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

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

Roberto Chiosa[a], Marco Savino Piscitelli[a*], Cheng Fan[b], Alfonso Capozzoli [a]

[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

Recently, the spread of IoT technologies has led to an unprecedented acquisition of energy-related data providing accessible knowledge on the actual performance of buildings during their operation. A proper analysis of such data supports energy and facility managers in 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 energy performance patterns enabling a better decision-making to reduce energy wastes and system inefficiencies. To this aim, this paper introduces a novel meter-level ADD process capable to identify energy consumption anomalies at meter-level and perform diagnosis by exploiting information at sub-load level. The process leverages supervised and unsupervised analytics techniques coupled with the distance-based contextual matrix profile (CMP) algorithm to discover infrequent subsequences in energy consumption timeseries considering specific boundary conditions. The proposed process has self-tuning capabilities and can rank anomalies at both meter and sub-load level by means of robust severity score. The methodology is tested on one-year energy consumption timeseries of a medium/low voltage transformation cabin of the university campus of Politecnico di Torino leading to the detection of 55 anomalous subsequences that are diagnosed by analysing a group of 8 different sub-loads.

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 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 data volume 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 of buildings and systems. Building-related data are heterogeneous and reflects the complex interactions occurring between occupants, energy systems, building envelope, and forcing conditions [4]. If properly managed, ingested and analysed, such 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) and energy domain knowledge proved to be effective in achieving relevant energy saving by exploiting a variety of energy management solutions [6]. The tools providing such capabilities are the so-called energy management and information systems (EMIS) which are employed to monitor,

* Corresponding author
Email address: marco.piscitelli@polito.it (Marco Savino Piscitelli)
URL: www.baeda.polito.it (Marco Savino Piscitelli)

analyse and control energy systems in buildings leveraging advanced data analytics technologies for supporting facility staff to enhance energy performance and efficiency [6]. 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 to the main sub-loads) while the second focuses on more detailed data pertaining to the operation of specific energy systems or components (e.g., component operation of air handling units in HVAC systems).

A subgroup of EMIS, conceived for the collection and analysis of meter-level data, are the so-called energy information systems (EIS). EIS typically focus on data not usually collected through building automation systems (BAS) providing visual and analytical insights also enabling predictive energy management strategies. According to Granderson et al. [9] EIS tools, such as energy consumption forecasting [7–10], anomaly detection and diagnosis (ADD) [11–15], advanced energy benchmarking [16–19], load profiling [20–22], and schedule optimization of building energy systems [23,24] proved to lead to systematic energy saving and, as reported in [6], to an average two-year simple payback period, thus demonstrating their high competitiveness as profitable investment option in the building sector.

Among these solutions, ADD has been the most underdeveloped for application on meter-level data [11,12]. ADD tools allow a prompt and automatic recognition of abnormal and non-optimal performance patterns of energy systems providing information for the identification 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 related to the influence of several factors (e.g., building features, climatic conditions, occupant behaviour, comfort requirements, operating schedules and management) it is of considerable value in real world case studies where the available measured variables are commonly related to the meter scale [15]. 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) the detection of infrequent/anomalous patterns and (iii) the diagnosis of the detected anomalies by inferring the occurrence of anomalous patterns at sub-load level.

According to these objectives, this paper introduces the conceptualization and development of a novel EIS tool capable of robustly performing a meter-level ADD on building electrical energy consumption timeseries. The proposed approach employs a pattern recognition technique derived from the matrix profile (MP) [25] algorithm, called contextual matrix profile (CMP) [26] for the automatic detection of energy consumption anomalies at the whole-building level and their diagnosis by analysing the group of associated sub-loads. The following sections review the main works related to ADD in the energy and buildings field and to the application of MP-based algorithms as anomaly detection methods in timeseries.

1.1. 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 [27]. 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 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 energy consumption and technical and non-technical energy losses [28]. Thus, the nature of energy timeseries data in buildings requires to carefully address the definition of anomaly, which can be classified as point, collective, or contextual [29]. 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. Eventually, a *context anomaly* is so defined only if considered in a certain context (i.e., boundary conditions) and may not be considered an anomaly in a different context [30].

ADD has been traditionally addressed by means of statistical analysis, however the increasing spread of advanced machine learning and data analytics techniques coupled with the large volume of available data have opened new possibility to develop more sophisticated and reliable data-driven processes [3].

In ADD, the detection phase is usually accomplished by estimating a reference baseline representing normal behaviour according to specific boundary conditions, and labelling each observation that diverges from it as anomalous [29]. 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., datasets with verified ground truth about the 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 [31]. Zhao et al. [32] developed an SVM-based process that proved to be effective in detecting anomalies with different severities in chiller operation. Regression and decision tree classifiers are other widely used supervised techniques to develop energy consumption reference models for discovering abnormal patterns [33]. An evolutionary tree model was employed as a detection method in [12] to effectively discover frequent and infrequent patterns in meter-level electrical loads. A hybrid neural net ARIMA model was employed in [34] to predict the energy consumption and then identify anomalies comparing the actual and predicted energy consumption using the two-sigma rule. In [11] 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 applications is still limited compared to unsupervised methods, mainly due to the absence of reliable annotated datasets [28,35] and the difficulty in obtaining high quality training data [23] for machine learning algorithms.

Unsupervised anomaly detection is more promising for practical ADD applications since it makes it possible to detect rare and unknown anomalous patterns, without any a-priori knowledge and does not require pre-labelled datasets [36]. Beside statistical unsupervised methods, such as principal component analysis (PCA) [37] or generalized extreme studentized deviate (GESD) [8], data mining methods, such as association rule mining (ARM) [27] and clustering analysis [38], gained popularity thanks to their capability to automatically extract significant relations between complex and massive data [5,39,40]. 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 in form of 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 subsequences related to discordant observations (i.e., discords) that diverges from the rest of the dataset [28,41,42]. Li et al. [43] employed GESD to identify anomalous observations in electricity usage timeseries referred to 40 buildings. Fan et al. [8] 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 applied on smart meter data stream was presented in [44], where ARM and categorical clustering were used to detect anomalies.

In some cases, a reduction and transformation of timeseries data, could enhance the performance and reduce the computational time of anomaly detection techniques, also helping to effectively extract useful information. Lin et. al [45] 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 (i.e., motifs and discords). This method, which introduces a simple and low-computational cost process 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 [46,47]. Miller et al. [13] 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 [14] 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 [15] 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 [12].

Despite SAX introduced a lot of opportunities in the field of ADD in timeseries, 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 [47]. 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 [13].

One of the most promising timeseries analytics techniques, that is not subjected to information loss, is matrix profile (MP). Introduced by [25], MP is a novel exact pattern recognition algorithm that performs all-similarity-join-search (i.e., full-join) among timeseries, i.e. given a collection of data objects it retrieves the nearest neighbour for every object (where the considered object is intended to be a timeseries subsequence). MP is an unsupervised distance-based anomaly detection algorithm that proposes a fast similarity search under the z-normalized Euclidean distance, does not reduce dimensionality, calculates the full-join among timeseries and eliminates the need of setting a threshold making the method almost parameter-free and exact. MP 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. [48] proposed a real time MP-based anomaly detection method which was tested on electro-cardiogram (ECG) timeseries. The method employs the concept of shaplet [49] 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 [50], where the MP was combined with the hamming distance to automatically detect intrusions in the network of a water processing facility. PanMP is a generalization of MP algorithm able to discover anomalous subsequences of different lengths that was introduced in [51], where it was employed to perform anomaly detection in automated pedestrian counting system developed in Taipei.

MP has been also largely employed to identify anomalies in IT field. Herath et al. [52] 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.[53] applied a noise elimination technique on real Yahoo! internet traffic timeseries and detected anomalous behaviours through MP; while in [54] was demonstrated how the elimination of noise can help in anomaly detection of noisy dataset by testing the algorithm on Numenta Benchmark [55].

In the energy field few implementation studies of MP algorithm are available. Nichiforov et al. [56] 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 length as a feature extraction method to identify unusual behaviours in energy consumption timeseries of a large academic building dataset with a length of one year. The process was tested on 422 buildings whose end-uses were classrooms, offices, laboratories and dormitory. Zhu et al. [57]

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. [21] applied MP as a part of an automated load profile discord identification (ALDI) based on statistic comparison between normal and anomalous energy consumption patterns in a large portfolio of buildings. The MP method was used to quantify the similarities of daily electrical load subsequences under z-normalized distance. The computed MP values were then compared with typical-day MP distribution and it was effective in the identification of unique load shape patterns and discords.

Despite MP proved to be effective in several application fields, in its full-join formulation, it may compare regions of timeseries characterized by different operating and boundary conditions (i.e., contexts) and, in specific domains of application, may conduct to misleading results in terms of pattern similarity. For this reason, in [26] was introduced a 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 subsequences 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 performs the same search within a reference context set by the analyst.

An application of the CMP method, for anomaly detection in buildings, is presented in [26], where the method was applied to a dataset including different variables related to indoor air quality (e.g., temperature, humidity, CO_2 etc.) measured in residential built environments. CMP was applied on CO_2 concentration timeseries, enabling the identification of six behavioural anomalies (i.e., subsequences) occurring only during weekend mornings. 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 of CMP in the building energy field, in which the definition of anomaly is strongly related to the expert definition of contexts and boundary conditions, also leading to the robust recognition of patterns not easy to be detected. 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 theoretical 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 ADD in energy and buildings.

1.2. Matrix profile and contextual matrix profile method for anomaly detection

As defined in [25], given a collection of data objects (e.g., timeseries), MP performs the so-called similarity-join-search which is able to find the nearest neighbour for every object (e.g., identification of nearest pairs of subsequences) from the two collections.

Given two timeseries and a 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 position index of the nearest neighbour in the second timeseries. By joining information of MP and MPI useful knowledge could be extracted. By finding the minimum value of the MP it is possible to identify the best matching subsequences 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 from its nearest match (i.e., discord discovery) [25]. In this sense, discord discovery may be interpreted as an anomaly detection procedure 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.

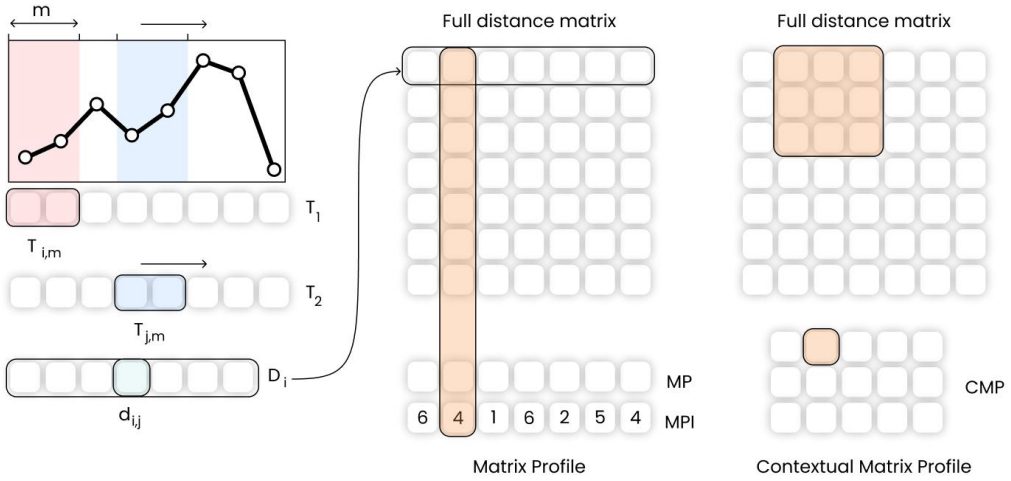

Figure 1. Description of MP and CMP calculation steps in case of self-join of a timeseries $T_1 = T_2$. From left to right is explained the calculation of the element $d_{i,j}$ of the distance vector D_i given the query $T_{\text{i.m}}$. By calculating the distance vector for the *all-subsequences-set* of T, 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.

First, a *timeseries* $T \in R^n$ is a sequence of real-valued numbers $t_i \in R : T = \{ t_1, t_2, \dots, t_n \}$ with $1 \leq i \leq n$ where n is the length of T. Since the focus is on local properties of timeseries (i.e., portions of timeseries), a *subsequence* $T_{i,m} \in R^m$ is defined as a continuous subset of values from T of length m starting in position i; formally defined as $t_i \in R : T_{i,m} = \{t_i, t_{i+1}, \dots, t_{i+m-1} \}$ with $1 \leq i \leq n - m + 1$.

An ordered set of all possible subsequences of T, obtained by sliding a window of length m across T, is called *all-subsequences-set* A of a timeseries T and is formally defined as follows: $A = \{T_{1,m}, T_{2,m}, \dots, T_{n-m+1,m} \}$ where m is a user-defined subsequence length.

By computing the distance between a given query (i.e., subsequence $T_{i,m}$) and each subsequence in the *all-subsequences-set* A, it is possible to define a vector of distances called *distance profile* D_i of a timeseries T. Formally, $D_i = [d_{i,1}, d_{i,2}, \dots, d_{i,n-m+1}]$, where $d_{i,j} = \text{dist}(T_{i,m}, T_{j,m})$ for all $j \in [1,2, \dots, n - m + 1]$, where $i \neq j$ and dist is the distance metric applied. It is possible to adopt different kind of distances to compute the *distance profile* as reported in [57–59] but the original MP method makes use of Euclidean distance between the z-normalized subsequences.

If the *distance profile* is calculated between a query in $T_{i,m}$ and the *all-subsequences-set* of T (i.e., self-join), by definition the i^{th} location of the *distance profile* D_i is zero since the distance is calculated between the query and itself, $d_{i,i} = \text{dist}(T_{i,m}, T_{i,m}) = 0$. Moreover, the distance is close to zero just before and after this position. Those matches are called *trivial matches* and are usually avoided during similarity-search by imposing an *exclusion zone* before and after this location (as function of m, usually set to $m/4$).

By calculating all the *distance profiles* for the *all-subsequences-set* of T, the *full distance matrix* is obtained. Formally defined as $\mathcal{M} = [D_1, D_2, \dots, D_{n-m+1}]$, it includes $n - m + 1$ *distance profiles* arranged in a square matrix.

Eventually it is possible to define MP as the vector that stores the z-normalized Euclidean distance values between each subsequence $T_{i,m}$ and its nearest neighbour. Formally, $MP = [\min(D_1), \min(D_2), \dots, \min(D_{n-m+1})]$, where D_i is the *distance profile* corresponding to query $T_{i,m}$ and timeseries T. In other words, it can be generated by extracting the smallest value in each row/column of the *full distance matrix* $\mathcal{M}$. 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 T_1 subsequence for any subsequence

in T_2. 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 have been proposed for the MP calculation to reduce time and dimensionality complexity such as STAMP, STAMPI and STOMP based on MASS algorithm [60], approximated AMPSA and AMP [61] and multidimensional mSTAMP [62].

MP represents a reference algorithm in the field of timeseries analytics and in particular among similarity-join-search algorithms. However, performing a similarity-join-search considering all the possible subsequence pairs in a 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 subsequence also exploiting user's knowledge. The concept of Annotation Vector (AV) can be then used to introduce domain knowledge in the process of motif and discord discovery [63], which allows to find results that respect user defined constraints and produce robust results, closer to expectations of the analyst. AV is a meta timeseries used to correct a-posteriori the values of the original MP, changing its shape and then manipulating the motif/discord search [64]. However, this method does not modify the MP calculation itself: *all-pairs-similarity-search* is always performed and then a downstream processing is conducted.

As a possible solution, in [26] was introduced the concept of contextual matrix profile (CMP), defined as the minimum over rectangular regions of the *full distance matrix* (see Figure 1), allowing then the search of the best matching subsequences over T_1 and T_2 only considering a-priori determined set of timeseries segments and excluding others.

This allows to group data in the timeseries in a custom way comparing only portions of T_1 with portions of T_2. The CMP calculation is based on the definition of contexts m_c, intended as lapse of time, during which a subsequence of length m may start. Figure 2 shows an example of context and subsequence length definition. Suppose that in an electrical load timeseries T with $1h$ timestep an analyst wants to find for every day, included in the timeseries, the nearest neighbour between the 3 subsequences of length 8h that could start at 04:00, 05:00 or 06:00 a.m., avoiding then similarity search in other time intervals. To this aim the subsequence length must be set to $m = 8$, (grey regions in Figure 2) and the context length $m_c = 3$ (yellow regions in Figure 2) and the resulting self-join CMP will have a row/column for each day. Each point of the CMP would display the distance between the best matching 8h-long subsequences occurring in two different days: lower the distance better the match and vice versa. While context is suitable to define a-priori the subsequence matching in the timeseries, once the CMP is calculated it is even possible to further divide the CMP in fragments, so-called *groups*. For instance, in the case of building energy consumption timeseries, it is possible to fragment the CMP by keeping only weekdays or weekends in order to consider the existence of different occupancy-based operational schedules in the discovery of motif and discords. Considering the main features and parameters (i.e., context and subsequence lengths, groups) 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.

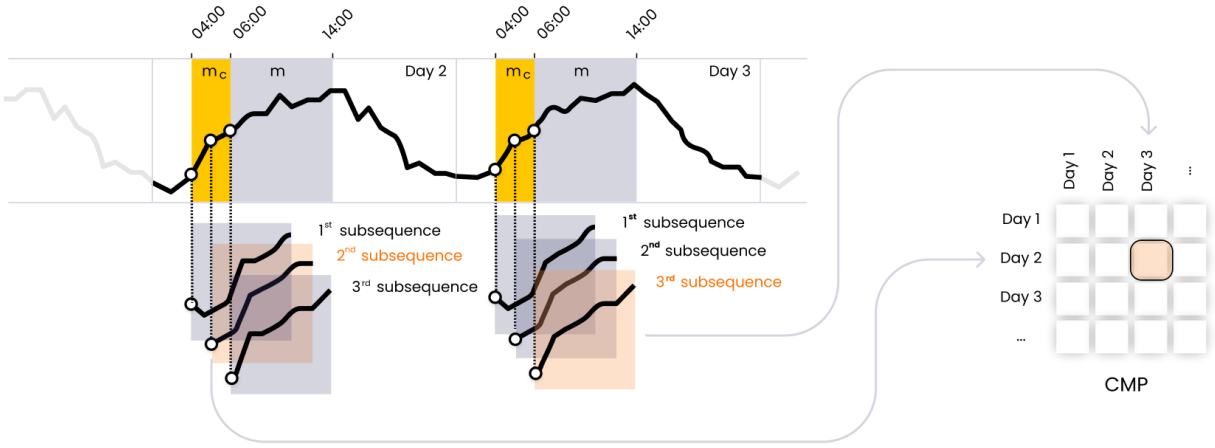

Figure 2. Focus on the definition of subsequence length m and contexts m_c in an electrical load timeseries T and their importance in the CMP calculation. The timeseries has one hour frequency and for each day a context of $m_c = 3$ (yellow) and a subsequence length of $m = 8$ (grey) are defined. When 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.

1.3. Research gap and contribution of the paper

The MP algorithm has been successfully employed in different anomaly detection applications and researchers have proposed different MP implementations according to specific constraints and boundary conditions that are characteristic of the domain of interest.

In buildings, anomalous subsequences in energy consumption timeseries are defined as unexpected behaviours that result in subsequences with atypical shape and/or magnitude. The original MP algorithm, through z-score normalization of the subsequences, searches the nearest neighbour based on shape similarity by losing 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 weekends when the load profile is almost flat. Applying the MP algorithm 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, Figure 3(c) shows that the not-normalized subsequences have very different amplitudes and are referred to distinct energy consumption patterns that are typical of weekdays and weekends respectively. This is a clear example of how the subsequence normalization could led to misleading results in terms of subsequence similarity.

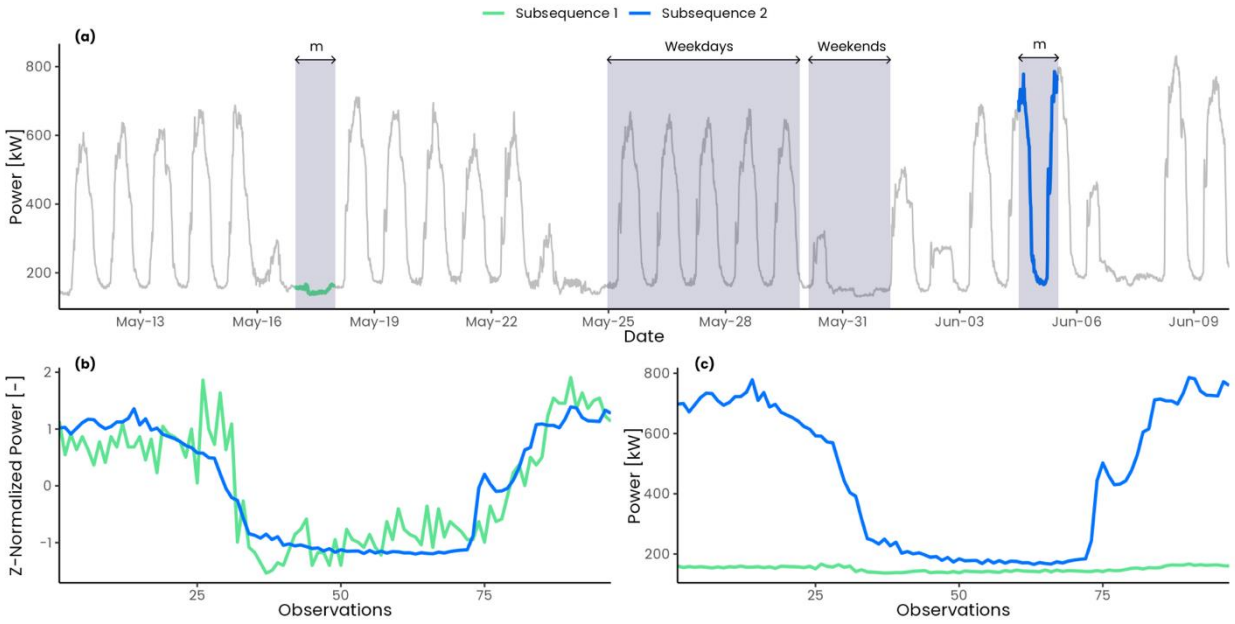

Figure 3. Effect of z-score normalization on two electrical load timeseries subsequences (blue, green) of length 24h and a timestep of 15-min (96 observations per day): (a) full electrical load timeseries; (b) comparison between z-score normalized subsequences; (c) comparison between not normalized subsequences.

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 z-normalized Euclidean distance is higher if compared to non-flat subsequences, leading to high 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, the normalization led to an amplification of the noise of the original subsequences (Figure 4(c)), resulting into a high Euclidean distance d = 9.25, while in the second case (Figure 4(d)), the Euclidean distance is much lower (d = 1.5). This issue has been largely analysed in [53], considering as possible solution to discard flat regions from the analysis, to change the subsequence length, or to smooth the timeseries. As a clear consequence of normalization, in an electrical load timeseries (Figure 3(a)), the MP method would identify the weekends as discords since they are typically associated to flat profiles compared to weekdays subsequences. 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.

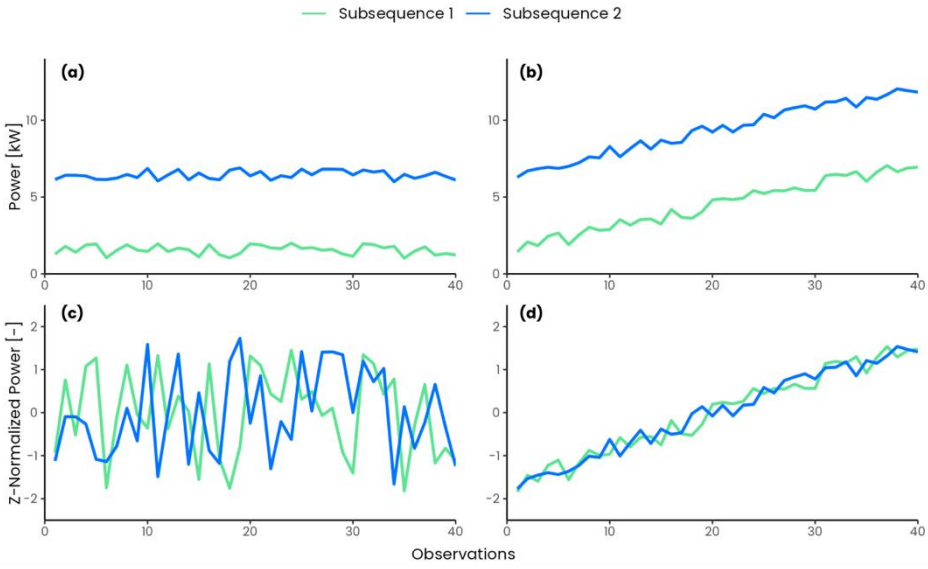

Figure 4. Effect of z-score normalization on noisy flat subsequences (a-c) and on noisy subsequences with positive slope (b-d). Adapted from [54]

In this perspective, comparing subsequences belonging to different energy patterns could be not so useful, making the introduction of domain knowledge, extremely important to drive the motif and discord discovery in timeseries. As previously discussed, a possible solution has been proposed by [26] introducing the CMP algorithm. In some applications CMP can be useful, since it allows to exclude some regions, or to split subsequences into different groups, and then perform the similarity search.

The application of CMP, coupled with domain knowledge, can be effectively exploited for the accurate discovery of anomalies in energy consumption timeseries, reducing energy wastes and enhancing energy management in buildings. Domain expertise should be used for an aware setting of the parameters of CMP algorithm (i.e., subsequence length, context length and groups) to avoid undesirable results and to make the ADD process as robust as possible. However, this is not a task that can be easily generalizable and still requires research effort for applications in the energy and building field.

To this purpose, according to the previous literature review and to the reasoning on MP and CMP methods, this paper intends to contribute as follows:

1) Develop a contextual ADD methodology by introducing a CMP-based process that employs supervised and unsupervised algorithms (i.e., clustering and decision tree) for the identification of parameters such as subsequence length, groups and contexts, systematically introducing domain expertise in the whole analytical framework.
2) Develop an ADD methodology based on the concept of *similarity-join* without any loss of information in the discovery of motifs and discords (e.g., avoiding dimensionality reduction and transformation of timeseries as in SAX algorithm) that is flexible enough to identify anomalous/typical patterns even if they are not perfectly aligned in time (i.e., time-tolerant analysis according to the length of the contexts).
3) Perform a discord discovery that is sensitive to both shape and magnitude of energy consumption patterns. To this purpose normalization aspects are considered upstream the computation of subsequence distances in CMP allowing a knowledge-based comparison between energy patterns in the analysed time series.
4) Introduce a robust severity score, based on different statistical methods and domain knowledge, that allows potential meter-level anomalies to be identified, ranked and diagnosed, within particular boundary conditions.

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 ADD results while Section 4 critically discusses the outcomes and includes the concluding remarks.

2. Methodology

In this section the methodological framework employed to conduct ADD in energy consumption timeseries of buildings 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 the tuning of CMP parameters (subsequence length, context length, groups). The procedure, depicted in Figure 5, unfolds over of six steps described in detail in the following paragraphs.

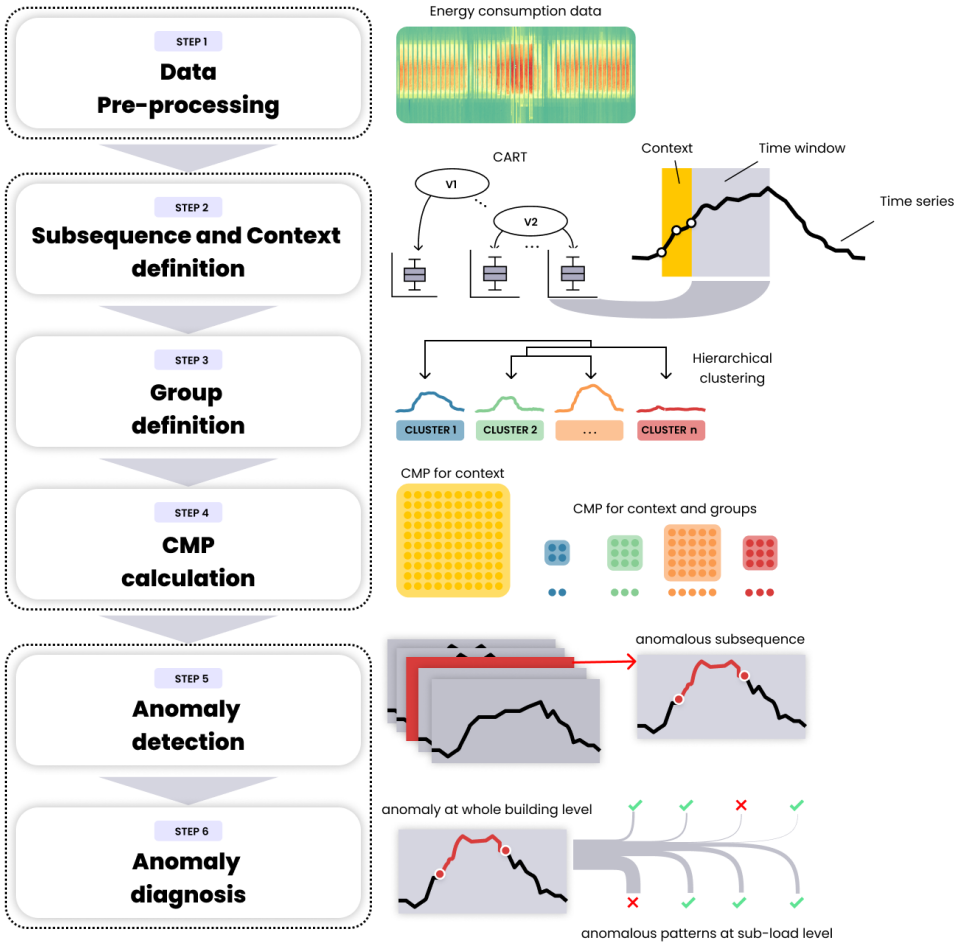

Figure 5. Graphical representation of the methodological framework.

2.1. Pre-processing.

The first step consists in data pre-processing and is 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.

2.2. Contextual matrix profile.

By following the methodological framework in Figure 5, the application of the CMP method goes through the following steps: subsequence and context definition (step 2), group definition (step 3) and CMP calculation (step 4).

Within the energy consumption timeseries of buildings, different characteristic sub-daily load patterns (e.g., base load, peak load, ramp-up and ramp-down period) can be identified and set as relevant subsequences to perform similarity-join-search by means of CMP algorithm. Such subsequences, with their relative lengths, can be statistically defined or inferred from the typical building operational and occupational schedule [65,66] The methodology proposed in this paper identifies sub-daily subsequences (t_w) through the implementation of recursive partitioning Classification and Regression Tree (CART) [15]. Starting from the root node (that contains all the available observations) this method proceeds through a set of binary decisions to split the instances in purer subsets (nodes) in a forward stepwise fashion, minimizing at each decision step the impurity/variance of each node [15,27,67], yielding local optimum [68] once a stopping condition is satisfied. To address this task, the decision tree was set as a regression model using the electrical load values 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 leaf nodes that group together hours with similar electrical load values in n non-overlapping sub-daily subsequences in the time domain $t_{w,i} \in \{t_{w,1}, t_{w,2}, \dots, t_{w,n}\}$. As a reference, the choice to not setting the time variable as a categorical one, allows CART to not indiscriminately group together electrical load values pertaining to the early morning with those referring to the night.

Thus, for each leaf node it is possible to extract a start and end-time of the subsequence and its length in terms of time duration ($m_i = t_{w,i}$). Once, all the interesting subsequences are identified, for each of them a context period is set to drive their *similarity-join-search* in the analysed time series of energy consumption. As reported in Figure 2 the context is defined as the time lapse during which can be located the start-time of a subsequence. This allows to investigate similarity of subsequences of the same length but not necessarily aligned in the time domain. Specifically, the context i ends at the beginning of the subsequence $t_{w,i}$, assuming as reference start time of $t_{w,i}$ the one extracted from the leaf node of the regression tree. Instead, the context length is set equal to the half of the smallest subsequence length ($m_c = \min(t_{w,i})/2$) identified through the regression tree. For instance, if the smallest subsequence is two hours long $t_w = 2h$ (e.g., from 6:00 to 8:00 a.m.) the context is defined as one hour long $m_c = 1\text{h}$ (i.e., from 5:00 to 6:00 a.m.).

The following step is then aimed to define groups in CMP. This step is based on daily load profile clustering. An expert-based approach is firstly applied to split working days, holidays/not working days and partially working days (e.g., Saturdays). Then a nested hierarchical clustering analysis is performed on the daily profiles of working days to better characterize them. The hierarchical clustering generates non-overlapping groups 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 (i.e., dendrogram). The grouping of daily load profiles in representative clusters of specific energy consumption patterns, allows the CMP results to be evaluated in groups with high internal similarity and to make the anomaly detection 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 n not overlapping contexts and the resulting CMP contains one row/column for each day. Eventually, the obtained overall CMP matrix 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.

2.3. Anomaly detection

The anomaly detection process, shown in Figure 6, is carried out in each group after performing the similarity-join-search of subsequences in their relative 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., V_{CMP} and V_{EN}. The first vector V_{CMP} contains the median value of each row/column of the CMP i.e., the median of the Euclidean distances between a specific subsequence and its nearest neighbours in the associated context considering only the days in the same group (i.e., cluster of daily load profiles). The second vector V_{EN}, contains for each subsequence the corresponding energy consumption. In this way it is possible to label as anomalous, subsequences with Euclidean distance and energy consumption values that are out-of-range respect to the domain that is considered normal for both V_{CMP} and V_{EN}. With reference to Figure 6, as a first step the CMP matrix is reduced into a vector V_{CMP} by calculating the median of each row/column, then the four outlier detection methods are applied to each component of V_{CMP} producing 4 new vectors that define whether an element is anomalous or not in a boolean form $B_i = \{0,1\}$. Then the severity score associated to a subsequence is calculated counting the number of positive detections $S_{CMP} = \sum_{i=1}^{4} B_i$. To make more robust the anomaly detection process and consider only positive anomalies (e.g., anomalies that resulted into an over-consumption of energy) the energy consumption for each subsequence is calculated and stored in a vector V_{EN} which undergoes to the same process described before: outlier methods are applied and then the severity is calculated. With reference to Figure 6, given $n = 4$ outlier detection methods the severity vector S_{CMP} and S_{EN} range from a minimum of 0 (four methods out of four did not tag the subsequence as out-of-range observation) to a severity of 4 (four methods out of four found tagged the subsequence as out-of-range observation). By summing S_{CMP} and S_{EN} the resulting overall severity S ranges from 0 (four methods out of four did not tag the subsequence as out-of-range observation for both energy and Euclidean distance vector) to 8 (four methods out of four tagged the subsequence as out-of-range observation for both energy and Euclidean distance vector).

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

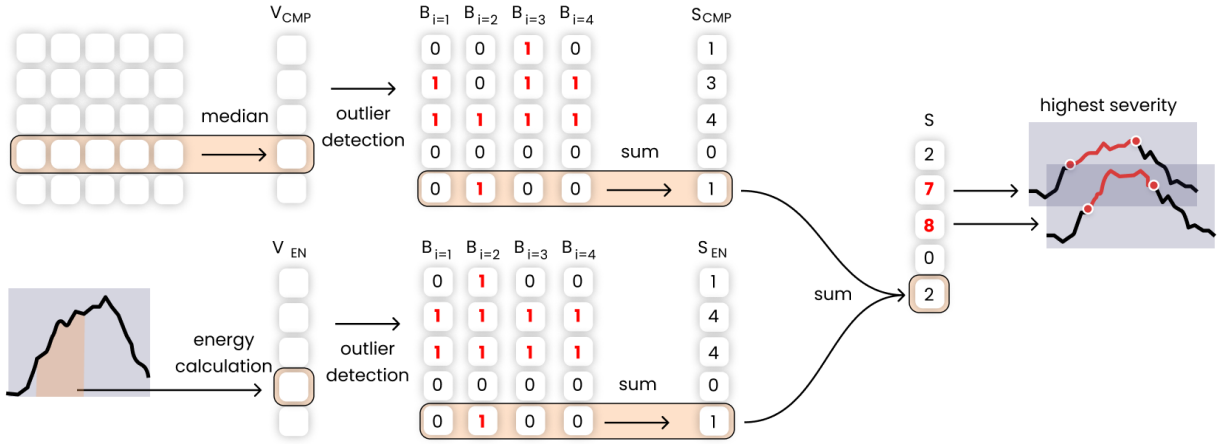

Figure 6. Graphical representation of the anomaly detection phase.

The four univariate outlier detection methods are: (i) inter quartile range analysis, (ii) Z-score, (iii) elbow method and (iv) Generalized Extreme Studentized Deviate. The basic principles behind these statistical methods are introduced and described below.

Inter quartile range analysis defines as outlier any observation that falls below $Q_1 - 1.5 * IQR$ and above $Q_3 + 1.5 * IQR$ where the interquartile range (IQR) is defined as the difference between the third quartile and the first quartile $IQR = Q_3 - Q_1$ of the population of observations. In this study only positive outliers are considered i.e., values above $Q_3 + 1.5 * IQR$.

Z-score, or standard score, is the output of data standardization process and is intended as the number of standard deviations by which an object, in a population of observations, is above or below the mean value. Observations above the mean have positive Z-scores, while those below the mean value have negative Z-scores. This method can be employed to identify extremal observations (i.e., outliers) by setting a z-score threshold value. For instance, define a Z-score threshold equal to 2 means that the all the observations that differ from the mean more than 2σ are tagged as outliers, and in the case of normally distributed observations they represent the 2.3% of the whole population.

Elbow method is a heuristic method that allows to find the elbow (or knee) of a curve. The implemented method employs the kneedle algorithm [69] 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 $T = \{ t_1, t_2, \dots, t_n \}$ through a statistical test [8]. The method initialization requires ($i = 1$) a presumed number of outliers r and confidence interval α, then the following statistical test is performed:

- H_0 There are no outliers in the timeseries
- H_a There are up to r outliers in the timeseries

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

$$R_i = \frac{\max|t_i - \mu_T)|}{\sigma_T} \quad \lambda_i = \frac{(n-i) * t_{p,n-i-1}}{\sqrt{\left(n-i-1+t_{p,n-i-1}^2\right) * (n-i+1)}}$$

Where μ_T and σ_T denote sample mean and sample standard deviation of the timeseries, n is the timeseries length, $i = \{1,2,\dots,r\}$ is the iteration number, $t_{p,\nu}$ is the 100p percentage point from the t distribution with ν degrees of freedom and $p = 1 - \frac{\alpha}{2(n-i+1)}$

2.4. Anomaly diagnosis phase

Once the anomaly at meter-level is identified (i.e., anomalous subsequence with at least a severity score S equal to 6), the diagnosis phase is enabled with the aim to identify if also sub-loads are anomalous in the same time interval and contribute to the diagnosis definition. The diagnosis process involves only those subsequences that resulted to be anomalous at meter-level and extend the analysis at sub-load level.

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 subsequence of the sub-loads, it is possible to assess if the sub-load is anomalous by calculating the severity score that allows to rank the sub-loads accordingly (i.e., from low to high severity). The sub-loads that present severity scores in the range (6 - 8) are assumed as anomalous and contribute to the diagnosis of the anomaly detected at higher level. Otherwise, if no sub-load, for a specific anomalous subsequence at meter level, reaches an anomaly score at least of 6, the anomaly is considered as undiagnosed.

3. Results

The analysis was carried out using the R statistical software [70] for the pre-processing, regression analysis, cluster analysis and visualization, and Python [71] for the CMP calculation and anomaly detection process.

The presented methodology has been tested on the energy consumption timeseries of a MV/LV transformer cabin (i.e., substation C) that serves part of the university campus of Politecnico di Torino located in the norther Italy. The monitoring infrastructure continuously provides both meter-level and sub-meter level measurements of the mean electrical load with a timestamp of 15 min. The hierarchical structure of the installed monitoring infrastructure is shown in Figure 7: 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 breakdown in terms of average annual energy consumption was provided for each sub-load.

The substation serves a bar and a staff canteen that are at disposal of students and campus staff and, on average, represent the 2.75% and 15.98% of the yearly electrical energy consumption of substation C respectively. The university data centre accounts for the 13.17% of the total energy consumption. The administration offices (Rectorate) correspond to the 3.84% of energy consumption while the mathematics department (DIMAT) to the 2.21%. A relevant amount of energy consumption (12.18%) is related to the mechanical room. The equipment located in this room includes hot and chilled water circuits with their associated 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 single synthetic sub-load tagged as “Not labelled” as showed in Figure 7. It accounted for 48.85% of the yearly energy consumption of the substation and since it is not directly measured cannot be labelled, as the other sub-loads, respect to the energy service provided or building served. The authors tested the presented methodology on a one-year dataset that spans from 01/01/2019 to 12/31/2019. The year 2019 was selected, even if more recent data were available, mainly because the pandemic COVID-19 completely changed the operational patterns due to the closure of the university campus from February 2020. The raw dataset includes 35040 observations with 15-min timestep. The dataset is characterised by 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.

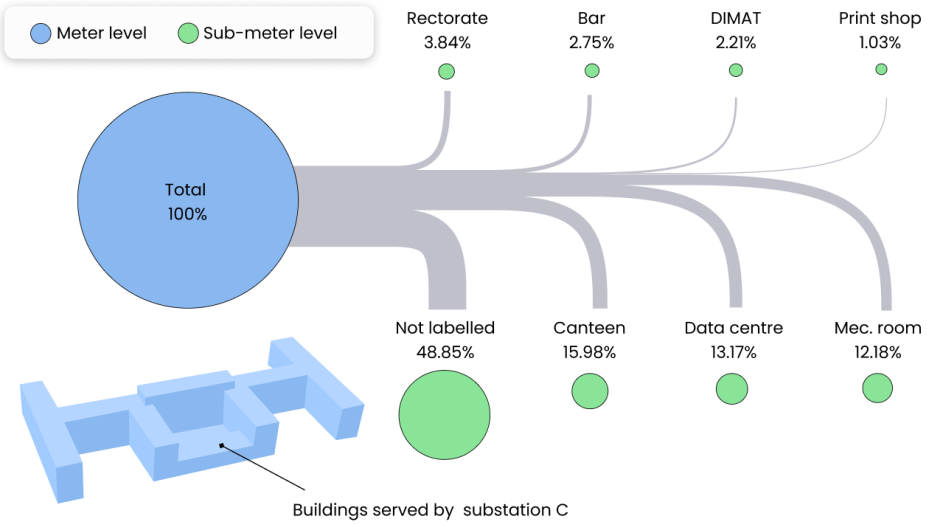

Figure 7. Hierarchical structure of the electrical load database under study.

3.1. Contextual matrix profile results

To identify the subsequences to be used for the *similarity-join-search* through CMP, not overlapping time periods were evaluated through CART algorithm using the meter-level 15-min electrical load as target variable, and hour of the day as numeric predictive variable. To preserve the accuracy of the model in the leaf nodes, pertaining the operation hours of the energy systems (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.

In fact, for days with a low standard deviation of the electricity demand, that are almost flat, the ADD results would be less sensitive to the subsequence definition. On the other hand, for days with high variability of electricity demand it is particularly important to extract the best patterns that can properly characterize the load variation over time. In particular the CART is extremely effective to identify leaf nodes (in this case associated with subsequence length) with low variable load and others that instead include all the load variability (i.e., morning ramp-up, evening ramp-down). Indeed, even though the accuracy is not the highest achievable, the results that can be obtained are exactly the ones of interest.

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 subsequences 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 are represented as sub-daily subsequences also in the Figure 8(b), where are reported with their duration on a daily scale.

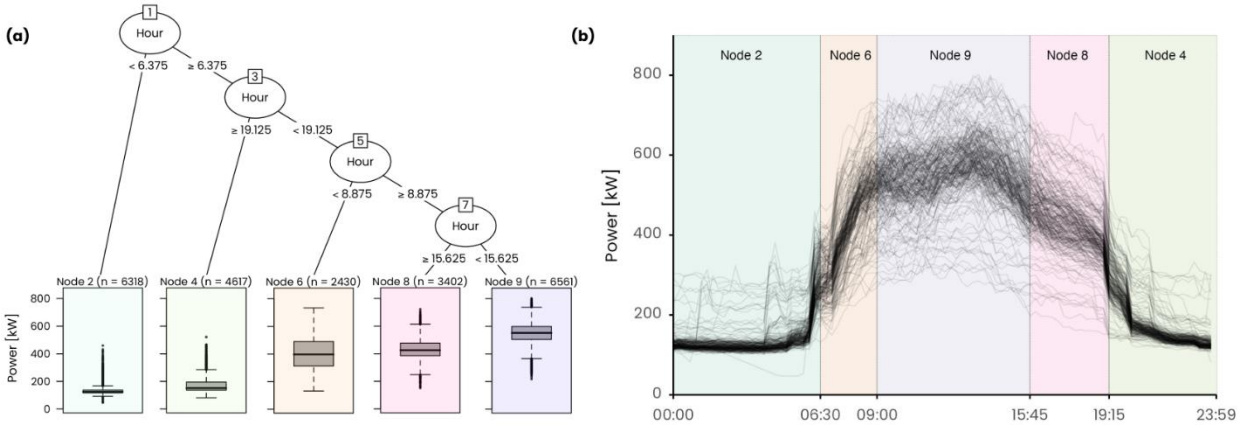

Figure 8. (a) Identification of sub-daily subsequence by means of the CART algorithm (b) graphical representation of the daily load profiles and the identified subsequences on daily scale.

The model was able to effectively separate the night hours (subsequences 1 and 5 extracted from leaf node 2 and 4 respectively of the tree in Figure 8(a)) from the diurnal operation and was able to identify the ramp-up (subsequence 2 extracted from leaf node 6 of the tree in Figure 8(a)), mid-day operation (subsequence 3 extracted from leaf node 9 of the tree in Figure 8(a)) and ramp-down (subsequence 4 extracted from leaf node 8 of the tree in Figure 8(a)). The smallest subsequence lasts 2.5 hours. Following the methodological process, once the reference sub-daily subsequences 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 subsequence length identified ($m_c = 2.5\mathrm{h}/2 \cong 1h$). The outcome of this preliminary step led to the definition of the lengths of 5 subsequences and 5 contexts (i.e., the number of observations in each subsequence and context) with their durations for the CMP calculation (Table 1). For instance, the third sub-daily subsequence, that ranges in [09:00 – 15:45) (as extracted from the tree in Figure 8(a)), has a length of 27 observations (i.e., $27 * 15\,min =$

6 h 45 min) and in the similarity-join-search with CMP all the subsequences of length 27 that starts in the 1h context [08:15 – 09:15) are compared to identify pairs of nearest neighbours among days in the same group.

Table 1. Context and subsequence lengths.

Subsequence				Context			
ID	Time interval	Duration	Length	ID	Time interval	Duration	Length
$t_{w,1} = m_1$	[00:00 - 06:30)	6 h 30 min	26	$m_{c,1}$	[23:15 - 00:15)	1 h	4
$t_{w,2} = m_2$	[06:30 - 09:00)	2 h 30 min	10	$m_{c,2}$	[05:45 - 06:45)	1 h	4
$t_{w,3} = m_3$	[09:00 - 15:45)	6 h 45 min	27	$m_{c,3}$	[08:15 - 09:15)	1 h	4
$t_{w,4} = m_4$	[15:45 - 19:15)	3 h 30 min	14	$m_{c,4}$	[15:00 – 16:00)	1 h	4
$t_{w,5} = m_5$	[19:15 - 24:00)	4 h 45 min	19	$m_{c,5}$	[18:30 - 19:30)	1 h	4

The second step consists in the definition of the groups and was performed using a semi-supervised approach applied on the set of 365 daily load profiles available in the dataset. Firstly 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 (i.e., 96 observations considering a 15-min timestep). Then a hierarchical clustering algorithm with Ward method was implemented on the not normalized daily load profiles. The silhouette index, implemented in the package NbClust [72], 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 were 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 well-defined groups of load profiles each one representing a typical load condition useful to split the CMP for a given context into homogeneous groups for conducting the anomaly detection process.

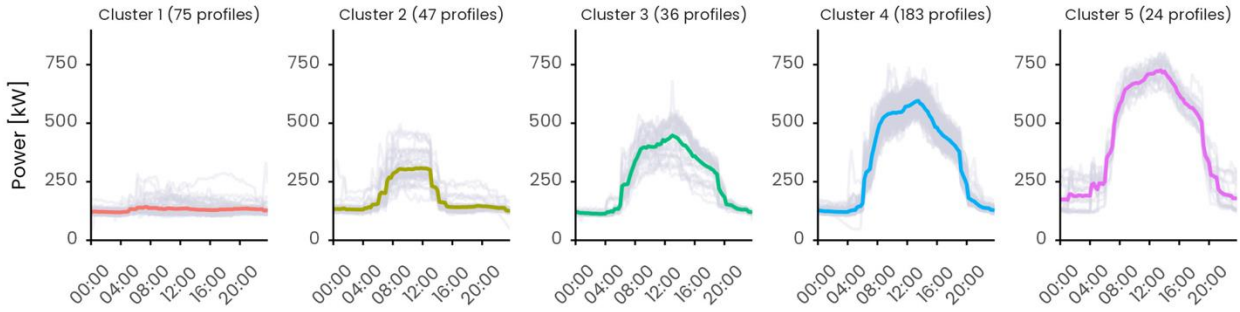

Figure 9. Clusters of daily electrical load profiles with the corresponding centroids.

The CMP was then 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 [26] 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 a specific context occurs one time per day the resulting CMP is a 365x365 symmetric matrix where each value represents the Euclidean distance between the best matching subsequences

among two days considering all the possible subsequences that start during the same context. The higher the distance value the higher the dissimilarity between a pair of nearest neighbours.

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 with 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 facilities of university campus. By further splitting the CMP in the previously defined groups (i.e., clusters), days expected to behave in a similar manner are grouped together to perform a more robust inspection of anomalous patterns. For instance, the day 10th of November 2019 (highlighted in Figure 10 at index 62 of cluster 1 sub-matrix of CMP) stands out to be remarkably different from all the other days in the same cluster, which is not so evident by only visualizing the full CMP. This highlights the impact that the group definition has, after the CMP computation, for the identification of contextual anomalies, given the relative importance that a pattern could have according to the group in which it is compared with its nearest neighbours. As a reference, performing clustering after the definition of the subsequences (and so contexts) makes it possible to extracted more general patterns (i.e., occurrence time of an anomaly) that can be compared among groups regardless to their composition (even in the case they change over time).

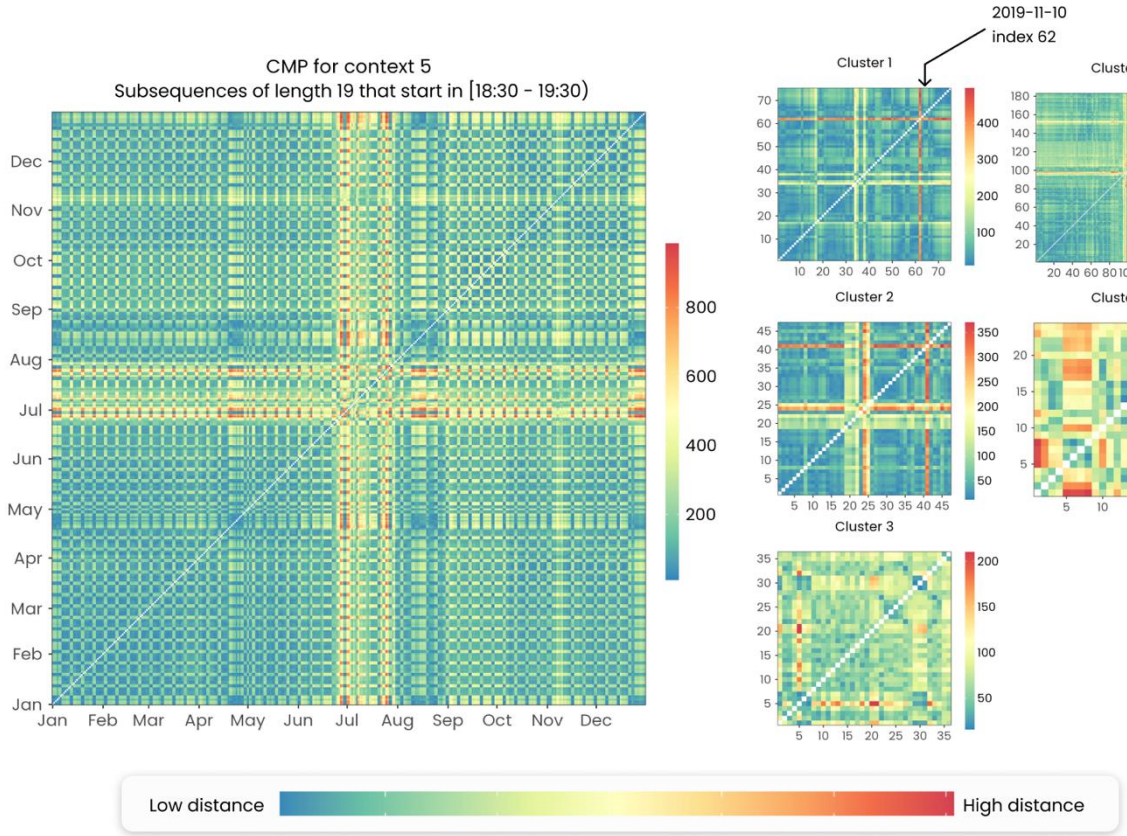

Figure 10. On the left the CMP result is reported for the similarity-join search of subsequences that start in context 5. Each value of the matrix shows the Euclidean distance between the best matching subsequences. Lower the distance better the match. The CMP is divided on the right side into submatrices according to the membership of each day to one of the predefined groups (i.e., clusters).

3.2. Anomaly detection results

The anomaly detection is performed among the subsequences that start in the same context considering a group at time. For each context-group CMP matrix, median values of columns/rows of the CMP and the energy consumption related to the subsequences are stored in the vector V_{CMP} and V_{EN} respectively, 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 + Q_3, the z-score method considers only the positive observations over a value of 2, GESD method is initialized with presumed number of outliers $r = 10$ and confidence interval $\alpha = 0.05$ and eventually, the elbow method, since it is a pure graphical method, considers as outliers all the observations above the evaluated knee point. Each method defines whether a subsequence is anomalous or not in a Boolean form $B_i = \{0,1\}$. Then, the severity is obtained summing up the number of positive detections. By summing the two resulting severity vectors S_{CMP} and S_{EN} an overall severity S ranging from 0 to 8 was calculated to robustly rank anomalies from the most to the least severe one.

Figure 11 reports the anomaly detection results obtained for subsequences pertaining the context 5 considering days in 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 consumption vector and this resulted into an overall severity score of 8.

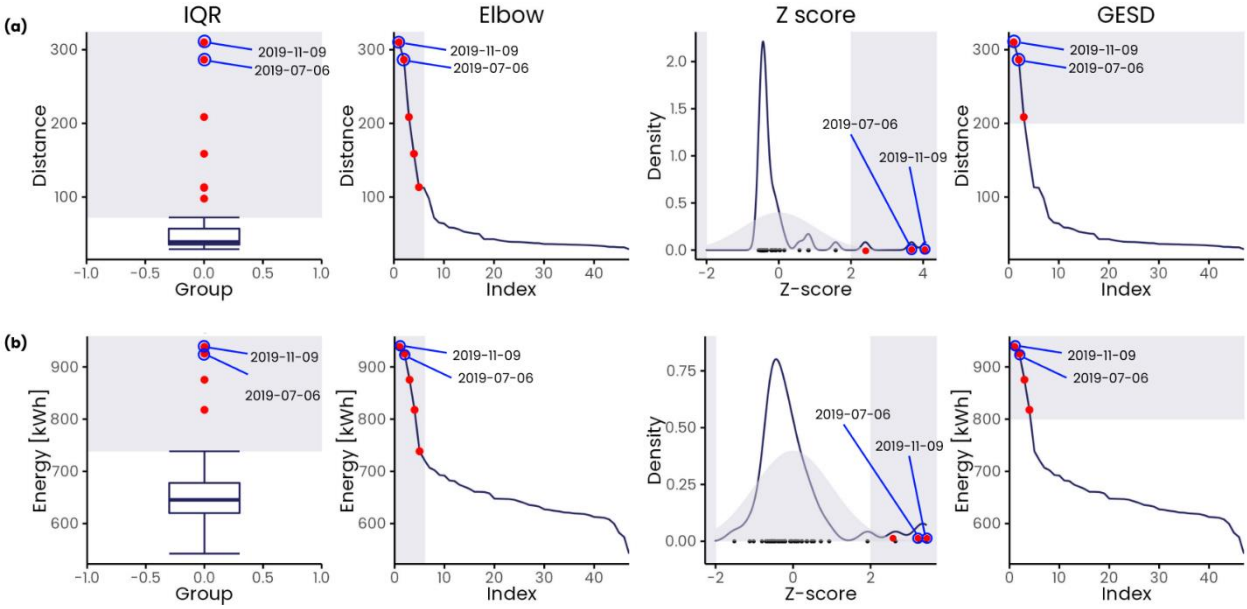

Figure 11. Example of severity calculation on (a) median Euclidean distance vector and (b) energy consumption vector using the four anomaly detection methods respectively IQR method, Z-score method, elbow method and GESD.

To reduce the number of spurious alerts, only the severities 6-7-8 were considered as relevant, resulting in 55 anomalous subsequences detected: 20 with severity 6, 7 with severity 7 and 28 with severity 8. A high severity denotes a significant difference in terms of shape and magnitude of a subsequence within the relative group and context.

Results are summarized in Figure 12 through a calendar plot which shows anomalies over the year for the 5 contexts considered in the CMP *similarity-join-search*. According to Figure 12, anomalies resulted to be more frequent at the beginning and end of the day: 13 anomalous subsequences started in the 1st context, 10 in the 2nd context, 7 in the 3rd context, 13 in the 4th context and 12 in the 5th and last context. Moreover, the calendar plot clearly shows as 44 anomalies out of 55 detected, occurred during summer, from June to August, compared to the rest of the year where only 11 anomalous subsequences were detected.

The definition of the 5 contexts and related subsequences, with diversified lengths among contexts, allowed the ADD process to be effective in finding anomalous subsequences that are limited to one context during the day (i.e., spot anomalies) and others that instead persisted among subsequent contexts and last until the end of the day (i.e., persistent anomalies).

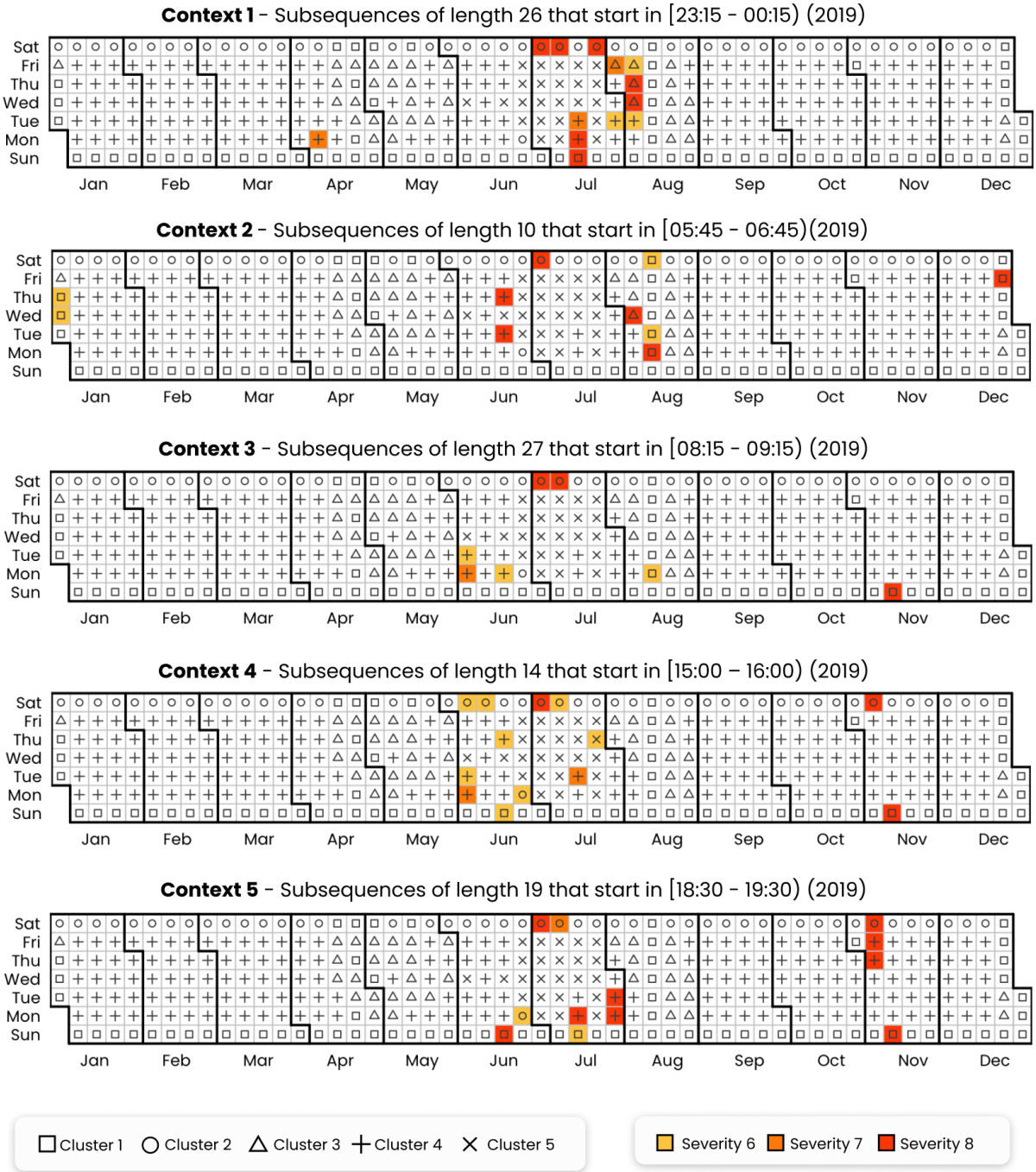

Figure 12. Calendar plot reporting the occurrence of the detected anomalies through the whole year dataset. Each row corresponds to a different context. For each day, the colour fill, represents the severity of the anomaly while the symbol represents the cluster membership.

An example of a spot anomaly is the subsequence starting at 05:45 in 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 belonging to cluster 1 where a flat profile is expected to be found. 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 about 285 kWh. The same pattern, shown in Figure 13 (b), was detected during summer season in the subsequence starting at starting at 05:45 of Monday 12th of August 2019, when despite the university campus was closed for the summer break, an abnormal increase in electrical load in the second context was detected resulting into a deviation of more than 280 kWh from the cluster centroid. These two examples are symptoms of a wrong schedule of the energy systems pertaining to the substation C.

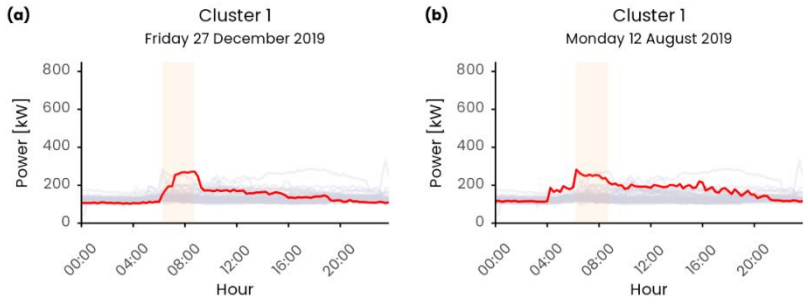

Figure 13. Anomalous subsequences identified as spot anomalies, respectively: (a) Monday 12th August 2019 (context 2, cluster 1); (b) Friday 27th December 2019 (context 2, cluster 1). The daily load profile which includes the anomalous subsequence is represented with the red line while the grey lines correspond to the daily load profiles included in the relative cluster. The vertical orange band denotes the anomalous subsequence while the band width is related to the subsequence length.

An example of persisting anomaly is Saturday 6th of July 2019 that presents anomalous subsequences 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 periods of day by keeping an average offset of almost 80 kW compared with the cluster centroid, leading to an overall deviation compared to the cluster centroid of about 2490 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 starting at 00:00 in contexts 1 and at 19:15 in context 5 leading to a deviation of about 800 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 10th of November 2019 an anomalous energy consumption involving subsequences that respectively start in contexts 3-4-5, is detected, leading to a deviation of about 1620 kWh from the reference behaviour of cluster centroid.

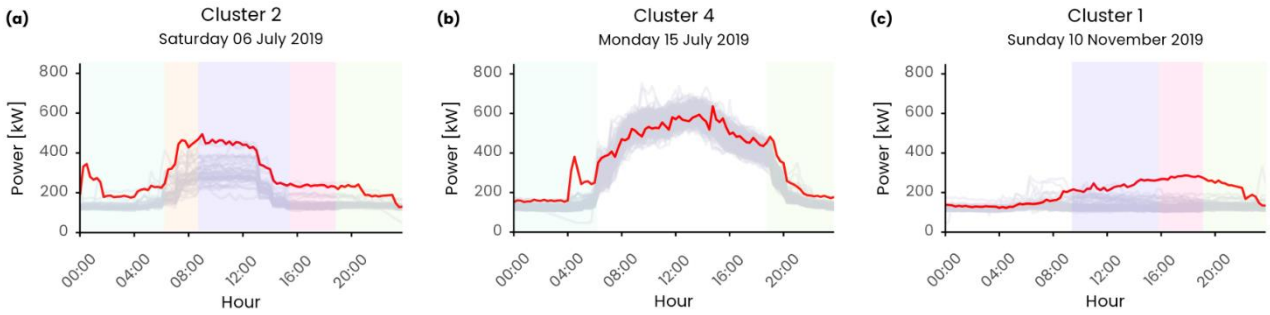

Figure 14. Anomalous daily load profiles identified as persisting anomalies, respectively: (a) Saturday 6th July 2019 (context 1, 2, 3, 4, 5 - cluster 2); (b) Monday 15th July 2019 (context 1, 5 - cluster 4); (c) Sunday 10th November 2019 (context 3, 4, 5 - cluster 1). The anomalous load profile is represented with the red line while the grey lines correspond to the load profiles contained in the relative cluster. The vertical band denotes the anomalous subsequences, the band colours are referred to the contexts considered, while the band widths are related to the subsequence lengths.

3.3. Anomaly diagnosis results

The diagnosis process aims to assess if during anomalous subsequences at meter-level also anomalies at sub-load level occurred, in order to provide an explanation to the detected infrequent energy consumption patterns. A sub-load with high severity score (i.e., higher than 6) is considered in the diagnosis of meter-level anomalies. The 55 anomalies discovered at meter-level are investigated and results are presented in the following. In particular, for 8 out 55 anomalous subsequences, any sub-load does not reach a severity score at least of 6. It means that those anomalies are considered as

undiagnosed given that it is not possible to clearly recognize abnormal patterns among sub-loads according to the developed severity score. However, for those subsequences, the "Mechanical room" resulted to be the sub-load with the highest severity score (ranging between 3 and 5) in any case supporting the end user in the process of interpreting the detected anomalous energy consumption pattern. For what concerns the 47 remaining anomalies, 30 times are associated with the occurrence of a single anomalous sub-load (i.e., with severity score at least of 6), 12 times with two anomalous sub-loads, and 5 times with three anomalous sub-loads. Additional 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 to the meter-level anomaly due to severity scores at least of 6. "Not labelled" and "Mechanical room" are most frequently included in the lists of sub-loads contributing to the diagnosis of the detected anomalies at meter level, occurring respectively 24 times and 26 times (also considering the co-occurrence with other sub-loads). Other sub-loads, such as "Canteen", "Data centre" were found to be anomalous 4 times while the "Print shop" exhibited an abnormal pattern 11 times. It is worth to note, that the severity score is formulated to reveal the occurrence of infrequent shapes and magnitudes of the subsequences by means of analysis that are internal to the timeseries of the single sub-load. It means that, even though some sub-loads have a relatively low impact in terms of yearly energy consumption on the total one of the whole substation C, the diagnosis process is able to highlight if they are normally behaving during an anomalous subsequence at meter-level regardless to their contribution to the total energy consumption.

Table 2. Summary of the diagnosis results.

Number of occurrences	Sub-loads contributing to the diagnosis
16	Not labelled
11	Mechanical room
5	Not labelled, Mechanical room
4	Mechanical room, Print shop
3	Not labelled, Mechanical room, Print shop
1	Canteen
1	Data centre
1	Print shop
1	Canteen, Print shop
1	Mechanical room, Data centre
1	Canteen, Data centre
1	Canteen, Mechanical room, Print shop
1	Data centre, Mechanical room, 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 the load profiles of Saturday 6th of July 2019 during which both the "Mechanical room" and "Not labelled" sub-load contributed to the anomalous trend detected on the total electrical load of substation C, starting in contexts 1-2-3-4-5. For subsequences starting in contexts 1-2 the "Mechanical room" and "Not labelled" are both responsible for the anomaly detected. In detail, to the "Mechanical room" sub-load, is associated a severity score of 8 in both contexts while the "Not labelled" sub-load has a severity score of 8 for the subsequence starting in context 1 and severity 6 for the one starting in context 2. 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 infrequent high night load of the "Not labelled", led to an abnormal behaviour at meter-level. Then, for the remaining subsequences starting in 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). Figure 15 shows that the anomalous load of "Mechanical room" after 12:00 led to a deviation of the "Total Power" load at meter-level until the end of the day.

Figure 15(b) shows the load profiles of Tuesday 30th of July 2019 in which a single sub-load contributed to the meter-level anomaly detected. An early morning start-up at 4:00 of the "Mechanical room" identified by a severity 7 contributed to the "Total power" anomaly of severity 6 starting in the first context. While an unexpected activity of the "Canteen" load after 19:00, tagged with severity 8, contributed to the occurrence of an anomalous subsequence of severity 8 at meter-level that started at 19:15 in context 5.

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 detected, diagnosed and rectified by facility managers by reviewing the system ON-OFF logics. On the other hand, the correction of energy management procedures pertaining the "Not labelled" load can be more challenging since there is no detail about the energy systems served by this sub-load of the substation C.

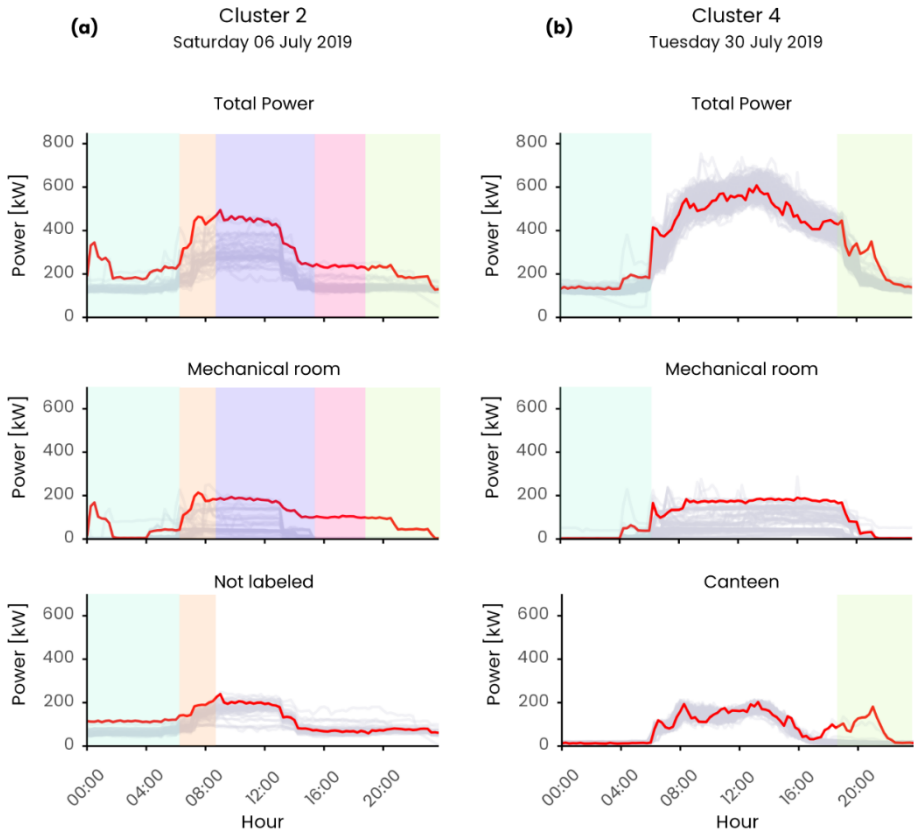

Figure 15. Comparison between anomalous meter-level daily electrical load profile and responsible sub-load identified though the diagnosis process. The figure shows the load profiles of (a) Saturday 6th of July 2019 belonging to cluster 2 and the load profiles of (b) Tuesday 30th of July 2019 belonging to cluster 4. The vertical bands denote the anomalous subsequences at meter and sub-load level, the band colours are referred to the contexts considered, while the band widths are related to the subsequence lengths.

4. Discussion and conclusion

In this paper a meter-level ADD process is proposed in order to demonstrate (i) the capabilities of the CMP algorithm as unsupervised method for detecting anomalies in energy consumption timeseries and (ii) the potential of integrating such tool in a building energy management system to gain insights on infrequent energy consumption patterns of sub-loads, promoting the correction of anomalies and reduction of energy wastes. The methodology is based on the coupling of unsupervised, supervised and timeseries analytics methods with domain knowledge to detect infrequent/anomalous patterns in energy consumption timeseries of building collected at meter-level.

The capability to identify anomalies that start in specific sub-daily periods (i.e., context) represents an opportunity for the early-stage identification and prompt correction of incorrect operations that can help prevent energy wastes over time. The recognized 55 anomalous subsequences start at different times in the identified contexts demonstrating the importance that the relaxation of temporal constraints (i.e., time alignment in the search of frequent/infrequent patterns) has in the detection of frequent and infrequent patterns in energy consumption timeseries of buildings.

As a matter of fact, in Figure 16 is reported the percentage of best matching subsequences that begin in each of the 4 possible start-positions of the considered 5 contexts with a duration of 1h, where for instance position 3 is located three 15-min timesteps apart from the beginning of the context. In particular, it is possible to observe that the context length was fully exploited in the *similarity-join-search*, allowing to identify pairs of similar subsequences that have approximately the same overall shape and magnitude even though they are slightly shifted between each other in the time axis.

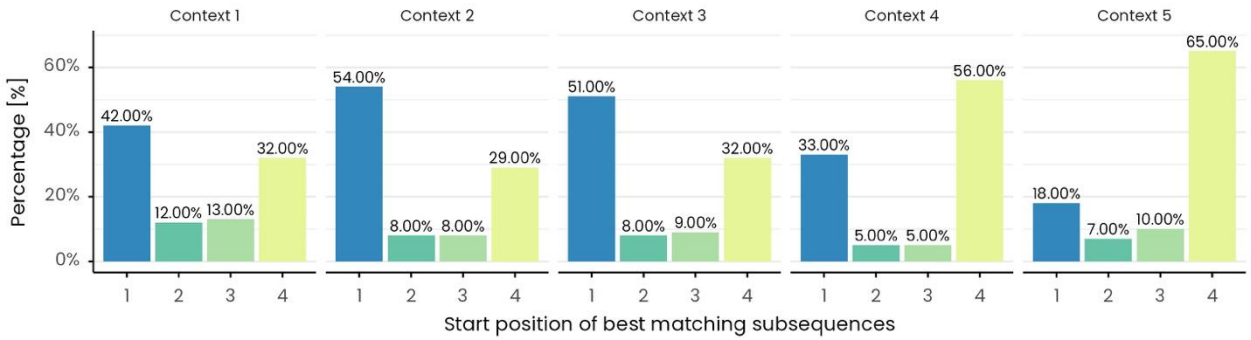

Figure 16. Relative frequency associated to the start positions of best matching subsequences identified through CMP for each context. The start positions corresponding to the four 15-min timesteps in a context are numbered from 1 to 4.

Correctly defining the number and the length of contexts and subsequences for tuning the *similarity-join-search* is a complex task and their wrong setting may negatively affect the capability to isolate important features of the daily load profile such as the periods related to the start-up or shut down of building energy systems. The regression tree model (developed using the CART algorithm) has proven to be flexible enough to adapt to different load conditions (also with reference to different building end-uses as reported in [15]) thanks to the easily generalizable criteria in the tuning of model parameters. The introduction of the concept of context, i.e., the time interval during which a subsequence of interest may start, allows to compare subsequences that are not perfectly aligned in the time domain and at the same time to avoid the computation of similarity-joins that are meaningless considering the operation schedule of the analysed building.

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 reference time interval of subsequences that were extracted from the regression tree (reported in Table 1). However, the practical implementation requires to tackle different technical challenges.

Firstly, the ADD process 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 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 reference time interval; where for batch process is intended a process in which the whole timeseries must be available before calculating the CMP [25]. In this perspective, the process execution is triggered only when the last subsequence, that may start in a context, has been fully collected (i.e., at the end of the reference time intervals). In the analysed case study, the minimum interval of time between subsequent triggers of the ADD process is 2.5 hours (i.e., the length of the smallest subsequence), 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, tested on 32 GB RAM Intel(R) Core(TM) i7-4790, is of 1 min and 45 seconds for each load which is far less than 2.5 hours proving that a semi-real-time batch approach is suitable for this kind of analysis and further scalable to a higher number of sub-loads. However, a streaming approach is more desirable and would enable an even earlier recognition of anomalies. The computational burden could be reduced performing the CMP calculation by using an incremental approach like the STAMPI algorithm [25] that adjusts the CMP rather than re-compute it once a new observation is collected.

The second challenge, that exists in the deployment phase of the ADD process is related the so-called cold-start problem, intended as scarcity of initial data useful to perform an accurate definition of subsequences, contexts, groups for the CMP calculation. One advantage of the MP algorithm is that, differently from CMP, it does not require a minimum length of the timeseries to be calculated. On the other hand, since the CMP objective is to perform a *similarity-join-search* to discover frequent and infrequent patterns within specific segments of timeseries that may be characterized by a certain periodicity, with a small dataset CMP may fail to recognize and highlight relevant patterns. A dataset containing at least 4 weeks of observations is desirable in order to have a minimum set of daily load profiles on which it is possible to perform a more robust discord/motif discovery. With respect to the hyperparameters of CMP (subsequences, contexts and groups), in the first deployment period, they can be easily defined based on domain knowledge, setting a reasonable context length and by grouping daily load profiles according to typical building operational schedules. As a first configuration it is possible to define the subsequences by simply split the 24h into N not overlapping time windows of fixed length. As a reference, this approach was followed by [13] in the definition of SAX parameters for the characterization of daily load profiles, where after a sensitivity analysis the suggested setting for N was between 3 and 4. Instead, the initial number of groups (i.e., clusters) can be based on the weekly operational calendar. For example, a possible hyperparameter setting may consist in 3 groups (weekday, Saturday and Sundays/holidays), 4 fixed-length subsequences of 6 hours with the corresponding context of 1h length.

To better discuss this scenario, the authors performed the same experiment on the entire one-year dataset, using this hyperparameter setting. The results obtained, although not comparable with the ADD results presented in the paper since no ground truth is available, shows that 40 anomalous sub-sequences were detected at meter-level (almost in the same periods of the presented results). However, it was possible to noticed that, despite the a-priori grouping and a-priori subsequence definition is a good compromise, it was not effective as the proposed process, since the context-group CMPs exhibited very high internal variability of distance values, meaning that the subsequences of daily load profiles within the same group were not so homogeneous. In this way, a wider distribution of the distances made the identification of isolated and extremal observations, by means of statistical methods, more difficult to be performed.

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 [25]. This implies that if a rare subsequence occurs more than once in the timeseries it may be considered as common or even frequent pattern [73] 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 subsequence set: this issue is recognized in the literature as the twin freak problem [73].

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 [73] 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 load conditions and their variability throughout the day due to multiple forcing factors such as the building occupancy, user behaviour, system failures and external boundary conditions it is necessary to assist energy managers during building operation by means of easily interpretable and effective informative tools. To this aim, the detection and diagnosis output of the proposed ADD process has been represented with an easy interpretable anomaly severity score, coupled with an effective visualization that allows to immediately compare the detected anomalous behaviour with the reference one for the same context and group (i.e., cluster centroid profile). Thanks to the proposed hierarchical methodological framework, the user is not overwhelmed by the stream of information and data but instead is supported in the ADD process and involved only in the case the sub-loads determine the occurrence of an anomalous energy trend at meter-level de facto rationalizing the quality and number of feedbacks he/she receive during the day-by-day energy management operations.

References

[1] Directorate-General for Energy (European Commission), Clean energy for all Europeans - Publications Office of the EU, Publ. Off. EU. 14 (2019) 3. https://doi.org/10.2833/9937.

[2] T. Ramesh, R. Prakash, K.K. Shukla, Life cycle energy analysis of buildings: An overview, Energy Build. 42 (2010) 1592–1600. https://doi.org/10.1016/j.enbuild.2010.05.007.

[3] L. Erhan, M. Ndubuaku, M. Di Mauro, W. Song, M. Chen, G. Fortino, O. Bagdasar, A. Liotta, Smart anomaly detection in sensor systems: A multi-perspective review, Inf. Fusion. 67 (2021) 64–79. https://doi.org/10.1016/j.inffus.2020.10.001.

[4] S. Rinaldi, P. Bellagente, A.L. Camillo Ciribini, L. Chiara Tagliabue, T. Poli, A. Giovanni Mainini, A. Speroni, J.D. Blanco Cadena, S. Lupica Spagnolo, A cognitive-driven building renovation for improving energy effciency: The experience of the elisir project, Electron. 9 (2020). https://doi.org/10.3390/electronics9040666.

[5] M. Molina-Solana, M. Ros, M.D. Ruiz, J. Gómez-Romero, M.J. Martin-Bautista, Data science for building energy management: A review, Renew. Sustain. Energy Rev. 70 (2017) 598–609. https://doi.org/10.1016/j.rser.2016.11.132.

[6] H. Kramer, G. Lin, J. Granderson, C. Curtin, E. Crowe, Synthesis of Year One Outcomes in the Smart Energy Analytics Campaign Building Technology and Urban Systems Division, Lawrence Berkeley National Laboratory: Berkeley, CA, USA. (2019). https://doi.org/10.2172/1545159

[7] C. Fan, Y. Sun, Y. Zhao, M. Song, J. Wang, Deep learning-based feature engineering methods for improved building energy prediction, Appl. Energy. 240 (2019) 35–45. https://doi.org/10.1016/j.apenergy.2019.02.052.

[8] C. Fan, F. Xiao, S. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Appl. Energy. 127 (2014) 1–10. https://doi.org/10.1016/j.apenergy.2014.04.016.

[9] N. Somu, G.R. M R, K. Ramamritham, A hybrid model for building energy consumption forecasting using long short term memory networks, Appl. Energy. 261 (2020) 114131. https://doi.org/10.1016/j.apenergy.2019.114131.

[10] T. Liu, Z. Tan, C. Xu, H. Chen, Z. Li, Study on deep reinforcement learning techniques for building energy consumption forecasting, Energy Build. 208 (2020). https://doi.org/10.1016/j.enbuild.2019.109675.

[11] M.S. Piscitelli, S. Brandi, A. Capozzoli, F. Xiao, A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings, Build. Simul. (2020) 1–17. https://doi.org/10.1007/s12273-020-0650-1.

[12] R. Chiosa, A. Capozzoli, M.S. Piscitelli, A Data Analytics-Based Energy Information System (EIS) Tool to Perform Meter-Level Anomaly Detection and Diagnosis in Buildings, Energies. 14 (2021) 1–28. https://doi.org/10.3390/en14010237.

[13] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Autom. Constr. 49 (2015) 1–17. https://doi.org/10.1016/j.autcon.2014.09.004.

[14] C. Fan, F. Xiao, H. Madsen, D. Wang, Temporal knowledge discovery in big BAS data for building energy management, Energy Build. 109 (2015) 75–89. https://doi.org/10.1016/j.enbuild.2015.09.060.

[15] A. Capozzoli, M.S. Piscitelli, S. Brandi, D. Grassi, G. Chicco, Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings, Energy. 157 (2018) 336–352. https://doi.org/10.1016/j.energy.2018.05.127.

[16] P. Arjunan, K. Poolla, C. Miller, BEEM: Data-driven building energy benchmarking for Singapore, Energy Build. 260 (2022) 111869. https://doi.org/10.1016/j.enbuild.2022.111869.

[17] K. Li, Y. Sun, D. Robinson, J. Ma, Z. Ma, A new strategy to benchmark and evaluate building electricity usage using multiple data mining

technologies, Sustain. Energy Technol. Assessments. 40 (2020) 100770. https://doi.org/10.1016/j.seta.2020.100770.
[18] A. Capozzoli, M.S. Piscitelli, F. Neri, D. Grassi, G. Serale, A novel methodology for energy performance benchmarking of buildings by means of Linear Mixed Effect Model: The case of space and DHW heating of out-patient Healthcare Centres, Appl. Energy. 171 (2016) 592–607. https://doi.org/10.1016/j.apenergy.2016.03.083.
[19] X. Luo, T. Hong, Y. Chen, M.A. Piette, Electric load shape benchmarking for small- and medium-sized commercial buildings, Appl. Energy. 204 (2017) 715–725. https://doi.org/10.1016/j.apenergy.2017.07.108.
[20] M.S. Piscitelli, S. Brandi, A. Capozzoli, Recognition and classification of typical load profiles in buildings with non-intrusive learning approach, Appl. Energy. 255 (2019) 113727. https://doi.org/10.1016/j.apenergy.2019.113727.
[21] J.Y. Park, E. Wilson, A. Parker, Z. Nagy, The good, the bad, and the ugly: Data-driven load profile discord identification in a large building portfolio, Energy Build. 215 (2020) 109892. https://doi.org/10.1016/j.enbuild.2020.109892.
[22] J.Y. Park, X. Yang, C. Miller, P. Arjunan, Z. Nagy, Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset, Appl. Energy. 236 (2019) 1280–1295. https://doi.org/10.1016/j.apenergy.2018.12.025.
[23] K. Nweye, Z. Nagy, MARTINI: Smart Meter Driven Estimation of HVAC Schedules and Energy Savings Based on WiFi Sensing and Clustering, (2021) 0–13. http://arxiv.org/abs/2110.08927.
[24] A. Capozzoli, M.S. Piscitelli, A. Gorrino, I. Ballarini, V. Corrado, Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings, Sustain. Cities Soc. 35 (2017) 191–208. https://doi.org/10.1016/j.scs.2017.07.016.
[25] C.-C.M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H.A. Dau, D.F. Silva, A. Mueen, E. Keogh, Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets, (2017) 1317–1322. https://doi.org/10.1109/icdm.2016.0179.
[26] D. De Paepe, S. Vanden Hautte, B. Steenwinckel, F. De Turck, F. Ongenae, O. Janssens, S. Van Hoecke, A generalized matrix profile framework with support for contextual series analysis, Eng. Appl. Artif. Intell. 90 (2020). https://doi.org/10.1016/j.engappai.2020.103487.
[27] P. Tan, M. Steinbach, V. Kumar, Introduction to data mining, Addison Wesley, 2011. https://doi.org/10.1007/978-3-642-19721-5_1.
[28] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, A. Amira, Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives, Appl. Energy. 287 (2021) 116601. https://doi.org/10.1016/j.apenergy.2021.116601.
[29] N.R. Prasad, S. Almanza-Garcia, T.T. Lu, Anomaly detection: A Survey, Comput. Mater. Contin. 14 (2009) 1–22. https://doi.org/10.1145/1541880.1541882.
[30] J.E. Zhang, D. Wu, B. Boulet, Time Series Anomaly Detection for Smart Grids: A Survey, 2021 IEEE Electr. Power Energy Conf. EPEC 2021. (2021) 125–130. https://doi.org/10.1109/EPEC52095.2021.9621752.
[31] H. Chen, X. Fei, S. Wang, X. Lu, G. Jin, W. Li, X. Wu, Energy Consumption Data Based Machine Anomaly Detection, Proc. - 2014 2nd Int. Conf. Adv. Cloud Big Data, CBD 2014. (2015) 136–142. https://doi.org/10.1109/CBD.2014.24.
[32] Y. Zhao, S. Wang, F. Xiao, Pattern recognition-based chillers fault detection method using Support Vector Data Description (SVDD), Appl. Energy. 112 (2013) 1041–1048. https://doi.org/10.1016/j.apenergy.2012.12.043.
[33] K. Kammerer, B. Hoppenstedt, R. Pryss, S. Stökler, J. Allgaier, M. Reichert, Anomaly detections for manufacturing systems based on sensor data—insights into two challenging real-world production settings, Sensors. 19 (2019). https://doi.org/10.3390/s19245370.
[34] J.S. Chou, A.S. Telaga, Real-time detection of anomalous power consumption, Renew. Sustain. Energy Rev. 33 (2014) 400–411. https://doi.org/10.1016/j.rser.2014.01.088.
[35] R. Wu, E. Keogh, Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress, IEEE Trans. Knowl. Data Eng. (2021) 1–9. https://doi.org/10.1109/TKDE.2021.3112126.
[36] C. Fan, F. Xiao, Y. Zhao, J. Wang, Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data, Appl. Energy. 211 (2018) 1123–1135. https://doi.org/10.1016/j.apenergy.2017.12.005.
[37] D.J. Bartholomew, Principal components analysis, Int. Encycl. Educ. (2010) 374–377. https://doi.org/10.1016/B978-0-08-044894-7.01358-0.
[38] P. Esling, C. Agon, Time-series data mining, ACM Comput. Surv. 45 (2012). https://doi.org/10.1145/2379776.2379788.
[39] A. Capozzoli, M.S. Piscitelli, S. Brandi, Mining typical load profiles in buildings to support energy management in the smart city context, Energy Procedia. 134 (2017) 865–874. https://doi.org/10.1016/j.egypro.2017.09.545.
[40] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, J. Li, A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis, Energy Built Environ. 1 (2020) 149–164. https://doi.org/10.1016/j.enbenv.2019.11.003.
[41] V. Chandola, D. Cheboli, V. Kumar, Detecting anomalies in a time series database, (2009). https://hdl.handle.net/11299/215791
[42] C. Zhang, 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. (2019) 54–59. https://doi.org/10.1109/I-SPAN.2018.00018.
[43] K. Li, Z. Ma, D. Robinson, J. Ma, Using Evidence Accumulation-Based Clustering and Symbolic Transformation to Group Multiple Buildings Based on Electricity Usage Patterns, Sustain. Energy Build., (2019) 61–67. https://doi.org/10.1007/978-981-32-9868-2.
[44] B. Rossi, S. Chren, B. Buhnova, T. Pitner, Anomaly Detection in Smart Grid Data: An Experience Report, (2016) 2313–2318.
[45] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: a Novel Symbolic Representation of Time Series, Cs.Gmu.Edu. 15 (2007) 107–144.
[46] E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotra, Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases, Knowl. Inf. Syst. 3 (2001) 263–286. https://doi.org/10.1007/pl00011669.
[47] H. Ren, M. Liu, Z. Li, W. Pedrycz, A Piecewise Aggregate pattern representation approach for anomaly detection in time series,

Knowledge-Based Syst. 135 (2017) 29–39. https://doi.org/10.1016/j.knosys.2017.07.021.

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

[49] L. Ye, E. Keogh, Time series shapelets: A new primitive for data mining, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (2009) 947–955. https://doi.org/10.1145/1557019.1557122.

[50] S.D.D. Anton, 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. (2020) 347–353. https://doi.org/10.1109/WoWMoM49955.2020.00065.

[51] F. Madrid, S. Imani, R. Mercer, Z. Zimmerman, N. Shakibay, 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. (2019) 175–182. https://doi.org/10.1109/ICBK.2019.00031.

[52] J. Dinal Herath, C. Bai, G. Yan, P. Yang, S. Lu, RAMP: Real-Time Anomaly Detection in Scientific Workflows, Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019. (2019) 1367–1374. https://doi.org/10.1109/BigData47090.2019.9005653.

[53] D. De Paepe, D.N. Avendano, S. Van Hoecke, Implications of Z-Normalization in the Matrix Profile, Lect. Notes Comput. Sci. (2020) 95–118. https://doi.org/10.1007/978-3-030-40014-9_5.

[54] D. De Paepe, O. Janssens, S. Van Hoecke, Eliminating noise in the matrix profile, ICPRAM 2019 - Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods. (2019) 83–93. https://doi.org/10.5220/0007314100830093.

[55] S. Ahmad, A. Lavin, S. Purdy, Z. Agha, Unsupervised real-time anomaly detection for streaming data, Neurocomputing. 262 (2017) 134–147. https://doi.org/10.1016/j.neucom.2017.04.070.

[56] C. Nichiforov, G. Stamatescu, I. Stamatescu, I. Fagarasan, Learning Dominant Usage from Anomaly Patterns in Building Energy Traces, IEEE Int. Conf. Autom. Sci. Eng. (2020) 548–553. https://doi.org/10.1109/CASE48305.2020.9216794.

[57] Y. Zhu, S. Gharghabi, D.F. Silva, H.A. Dau, C.C.M. Yeh, N. Shakibay Senobari, A. Almaslukh, K. Kamgar, Z. Zimmerman, G. Funning, A. Mueen, E. Keogh, The Swiss army knife of time series data mining: ten useful things you can do with the matrix profile and ten lines of code, Springer US, 2020. https://doi.org/10.1007/s10618-019-00668-6.

[58] S. Alaee, K. Kamgar, E. Keogh, Matrix profile XXII: Exact discovery of time series motifs under DTW, Proc. - IEEE Int. Conf. Data Mining, ICDM. 2020-Novem (2020) 900–905. https://doi.org/10.1109/ICDM50108.2020.00099.

[59] S. Gharghabi, S. Imani, A. Bagnall, A. Darvishzadeh, E. Keogh, An ultra-fast time series distance measure to allow data mining in more complex real-world deployments, Springer US, 2020. https://doi.org/10.1007/s10618-020-00695-8.

[60] A. Mueen, Y. Zhu, M. Yeh, K. Kamgar, K. Viswanathan, C. Gupta, E. Keogh, The Fastest Similarity Search Algorithm for Time Series Subsequences under Euclidean Distance, (2017). http://www.cs.unm.edu/~mueen/FastestSimilaritySearch.html

[61] C. Onwongsa, C. Ratanamahatana, An enhanced time series motif discovery using approximated matrix profile, ACM Int. Conf. Proceeding Ser. (2020) 180–189. https://doi.org/10.1145/3421558.3421586.

[62] C.C.M. Yeh, N. Kavantzas, E. Keogh, Matrix profile VI: Meaningful multidimensional motif discovery, Proc. - IEEE Int. Conf. Data Mining, ICDM. 2017-Novem (2017) 565–574. https://doi.org/10.1109/ICDM.2017.66.

[63] H.A. Dau, E. Keogh, Matrix profile V: A generic technique to incorporate domain knowledge into motif discovery, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. Part F1296 (2017) 125–134. https://doi.org/10.1145/3097983.3097993.

[64] G.E.A.P.A. Batista, E.J. Keogh, O.M. Tataw, V.M.A. De Souza, CID: An efficient complexity-invariant distance for time series, Data Min. Knowl. Discov. 28 (2014) 634–669. https://doi.org/10.1007/s10618-013-0312-3.

[65] J.L. Mathieu, P.N. Price, S. Kiliccote, M.A. Piette, Quantifying changes in building electricity use, with application to demand response, IEEE Trans. Smart Grid. 2 (2011) 507–518. https://doi.org/10.1109/TSG.2011.2145010.

[66] J. Zhu, Y. Shen, Z. Song, D. Zhou, Z. Zhang, A. Kusiak, Data-driven building load profiling and energy management, Sustain. Cities Soc. 49 (2019) 101587. https://doi.org/10.1016/j.scs.2019.101587.

[67] R. Yan, Z. Ma, Y. Zhao, G. Kokogiannakis, A decision tree based data-driven diagnostic strategy for air handling units, Energy Build. 133 (2016) 37–45. https://doi.org/10.1016/j.enbuild.2016.09.039.

[68] T. Grubinger, A. Zeileis, K.P. Pfeiffer, Evtree: Evolutionary learning of globally optimal classification and regression trees in R, J. Stat. Softw. 61 (2014) 1–29. https://doi.org/10.18637/jss.v061.i01.

[69] V. Satopää, J. Albrecht, D. Irwin, B. Raghavan, Finding a “kneedle” in a haystack: Detecting knee points in system behavior, Proc. - Int. Conf. Distrib. Comput. Syst. (2011) 166–171. https://doi.org/10.1109/ICDCSW.2011.20.

[70] R Core Team, R: A Language and Environment for Statistical Computing, (2017). https://www.r-project.org/.

[71] G. Van Rossum, F.L. Drake Jr, Python reference manual, Centrum voor Wiskunde en Informatica Amsterdam, (1995).

[72] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, NbClust : An R Package for Determining the, J. Stat. Softw. 61 (2014) 1–36. https://doi.org/10.18637/jss.v061.i06.

[73] Y. He, X. Chu, Y. Wang, Neighbor profile: Bagging nearest neighbors for unsupervised time series mining, Proc. - Int. Conf. Data Eng. (2020) 373–384. https://doi.org/10.1109/ICDE48307.2020.00039.