Towards a self-tuned data analytics-based process for an automatic context aware detection and diagnosis of anomalies in building energy consumption timeseries

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 Department of Construction Management and Real Estate, College of Civil and Transportation Engineering, Shenzhen University, Shenzhen, China

Abstract

The spread of IoT sensors in buildings has led to an unprecedented acquisition of reliable energy-related data providing accessible knowledge related to the actual performance of buildings during their operation. A proper analysis of such data could assist energy and facility managers in detecting anomalous energy consumption in their buildings, understanding the cause of such anomalies and thereby spotting valuable energy saving opportunities. In this context, anomaly detection and diagnosis (ADD) tools allow a prompt and automatic recognition of abnormal and non-optimal performance patterns of energy systems de facto enabling a better decision-making to reduce energy wastes and energy systems inefficiencies during operation. This paper proposes an automatic meter-level ADD framework which is able to identify energy consumption anomalies at the whole-building level and perform diagnosis at the sub-load level, revealing which sub-loads impacted the most. The multi-step procedure leverages on supervised and unsupervised data analytics techniques coupled with the distance-based Contextual Matrix Profile (CMP) algorithm to perform pattern similarity search among subsequences of electrical load timeseries. This novel approach allows to discover the most unique patterns (i.e., discords) considering specific boundary conditions, enabling a more accurate and context aware anomaly detection and diagnosis. Thanks to the perfect integration among the employed algorithms and the introduction of a robust anomaly score, the ADD process has self-tuning capabilities and is able to filter and properly rank anomalies considering only those that negatively impact the energy use. To demonstrate the effectiveness of the proposed approach, the ADD methodology is tested on one-year monitored data of a medium voltage/low voltage (MV/LV) transformation cabin of the university campus of Politecnico di Torino.

*Keywords:* building energy consumption, anomaly detection and diagnosis, contextual matrix profile, timeseries analytics

1. Introduction

In the last few years, the increasing widespread use of IoT sensors in buildings for the pervasive monitoring of energy-related data, has led to an unprecedented acquisition of reliable and accessible knowledge related to the actual performance of buildings during their operation. Considering that in Europe the building sector accounts for 40% of final energy use [1] and 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 monitoring data is of paramount importance to i) reduce energy consumption, ii) increase system efficiency, iii) prevent energy wastes and vi) operate their buildings more efficiently.

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 clearly spot energy reduction opportunities, find the root cause of anomalies or simply be aware of energy usage in buildings and systems. Building related data are heterogeneous and reflects the complex interaction occurring between occupants, energy systems, building envelope, and forcing conditions [4]. If properly managed, ingested and analysed building related-data provide the opportunity to gain insight on the building operational behaviour discovering valuable and ready-to-implement energy conservation measures [5].

A robust coupling of IoT sensors data, artificial intelligence (AI) approaches and energy domain knowledge proved to be effective in achieving high energy saving potential by leveraging a variety of energy management solutions [6]. 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 in buildings leveraging advanced data analytics technologies and are intended to support facility staff to enhance building systems performance and efficiency [7]. Depending on the level of detail of the measured data, EMIS solutions can be classified as meter-level or system-level: the first includes analysis of high-level measurements (e.g., data related to the whole-building electrical load or of the main sub-loads) while the second focuses on more detailed data related to the operation of specific energy systems or components (e.g., component operation of air handling units in HVAC systems).

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 visual and analytical insights enabling predictive energy management solutions such as energy consumption forecasting, anomaly detection and diagnosis, advanced energy benchmarking, load profiling, and schedule optimization of building energy systems[8], [9], [6]. Among these solutions, anomaly detection and diagnosis (ADD) has been the most underdeveloped for application on meter-level data.

﻿ADD tools allow a prompt and automatic recognition of abnormal and non-optimal performance patterns of energy systems providing information for the identification of potential of energy waste and for the prioritization of corrective interventions. While fault detection and diagnosis (FDD) tools analyse system/component-level data to detect faults and anomalies, ADD tools generally 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 available measured variables are commonly related to the meter scale. 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 by inferring the main responsible sub-loads.

In accordance with these objectives, this paper introduces the conceptualization and development of a novel EIS tool capable of performing meter-level ADD on building timeseries of electrical energy consumption. The proposed approach employs a pattern recognition technique derived from the Matrix Profile [10] algorithm (called Contextual Matrix Profile [11]) 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 section review the main works related to ADD in the energy and buildings field with a particular focus on the capabilities that Matrix Profile-based algorithms have as anomaly detection methods and how the authors tailored its application in the building field.

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

Generally speaking an anomaly is a region of data with significantly different behaviour from other data and that do not conform to expected values [12]. It can be referred as discord, 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, which mainly involves univariate timeseries data (e.g., electrical load or energy consumption), the definition of anomaly is very domain-specific and may include abnormal behaviour of occupants, faulty operations of appliances, incorrect management of energy systems, anomalous sub-load consumption and technical and non-technical energy losses [13]. Thus, the nature of energy timeseries data in buildings requires to carefully address the definition of anomaly, which according to [14], 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. A *collective anomaly* is a collection of anomalous instances with respect to the entire dataset. Finally, *context anomaly* 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 [15].

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

In ADD, the detection phase is usually accomplished by estimating a reference baseline representing normal behaviour, according to specific boundary conditions, and label each observation that diverges from it as anomalous [14]. The discrimination between normal and abnormal behaviour is essential, so both the robust development of a reference model and the proper selection of the features used to define anomalies are of paramount importance.

Supervised methods have been used to train machine learning algorithms using labelled dataset (i.e., ground truth about the verified presence of an anomaly) to create reference models able to distinguish anomalous from normal energy consumption. Support vector machines (SVM) and multi-layer perceptron are largely used to perform model based anomaly detection [16]. Zhao et al. [17] developed an SVM-based method that was proven to be effective in detecting chiller operation anomalies of different severities. Regression and decision tree classifiers are other widely used supervised techniques used to develop energy consumption reference models for discovering abnormal behaviours [19]. An evolutionary tree model was employed as a detection method in [20] to effectively discover frequent and infrequent patterns in meter-level electrical loads. An hybrid neural net ARIMA model was employed in [21] to predict the energy consumption and then identify anomalies comparing the actual and predicted energy consumption using the two-sigma rule. In [22] 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 ADD is still limited compared to unsupervised methods, mainly due to the difficulty in obtaining high quality datasets [23], [24] and the absence of reliable pre-labelled datasets to train machine learning algorithms [13], [25].

*Unsupervised* anomaly detection is more promising for practical applications since it makes possible to detect rare and unknown anomalous patterns without any a-priori knowledge. Beside statistical unsupervised methods like Principal Component Analysis (PCA) [26] or Generalized Extreme Studentized Deviate (GESD) [27] the data mining methods such as association rule mining (ARM) [12] and clustering analysis [28] gained popularity thanks to the capability to automatically extract significant relations between complex and massive data [5], [29], [30]. In this context, timeseries analytics has been stimulating a great deal of interest in the scientific literature in recent years, since building related data are often stored in datasets as timeseries on which the anomaly detection is performed also considering the time domain. A typical timeseries data mining approach for anomaly detection involves the extraction of infrequent patterns (i.e., discords) that diverges from the rest of the dataset [31], [24], [32]. Li et al. [33] employed GESD to identify anomalous observations in electricity usage timeseries data collected from 40 buildings. Fan et al. [27] 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 was presented in [34] where it was tested in a smart grid context.

In some cases, a reduction and transformation of data, with respect to the original data set, could enhance the performance and reduce the computation time of anomaly detection techniques, also helping to effectively extract specific information from the timeseries. Lin et. al [35] proposed Symbolic Aggregate approXimation (SAX) as a method for the reduction of a timeseries and its transformation in a symbolic sequence for an easy detection of relevant symbolic strings (motifs and discords). This method, which introduces a simple and low-computational cost method to reduce a timeseries, while preserving the key information, is an extension of the Piecewise Aggregate Approximation (PAA) technique and was employed in the literature also for the recognition of frequent/infrequent patterns in energy consumption timeseries of buildings [36], [37]. Miller et al. [38] through a SAX-based analysis identified the most infrequent symbolic sequences referred to daily load profiles of non-residential buildings and furtherly characterized those patterns carrying out a cluster analysis. SAX and temporal association rule mining (TARM) were employed in [39] to extract discords in the energy consumption timeseries, assess building system performance and suggest the implementation of possible energy conservation measures. An adaptive SAX method (aSAX) was employed in [40] to optimize the dimensionality reduction of an energy consumption timeseries and to enhance the detection of frequent and infrequent patterns through a classification tree model [20]. Despite SAX introduced a lot of opportunities in the field of ADD in timeseries, it is worth to say that it is an approximation-based pattern recognition technique and the dimensionality reduction provided by PAA method coupled with the symbolic encoding always lead to information loss from the original timeseries [37], [41]. In addition, the information loss is particularly sensitive respect to the setting of input parameters such as the time window length and the number of symbols for the timeseries encoding [38].

One of the most promising timeseries analytics techniques that is not sensitive to information loss issues is Matrix Profile (MP). Introduced by [10], Matrix Profile (MP) is a novel exact pattern recognition algorithm that performs all-similarity-join-search among timeseries, i.e. given a collection of data objects, retrieve the nearest neighbour for every object (where the considered object is intended to be a timeseries sub-sequence). MP is an unsupervised distance-based anomaly detection algorithm that proposes an ultra-fast similarity search under the z-normalized Euclidean distance that does not reduce dimensionality, but calculates the full join among timeseries, eliminating the need of setting a threshold making the method almost parameter-free and exact. This exact and scalable algorithm allows the method to be incrementally maintainable, deterministic in time and so parallelizable on multicore processor and distributed systems.

The MP method has been successfully applied in different fields of anomaly detection. Alshaer at al. [42] proposed a real time MP-based anomaly detection method which was tested on ﻿electro-cardiogram (ECG) timeseries (﻿MIT-BIH database [32]). The method employs the concept of shaplet [43] to perform continuous learning of anomalies which are extracted though MP algorithm, stored in an anomaly library and then used for sliding-window based anomaly detection. An industrial application was reported in [44] where the MP was combined with the hamming distance to automatically detect intrusions in the network of a water processing facility. A generalization of MP algorithm called PanMP was proposed in [45] to find different length anomalies in ﻿automated pedestrian counting system developed in Taipei. MP has been also largely employed to identify anomalies in IT field. Herath et al. [46] introduced ﻿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.[47] applied a noise elimination technique based on MP on real Yahoo! internet traffic metrics to detect anomalous behaviours; In [48] was demonstrated how the elimination of noise can help in anomaly detection of noisy date by testing the algorithm on Numenta Benchmark [49].

In the energy field very few implementations of MP algorithm are reported. ﻿Nichiforov et al. [50] used the MP to provide insights about the dominant energy usage patterns in large academic buildings . The authors applied the 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 academic building dataset on one year period. The process was tested on 422 buildings whose end-uses were classrooms, offices, laboratories and dormitory. Zhu et al. [51] demonstrated how MP can be useful in detecting rare anomalous electricity consumption occasionally produced by a meter swapping events. The algorithm was tested on a synthetic meter swapping event built on top of two timeseries of household electrical demand and was proven to be effective to discover the suspicious similarity between the two timeseries. Park et al. [52] 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.

Despite MP proved to be effective in several application fields, in its full join formulation, it may compare regions of timeseries belonging to different contexts, operating conditions or different boundary conditions and may lead to misleading results in terms of pattern similarity. To address this issue, in [11] was introduced the Matrix Profile-based algorithm called Contextual Matrix Profile (CMP). CMP allows to identify contexts in the timeseries in which it is possible to compare between each other only patterns that are characterized by homogeneous boundary conditions avoiding the identification of meaningless similarity matches. In brief, while MP searches for unique patterns (i.e., discords) in the whole timeseries the CMP finds, within a reference context set by the analyst, anomalous patterns that differ the most.

An application of the CMP method for anomaly detection in buildings is presented in [11] 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, enabling the identification of 6 behavioural anomalies only during weekend morning subsequences. Moreover, the authors were able to identify periodic behavioural patterns even if there was not a time alignment among them. The 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. To the best of author knowledge, other applications of CMP in buildings, including ADD in energy consumption timeseries, are so far missing in the literature. In the next section, the main technical aspects related to the MP and CMP algorithms are introduced and discussed to better highlight the contributions brought by this study in the field of energy anomaly detection in buildings.

* 1. Matrix profile and Contextual Matrix Profile method for anomaly detection

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 values 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 it is possible to identify 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 identify 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.

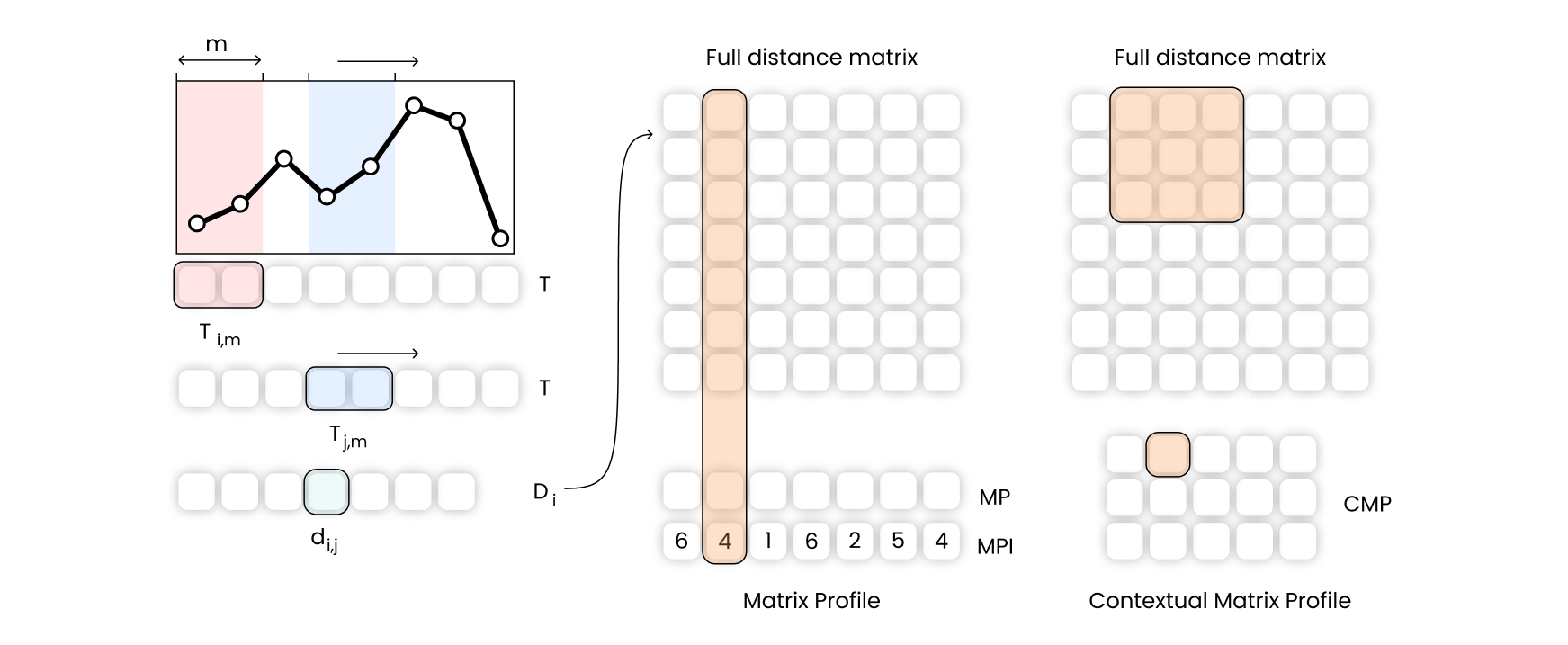
With reference to Figure 1 some fundamental concepts and definitions need to be introduced before going deeper into the topic. First, a *timeseries* is a sequence of real-valued numbers with where is the length of . Since the focus is on local properties of timeseries (i.e., portion of timeseries) a *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 [53], [51], [54] 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 values 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 it finds the best matching subsequence 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 [55], approximated AMPSA and AMP [56] and multidimensional mSTAMP [57].



**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.

MP represents a reference algorithm in the field of timeseries analytics and in particular among similarity-join-search algorithms. As defined in [10], given a collection of data objects, perform a similarity-join-search means to retrieve the nearest neighbour for every object. Thus, given the all-subsequences-set of a timeseries and the all-subsequences-set of a timeseries similarity join finds all the nearest pairs of subsequences from the two sets. However, extending the search considering all the possible subsequence pairs in the timeseries is not always the most robust way to discover motifs and discords. In fact, when the domain of application constraints the meaningfulness of the similarity search in a timeseries, it could be useful to retrieve the most similar pattern of each data object allowing the search only into a-priori determined set of subsequences of the timeseries and excluding others.

As a solution to this problem, in [11] 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 the timeseries 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 04:00 and 06:00 AM. To this aim the subsequence length must be set to , (grey regions in Figure 2) and the context length (yellow regions 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 sub-sequences 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. Considering the main features of the CMP algorithm, in the next section are introduced the contributions of this study and the research gaps that this research aims to bridge.

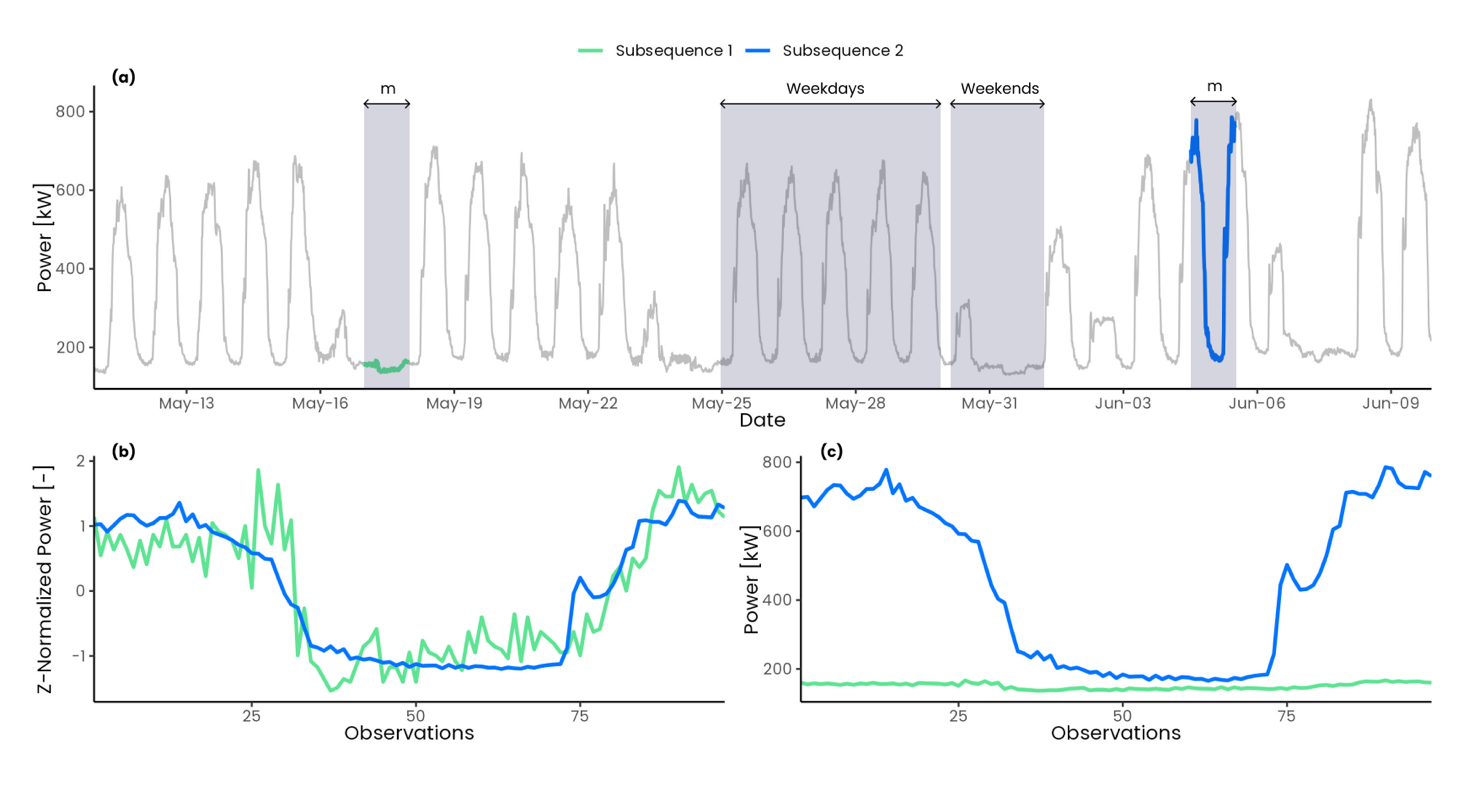


**Figure 2.** Focus on the definition of subsequence length and contexts in an electrical load timeseries and their importance in the CMP calculation. The timeseries has one hour frequency and for each day a context of (yellow) and a subsequence length of (grey) are defined. When calculating the CMP, 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.

* 1. Research gap and contribution of the paper

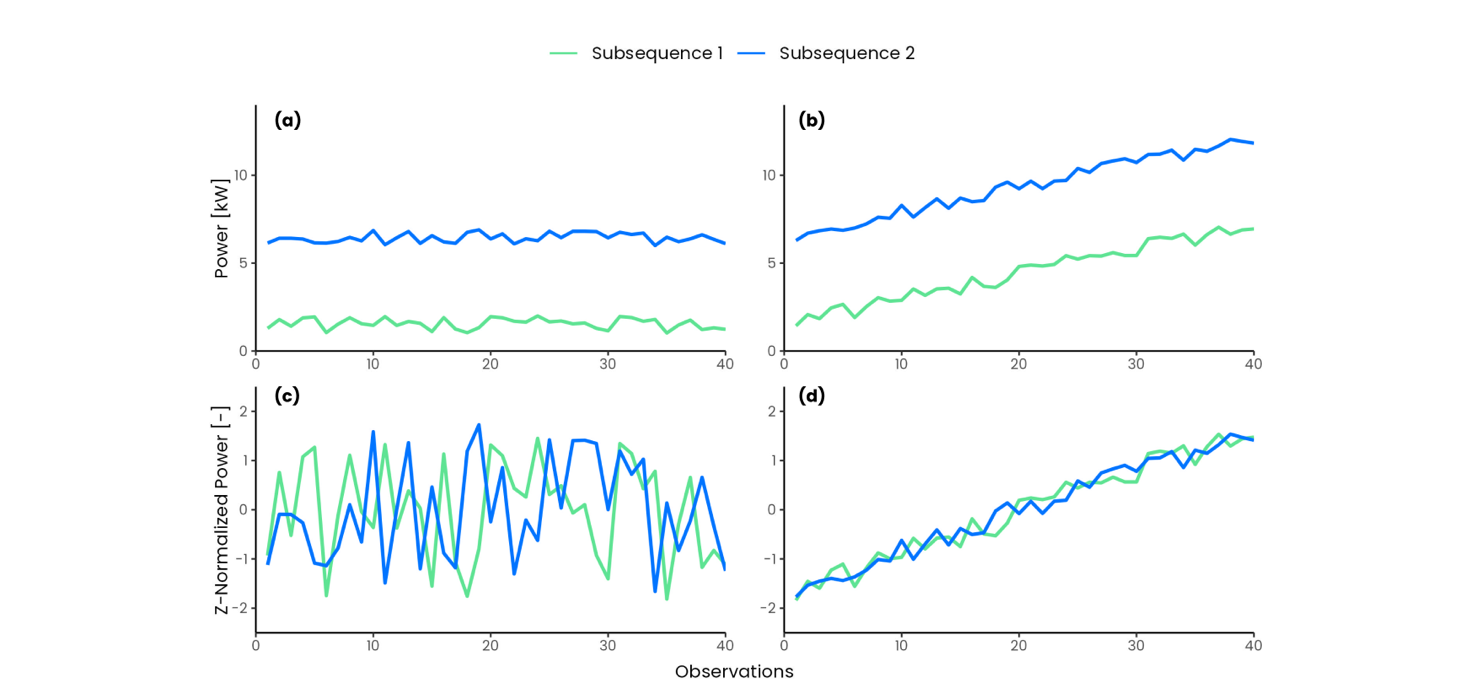
From the previous literature review it emerges that the MP method has been employed successfully in different fields of anomaly detection and several researchers have proposed different implementations according to the field of interest. In fact, even if MP is an unsupervised method useful for discord/motif 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 considering forcing variables such as occupancy, weather conditions, energy system configuration and so on.

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 (i.e., 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 neighbours. 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 sub-sequences 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 could 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 to 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 [47] where different solutions were proposed: discard flat regions from the analysis, change the subsequence length or smooth the timeseries. A clear consequence of this issue is that, referring to Figure 3(a), the MP method would identify the weekends as discords since they are almost associated to 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). Adapted from [48]

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 become 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 [58], 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 [59]. 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 regions or to split sub-sequences 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 [11] introducing the Contextual Matrix Profile (CMP) algorithm.

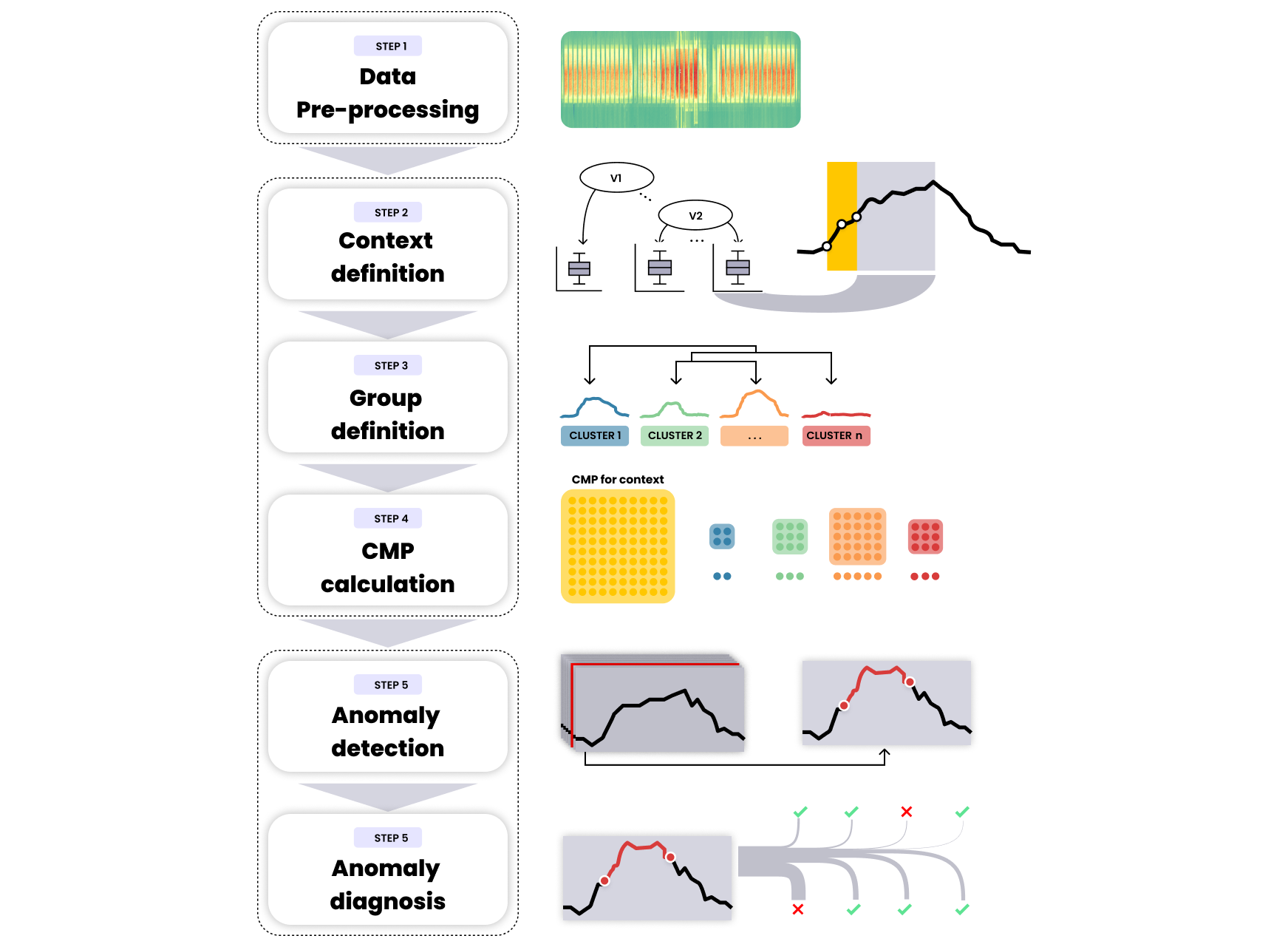
The application of unsupervised methods coupled with expert knowledge is crucial for the accurate discovery of anomalies in building electrical load and it is the key to reduce energy wastes and enhance energy management in buildings. To this aim the objective of this work is the introduction of an unsupervised anomaly detection procedure based on CMP algorithm to detect anomalous electrical load at building level in quasi real-time and identify the responsible sub-loads through 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 CMP-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. Development of an anomaly detection and diagnosis methodology that performs the *similarity join* within a timeseries without any loss of information (e.g., avoiding dimensionality reduction and symbol encoding as in SAX algorithm) and flexible enough to discover anomalous/typical patterns even if they are not being aligned in time (i.e., time-tolerant analysis according to the length of the contexts)
3. 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.
4. 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 methodological framework introduced. Section 3 presents the case study and the obtained anomaly detection and diagnosis results while Section 4 critically discusses the outcomes and contains the concluding remarks.

1. Methodology

In this section the methodological framework employed to conduct the ADD process is presented. The framework 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 buildings. The procedure, reported in Figure 6 consists of 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 was performed through univariate statistical approaches. Punctual statistical outliers and inconsistencies (e.g., negative values of electrical load) in the timeseries are identified and removed by means of boxplot analysis. Then all the missing values are replaced by means of linear interpolation.

* 1. Contextual matrix profile.

The application of the CMP method goes through the following steps: (a) context definition (b) group definition (c) CMP calculation.

Within the daily electrical load timeseries can be identified different regions and different load patterns (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 [60], [61]. By identifying typical electrical load sub-sequences during a day, information of particular interest can be extracted for building energy management. The methodology proposed in this paper identifies sub-daily time windows () through the recursive partitioning Classification and Regression Tree (CART) [40], [41]. Starting from the root node (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 [12], [40], [62], yielding local optimum [63] once a stopping condition is satisfied. The identification of these region through CART has a twofold advantage: (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 subsequences 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 (). For instance, if the smallest time window is two hours long (e.g., from 6:00 to 8:00) the context is defined as one hour long (e.g., from 5:00 to 6:00).

As a second step, a group definition based on daily load profiles is performed. A supervised expert approach is first applied to split working days, holidays/non-working days and partially working days (e.g., Saturdays). Then a nested hierarchical clustering, with Euclidean distance, is performed on the daily profiles of working days. The hierarchical clustering generates non-overlapping clusters by splitting instances based on a geometrical distance measure and each cluster can be further divided into subclusters and so on, creating a tree structure. By grouping daily load profiles in clusters which are representative of specific energy consumption patterns allows the anomaly detection process to be performed on groups with high internal similarity, making the process more robust.

Once contexts and subsequence lengths are defined, the CMP is calculated for each context under the hypothesis of not-normalized Euclidean distance. According to previous definitions, each day has not overlapping contexts and the resulting CMP contains one row/column for each day. Eventually, the overall CMP is split into different sub-matrices according to the membership of each day to a specific group identified by means of the cluster analysis. After this, the anomaly detection process can be carried out for each context in each group.

* 1. Anomaly detection

The anomaly detection process, shown in Figure 7, is performed for a given group within a given context, and consists in the definition of a severity score through the use of four statistical univariate outlier detection methods. The considered four methods are applied on two vectors i.e., and . The first vector contains the median of each row/column of the CMP i.e., the median of the Euclidean distance between a specific subsequence and its nearest neighbours in the same context identified for all the days in the same cluster. The second vector , contains for each electrical load subsequence the underlying energy consumption. In this way it is possible to label as anomalous, sub-sequences with Euclidean distance and energy consumption values that are out-of-range respect to the domain that is considered normal for both and . With reference to Figure 7, as a first step the CMP is reduced into a vector by calculating the median of each row/column, then the four outlier detection methods are applied producing 4 new vectors that define 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: outlier methods are applied and then the severity is calculated. With reference to Figure 7, given outlier 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).

To avoid spurious alerts and reduce the number of feedbacks only the sub-sequences with severity 6-7-8 were considered and further analysed in the diagnosis phase.



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

The four univariate outlier detection methods are introduced and described below.

*Inter quartile range analysis* defines outliers any of the observations that fall below and above where the interquartile range () is defined as the difference between the third quartile and 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 defines meaning that the probability that an observation outlies that range is equal to 2.3%. To apply this method to non-normal distribution data, z-score standardization is needed.

*Elbow method*: is a heuristic method that allows to find the elbow (or knee) of a curve. The implemented method employs the kneedle algorithm [64] to identify the point of maximum curvature and thus locate the elbow. 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, where observations 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 [27]. The method initialization requires () a presumed number of outliers and confidence interval , then 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

Once the anomaly at meter-level is identified, the diagnosis phase is enabled with the aim to identify which sub-loads are the most responsible. The diagnostic process consists in selecting only those sub-sequences that resulted to be anomalous at meter level and analyse the sub-load impact.

By keeping the same hyperparameter settings (i.e., contexts, subsequence length and group) as previously described, each sub-load timeseries undergoes through CMP calculation and anomaly detection process using the 4 outlier methods previously presented. In this way for each sub-sequence of the sub-loads it is possible to assess the impact by calculating the severity score that allows to rank the sub-loads from the least impacting (i.e., low severity) to the more impacting (i.e., high severity). The sub-load that presents the highest severity is the one that has affected the anomalous behaviour identified at meter level.

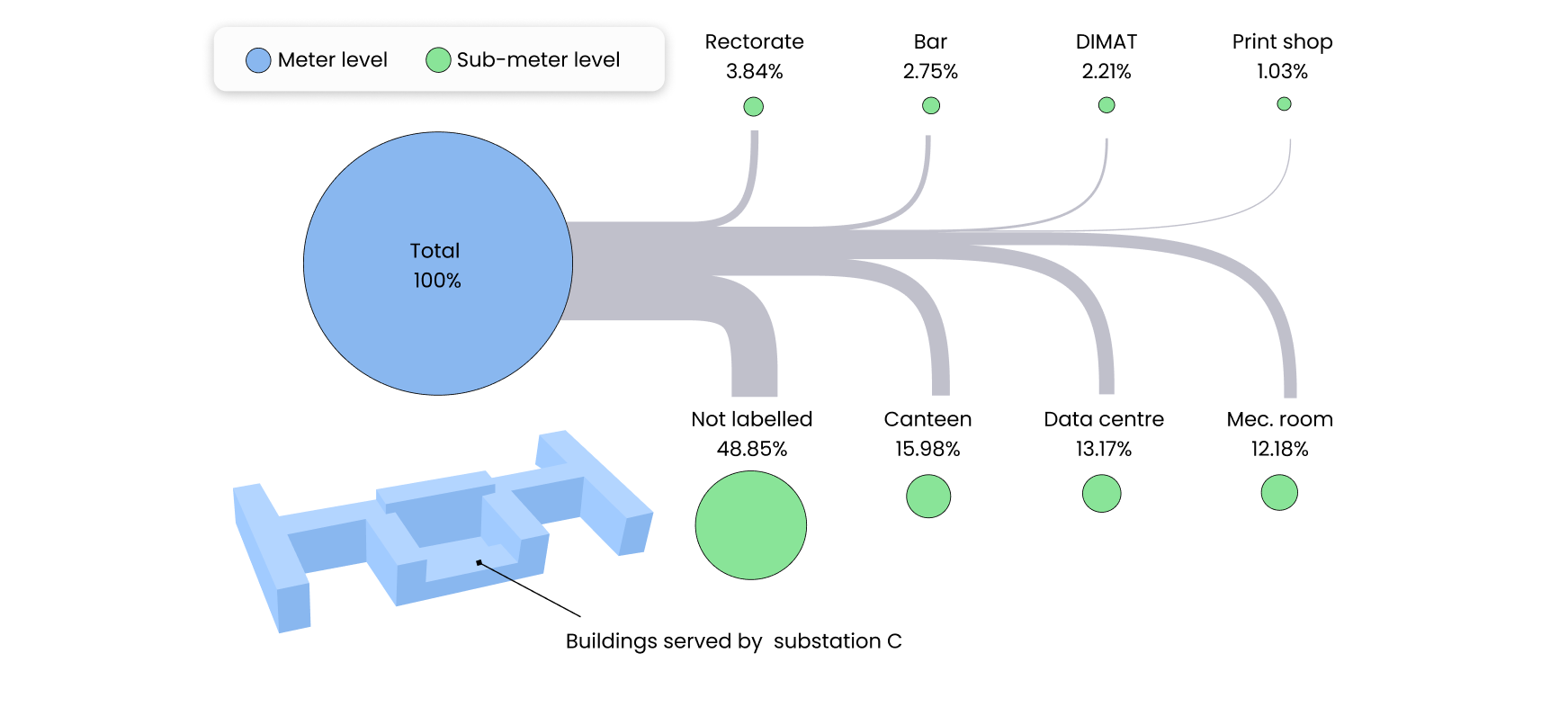
1. Results

The analysis was carried out using the R statistical software [65] for the pre-processing, CART, clustering and visualization, and Python [66] 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 (i.e., substation C) that serves a part of the Italian university campus of Politecnico di Torino (PoliTo). The measurement infrastructure continuously provides both meter level and sub-meter level measurements of the electrical load with 15 min timestamps. 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 associated sub-loads. In addition, the load breakdown in terms of average annual energy consumption was provided.

In a bar and a staff canteen are at the disposal of students and campus staff and on average, represent the 2.75% and 16.03% of the yearly electrical energy consumption of substation C respectively. The university data centre accounts for the 13.16% of the total energy consumption. The administration offices (Rectorate) correspond to the 3.83% of energy consumption while the mathematics department (DIMAT) to the 2.21%. A relevant amount of consumed energy (12.22%) is related to the mechanical room. The equipment located in this room includes hot and chilled water circuits with their respective auxiliaries such as recirculation pumps. The chilled water is produced by two chillers with 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.

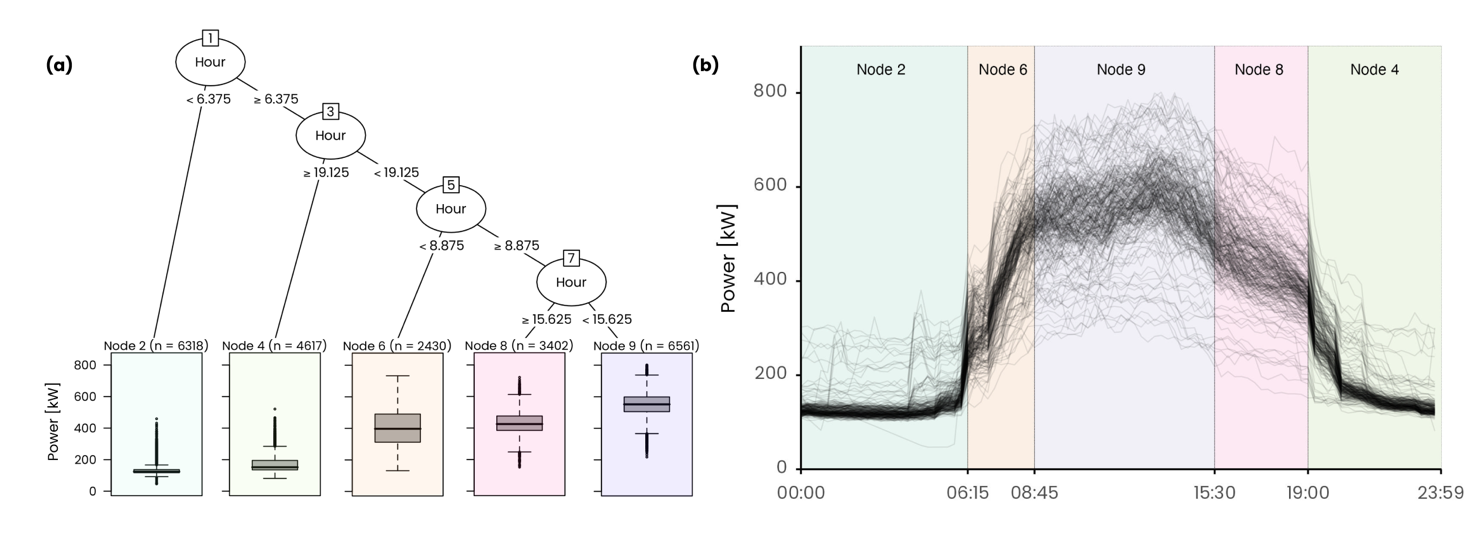
The remaining energy consumption is aggregated under a unique synthetic sub-load tagged as “Not labelled” as showed in Figure 5. It accounted for 48.76% of the total energy consumption, and since it was not directly measured, cannot be characterized, as the other sub-loads, respect to the energy service or building served. The authors tested the presented methodology on a dataset that spans from 01.01.2019 to 31.12.2019 even if more recent data were available, mainly because the pandemic COVID-19 completely changes operational patterns and caused a closure of the university from February 2020. The raw dataset contained 35040 15-min observations with a missing value ratio of less than 0.1%. Inconsistences were removed and missing values replaced through linear interpolation following the first step of the methodology.



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

* 1. Contextual matrix profile results

To identify homogeneous electricity consumption regions within the daily load profile, non-overlapping time windows were evaluated through CART algorithm using the meter-level 15-min electrical load as target variable and time of the day as numerical predictive variable. To preserve the accuracy of the model in the leaf nodes during the operation hours of the system (e.g., from 07:00 to 19:00), only working days were considered, and days with a low standard deviation of the electricity demand (e.g., Sundays, holidays) were excluded. The adopted stopping criterion, to avoid overfitting of the regression tree, was based on the minimum number of objects in a child node in order to identify time windows with a length of at least 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 leaf nodes resulting from the tree have been represented also in the Figure 8(b), where are reported the actual amplitudes of the temporal windows on a daily scale.



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

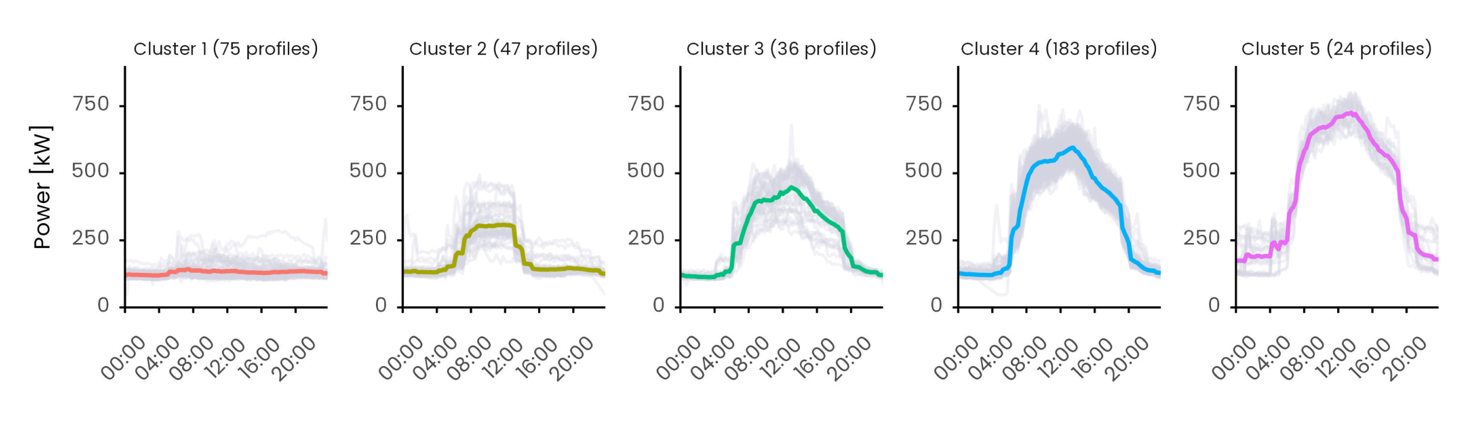
The model was able to effectively separate the night hours (time window 1 and 5) from the diurnal operation and was able to identify the ramp-up (time window 2), mid-day operation (time window 3) and ramp-down (time window 4).The smallest time window lasts 2.5 hours. Following the methodological process, once the sub-daily time windows are evaluated the contexts for the definition of the CMP can be identified. In this study the context length was defined as the half of the smallest time window length identified (). The outcome of this preliminary step led to the definition of 5 time window durations (e.g., subsequence length) and 5 contexts for the CMP calculation, summarized in Table 1. For instance, the first time window, that ranges from 00:00 to 06:15, defines a sub-sequence length of 25 observations (i.e., ) that starts in the context between 23:00 and 00:00 of the previous day.

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Time Window | | | | Context | | | |
| ID | Interval | Duration | Subsequence length | ID | Interval | Duration | Subsequence length |
| *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, consisting in the definition of the groups, was performed using a semi-supervised approach applied on the 365 daily load profiles available in the dataset. As a first step the 75 daily load profiles corresponding to public holidays, university closures and Sundays were extracted and grouped together in a cluster labelled as Cluster 1. Secondly, the 47 daily load profiles of half working days and Saturdays were extracted and assigned to Cluster 2. The remaining 243 daily load profiles, corresponding to working days, were organized into a MxN matrix 243x96 where M is the number of days considered and N the number of observations per day. 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 [67], was used to search the optimal number of clusters in a range between 2 and 6. Three clusters were identified as the optimal partition, and labelled as follows: Cluster 3 (36 profiles), Cluster 4 (183 profiles), Cluster 5 (24 profiles).

Figure 9 shows the five final clusters identified: the grey lines represent the daily load profiles belonging to the cluster while the coloured line represents the cluster centroids. The clustering process led to a well-defined set of clusters each one representing a typical energy consumption behaviour useful 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 [11] 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 symmetric matrix. The higher the distance value the higher the dissimilarity. The overall CMP (on the left of Figure 10) shows a weekly regularity: there are typically 5 days with the same behaviour (green) followed by two days of different behaviour (yellow). Moreover, a change of typical patterns during summer can be observed, especially during July and August, corresponding to holidays and summer closures of the university facilities. By further splitting the CMP in the previously defined clusters, days expected to behave in similar manner are grouped together to perform a more robust inspection. For instance, the day 10th of November 2019 (highlighted in Figure 10 at index 63 of cluster 1 submatrix) stands out to be remarkably different from all the others day in the same cluster, which is not so evident by only visualizing the global CMP. This highlights the importance of groups for the identification of contextual anomalies.

Immagine che contiene testo, elettronico, screenshot

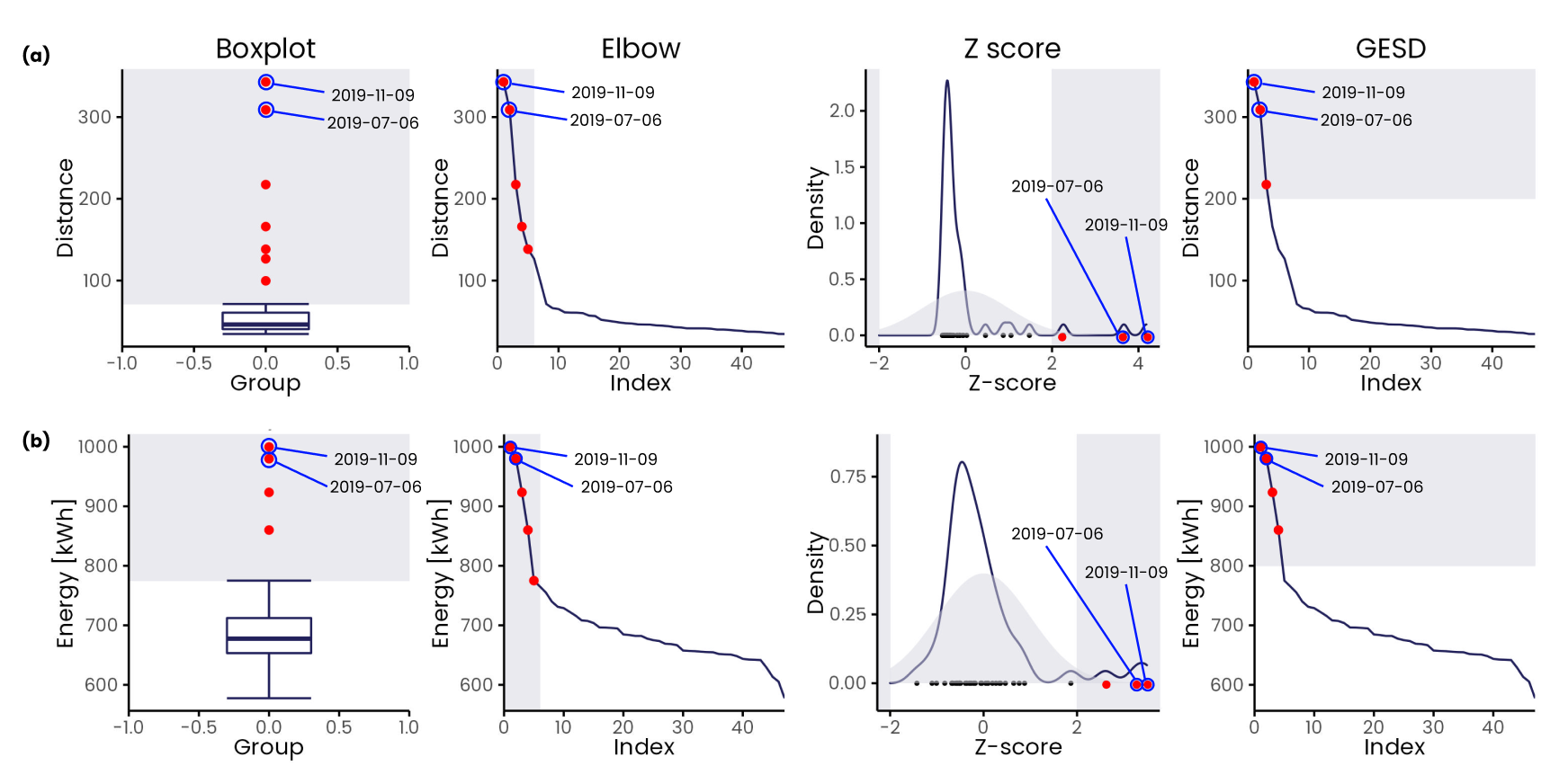
Descrizione generata automaticamente

**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. The CMP is divided on the right side into submatrices according to the clusters (i.e., groups).

* 1. Anomaly detection results

The anomaly detection is performed for each context in each group. For each daily subsequence within a group both the median of the Euclidean distance (i.e., median of column/row of the CMP) and the energy consumption were calculated, and the four univariate outlier 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 method considers only the positive observations over 2, GESD method is initialized with presumed number of outliers and confidence interval and finally the elbow method, since it is a pure graphical method, considers as outliers all the data points 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 an overall severity ranging from to is calculated to robustly rank anomalies from the most to the least severe one.

Figure 11 reports the anomaly detection results obtained for the context 5 within cluster 2. In detail, Figure 11(a) shows the case of the outlier methods applied on the vector of the median Euclidean distances while the Figure 11(b) 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 score of 8.

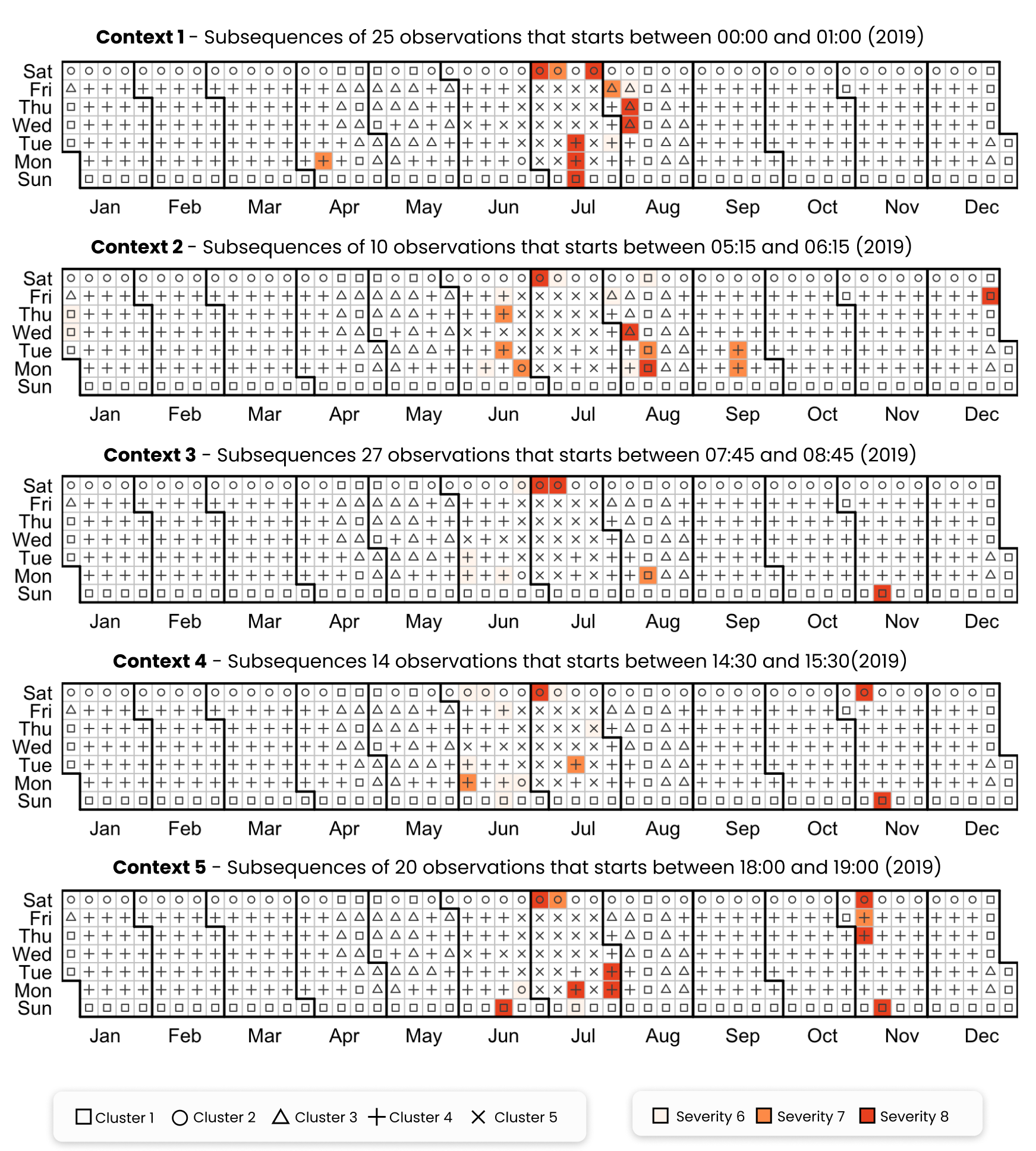


**Figure 11.** Example of severity calculation on (a) median Euclidean distance vector and (b) energy consumption vector using the four anomaly detection methods respectively inter quartile, Z-score standardization, elbow method and Generalized Extreme Studentized Deviate (GESD).

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

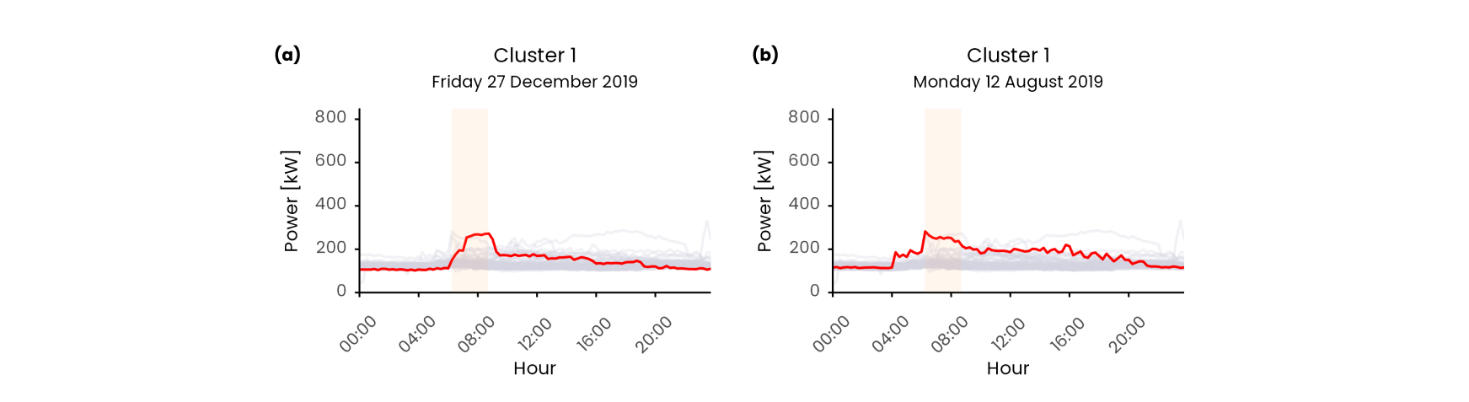
Results are summarized in Figure 12 through a calendar chart which shows anomalies over the year for the different contexts. According to Figure 12, anomalies resulted more frequent at the beginning and end of the day: 12 in the 1st context, 18 in the 2nd context, 8 in the 3rd context, 14 in the 4th context and 12 in the 5th and last context. Moreover, the calendar chart clearly shows as 51 anomalies out of 64 happened during summer from June to August, compared to the rest of the year where only 13 anomalies were detected.

The definition of context and time window allowed the ADD process to be effective in finding anomalies that occasionally occurred and were limited within one time window (i.e., spot anomalies) and others that instead persisted among subsequent contexts and are likely to last until the end of the day (i.e., persistent anomalies).



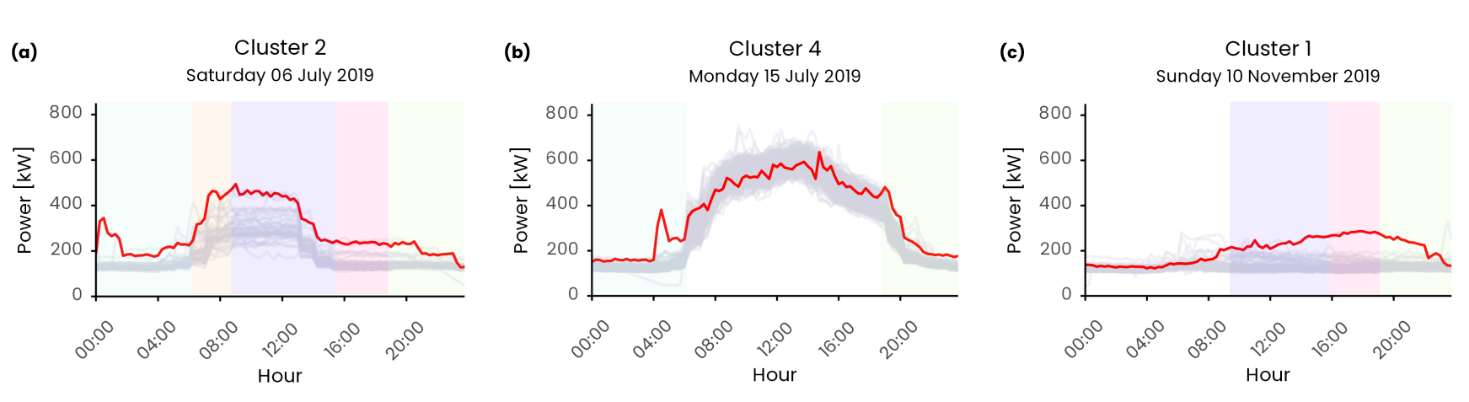
**Figure 12.** Calendar chart representing the distribution of anomalies through the whole year dataset. Each row corresponds to a different context. For each day the colour fill represents the severity of the anomaly while the symbol represents the cluster membership

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 was expected, however a rise of the electrical load after 6:00 and an abrupt switch off at 9:00 was detected, resulting in a deviation from the centroid of 260 kWh. The same pattern, shown in Figure 13 (b), was detected during summer season on Monday 12th of August 2019, when despite the university was closed for the summer break, an abnormal increase in electrical load in the second context was detected resulting into a deviation of 314 kWh from the cluster centroid. These two examples are symptoms of a wrong schedule of the energy systems pertaining to the substation C, that will be further analysed 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 load profile was anomalous also during the following time windows by keeping an offset of almost 80 kW compared with the cluster centroid, leading to an overall deviation 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 was observed on Monday the 15th of July 2019 where a wrong schedule of start-up and switch-off resulted in anomalies during context 1 and 5 leading to a deviation of 801 kWh compared to the average cluster energy consumption, see Figure 14 (b). Another example of persisting anomaly is shown Figure 14 (c) where during winter season on Sunday the 11th of November 2019 an anomalous energy consumption from context 3-4-5 was detected, from the morning start up to midnight, leading to a deviation 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 aims to spot the sub-loads that are responsible of the anomalies calculated at meter level with the severity score that ranges from 6 to 8. A sub-load with higher severity score is likely to impact more on the meter-level anomaly that a sub-load with low severity score. The 64 anomalies discovered at meter level were investigated and results are reported in Table 2. The table reports the number of times in which a sub-load or group of sub-loads was identified as a potential root cause for the anomaly due to anomaly scores with high severities. “Mechanical room” and “Not labelled” loads were most frequently responsible for the detected anomalies, respectively 26 times and 18 times out of 64. However, for 5 times both sub-loads were identified as co-responsible for the detected anomaly at meter-level. The sub-load “Mechanical Room” contributed to the anomalous behaviour of the total energy consumption of the substation C, in concurrence with other sub-loads five more times: 2 times together with “Print shop”, 1 time together with “Data centre” and 1 time together with both “Data centre” and “DIMAT”. Other interesting sub-loads that contribute several times to the meter-level anomaly are the “Canteen” and “Print shop” despite their relatively low impact in terms of yearly energy consumption on the total one of the whole substation.

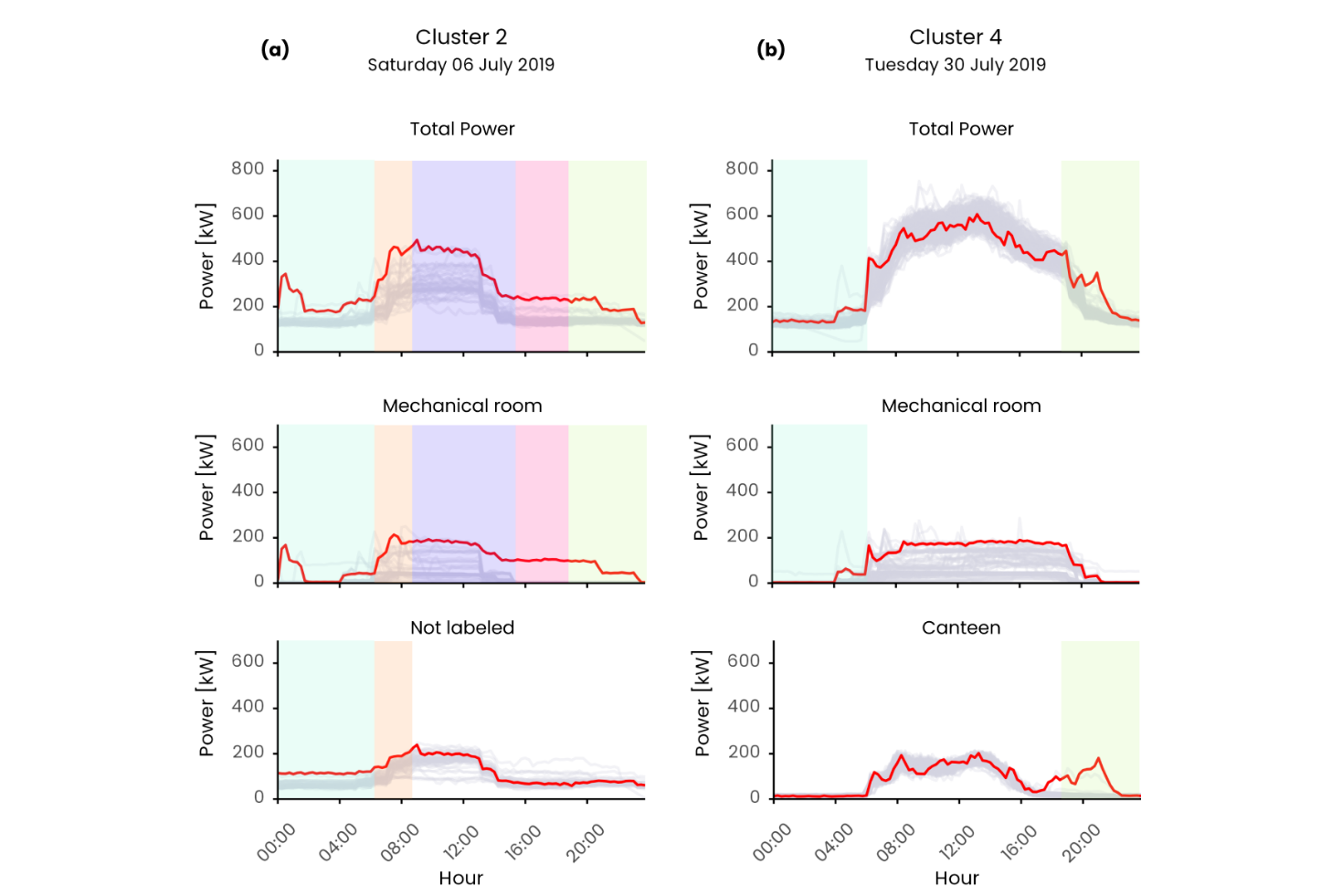
**Table 2.** Summary of the diagnosis results.

|  |  |
| --- | --- |
| Number of occurrences | Responsible sub-load |
| *26* | Mechanical room |
| *18* | Not labelled |
| *5* | Not labelled, Mechanical room |
| *3* | Canteen |
| *3* | Print shop |
| *2* | Canteen. Print shop |
| *2* | Mechanical room, Print shop |
| *1* | Canteen, Data centre |
| *1* | Data centre |
| *1* | Mechanical room, Data centre |
| *1* | Mechanical room, Data centre, DIMAT |
| *1* | Not labelled , Print shop |

Figure 15 reports two representative examples in which the diagnosis approach was able to clearly identify and isolate the anomalous sub-loads responsible for the anomaly detected at meter level. Figure 15(a) shows Saturday 6th of July 2019 during which both the “Mechanical room” and “Not labelled” loads contributed to the anomalous trend 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 have a severity score of 8. The “Mechanical room” was characterized by 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”, led 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 showed no impact at meter-level (i.e., very low severity, less than 1). From Figure 15 it can be see that the anomalous load of “Mechanical room” after 12:00 led 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 contributed 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, contributed 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 were proposed in order to demonstrate (i) the flexibility of the Contextual 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 anomalies and reduction of wastes. The methodology was based on unsupervised machine learning methods and timeseries analytics coupled with domain knowledge 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 [40] 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 starts, adds an additional degree of freedom that allows for the comparison of subsequences that are not perfectly aligned in the time domain.

The methodology was tested off-line 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 subsequent triggers 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 can be 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 CMP [10]. 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), meaning that the execution time of the whole ADD procedure including CMP calculation on all the timeseries (meter level and sub-meter level) must be lower than 2.5 hours. The actual execution time on the offline one-year dataset presented in the case study is far less than 2.5 hours proving that a semi-real time batch approach is suitable for this kind of analysis. However, a 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 algorithm [10] that adjusts the CMP rather than re-compute it. Following this approach at the end of each time window, the observation would be 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 CMP 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. A dataset containing at least 4 weeks of observations is desirable in order to have 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 following a knowledge-driven approach. 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 [38] 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 defined 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 CMP 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 CMP are defined as the ones with smallest/largest 1st nearest neighbour distance [10]. This implies that if a rare subsequence occurs more than once in the timeseries it may be considered as common or even frequent pattern [68] and tackle this issue 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 motif. A repeated anomaly would cause false negatives due to the previous anomaly instance being part of all sub-sequence set: this issue is recognized in the literature as the twin freak problem [68].

To address the twin freak problem, in the future, the implementation of the kNN distance instead of 1st Nearest Neighbour distance in the CMP calculation presented in this paper can be included as a process improvement. One of the possible algorithmic implementation was discussed in [68] where the authors proposed a density based approach for the kNN calculation applied in the MP algorithm. A further step would be to enhance the robustness of the ADD methodology, through the dynamic adjustment of 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 an easy interpretable severity score, ranging from values between 0 and 8, coupled with an effective visualization that allows immediately to compare the detected anomalous behaviour with the expected in the given context and boundary conditions. Thanks to the proposed hierarchical ADD approach, the user is supported in the ADD process and has an immediate perception of anomalous trends that deviates from the normal ones enabling prompt interventions thus avoiding further energy waste over time.

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] S. Rinaldi *et al.*, “A cognitive-driven building renovation for improving energy effciency: The experience of the elisir project,” *Electron.*, vol. 9, no. 4, 2020, doi: 10.3390/electronics9040666.

[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] 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.

[8] C. Fan, D. Yan, F. Xiao, A. Li, J. An, and X. Kang, “Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches,” *Build. Simul.*, 2020, doi: 10.1007/s12273-020-0723-1.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] D. J. Bartholomew, “Principal components analysis,” *Int. Encycl. Educ.*, pp. 374–377, 2010, doi: 10.1016/B978-0-08-044894-7.01358-0.

[27] 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.

[28] P. Esling and C. Agon, “Time-series data mining,” *ACM Comput. Surv.*, vol. 45, no. 1, 2012, doi: 10.1145/2379776.2379788.

[29] 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.

[30] 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.

[31] V. Chandola, D. Cheboli, and V. Kumar, “Detecting anomalies in a time series database,” *… Dep. Univ. Minnesota, Tech. Rep*, p. 12, 2009, [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Detecting+Anomalies+in+a+Time+Series+Database#3.

[32] 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.

[33] 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.

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

[35] J. Lin, E. Keogh, L. Wei, and S. Lonardi, “Experiencing SAX: a Novel Symbolic Representation of Time Series,” *Cs.Gmu.Edu*, vol. 15, pp. 107–144, 2007.

[36] 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.

[37] 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.

[38] 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.

[39] C. Fan, F. Xiao, H. Madsen, and D. Wang, “Temporal knowledge discovery in big BAS data for building energy management,” *Energy Build.*, vol. 109, pp. 75–89, 2015, doi: 10.1016/j.enbuild.2015.09.060.

[40] 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.

[41] 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.

[42] 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.

[43] L. Ye and E. Keogh, “Time series shapelets: A new primitive for data mining,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 947–955, 2009, doi: 10.1145/1557019.1557122.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] V. Satopää, J. Albrecht, D. Irwin, and B. Raghavan, “Finding a ‘kneedle’ in a haystack: Detecting knee points in system behavior,” *Proc. - Int. Conf. Distrib. Comput. Syst.*, pp. 166–171, 2011, doi: 10.1109/ICDCSW.2011.20.

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

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

[67] 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.

[68] 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)