hierarchical clustering generates non-overlapping clusters by splitting instances based on a distance metrics, and each cluster can be further divided into subclusters and so on, creating a tree structure. The aim of grouping daily profiles in clusters which are representative of specific energy consumption patterns is to create homogeneous groups in which the CMP can be further split and in which anomaly detection process is able to detect anomalous subsequences, by considering as reference that specific energy pattern.

Once contexts and subsequence length are defined the CMP is calculated for each context under not normalized Euclidean distance. Given that by previous definition each day have not overlapping contexts, the resulting CMP contains one row/column for each day. Then, the overall CMP is split into different groups according to the groups defined previously. On each group CMP the anomaly detection process is performed.

* 1. Anomaly detection

In this work four statistical model-based outlier detection methods used for outlier identification in univariate timeseries. All those methods accept as input a timeseries and annotates each observation of the timeseries with Boolean value: if the observation is not an outlier, if it is an outlier.

*Inter quartile* defines outliers any of the observations that fall below and above where is the interquartile range () is defined as the difference between the third quartile as the first quartile .

*Z-score standardization* is a model-based outlier detection method which defines an outlier based on the gaussian normal distribution . This method defines outlier any of the observations outside the interval where is a user defined constant in z-score. The normal probability distribution usually defined meaning that the probability to find an observation outlies that range is equal to 2.3%. To apply this method to a not normal distribution z-score standardization is needed.

*Elbow method*: is a graphical method that allows to find the elbow of a curve. By finding the elbow of a univariate vector ordered in descending values it is possible to identify two different regions: the region below the elbow and the one above the elbow, in which observation are tagged as outliers.

*Generalized Extreme Studentized Deviate (GESD):* is an iterative method that progressively evaluates the presence of outliers in a univariate timeseries through a statistical test. The method initialization requires () a presumed number of outliers and confidence interval is set, then for a given the following statistical test is performed:

* There are no outliers in the timeseries
* There are up to outliers in the timeseries

The hypotheses test is performed by calculating the statistic and the critical value as follows:

Where and denote sample mean and sample standard deviation of the timeseries, is the timeseries length, is the iteration number,  is the 100p percentage point from the [t distribution](https://www.itl.nist.gov/div898/handbook/eda/section3/eda3664.htm) with ν degrees of freedom and

The anomaly detection, described in Figure 7, is performed for a given group within a given context, by applying the four methods previously described to identify anomalies and define a severity. As a first step the CMP is reduced into a vector by calculating the median of each row/column, then 4 anomaly detection methods are applied producing 4 new vectors that defines whether an element is anomalous or not in a Boolean form . Then the severity is calculated by counting by the number of positive detections . To make more robust the anomaly detection method and consider only positive anomalies (e.g., anomalies that result into a higher energy consumption) the energy consumption for each subsequence is calculated and stored in a vector which undergoes to the same process described before: methods are applied and then severity is calculated. By summing the two resulting severity vectors and it is possible to obtain an overall severity ranging from to that robustly ranks anomalies, respectively, from the least severe to the most severe. Finally, the most severe subsequences () are tagged and further analysed in the diagnosis phase. With reference to Figure 7, given anomaly detection methods the severity vector and range from a minimum of 0 (four methods out of four did not found an anomaly) to a severity of 4 (four methods out of four found an anomaly). By summing and the resulting overall severity ranges from 0 (four methods out of four did not found an anomaly for both energy and CMP severity vector) to 8 (four methods out of four found an anomaly for both energy and CMP severity vector).



**Figure 7.** Graphical description of the anomaly detection phase framework.

* 1. Anomaly diagnosis phase

Once the anomaly at meter-level is identified, the detection phase is enabled with the aim to identify which sub-loads are the most responsible. By keeping the same hyperparameter settings as previously described each sub-load timeseries undergo the CMP calculation and anomaly detection process through the 4 methods presented. The diagnostic process consists in selecting only those days that resulted anomalous at meter level and analyse the sub-load severity. The sub-load that presents the higher severity is the one that with great probability has affected the anomalous behaviour identified to meter level.

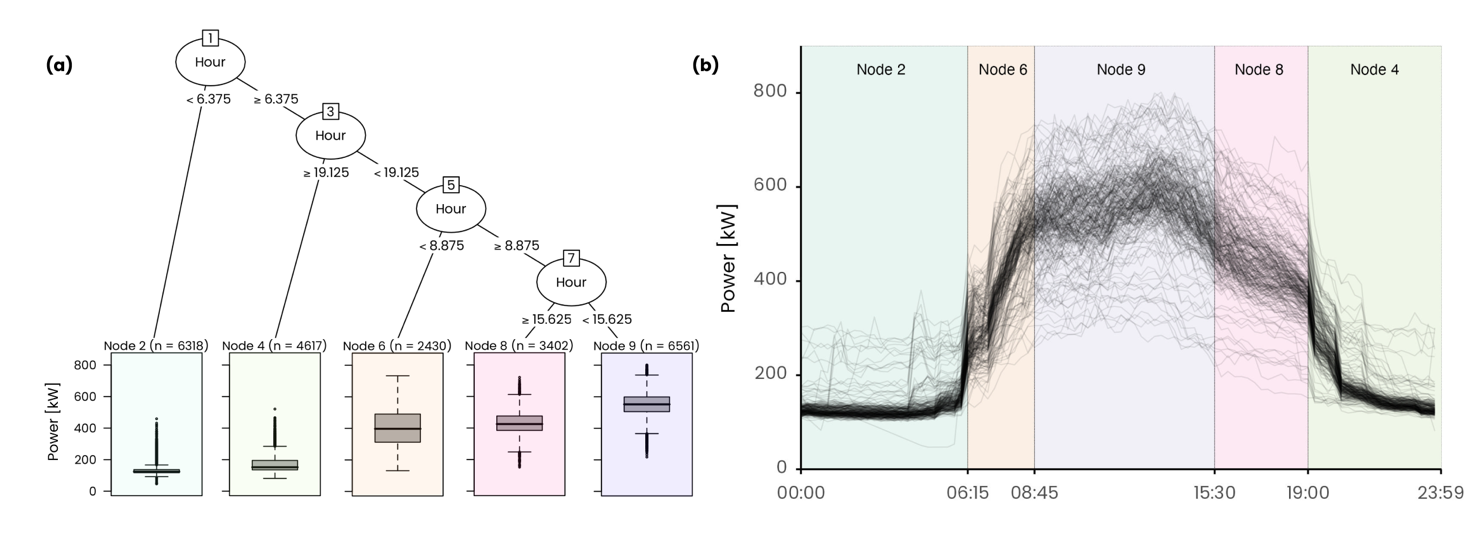
1. Results

The analysis was carried out using the R statistical software [58] for the pre-processing, CART, clustering and visualization and Python [59] for the CMP calculation and anomaly detection process.

The presented methodology has been tested on the electrical load timeseries of a MV/LV transformer cabin that serves a part of the Italian university campus of Politecnico di Torino (PoliTo). The measurement infrastructure continuously provides the total electrical load with 15 min timestamps. The authors decided to test the presented methodology on a dataset that spans from 01.01.2019 to 31.12.2019 even if more recent data are available, mainly because the pandemic COVID completely changes operational patterns and caused a closure of the university from February 2020. The raw dataset contained 35040 observations with a missing value ratio of less that 0.1%. Inconsistences were removed and missing values imputed through linear interpolation.

* 1. Contextual matrix profile.

To identify homogeneous electricity consumption regions within the daily load profile, non-overlapping time windows were evaluated through a CART using total electrical load as target variable and time of the day as numerical predictive variable. Daily profiles with low standard deviation of the daily electricity demand (i.e., weekends and holidays) were excluded to make the model more robust especially in the operating hours. The stopping criterion adopted was calculated such that the minimum number of objects in each leaf node would correspond to a minimum time window of 2.5 hours. The tree was subjected to cross validation and cost-complexity pruning, resulting in the five-leaf tree shown in Figure 8(a). For completeness the nodes resulting from the tree have been reported also in the Figure 8(b), where it is possible to observe the actual amplitude of the temporal window on a daily scale.



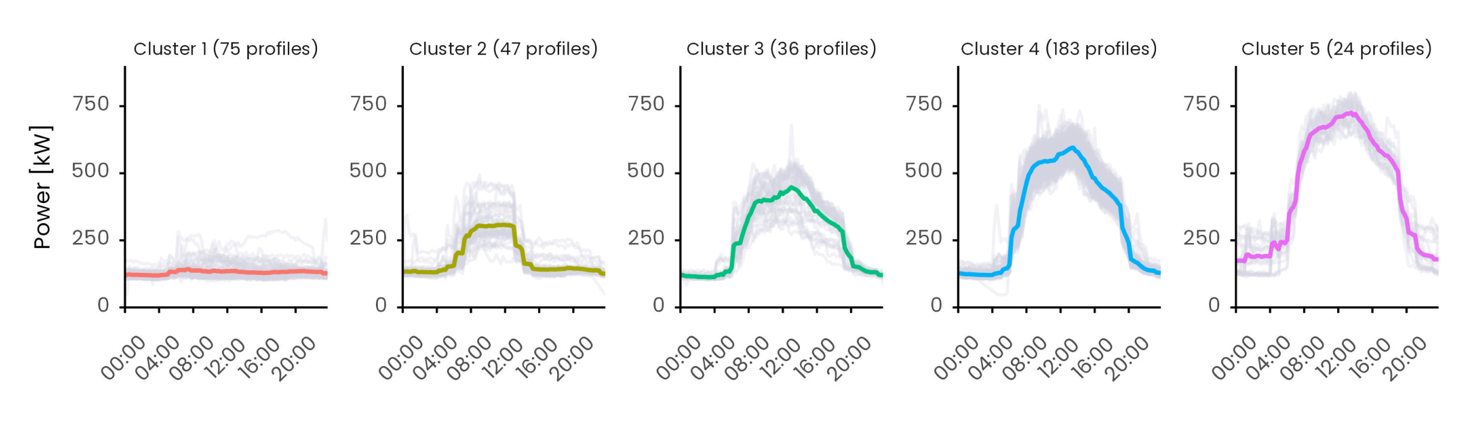
**Figure 8.** Daily electrical load profile clusters with the relative centroid.

It is possible to see how the model was able to effectively separate the night hours (time window 1 and 5) from the diurnal operation and was able to identify the ramp-up (time window 2), mid-day operation (time window 3) and ramp down (time window 4).The smallest time window corresponds to a duration of 2.5 hours, so the context length was defined as the half of the smallest time window length (). The outcome of this preliminary step was the definition of 5 time windows duration (e.g., subsequence length) and 5 contexts for the CMP calculation, summarized in Table 1. For instance, the first time window ranging from 00:00 to 06:15 defines a subsequence length of 25 observations (i.e., that starts in the context between 23:00 of the previous day and 00:00.

**Table 1.** Summary of resulting time windows and subsequence length.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Time Window | | | | Context | | | |
| ID | Interval | Duration | Observations | ID | Interval | Duration | Observations |
| *tw,1 = m1* | [00:00 - 06:15) | 6 h 15 min | 25 | *mc,1* | [23:00 - 00:00] | 1 h | 4 |
| *tw,2= m2* | [06:15 - 08:45) | 2 h 30 min | 10 | *mc,2* | [05:15 - 06:15] | 1 h | 4 |
| *tw,3= m3* | [08:45 - 15:30) | 6 h 45 min | 27 | *mc,3* | [07:45 - 08:45] | 1 h | 4 |
| *tw,4= m4* | [15:30 – 19:00) | 3 h 30 min | 14 | *mc,4* | [14:30 - 15:30] | 1 h | 4 |
| *tw,5= m5* | [19:00 – 24:00) | 5 h 00 min | 20 | *mc,5* | [18:00 - 19:00] | 1 h | 4 |

The second step consists in the definition of the groups was performed using a semi-supervised approach. As a first step the 75 profiles corresponding to public holidays, university closures and Sundays were extracted and grouped in Cluster 1. Secondly 47 profiles belonging to half working days and Saturdays were extracted and assigned to Cluster 2. The remaining 243 profiles, corresponding to working days, were organized into a matrix 243x96 where each row corresponds to a daily load profile. Then hierarchical clustering algorithm with ward.D2 method was implemented on the not normalized daily load profiles. The silhouette index, implemented in the package NbClust [60], was used to search the optimal number of clusters in a range 2-6. Three clusters were identified and labelled as follows: Cluster 3 (36 profiles), Cluster 4 (183 profiles), Cluster 5 (24 profiles). The results of the group definition are shown in Figure. It is possible to see that the grouping process led to a well-defined set of clusters each one representing a typical behaviour of the load profile and will be used to split the CMP for a given context into homogeneous groups for the anomaly detection process.



**Figure 9.** Daily electrical load profile clusters with the relative centroid.

The CMP was calculated by self-joining the data for each of the 5 contexts using the Euclidean distance between not normalized subsequences. The calculation was performed using the open source Python code [48] implementing the Series Distance Matrix framework to calculate the CMP. As a representative example in Figure 10 is reported the CMP for context 5. Since the dataset contains 365 days and there is a context per day the resulting CMP is a 365x365 symmetrical matrix. The higher the distance value (gradient fill) the higher the dissimilarity. It is possible to see a weekly regularity in the overall CMP, there are typically 5 days with the same behaviour (green) followed by two days of different behaviour (yellow). Moreover, there is a change of typical patterns during summer, especially July and August, corresponding to holidays and summer closure of the university facilities. By further split the CMP in groups, previously defined with the cluster analysis, it is possible to group days that we expect to behave in similar manner and perform a more robust inspection. For instance, the day 10th of November 2019 (at index 63 of cluster 1) stands out to be remarkably different from all the others day in the cluster, which is not so evident by only visualizing the global CMP. This highlights the importance of grouping for the identification of contextual anomalies.

Immagine che contiene testo, elettronico, circuito

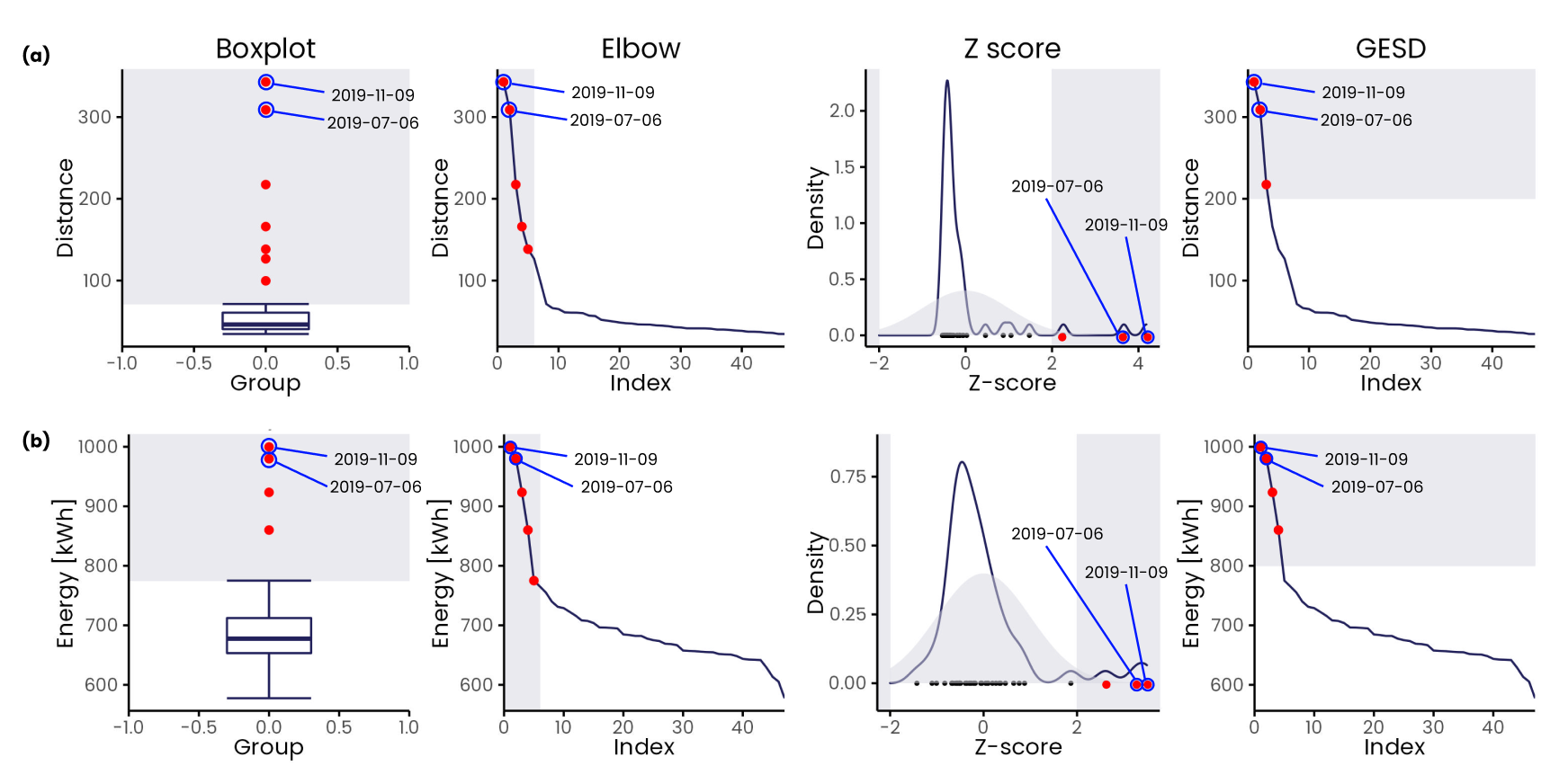
Descrizione generata automaticamente

**Figure 10.** On the left the contextual matrix profile result for context 5. Each point of the matrix shows the Euclidean distance between the best matching subsequences of the two days. Lower the distance better the match. On the right the CMP divided into clusters.

* 1. Anomaly detection results

The anomaly detection is performed for each group of each context. For each day within a group both the median of the Euclidean distance (i.e., median of column/row of the CMP) and the energy consumption in a time window are calculated, and the four univariate anomaly detection methods are applied. The methods are tuned as follows: the IQR method is calibrated to consider only the positive outliers over 1.5 IQR, the z-score only the positive observations over 2, GESD observations considered outliers with a 0.05 tolerance and the elbow method, since it is a pure graphical method, considers outliers above the knee. Each method defines whether an observation is anomalous or not in a Boolean form . Then the severity is obtained by counting by the number of positive detections. By summing the two resulting severity vectors and and it is possible to obtain an overall severity ranging from to that robustly ranks anomalies from the most severe to the least severe.

In Figure 11 are presented the anomaly detection results for the context 5 cluster 2, in Figure 11(a) methods are applied on the vector of the median Euclidean distances and in Figure 11(b) are applied on the vector of the energy consumption. It is possible to easily verify that two days, 9th of November 2019 and 6th of July 2019, were detected as anomalous by all the methods in both the Euclidean distance and energy vector and this resulted into an overall severity of 8.

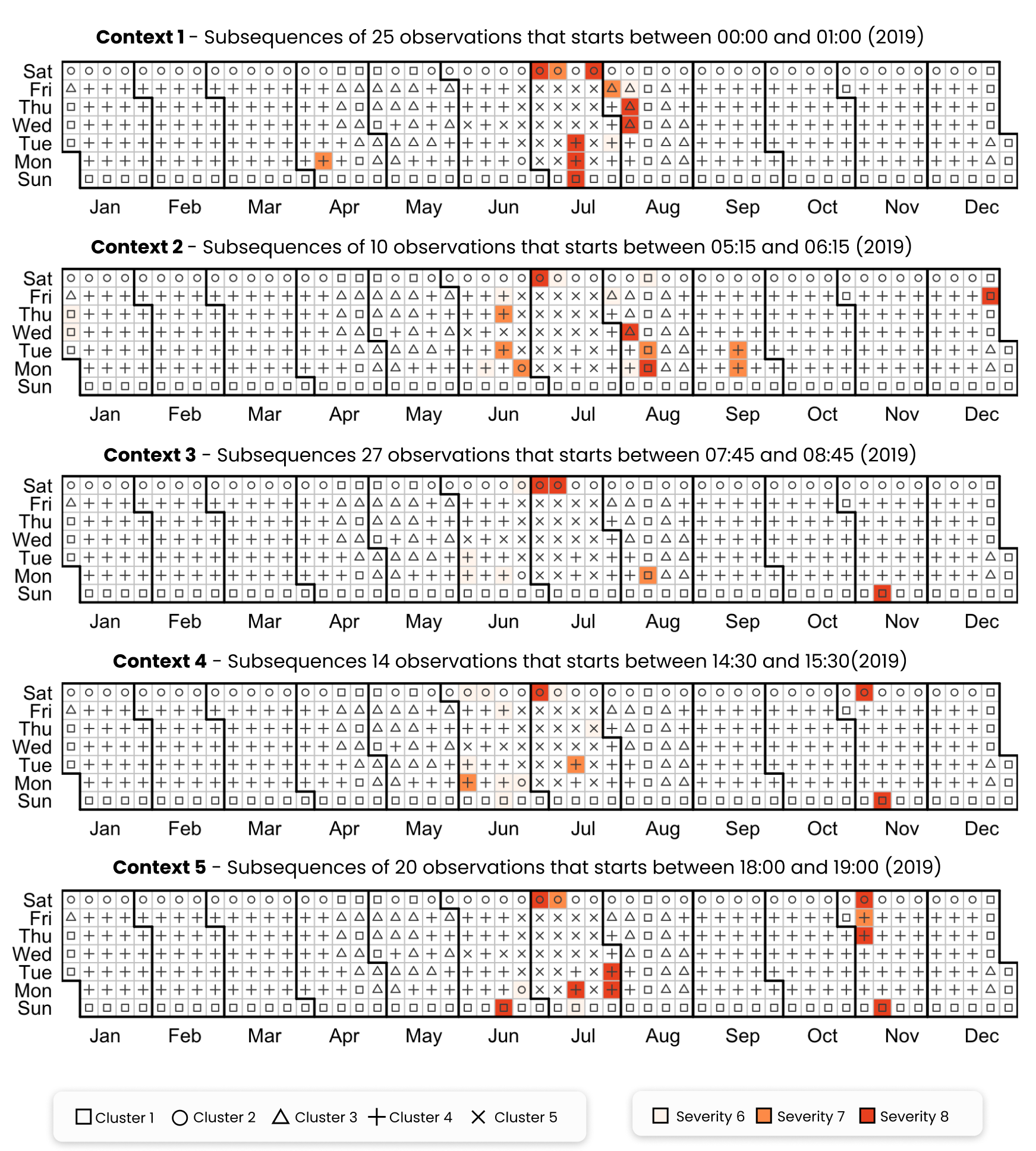


**Figure 11.** Example of severity calculation on (a) median Euclidean distance vector and (b) energy consumption vector using the four anomaly detection methods respectively inter quartile, Z-score standardization, elbow method and Generalized Extreme Studentized Deviate (GESD).

To reduce the number of spurious alerts and feedbacks only the severity 6-7-8 are considered as relevant and total of 64 anomalies were detected: 25 of severity 6, 14 of severity 7 and 25 of severity 8. A higher severity denotes a significant difference in terms of shape and energy from most of the days within the relative cluster and context. Furthermore, given an anomalous profile, it is possible to estimate the overconsumption by calculating the difference between the anomalous energy consumption and the energy consumption of the centroid of the respective cluster.

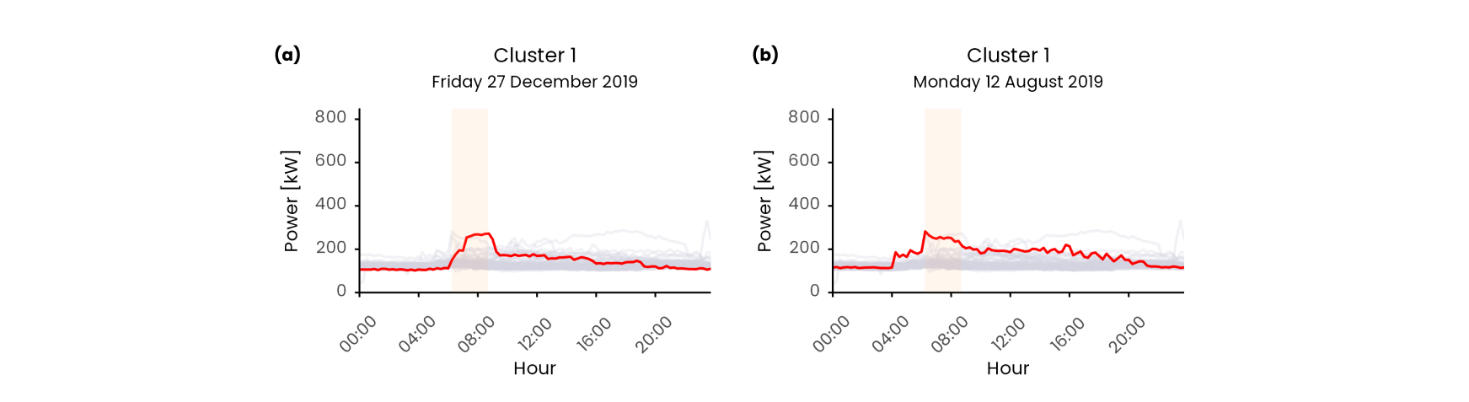
Results are summarized in Figure 12 through a calendar chart which is an effective way to visualize anomalies over the year for the different contexts. According to the Figure anomalies are more frequent at the beginning and end of the day: 12 in the 1st context, 18 in the 2nd context, 8 in the 3rd context, 14 in the 4th context and 12 in the 5th and last context. Another interesting pattern that can be inferred from the Figure is that 51 anomalies out of 64 are concentrated during summer from June to August, compared to the rest of the year where only 13 anomalies are detected.

A demonstration of the effectiveness of the context and time window definition is the ability of ADD process to find isolated anomalies that occasionally occur and are limited to within one time window (i.e., spot anomalies) and others that once arouse persist in subsequent contexts and are likely to last until the end of the day (i.e., persistent anomalies).



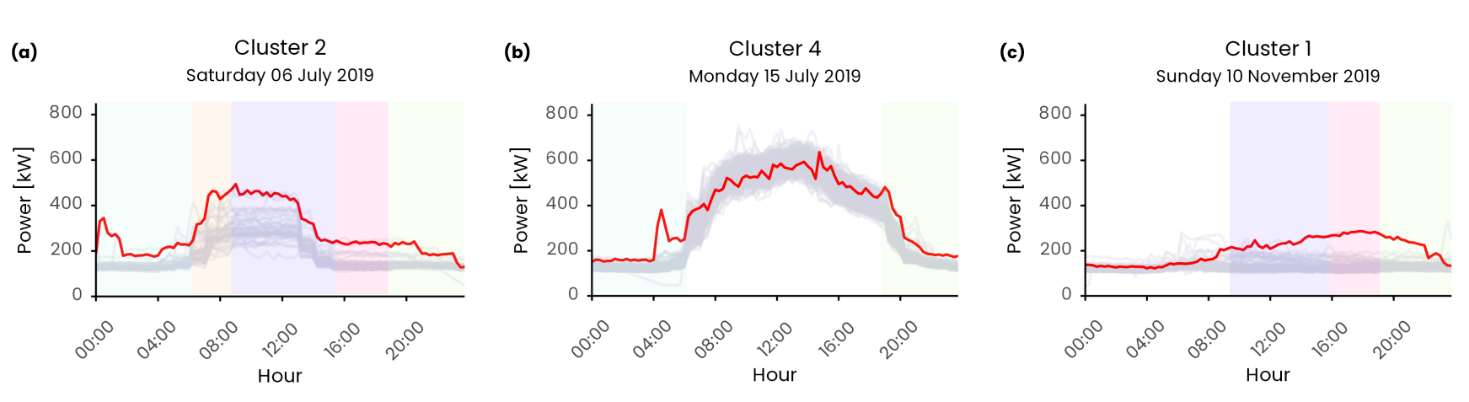
**Figure 12.** Calendar chart representing the distribution of anomalies through the whole year dataset. Each row corresponds to a different context. For each day the colour fill represents the severity of the anomaly while the symbol represents the cluster membership

An example of a spot anomaly is the second context of Friday 27th of December 2019 that was tagged as anomalous with a severity score of 8. Referring to Figure 13(a) this is a holiday day belonging to cluster 1 and a flat profile is expected, however a rise of the electrical load after 6:00 and an abrupt switch off at 9:00 is detected, resulting in an overconsumption of 260 kWh. The same pattern, shown in Figure 13 (b), is detected during summer season on Monday 12th of August 2019, where despite being a public vacation, as the university was closed for the summer break, it shows an abnormal increase in electrical load in the second context resulting into an overconsumption of 314 kWh. These two examples are symptoms of a wrong schedule of the energy systems pertaining to the substation C, that will be further discussed in the diagnosis phase.



**Figure 13.** Anomalous daily load profiles identified as spot anomalies, respectively: (a) Monday 12th August 2019 (context 2 cluster 1); (b) Friday 27th December 2019 (context 2 cluster 1). The anomalous load profile is represented with the red line while the grey lines correspond to the load profiles contained in the relative cluster. The vertical orange band denotes the second time window in which the anomaly was detected.

An example of persisting anomaly is Saturday 6th of July 2019 that presents an anomaly of severity 8 for the whole day. With reference to Figure 14 (a), starting from an unexpected peak during night hours, the daily load profile remains anomalous during the following time windows by keeping an offset of almost 80 kW compared with the cluster centroid, leading to an overall overconsumption compared to the cluster centroid of 2467 kWh at the end of the day. This behaviour is a symptom of energy systems running under unusual conditions, that may be related to a fault, a wrong schedule, or an exceptional outdoor boundary condition (e.g., high external temperature). A similar behaviour can be seen on Monday the 15th of July 2019 where a wrong schedule of start-up and switch-off results in anomalies during context 1 and 5 led to an energy surplus of 801 kWh compared to the average cluster energy consumption, see Figure 14 (b). Another example of persisting anomaly can be seen in Figure 14 (c) where during winter season on Sunday the 11th of November 2019 an anomalous energy consumption from context 3-4-5 is detected, from the morning start up to midnight, leading to an overconsumption of 1641 kWh.



**Figure 14.** Anomalous daily load profiles identified as spot anomalies, respectively: (a) Saturday 6th July 2019 (context 1-2-3-4-5 cluster 2); (b) Monday 15th July 2019 (context 1-5 cluster 2); (c) Sunday 16th June 2019 (context 3-4-5 cluster 1). The anomalous load profile is represented with the red line while the grey lines correspond to the load profiles contained in the relative cluster. The vertical band denotes the time window in which the anomaly was detected.

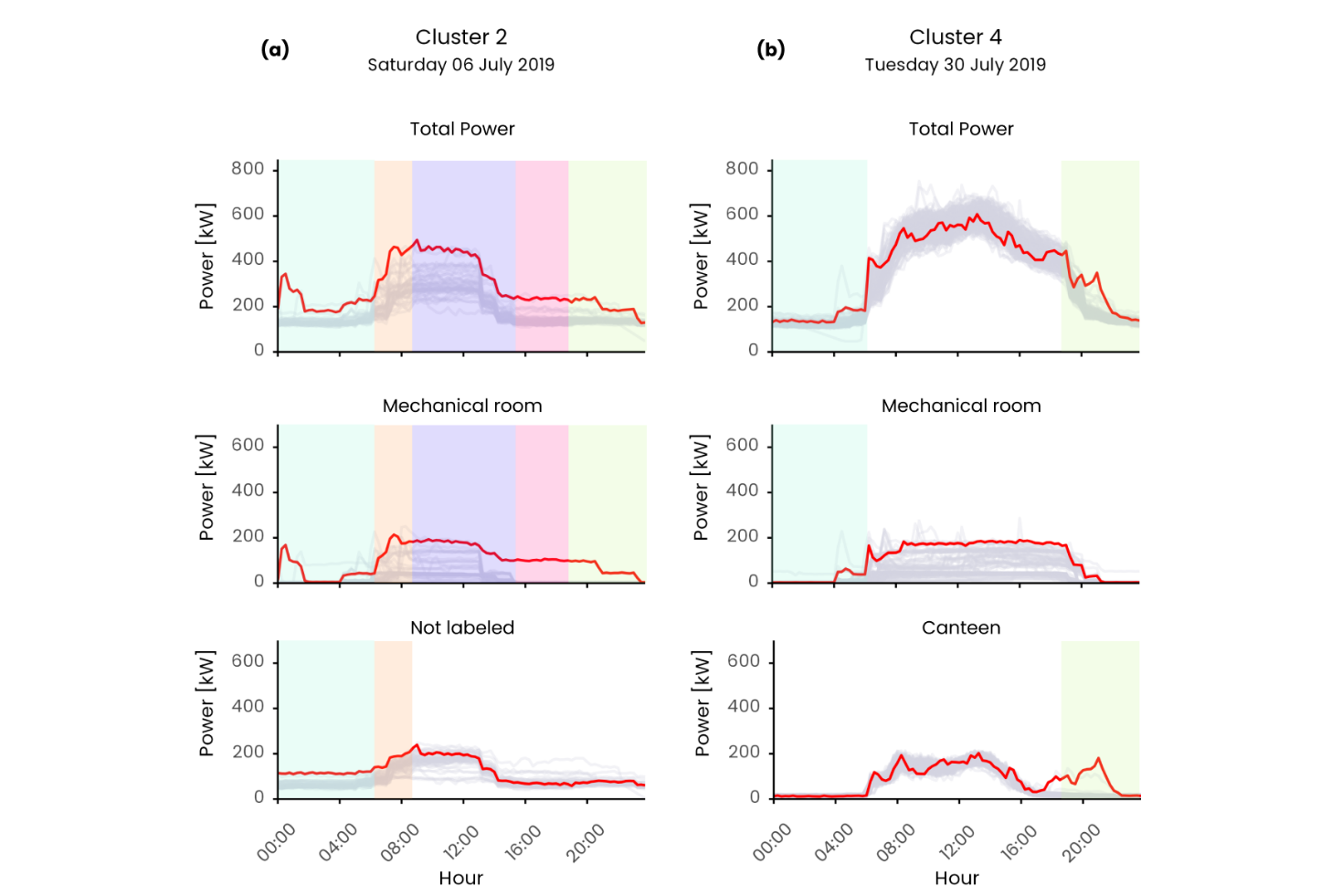
* 1. Anomaly diagnosis results

The diagnosis process allows to spot the sub-loads that are responsible of the anomalies calculating the severity score that ranges from 0 to 8. A sub-load with higher severity is likely to impact more on the meter-level anomaly that a sub-load with low severity. By investigating the diagnosis results obtained in the case study it is possible to see that “Mechanical room” and “Not labelled” loads were most frequently responsible for the detected anomalies, respectively 26 times and 18 times out of 64. Moreover, there are 5 times in which both sub-loads are equally responsible for the detected fault. In Figure 15 are shown two results in which a clear correlation between the anomaly detected at meter-level and the responsible sub-load identified by the diagnosis process.

Figure 15(a) shows Saturday 6th of July 2019 in which both the “Mechanical room” and “Not labelled” loads contribute to the fault detected on the total electrical load in context 1-2-3-4-5. For context 1-2 the “Mechanical room” and “Not labelled” are both responsible for the anomaly detected and they present severity of 8. The “Mechanical room” presents an unexpected switch-on/off during night hours (00:00 to 01:30) followed by a second switch-on at 04:00 that, superposed to the unusually high night load of the “Not labelled” lead to an abnormal behaviour at higher level. Then, for the remaining contexts 3-4-5 only the “Mechanical room” was identified as responsible with a severity of 8 while the “Not labelled” load shows no impact at meter-level (i.e., very low severity, less than 1). From the figure it can be see that the “Mechanical room” anomalous load after 12:00 lead to a positive bias of the “Total Power” load until the end of the day.

Figure 15(b) shows Tuesday 30th of July 2019 in which a single load contributes to the meter level anomaly detected. An early morning start-up at 4:00 of the “Mechanical room” identified by a severity 8 contributes to the “Total power” anomaly of severity 6 in the first context. While in context 5 an unexpected activity of the “Canteen” load after 19:00, tagged with severity 8, contributes to an anomaly of severity 8 at meter level.

It is very likely that the anomalous behaviour of the “Mechanical room” during night hours, the early start-up and late switch-off was caused by both incorrect operation and wrong schedule of the chillers. This kind of anomaly can be easily fixed and resolved by facility managers by reviewing the operational schedule. On the other hand, the correction of the “Not labelled” load can be more challenging since there is no detail on the electrical loads that contribute; it is therefore difficult to reduce energy consumption and further investigation is required.



**Figure 15.** Comparison between anomalous meter-level daily electrical load profile and responsible sub-load detected by the diagnostic process. The figure shows (a) Saturday 6th of July 2019 belonging to cluster 2 and (b) Tuesday 30th of July 2019 belonging to cluster 4. The different time windows (i.e., contexts) are represented as vertical bands and are present only if the relative load presents a severity of 8 in the relative context.

1. Discussion and conclusion

In this paper a meter-level anomaly detection and diagnosis process was proposed in order to demonstrate (i) the flexibility of the Matrix Profile algorithm in detecting anomalies and (ii) the potential of the integration of such tool in a building energy management process to gain insights on sub-loads unusual energy consumption, promoting the correction of faults and reduction of wastes. The methodology was based on unsupervised machine learning methods and timeseries analytics coupled with domain knowledge in order to detect unusual energy consumption patterns at whole building scale.

The capability to identify anomalies in specific sub-daily periods (i.e., time windows) represents an opportunity for the early-stage identification and prompt correction of incorrect operation that can help prevent energy wastes over time. Correctly defining the number of time windows and its length is a complex task and its wrong setting may negatively affect the capability to isolate important feature of the daily load profile such as the start-up or shut down. The CART model has proven to be flexible enough to adapt to different operational conditions of different buildings [55] thanks to the automatic self-tuning of the model. The introduction of the concept of context, i.e., the time interval before the time window in which a subsequence may start, adds an additional degree of freedom that allows for the comparison of subsequences that are delayed in time.

The methodology was tested on a static dataset but was conceptualized to work in semi-real time, by enabling the ADD process at the end of each time window. However, the practical implementation requires to tackle many technical challenges.

At first, the algorithm must follow the time constraint of the data stream, meaning that the execution time must be lower that the interval between two invocation of the procedure itself. The proposed methodology was not intended to be a pure real-time streaming process preformed upon every new incoming datapoint, rather it was conceived as a semi real- time batch process that is performed at the end of each time window; where for batch process is intended a process in which the whole timeseries must be available before calculating the MP [28]. Under this perspective, the methodology execution is triggered only when the whole subsequence is present (i.e., at the end of the time window). In the analysed case study the minimum interval of time between subsequent triggers is 2.5 hours (i.e., the length of the smallest time window) which is far more that the actual execution time, that considering the CMP calculation on all the timeseries (meter level and sub-meter level) on the offline one-year dataset is estimated to be less than 10 minutes. However, we recognize that the streaming approach is more desirable and would enable an earlier recognition of anomalies. The computational burden could be reduced performing the CMP calculation by using an incremental approach like the STAMPI [28] algorithm that adjusts the CMP rather than re-compute it. Following this approach at the end of each time window, the observation is added to the corresponding CMP and the CMP values would be adjusted accordingly.

The second challenge is the cold start problem, intended as scarcity of initial data useful to perform an accurate definition of time windows, contexts, clusters and CMP calculation. One advantage of the matrix profile algorithm is that it does not require a minimum length of the timeseries to be calculated. However, since the MP objective is to perform similarity join to discover frequent and infrequent patterns within the timeseries, with a small dataset it may fail to recognize patterns. We would recommend performing the analysis on a dataset containing at least 4 weeks of observations which represents a minimum set of daily load profiles on which it is possible to perform a more robust discord/motif discovery. With respect to the initial hyperparameter definition, it can be easily performed based on domain knowledge, defining a reasonable context length and by defining clusters in a supervised way. As a first approximation it is possible to define time windows by simply split the 24h into N non overlapping time windows of fixed length. This approach was followed by [61] in the definition of SAX parameters for the daily load pattern filtering methodology, where after a sensitivity analysis the suggested N was between 3 and 4. Moreover, the initial number of clusters can be defines in a supervised way based on the weekly operational calendar. For example, a possible initial configuration may consist in 3 clusters (weekday, Saturday and Sundays/holidays), 4 fixed length time windows of 6 hours and a context of 1 hour.

Along with the data scarcity for the matrix profile algorithm even the abundance of data may represent a critical issue, not only under the computational point of view but mainly from the conceptual one. Frequent and rare subsequences in the original concept of MP are defined as the ones with smallest/largest 1st nearest neighbour distance [28]. This implies that if a rare subsequence occurs more than once in the timeseries it may be considered as common or even frequent [62]. A repeated anomaly may would cause false negatives due to the previous anomaly instance being part of all sub- sequence set. And this issue is recognized as the *twin freak* problem. To address the twin freak problem is of paramount importance in building energy management since an anomaly if not promptly detected may persists in time and must not be considered as a frequent pattern. A possible nontrivial solution is the introduction of the kNN distance instead of 1NN distance. The algorithm implementation was discussed in [62] where the authors proposed a density based approach for the kNN calculation applied on the MP algorithm. Another way to address the twin freak problem is to adjust the anomaly detection capability of the methodology by dynamically adapt parameters and weights, based on human users’ feedbacks to report anomaly using a human-in-the-loop training scheme.

Given the variety of building operation conditions and its variability throughout the day according to multiple variables such as the buildings occupancy, users behaviour and external boundary conditions it is necessary to assist buildings energy managers during operation through an easy and effective tool. To this aim the output of the process at each time window has been represented with a severity ranging from 6 to 8 coupled with an effective visualization that allows at a glance to compare the detected behaviour to the expected in the given context and boundary condition and thanks to this holistic view to have an immediate perception of anomalous trends that deviates to the expected behaviour.

References

[1] Directorate-General for Energy (European Commission), “Clean energy for all Europeans - Publications Office of the EU,” *Publ. Off. EU*, vol. 14, no. 2, p. 3, 2019, doi: 10.2833/9937.

[2] T. Ramesh, R. Prakash, and K. K. Shukla, “Life cycle energy analysis of buildings: An overview,” *Energy Build.*, vol. 42, no. 10, pp. 1592–1600, 2010, doi: 10.1016/j.enbuild.2010.05.007.

[3] L. Erhan *et al.*, “Smart anomaly detection in sensor systems: A multi-perspective review,” *Inf. Fusion*, vol. 67, no. October 2020, pp. 64–79, 2021, doi: 10.1016/j.inffus.2020.10.001.

[4] M. Molina-Solana, M. Ros, M. D. Ruiz, J. Gómez-Romero, and M. J. Martin-Bautista, “Data science for building energy management: A review,” *Renewable and Sustainable Energy Reviews*, vol. 70. pp. 598–609, 2017, doi: 10.1016/j.rser.2016.11.132.

[5] H. Kramer, G. Lin, J. Granderson, C. Curtin, and E. Crowe, “Synthesis of Year One Outcomes in the Smart Energy Analytics Campaign Building Technology and Urban Systems Division,” 2019.

[6] J. Granderson and G. Lin, “Building energy information systems: synthesis of costs, savings, and best-practice uses,” *Energy Effic.*, vol. 9, no. 6, pp. 1369–1384, 2016, doi: 10.1007/s12053-016-9428-9.

[7] A. Capozzoli, T. Cerquitelli, and M. S. Piscitelli, *Enhancing energy efficiency in buildings through innovative data analytics technologies*. 2016.

[8] P. Tan, *Introduction to data mining*, vol. 12. 2011.

[9] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, “Anomaly detection of energy consumption in buildings: A review, current trends and new perspectives,” no. ii, 2020, [Online]. Available: http://arxiv.org/abs/2010.04560.

[10] J. E. Zhang, D. Wu, and B. Boulet, “Time Series Anomaly Detection for Smart Grids: A Survey,” *2021 IEEE Electr. Power Energy Conf. EPEC 2021*, pp. 125–130, 2021, doi: 10.1109/EPEC52095.2021.9621752.

[11] N. R. Prasad, S. Almanza-Garcia, and T. T. Lu, “Anomaly detection: A Survey,” *Comput. Mater. Contin.*, vol. 14, no. 1, pp. 1–22, 2009, doi: 10.1145/1541880.1541882.

[12] H. Chen *et al.*, “Energy Consumption Data Based Machine Anomaly Detection,” *Proc. - 2014 2nd Int. Conf. Adv. Cloud Big Data, CBD 2014*, no. February, pp. 136–142, 2015, doi: 10.1109/CBD.2014.24.

[13] Y. Zhao, S. Wang, and F. Xiao, “Pattern recognition-based chillers fault detection method using Support Vector Data Description (SVDD),” *Appl. Energy*, vol. 112, pp. 1041–1048, 2013, doi: 10.1016/j.apenergy.2012.12.043.

[14] J. Pereira and M. Silveira, “Unsupervised Anomaly Detection in Energy Time Series Data Using Variational Recurrent Autoencoders with Attention,” *Proc. - 17th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2018*, pp. 1275–1282, 2019, doi: 10.1109/ICMLA.2018.00207.

[15] K. Kammerer, B. Hoppenstedt, R. Pryss, S. Stökler, J. Allgaier, and M. Reichert, “Anomaly detections for manufacturing systems based on sensor data—insights into two challenging real-world production settings,” *Sensors (Switzerland)*, vol. 19, no. 24, 2019, doi: 10.3390/s19245370.

[16] R. Chiosa, A. Capozzoli, and M. S. Piscitelli, “A Data Analytics-Based Energy Information System ( EIS ) Tool to Perform Meter-Level Anomaly Detection and Diagnosis in Buildings,” *Energies*, vol. 14, no. 237, pp. 1–28, 2021, [Online]. Available: https://doi.org/10.3390/en14010237.

[17] J. S. Chou and A. S. Telaga, “Real-time detection of anomalous power consumption,” *Renew. Sustain. Energy Rev.*, vol. 33, pp. 400–411, 2014, doi: 10.1016/j.rser.2014.01.088.

[18] M. S. 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,” *Build. Simul.*, pp. 1–17, 2020, doi: 10.1007/s12273-020-0650-1.

[19] C. Fan, F. Xiao, Y. Zhao, and J. Wang, “Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data,” *Appl. Energy*, vol. 211, no. December 2017, pp. 1123–1135, 2018, doi: 10.1016/j.apenergy.2017.12.005.

[20] R. Wu and E. Keogh, “Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress,” *IEEE Trans. Knowl. Data Eng.*, pp. 1–9, 2021, doi: 10.1109/TKDE.2021.3112126.

[21] A. Capozzoli, M. S. 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, 2017, doi: 10.1016/j.egypro.2017.09.545.

[22] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, “A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis,” *Energy Built Environ.*, vol. 1, no. 2, pp. 149–164, 2020, doi: 10.1016/j.enbenv.2019.11.003.

[23] C. Fan, F. Xiao, and S. Wang, “Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques,” *Appl. Energy*, vol. 127, pp. 1–10, 2014, doi: 10.1016/j.apenergy.2014.04.016.

[24] K. Li, Z. Ma, D. Robinson, and J. Ma, “Using Evidence Accumulation-Based Clustering and Symbolic Transformation to Group Multiple Buildings Based on Electricity Usage Patterns,” in *Sustainability in Energy and Buildings*, 2019, pp. 61–67, doi: 10.1007/978-981-32-9868-2.

[25] B. Rossi, S. Chren, B. Buhnova, and T. Pitner, “Anomaly Detection in Smart Grid Data: An Experience Report,” no. 1, pp. 2313–2318, 2016.

[26] 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,” *Appl. Energy*, vol. 287, no. November 2020, p. 116601, 2021, doi: 10.1016/j.apenergy.2021.116601.

[27] C. Zhang and F. Wang, “Multi-feature fusion based anomaly electro-data detection in smart grid,” *Proc. - 2018 15th Int. Symp. Pervasive Syst. Algorithms Networks, I-SPAN 2018*, pp. 54–59, 2019, doi: 10.1109/I-SPAN.2018.00018.

[28] C.-C. M. Yeh *et al.*, “Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets,” pp. 1317–1322, 2017, doi: 10.1109/icdm.2016.0179.

[29] C. C. M. Yeh, “Towards a Near Universal Time Series Data Mining Tool: Introducing the Matrix Profile,” University of California, Riverside, 2018.

[30] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, “Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases,” *Knowl. Inf. Syst.*, vol. 3, no. 3, pp. 263–286, 2001, doi: 10.1007/pl00011669.

[31] H. Ren, M. Liu, Z. Li, and W. Pedrycz, “A Piecewise Aggregate pattern representation approach for anomaly detection in time series,” *Knowledge-Based Syst.*, vol. 135, pp. 29–39, 2017, doi: 10.1016/j.knosys.2017.07.021.

[32] S. Alaee, K. Kamgar, and E. Keogh, “Matrix profile XXII: Exact discovery of time series motifs under DTW,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2020-Novem, no. Icdm, pp. 900–905, 2020, doi: 10.1109/ICDM50108.2020.00099.

[33] S. Gharghabi, S. Imani, A. Bagnall, A. Darvishzadeh, and E. Keogh, *An ultra-fast time series distance measure to allow data mining in more complex real-world deployments*, vol. 34, no. 4. Springer US, 2020.

[34] Y. Zhu *et al.*, *The Swiss army knife of time series data mining: ten useful things you can do with the matrix profile and ten lines of code*, vol. 34, no. 4. Springer US, 2020.

[35] A. Mueen *et al.*, “The Fastest Similarity Search Algorithm for Time Series Subsequences under Euclidean Distance.” Aug. 2017.

[36] C. Onwongsa and C. Ratanamahatana, “An enhanced time series motif discovery using approximated matrix profile,” *ACM Int. Conf. Proceeding Ser.*, pp. 180–189, 2020, doi: 10.1145/3421558.3421586.

[37] C. C. M. Yeh, N. Kavantzas, and E. Keogh, “Matrix profile VI: Meaningful multidimensional motif discovery,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2017-Novem, pp. 565–574, 2017, doi: 10.1109/ICDM.2017.66.

[38] M. Alshaer, S. Garcia-Rodriguez, and C. Gouy-Pailler, “Detecting Anomalies from Streaming Time Series using Matrix Profile and Shapelets Learning,” *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 2020-Novem, pp. 376–383, 2020, doi: 10.1109/ICTAI50040.2020.00066.

[39] S. D. D. Anton and H. D. Schotten, “Intrusion Detection in Binary Process Data: Introducing the Hamming-distance to Matrix Profiles,” *Proc. - 21st IEEE Int. Symp. a World Wireless, Mob. Multimed. Networks, WoWMoM 2020*, pp. 347–353, 2020, doi: 10.1109/WoWMoM49955.2020.00065.

[40] F. Madrid, S. Imani, R. Mercer, Z. Zimmerman, N. Shakibay, and E. Keogh, “Matrix profile XX: Finding and visualizing time series motifs of all lengths using the matrix profile,” *Proc. - 10th IEEE Int. Conf. Big Knowledge, ICBK 2019*, pp. 175–182, 2019, doi: 10.1109/ICBK.2019.00031.

[41] J. Dinal Herath, C. Bai, G. Yan, P. Yang, and S. Lu, “RAMP: Real-Time Anomaly Detection in Scientific Workflows,” *Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019*, pp. 1367–1374, 2019, doi: 10.1109/BigData47090.2019.9005653.

[42] D. De Paepe, D. N. Avendano, and S. Van Hoecke, “Implications of Z-Normalization in the Matrix Profile,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11996 LNCS, no. 768869, pp. 95–118, 2020, doi: 10.1007/978-3-030-40014-9\_5.

[43] D. De Paepe, O. Janssens, and S. Van Hoecke, “Eliminating noise in the matrix profile,” *ICPRAM 2019 - Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods*, pp. 83–93, 2019, doi: 10.5220/0007314100830093.

[44] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, “Unsupervised real-time anomaly detection for streaming data,” *Neurocomputing*, vol. 262, pp. 134–147, 2017, doi: 10.1016/j.neucom.2017.04.070.

[45] C. Nichiforov, G. Stamatescu, I. Stamatescu, and I. Fagarasan, “Learning Dominant Usage from Anomaly Patterns in Building Energy Traces,” *IEEE Int. Conf. Autom. Sci. Eng.*, vol. 2020-Augus, pp. 548–553, 2020, doi: 10.1109/CASE48305.2020.9216794.

[46] C. Miller and F. Meggers, “The Building Data Genome Project: An open, public data set from non-residential building electrical meters,” *Energy Procedia*, vol. 122, pp. 439–444, 2017, doi: 10.1016/j.egypro.2017.07.400.

[47] J. Y. Park, E. Wilson, A. Parker, and Z. Nagy, “The good, the bad, and the ugly: Data-driven load profile discord identification in a large building portfolio,” *Energy Build.*, vol. 215, p. 109892, 2020, doi: 10.1016/j.enbuild.2020.109892.

[48] D. De Paepe *et al.*, “A generalized matrix profile framework with support for contextual series analysis,” *Eng. Appl. Artif. Intell.*, vol. 90, 2020, doi: 10.1016/j.engappai.2020.103487.

[49] H. A. Dau and E. Keogh, “Matrix profile V: A generic technique to incorporate domain knowledge into motif discovery,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. Part F1296, pp. 125–134, 2017, doi: 10.1145/3097983.3097993.

[50] G. E. A. P. A. Batista, E. J. Keogh, O. M. Tataw, and V. M. A. De Souza, “CID: An efficient complexity-invariant distance for time series,” *Data Min. Knowl. Discov.*, vol. 28, no. 3, pp. 634–669, 2014, doi: 10.1007/s10618-013-0312-3.

[51] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, “A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data,” *Front. Energy Res.*, vol. 9, no. March, pp. 1–17, 2021, doi: 10.3389/fenrg.2021.652801.

[52] J. L. Mathieu, P. N. Price, S. Kiliccote, and M. A. Piette, “Quantifying changes in building electricity use, with application to demand response,” *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 507–518, 2011, doi: 10.1109/TSG.2011.2145010.

[53] J. Zhu, Y. Shen, Z. Song, D. Zhou, Z. Zhang, and A. Kusiak, “Data-driven building load profiling and energy management,” *Sustain. Cities Soc.*, vol. 49, no. March, p. 101587, 2019, doi: 10.1016/j.scs.2019.101587.

[54] R. Chiosa, M. S. Piscitelli, and A. Capozzoli, “Detection and diagnosis of anomalous energy consumption patterns in buildings through a data analytics based approach The case of Politecnico di Torino,” Politecnico di Torino, 2020.

[55] A. Capozzoli, M. S. 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, 2018, doi: 10.1016/j.energy.2018.05.127.

[56] R. Yan, Z. Ma, Y. Zhao, and G. Kokogiannakis, “A decision tree based data-driven diagnostic strategy for air handling units,” *Energy Build.*, vol. 133, pp. 37–45, 2016, doi: 10.1016/j.enbuild.2016.09.039.

[57] T. Grubinger, A. Zeileis, and K. P. Pfeiffer, “Evtree: Evolutionary learning of globally optimal classification and regression trees in R,” *J. Stat. Softw.*, vol. 61, no. 1, pp. 1–29, 2014, doi: 10.18637/jss.v061.i01.

[58] R Core Team, “R: A Language and Environment for Statistical Computing.” Vienna, Austria, 2017, [Online]. Available: https://www.r-project.org/.

[59] G. Van Rossum and F. L. Drake Jr, “Python reference manual.” Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[60] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, “NbClust : An R Package for Determining the,” *J. Stat. Softw.*, vol. 61, no. 6, pp. 1–36, 2014, doi: 10.18637/jss.v061.i06.

[61] C. Miller, Z. Nagy, and A. Schlueter, “Automated daily pattern filtering of measured building performance data,” *Autom. Constr.*, vol. 49, no. PA, pp. 1–17, 2015, doi: 10.1016/j.autcon.2014.09.004.

[62] Y. He, X. Chu, and Y. Wang, “Neighbor profile: Bagging nearest neighbors for unsupervised time series mining,” *Proc. - Int. Conf. Data Eng.*, vol. 2020-April, no. April, pp. 373–384, 2020, doi: 10.1109/ICDE48307.2020.00039.

1. \* Corresponding author

   Email address: alfonso.capozzoli@polito.it (Alfonso Capozzoli)

   URL: www.baeda.polito.it (Alfonso Capozzoli) [↑](#footnote-ref-2)