Meter Level Electrical Load Anomaly Detection using Contextual Matrix Profile

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

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 increasingly 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 operations. 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 the building operation [2], supporting building owners and energy managers to extract valuable knowledge from collected data to reduce energy consumption, increase appliances efficiency, prevent energy wastes and in general operate their buildings more efficiently is of paramount importance.

Although a great deal of research has been done, the increasing volume of collected building energy data still overwhelms end-users, 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. In the last few years data gathered in the building sector reached the order of zettabyte [3] making buildings not only energy intensive but information intensive [4]. 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, provide the opportunity to gain insight on the building operational behaviour discovering opportunities for savings [5].

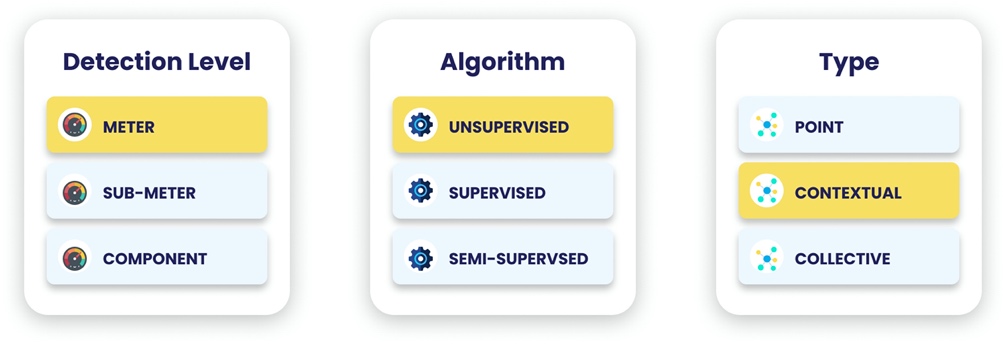
A robust coupling of IoT sensors data, machine learning approaches and energy domain has been proved to be effective in terms of energy savings opportunities to variety of tasks: pattern recognition, energy consumption forecasting, anomaly detection and diagnosis, advanced benchmarking, load profiling, and schedule optimization of building energy systems [6].

In this paper, we focus anomaly detection of electrical loads in buildings, which is a key application to aid decision makers to reduce wastes and promote sustainable behaviour of end users.

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

In the energy and buildings field, anomaly detection can be employed to detect abnormal behaviour of end users, to detect the faulty operation of appliances or energy systems and spotting technical and non-technical energy loss [7]. Strictly speaking, 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 it strongly depend on the analysis performed and the intended application. When performing anomaly detection on buildings energy consumptions to detect abnormal occupation patterns, wrong occupants behaviour, incorrect functioning of energy systems, abnormal sub loads consumption and so on [7], is of paramount importance to take into consideration other information sources related to the internal and external environmental conditions, level of detection, occupancy patterns, and join the domain knowledge.

Many categorization have been proposed in literature [9] and some are specific for building environments [3], [7]. The scope of this paper is not to go deep into categorization; thus, we adopted an anomaly classification based on type, level and algorithm as reported in Figure 1.



**Figure 1.** Classification of anomaly detection method depending on: (a) detection level (b) algorithm (c) anomaly type.

Classification based on type implies a comparison between the observation and the rest of the data. A *point anomaly* is one individual instance or observation that can be considered anomalous when compared to the remaining data. On the other side, a *collective anomaly* is an instance does not represent an anomaly per se, but only if considered within the collection of all the other events instances. 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.

Depending on the detail of electrical load monitored the anomaly detection can be performed at different levels. The *meter level* detection analyses the whole building electrical load, without having any information on the disaggregation of that load among the different sub loads or appliances. *Sub-meter level* detection analyses the disaggregated total electrical load and is usually referred to a specific energy system. Finally, *component level* detection consists in identifying anomalies referring to a given appliance/sensor.

The third is an algorithmic centric classification is based on data-driven anomaly detection techniques. *Supervised* anomaly detection requires to train a machine learning algorithm using labelled dataset (i.e., ground truth) to classify anomalous consumption or not. Although supervised anomaly detection can achieve high-accuracy identification results as demonstrated in academic frameworks, its adoption in real-world is still limited compared to unsupervised methods, mainly due to the absence of a reliable power consumption annotated datasets [7], [10]. Examples of supervised algorithms are deep learning, ANN, Regression, Probabilistic models, Traditional classification. On the other side, *unsupervised* anomaly detection consists in detecting rare and unknown anomalous energy patterns without any a priori knowledge. It usually consists in modelling the normal behaviour and then identify patterns that deviates, under the assumption that the number of anomalies is low compared to the observations. Examples of unsupervised algorithms are: … clustering, [11] performs anomaly detection on smart grid though the use of clustering. Finally, there are some semi-supervised algorithms that.

* 1. Matrix profile method for anomaly detection

One of the most promising techniques for unsupervised anomaly detection in timeseries is Matrix Profile (MP). Introduced by [12] it 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 [13]. Others perform similarity search by reducing the dimensionality of dataset through PAA ﻿[14], [15] to speed up computation, however, this method causes loss of valuable information.

Conversely, MP proposes an ultra-fast similarity search under the z-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 side 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 discovers the most unique subsequences in a dataset. Discord discovery using MP as anomaly detection method has been employed with success in different fields.

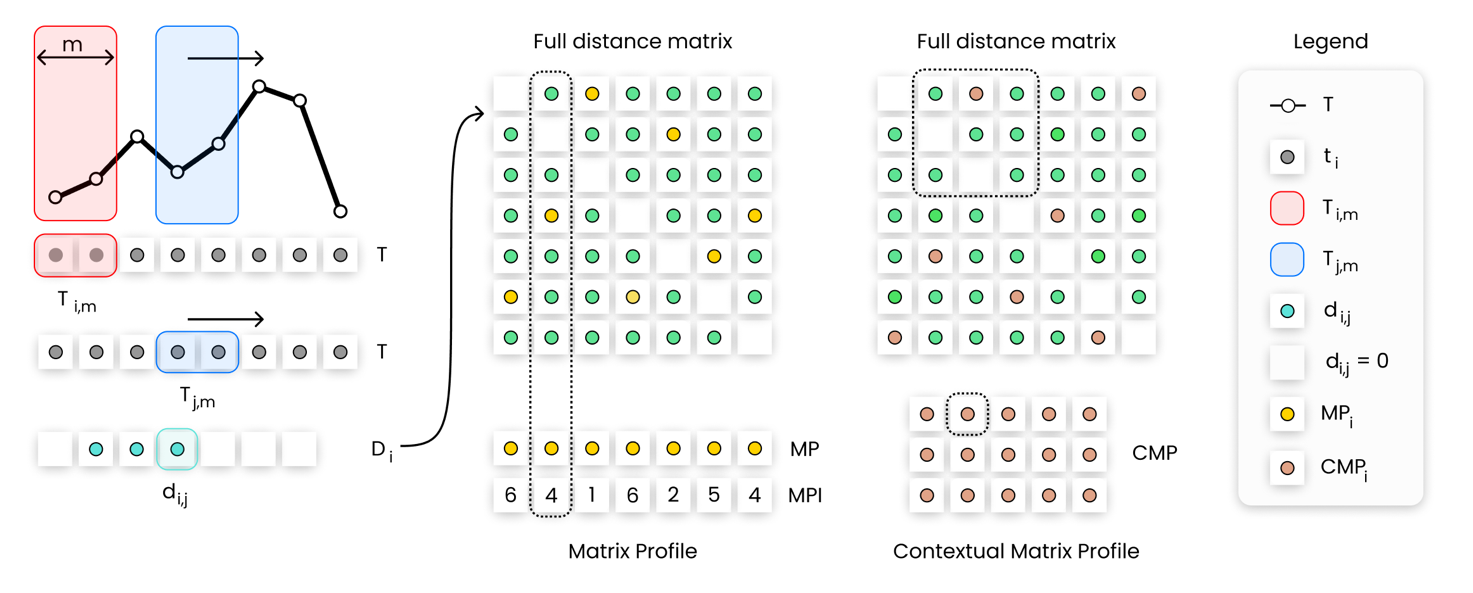
With reference to Figure 4 some fundamental concepts and definitions need to be introduced before going deeper into the topic. First, a *timeseries* is a sequence of real-valued numbers with where is the length of . Since the focus is on local properties of timeseries (i.e., portion of timeseries) a *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 [16], [17], [18] 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 m, 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 4, 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 [19], approximated AMPSA and AMP [20] and multidimensional mSTAMP [21].



**Figure 2.** 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 have been successfully applied in different fields for 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 [22]. 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 [23] 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 Pan MP is proposed in [24] to find different length anomalies in ﻿automated pedestrian counting system developed in Taipei. MP has been largely employed to identify anomalies in IT field. [25] introduces ﻿a real time anomaly detection framework based on MP called Real-Time Aggregated Matrix Profile (RAMP), that can identify anomalies in scientific workflows. [26] Applies a noise elimination technique based on MP on real Yahoo! internet traffic metrics to detect anomalous behaviours; [27] demonstrate how the elimination of noise can help in anomaly detection of noisy date by testing the algorithm on Numenta Benchmark [28].

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 [29]. The authors applied the classical MP approach with daily, weekly, and monthly time window lengths to identify top discords in energy consumption timeseries of a large commercial building dataset on one year period. The process was tested on 422 buildings [30] with primary use type classrooms, offices, laboratories and dormitory. The classic MP method is applied and resulted to be effective to get insights and label the unusual behaviour by providing a sufficient differentiation between dominant usage patterns of the analysed dataset. In [18] the authors 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. [31] applies 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 is used to quantify the similarities of daily subsequences in timeseries meter data under z-normalized distance. The computed MP values are then compared with typical-day MP distribution and was proven to be effective to identify unique load shapes patterns and discords.

Despite being proven effective; the original MP method compares regions of timeseries that belongs to different context or operating conditions or different boundary conditions and may result into misleading results. To address this issue, [32] introduced the Contextual Matrix Profile (CMP), defined as the minimum over rectangular regions of the full distance matrix (see Figure 4), allowing to find the best matching subsequence in ranges over and allowing 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. For example, given a timeseries of 365 days, with 15-min frequency, by setting a context of from 5:00 to 6:00 and a subsequence length , when computing a row/column of the CMP, the distance between the nearest neighbour between five subsequences starting in the given context (i.e., starting at 5:00, 5:15, 5:30, 5:45, 6:00) of a given day with all the subsequences of the context of another day is calculated. The resulting CMP will have 365 rows/columns﻿ where each point displays the distance between the best matching 2h 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.

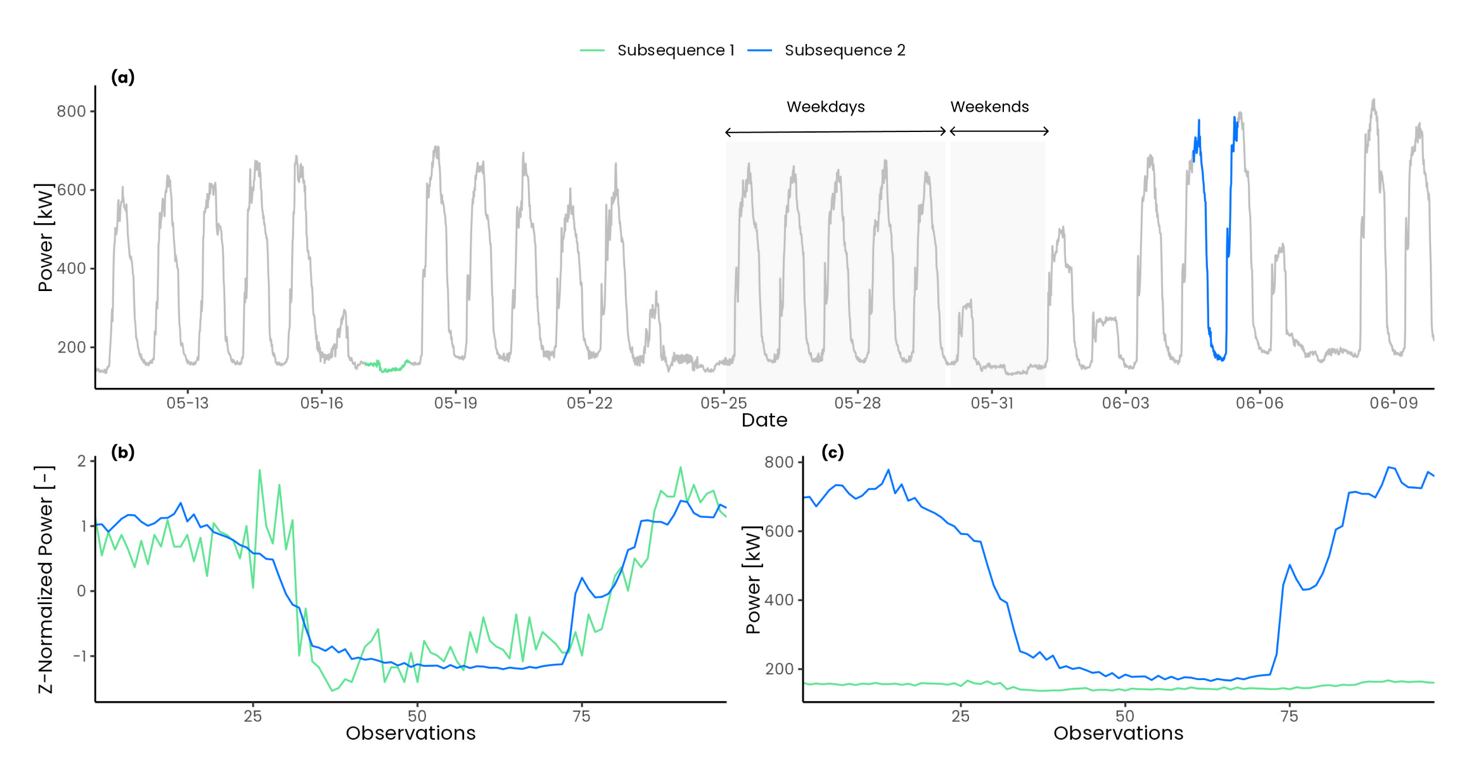
A remarkable application of the CMP method for anomaly detection in indoor buildings environment is presented in [32] where the method is applied to a dataset containing different indoor air quality measurement (temperature, humidity, CO2 etc.) referring to residential buildings. CMP is applied on the CO2 concentration timeseries, defining a 2-hour context and subsequence length of 3 hours and weekends were grouped. In this way it was possible to compare only weekend morning subsequences and identify 6 anomalies through the elbow method. Moreover, the authors were able to identify a periodicity in behavioural patterns even if not aligned in time.

This case study demonstrates 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

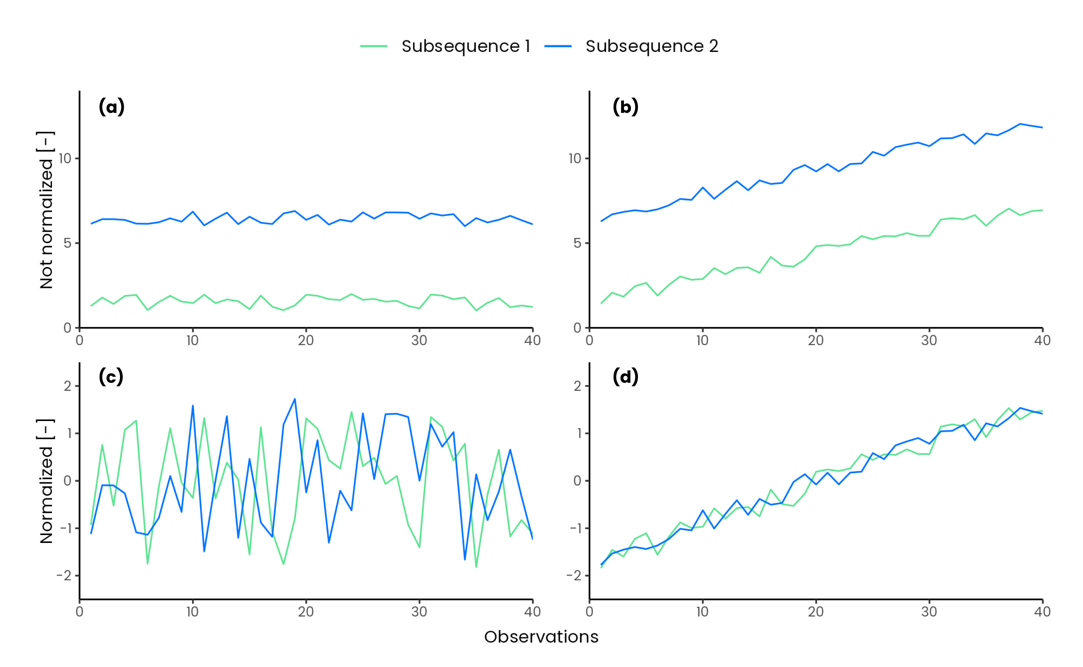
From the previous literature review it emerges that the MP method have been employed successfully in different fields for anomaly detection and the authors have proposed different implementations according to the field of interest. In fact, even if MP is an unsupervised method useful for discord discovery every field have constraint and peculiar boundary conditions that cannot be overlooked. In the field of energy and buildings energy consumption timeseries are strictly correlated to many different variables such as occupation, weather conditions, energy systems and so on. A completely unsupervised method may fail to consider the relation with those variables and extract ineffective or trivial results, not useful for anomaly detection. In the following paragraphs some examples are presented

In buildings, anomalies are defined as unexpected behaviours that result in an atypical energy consumption. The classic MP, by performing with z-score normalization, searches for each subsequence the nearest neighbour based on shape similarity, however, anomalous shapes not always correspond to anomalous energy consumption, as well as similar shapes in z-score not always reflect similar behaviour. Figure 2(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 dramatically from weekdays to weekend days when the load profile is almost flat. Applying the classic MP method with a subsequence length of one day (i.e., 96 observations), the two sub-sequences highlighted respectively in blue and green are identified as nearest neighbour. As shown in Figure 2(b) under z-score normalization they are almost overlapping. However, from Figure 2(c) it is possible to see that the not normalized sub-sequences have very different amplitudes, finally they also refer to completely distinct energy consumption patterns the first to a weekend day and the second to a weekday. This is a clear example of how the unsupervised and algorithm and the subsequences normalization led to completely wrong results.



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

Z-normalization not only ignores magnitude effects of subsequences but also tends to enhance any fluctuation and noise of the timeseries data. By comparing two relatively flat subsequences under z-score normalization the resulting Euclidean distance is higher compared to non-flat subsequences, these results into higher values of MP in flat regions of the timeseries. In Figure 3 a comparison between two synthetics random timeseries is shown. In Figure 3(a) the two timeseries are relatively flat and noisy while in Figure 3(b) the two timeseries present a positive slope. While calculating the Euclidean distance between the z-normalized subsequences in the first case shown in Figure 3(c) the effect of noise is enhanced resulting into a higher Euclidean distance (d = 9.25) while in the second case shown in Figure 3(d) the Euclidean distance is much lower (d = 1.5). This issue have been largely analysed in [26] where a smoothing is proposed as possible solution to this issue, beside the trivial solutions of discard flat regions or change the subsequence length. Referring to Figure 2(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 issue when dealing with electrical load timeseries that by their nature present different patterns.



**Figure 4.** Effect of z-score normalization of relatively flat subsequences.

Comparing two subsequences belonging to different energy pattern would be unfair and misleading, therefore, introducing domain knowledge to find discords only in sone subgroups of the timeseries became of paramount importance. The concept of Annotation Vector (AV) [33] is used to introduce domain knowledge in the process of motif and discord discovery, which allows to find results that follows users defined constraint and produce better results, closer to expectations of the analyst. Annotation vector is a meta timeseries used to correct a posteriori the values of the original matrix profile 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 [34]. 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 [32] where Contextual Matrix Profile (CMP) algorithm permits to define ranges along two timeseries and look for the best matching subsequence among these ranges. This permits different a priori grouping of the timeseries observations so that MP calculation can provide novel and more interesting insights.

The prompt and accurate discovery of anomalies in building electrical load 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. 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 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 related to 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 fairer comparison between subsequences.
3. Introduction of a robust anomaly score definition based on four different statistical methods, majority voting and domain knowledge that permits to discriminate and rank potential anomalies, within a given group and context, by considering only those that negatively impact the energy use (i.e., high energy consumption)

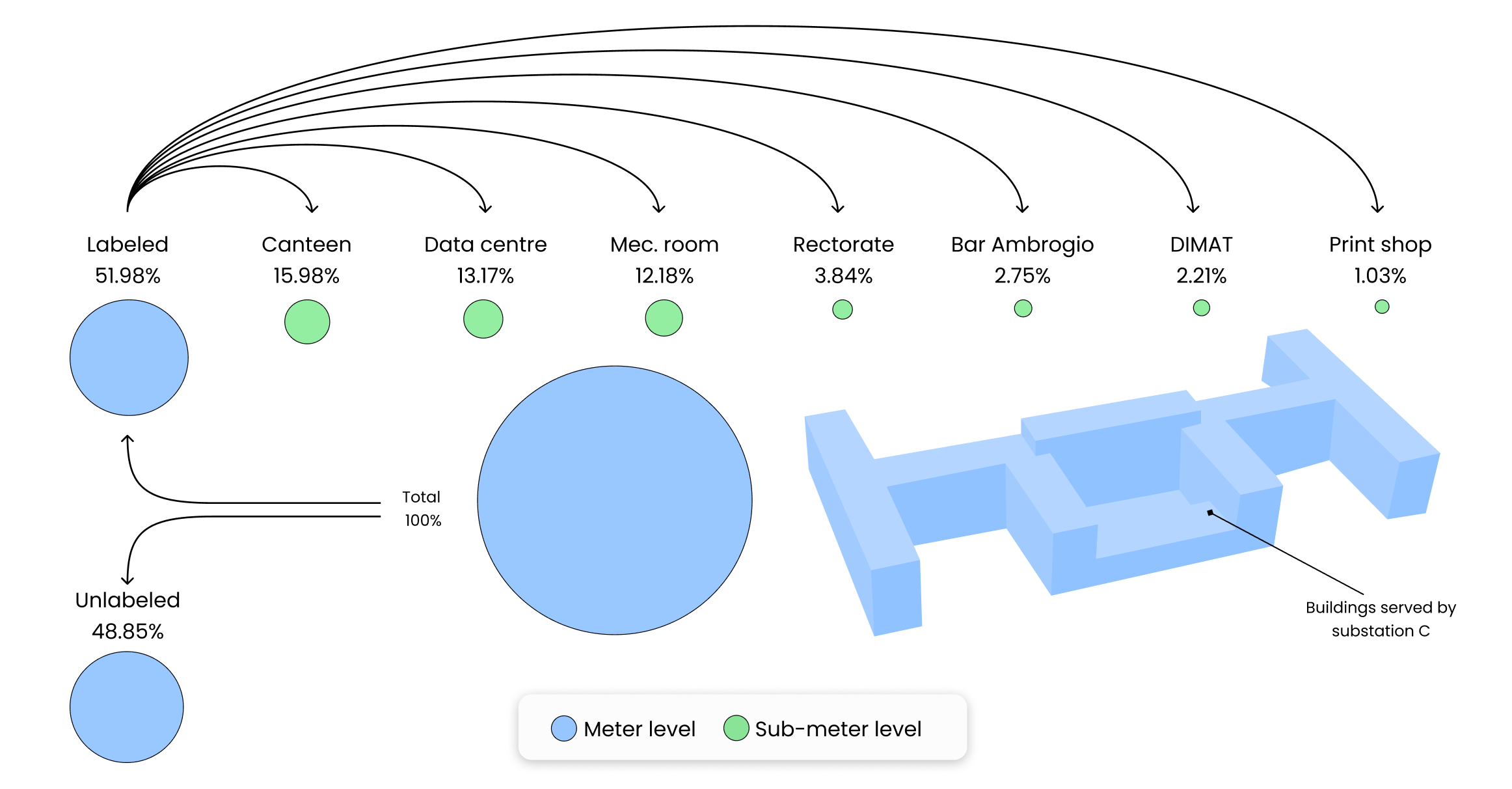
The rest of the paper is organized as follows. Section 2 provides the description of the case study. Section 3 presents the methodology adopted. Finally, Section 4 presents the results and Section 5 critically discusses the outcomes and contains the concluding remarks.

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 FigureXXX: The first level refers to the total electrical load of substation C, while the second level 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 canteen were at the disposal of students and campus staff and accounted for 2.75% and 16.03%, respectively, of the total electrical energy consumption of substation C. The university data centre accounted for 13.16% of the total energy consumption. The administration offices (Rectory) 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.

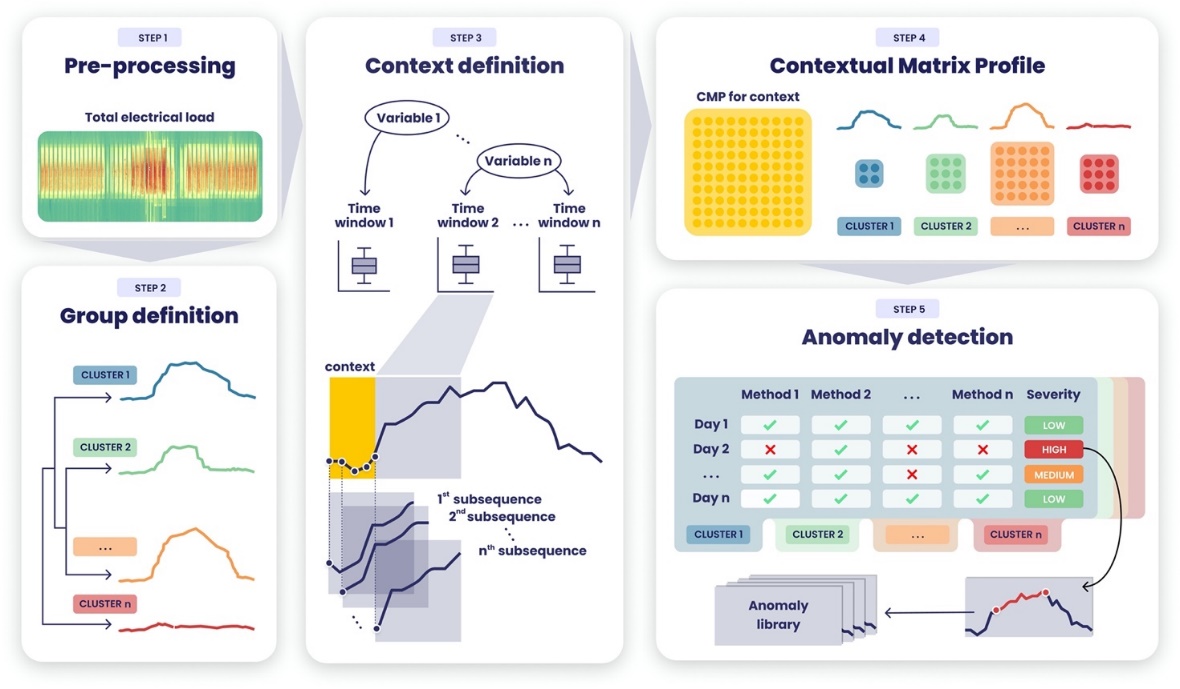
The remaining energy consumption was aggregated under a unique instance tagged as “Unlabelled\_load” as showed in FigureXXX. It accounted for 48.76% of the total energy consumption, and since it was not directly measured, cannot be assigned to a specific sub-load.



**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 techniques such as clusters and CART to perform anomaly detection on electrical load timeseries in the most parameter-free and automatic way. The multi-step procedure, reported in Figure 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 is a crucial task for the data analysis workflow. The proposed methodology does not focus on advanced pre-processing techniques since the dataset is assumed to have a good quality with a missing values and inconsistence ratio less than 5% on the overall observations [35]. Thus, the pre-processing is performed through univariate statistical approaches, in particular inconsistences removal and missing values imputations through linear interpolation.

* 1. Contextual matrix profile.

The application of the contextual matrix profile method is methodologically divided into (a) context definition (b) group definition (c) CMP calculation and its further splitting according to previously defined groups.

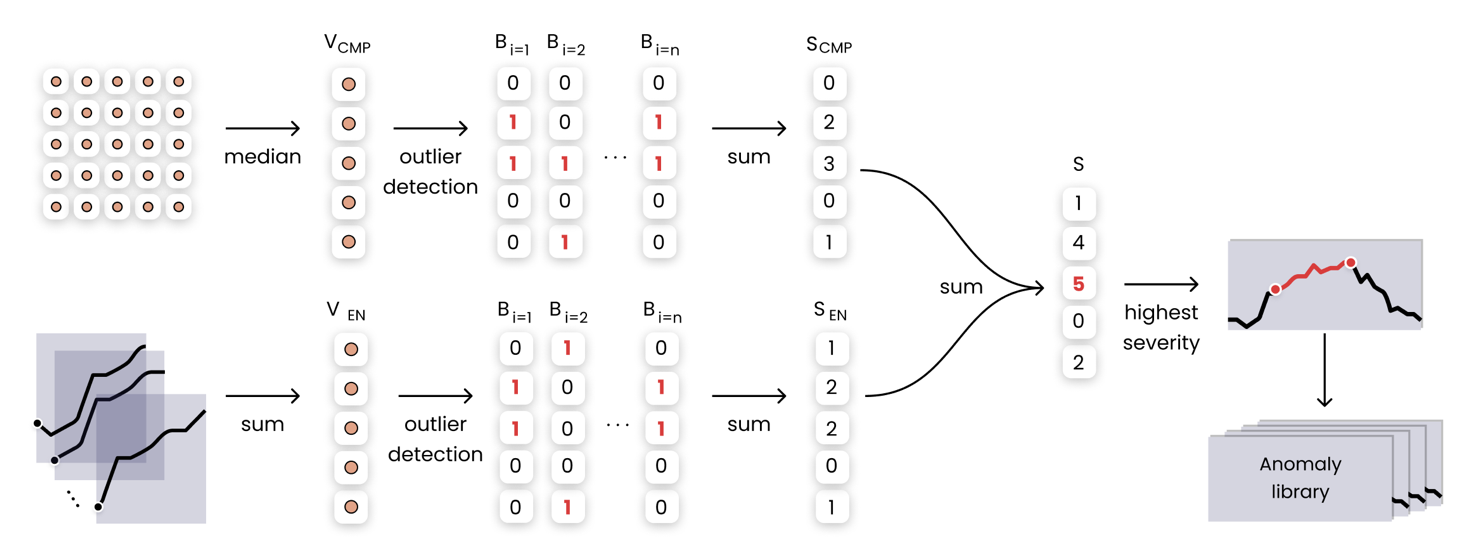
Within the daily electrical load timeseries it is possible to identify different regions and different behaviour (e.g., base load, peak load, ramp-up ramp down) whose relative time window length can be defined statistically or inferred from the typical building operational and occupational schedule [36], [37]. 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) [38], [39]. Starting from the root (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], [39], [40], yielding local optimum [41] once a stopping condition is satisfied. The identification of these region in an unsupervised way has a twofold meaning: (a) automatically identify time windows based on historical operational data, (b) define the two CMP parameters, subsequence length () and context length () that usually are set a priori based on domain knowledge. The regression tree is constructed using the electrical load as numeric target attribute and the hour of the day as explanatory attribute. This permits 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. Clustering is the process of creating groups (i.e., clusters) based on similarity within some attributes. 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 sub-sequences, 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

The anomaly detection is performed for a given group of as given context, by applying methods to identify anomalies and the defining the presence and severity of an anomalous through majority voting. The process is described with reference to Figure. As a first step the CMP is reduced into a vector by calculating the median of each row/column, then anomaly detection methods are applied producing new vectors that defines whether an element is anomalous or not in a Boolean form . Then the severity is then 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 and it is possible to obtain a overall severity ranging from to that robustly ranks anomalies from the most severe to the least severe. Finally, the most severe subsequences are saved and stored in an anomaly library.



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

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 point of the timeseries with Boolean value: if the observation is 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 permits 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 above the elbow and the one below the elbow, the region above contains the 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

* 1. Anomaly diagnosis phase…

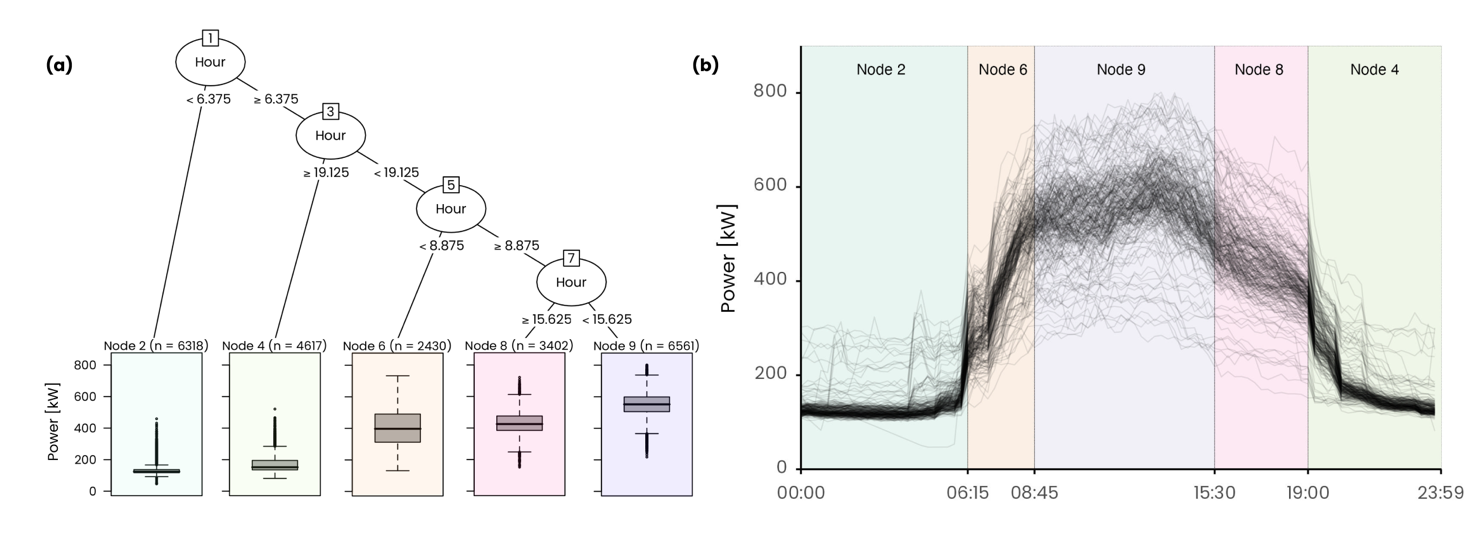
1. Results

The analysis was carried out using the R statistical software [42] for the pre-processing, CART, clustering and visualization and Python [43] 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.



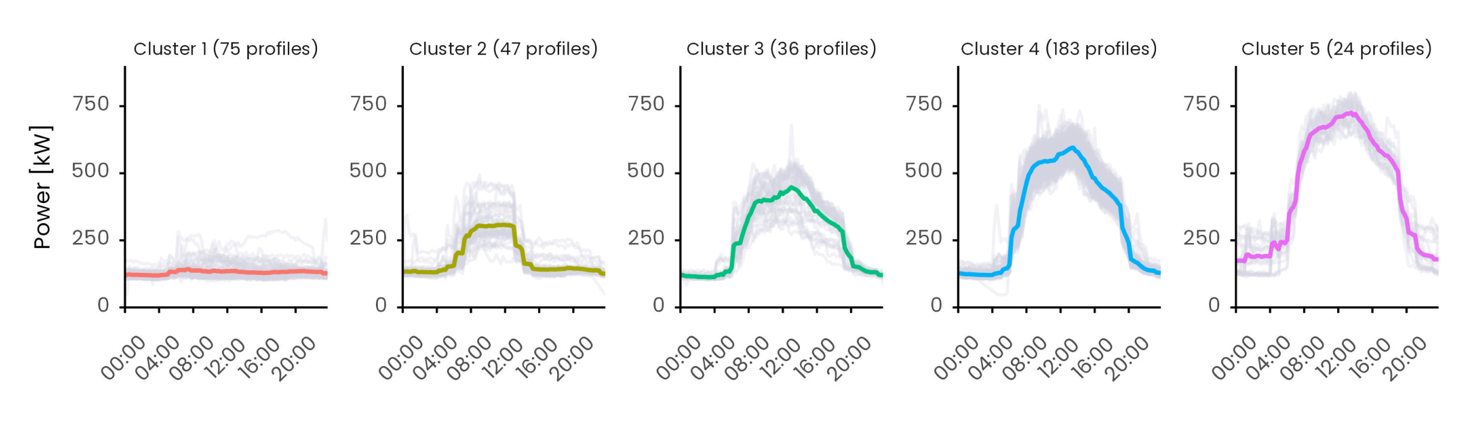
**Figure 8.** Daily electrical load profile clusters with the relative centroid.

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.

**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* | [00:00 - 01:00] | 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 [44], 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 sub-sequences. The calculation was performed using the open source Python code [32] implementing the Series Distance Matrix framework using Euclidean() as generator and ContextualMartixProfile() as consumer. 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 tuned to considers only the positive outliers over 1.5IQR, 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.

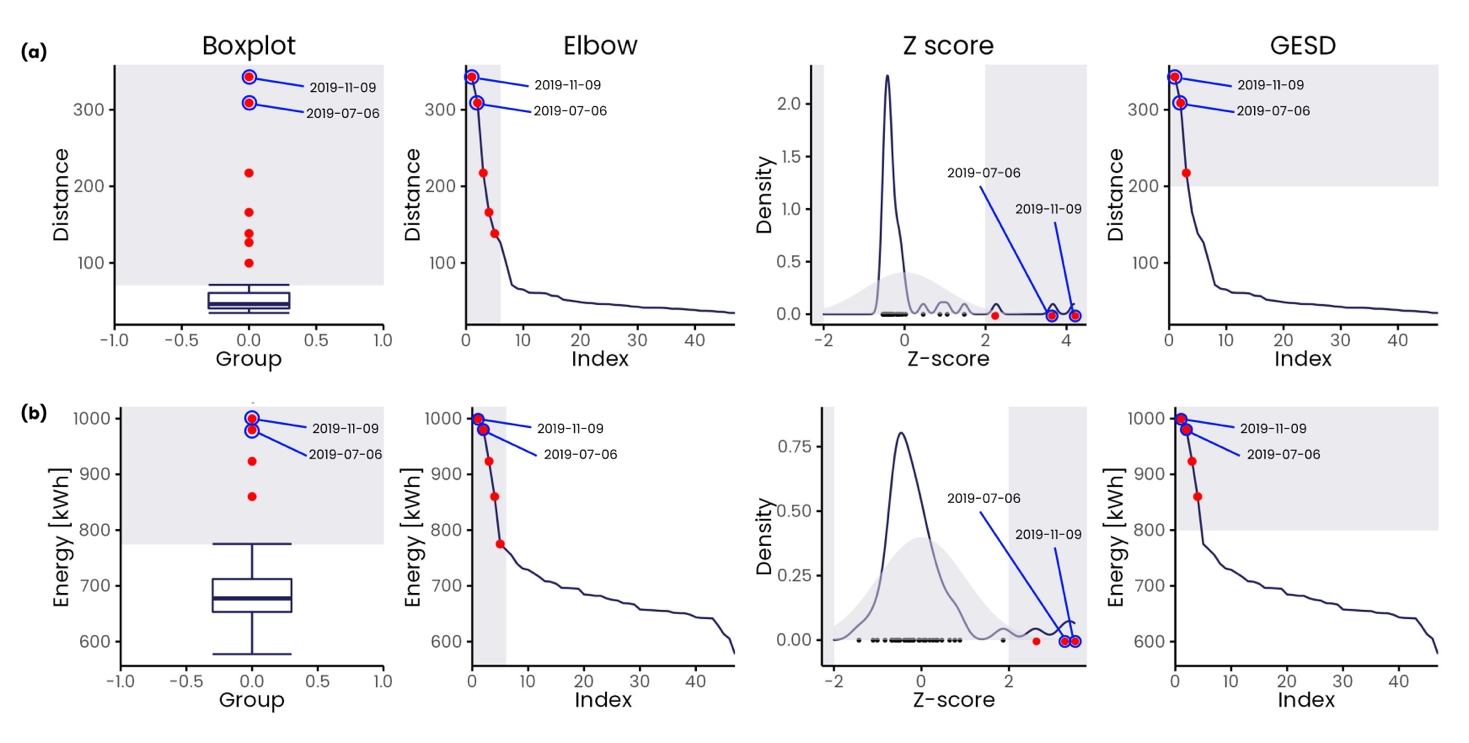
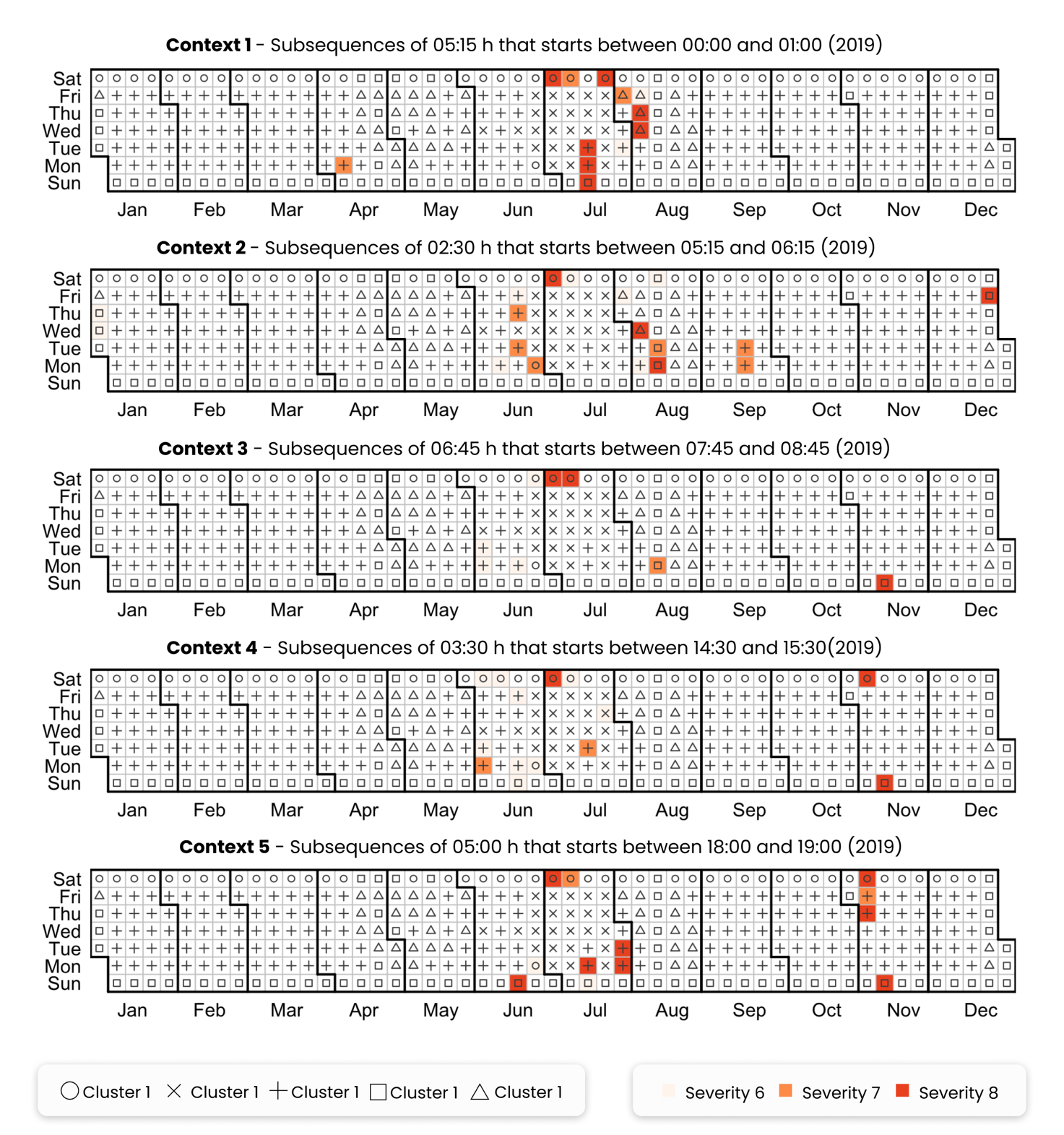


Figure 11. (a) first picture; (b) second picture.

To reduce the number of spurious alerts and feedbacks only the severity 6-7-8 are considered as relevant.

The summary of the results obtained is shown in Figure 12 with a calendar plot representation. According to the figure most of the anomalous behaviors are concentrated during summer from June to August. An interesting insight is that when an anomaly is detected in a given context it is likely that the anomaly persists for the following contexts until the end of the day. For instance, Saturday 6th of July 2019 presents an anomaly of severity 8 for the whole day, and the 10th of November 2019 shows an anomalous energy consumption from context 3-4-5 meaning that the anomaly persists from the morning start up to the midnight. This behavior 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). Moreover, the developed method can find spot-anomalies that refers to a given context and does not persist during the day, for instance the day Friday 27th of December 2019 where a severity 8 is detected for context 2. 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. This clearly shows a wrong scheduling of the energy systems that lead to a consumption. The same behavior was detected in 12th of July 2019 where the same pattern was identified.



* 1. A anomaly diagnosis results

Profili belli come immagini di diagnosi e questo è il sottocarioc che ne ha influenzato il caricamento

Those anomalies are then stored in an anomaly library and diagnosed

Figure 1. (a) first picture; (b) second picture.

1. Discussion

Twin freak

﻿For a given subsequence, Matrix Profile computes the Euclidean distance with respect to all other sub- sequences and identifies the minimum distance. Therefore, a repeated anomaly instance would cause false negatives due to the previous anomaly instance being part of all sub- sequence set.

Specifically, frequent/rare subsequences are defined as the ones with the smallest/largest 1-nearest neighbour distance, which are also known as motif/discord. However, discord fails

the ones with the smallest/largest 1-nearest neighbour distance, which are also known as motif/discord. However, discord fails to identify rare subsequences when it occurs more than once in the timeseries, which is widely known as the twin freak problem.

[25] through a semi-supervised model permits to limits the number of subsequences compared, considering for comparison only references with no anomalies.

[45] proposes a method called “Neighbour Profile” based on sampling and density estimation to perform anomaly detection and overcame the issue of twin freak.

1. Conclusion

# Nomenclature

MP Matrix Profile

CMP Contextual Matrix Profile

CART Classification and Regression Tree

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] C. Fan, F. Xiao, and D. Yan, “Advanced data analytics for building energy modeling and management,” *Build. Simul.*, vol. 14, no. 1, pp. 1–2, 2021, doi: 10.1007/s12273-020-0733-z.

[5] 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.

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

[7] 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.

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

[9] X. Xu, H. Liu, and M. Yao, “Recent Progress of Anomaly Detection,” *Complexity*, vol. 2019, 2019, doi: 10.1155/2019/2686378.

[10] 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.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

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

[44] 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.

[45] 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)