securityandtrust.lu

Towards a Plug-and-Play and Holistic Data Mining Framework for Understanding and Facilitating Operations in Smart Buildings

Daoyuan Li, Tegawendé F. Bissyandé, Jacques Klein, Yves Le Traon

Paul Schummer, Ben Muller, Anne-Marie Solvi

Interdisciplinary Centre for Security, Reliability and Trust

University of Luxembourg

TR-SNT-2017-5

ISBN: 978-99959-58-01-5

October 18, 2017

Version 1.0

UNIVERSITÉ DU LUXEMBOURG

Towards a Plug-and-Play and Holistic Data Mining Framework for Understanding and Facilitating Operations in Smart Buildings

Daoyuan Li, Tegawendé F. Bissyandé, Jacques Klein, Yves Le Traon,
Paul Schummer, Ben Muller, Anne-Marie Solvi

***Abstract*—Nowadays, a significant portion of the total energy consumption is attributed to the buildings sector. In order to save energy and protect the environment, energy consumption in buildings must be more efficient. At the same time, buildings should offer the same (if not more) comfort to their occupants. Consequently, modern buildings have been equipped with various sensors and actuators and interconnected control systems to meet occupants' requirements. Unfortunately, so far, Building Automation Systems data have not been well-exploited due to technical and cost limitations. Yet, it can be exceptionally beneficial to take full advantage of the data flowing inside buildings in order to diagnose issues, explore solutions and improve occupant-building interactions. This paper presents a plug-and-play and holistic data mining framework named PHoliData for smart buildings to collect, store, visualize and mine useful information and domain knowledge from data in smart buildings. PHoliData allows non technical experts to easily explore and understand their buildings with minimum IT support. An architecture of this framework has been introduced and a prototype has been implemented and tested against real-world settings. Discussions with industry experts have suggested the system to be extremely helpful for understanding buildings, since it can provide hints about energy efficiency improvements. Finally, extensive experiments have demonstrated the feasibility of such a framework in practice and its advantage and potential for buildings operators.**

***Index Terms*—Data mining for smart buildings, time series mining, outlier detection**

I. Introduction

It has been well widely acknowledged that buildings add up to approximately 40% [1] – a surprisingly large portion – of total energy consumption by all sectors, due to the fact that citizens in a modern society spend a vast majority of their daily life inside buildings either working or relaxing. For example, the International Energy Agency[1] reports that residential and commercial buildings account for 32% of energy consumption [1]; and 41% of energy consumption is attributed to buildings in the US, while buildings consume more energy than the industry sector in the EU [2]. Many initiatives have thus been proposed to address the energy consumption issue for the building sector. These initiatives include proposing more affordable and greener energy sources such as solar and wind energy, inventing better isolation materials, devising more energy-efficient building automation systems (BAS) and so on. BAS form a promising research direction since they try to fulfill occupants' comfort requirements while reducing energy footprints for building operations, including heating, ventilation, and air conditioning (HVAC), lighting and plug loads. Indeed, BAS can help improving buildings' energy performances [3]: it can help reducing up to 58% of lighting energy consumption in simulated experiments [4]; in other real-world evaluations, BAS can save 25% of energy for lighting [5] and 13% of electricity consumed by lights and air-conditioning system in offices [6].

A BAS often involves taking advantage of heterogeneous sensors, such as passive infrared sensors, cameras, motion and presence detectors, and environmental sensors like temperature, humidity, CO_2, etc., to monitor the status of the building and occupant activities. As the concept of Internet of Things (IoT) develops and increasingly more sensors/actuators are integrated in buildings, a tremendous amount of data are being generated by components installed in buildings. These advances have brought up many novel application scenarios, including predicting energy demand based on environmental conditions and occupant activities [7], [8], [9], monitoring building operations and detecting equipment failures [10], [11], [12], identifying energy usage patterns to fight against energy theft and fraud (also known as Non-Technical Losses, or NTL) [13], and so on.

Undoubtedly, mining BAS data can be extremely beneficial to improve energy efficiency of buildings. In recent literature, Xiao and Chen [11] have applied data mining techniques to the BAS database of the tallest building in Hong Kong and proven data mining to be helpful for knowledge discovery in BAS and for improving the operational performance of buildings. However, despite the advantages of data mining for buildings – including improving the energy efficiency, user comfort and operational reliability, data mining in buildings have not yet become mainstream due to several technical limitations of current approaches [1]. First of all, it is usually difficult to collect real-time data from buildings since different infrastructures come with heterogeneous technical components, making it a challenging task to set up a generic data collection system that can be reused across different buildings. Secondly, there exists a plethora of data mining algorithms, the variety and complexity of which make it difficult for normal users to

Daoyuan Li, Tegawendé F. Bissyandé, Jacques Klein and Yves Le Traon are with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, 29, Avenue J.F Kennedy, L-1855 Luxembourg (e-mail: firstname.lastname@uni.lu)

Paul Schummer, Ben Muller and Anne-Marie Solvi are with Paul Wurth Geprolux S.A., 32, rue d'Alsace, L-1122 Luxembourg (e-mail: firstname.lastname@paulwurth.com)

[1]https://www.iea.org/

use such systems. For instance, different algorithms require different data formats and need to be configured with varying algorithmic parameters, which makes it difficult for non-experts to use such systems. Finally, data mining approaches often produce results that are not straightforward – only data scientists may be able to interpret such results.

Besides technical challenges, BAS mining solutions also suffer from cost limitations. Many researchers propose new BAS control mechanisms and infrastructures to improve energy efficiency and user comfort in buildings. For instance, Shaikh et al. [14] have surveyed 121 relevant works on smart BAS and predicted that *"data management in building control systems is turning out to be a gigantic challenge for the near future"*. Furthermore, these approaches are generally disruptive and not cost-effective enough, since BAS are normally instigated with mechanical, electrical and plumbing (MEP) system controls. Unlike consumer electronics, components in buildings have much longer life spans and changes to the infrastructures are usually expensive and difficult (if feasible).

To tackle both technical and cost challenges, this work seeks to devise a cost-effective approach that takes advantage of existing BAS and investigates the feasibility of gaining extra intelligence about corresponding buildings and their occupants from such systems. This study proposes a generic **p**lug-and-play and **holi**stic **data** mining framework (PHoliData) for collecting and storing data from BAS systems, providing visualizations for exploratory data mining tasks and a data mining architecture that is easy to implement and integrate in a plug-and-play manner. PHoliData's data-centric architecture allows data to be stored in a document-based database that does not require predefined schemas, which makes it easy for maintenance and adaption to new application scenarios. PHoliData will thus offer non technical experts an easy access to exploring and understanding their buildings without much technical burden. The architecture of this framework has been designed and a prototype has been implemented and tested against real-world settings. Discussions with industry experts have proven the system to be extremely helpful for understanding buildings and providing hints about energy efficiency improvements and extensive experiments has demonstrated the feasibility of such a framework and its advantage and potential for buildings operators.

The remainder of this paper is organized as follows. Section II presents the methodology for modeling data in BAS and provides readers with the necessary technical background for relevant BAS data mining tasks. Section III introduces a prototype implementation of the PHoliData framework and offers straightforward data mining examples with real-world data. To evaluate the plug-and-play capability of the proposed framework, Section IV provides some evidences of how easy it is to adapt the prototype to a totally different setting. Finally, this paper is concluded in Section V with directions for future research and development.

II. Methodology

Data mining systems generally consist of two parts: (1) a data collector that harvests raw data and (2) a mining engine that extracts meaningful information from the collected data. A majority of research papers in this domain focus on applying well known algorithms to data generated from buildings, without mentioning the data sources, data formats and representations. However, the data mining community emphasizes that, in real-world cases, preparing and preprocessing data can be more time-consuming than the data mining process itself [15]. As a result, this paper seeks to provide a holistic view of how data mining would work within realistic settings. Fig. 1 presents this holistic approach. As illustrated, this work treats BAS as a blackbox and only interacts with BAS in the data collection phase. When data flows into the holistic data mining part, they are stored in a database to provide persistent access for subsequent phases. Data can be preprocessed to reduce noises and be stored back to the database when necessary. The visualization module (e.g., a dashboard) may access data either directly from the database for exploratory purposes or from the data mining component for more in-depth views. Data mining algorithms also access data from the database and can be monitored and controlled by administrators. The rest of this section describes in detail how to collect data from BAS and store these data, how to represent data and preprocess them, and finally how to conduct data mining tasks – extracting meaningful and interpretable domain knowledge.

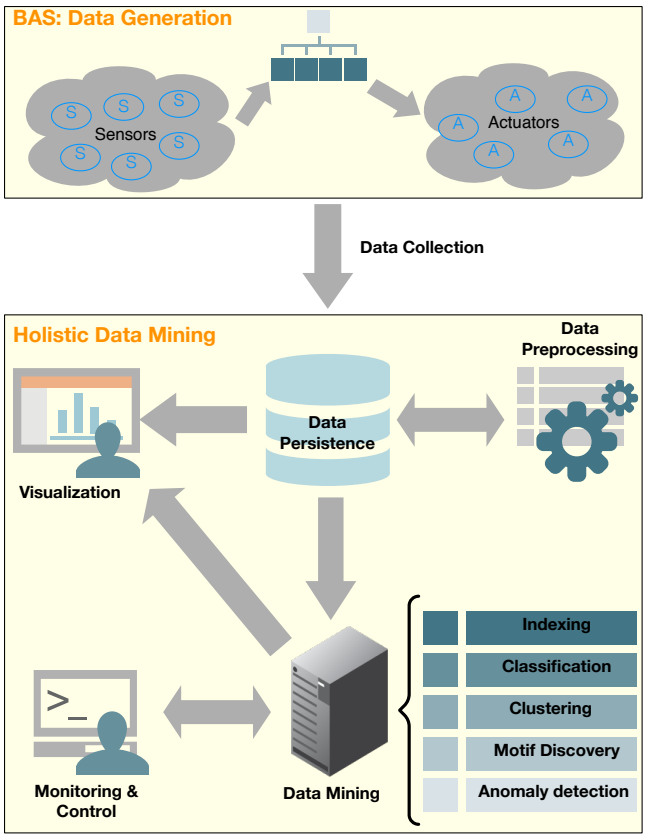

Fig. 1. Overview of the proposed approach. Arrows indicate data flows.

A. Data collection

It is well known that buildings have a wide range of control mechanisms and internal communication protocols, leading to interoperability issues that make it difficult to devise a single system that is able to collect data from any building. There have been mainly two ways to combat interoperability issues: devising intermediate gateways/hubs or promoting universal standards [16]. Both approaches are currently helping the building industry to reduce segmentation and improve interoperability. Furthermore, the majority of BAS are based on open communication standards such as KNX[2], BACnet (ISO 16484-5) and LonTalk (ISO/IEC 14908-1).

As presented in Fig. 1, this study does not attempt to interconnect different buildings. Instead, the main focus of the PHoliData framework is to harvest and mine data from buildings with heterogeneous communication protocols. As a result, a case-by-case strategy to treat interoperability issues can be more pragmatic and cost-effective than generic approaches in two aspects. Firstly, there have been great success in promoting universal standards in the building industry and support of prevalent standards enables collecting data from a vast majority of buildings. For instance, KNX as a worldwide standard for home and building control "is the most popular protocol in European markets sharing more than 70% of the total market value" in 2012 [17]. In this case, a KNX adaptor that bridges communication between KNX and the database would allow collecting data from many buildings. It is also easy to get a compatible watchdog module and simply attach it to the control bus of other standards and start collecting data. Secondly, it is theoretically impossible to incorporate all communication systems, especially when new communication mechanisms are under active development. For example, recent years have seen many wireless protocols (such as ZigBee and Z-Wave) proposed for smart buildings. Building a generic framework to only solve the interoperability issue can be both impractical and costly.

To sum up, one of the key challenges of allowing plug-and-play data mining framework is how to effectively collect data from BAS. This work prefers a more pragmatic approach since it can simplify system design, modularize components, lower bootstrapping and maintenance cost and offer more flexible solutions to data collection from BAS. As shall be discussed in subsequent sections, the proposed approach is able to collect data from two building with completely different data types and communication standards.

B. Data representation

Measurements and events recorded in buildings and BAS are almost always strictly related to time. For instance, it may be meaningless to tell the temperature of a room without indicating the timestamp. In order to persist the information related to time, data from buildings can be best represented by time series. In the data mining community, one time series is modeled as a sequence of data points that are ordered by their corresponding timestamps, that is, $S = \{P_1, P_2, ..., P_n\}$ where each data point P_i consists of a corresponding timestamp and a value, denoted correspondingly as t_i and v_i, i.e., $P_i = (t_i, v_i)$. Time series mining is a well-researched area since there are an abundant of time series data available in various sectors including IoT, medical and health care, and financial domains [18]. More generally, when comparing a set of time series with the same timestamp range and sampling interval, these timestamps can be omitted for the ease of processing and similarity comparison [19]. In this case, a time series S is generalized as a sequence of real-valued numbers, i.e., $S = \{v_1, v_2, ..., v_n\}$.

Data mining tasks can be divided into different categories, including classification, clustering, anomaly detection and trend prediction. However, a majority of the data mining tasks involve solving the most fundamental problem: how to describe the similarity or dissimilarity between different samples or instances. This is especially true for time series mining. When evaluating the similarity or dissimilarity of two time series instances S_x and S_y with $|S_x| = |S_y| = n$, a distance measure $D(S_x, S_y)$ can be defined. There are a plethora of distance measures and arguably there is not a single measure that suits for every scenario. Popular distance measures for time series include Euclidean distance $D_{Euclidean}(S_x, S_y) = \sqrt{\sum_{i=1}^{n}(S_{xi} - S_{yi})^2}$ that maps the i-th point in S_x to the i-th point in S_y in an one-to-one mapping scheme, and Dynamic Time Warping (DTW) distance, which tries to find the best way to warp the time axis and as a result aligns S_x and S_y differently. As shown in Fig. 2, an i-th point in S_x can be mapped to a j-th point (it is possible that $i \neq j$) in S_y, and one point in S_x can be mapped to zero or multiple points in S_y. For Euclidean distance, the gray dotted lines indicating the data point alignment are all vertical. Moreover, thanks to the time-axis warping feature of DTW, it can also be used to evaluate the similarity of two series with different lengths.

Previously, Fan et al. [20] have modeled BAS data as multivariate time series. They have taken advantage of Symbolic Aggregate approXimation (SAX) [21] to convert real-valued time series data into alphabet strings and conducted motif discovery and temporal association rule mining. Unlike [20], this study does not fully rely on SAX for data representation, due to the fact that SAX's alphabetic representation makes it difficult to apply numeric calculations in many applications other than motif discovery.

C. Data persistence

It is surprising that many BAS (especially those in medium-sized buildings) do not store historical records. What these BAS offer is at best the overall status of a building in realtime, while no operational data are stored. As one of the most important players in the building sector in Luxembourg, Paul Wurth Geprolux S.A. has not seen many BAS deployed in buildings that retains historical operation data. There can be different interpretations of the status quo. Besides interoperability issues discussed previously, the biggest obstacle may be the lack of IT expertise in both the BAS companies and the building operation teams. From the perspective of an IT expert, data storage can at most be a trivial problem.

[2]http://www.knx.org/

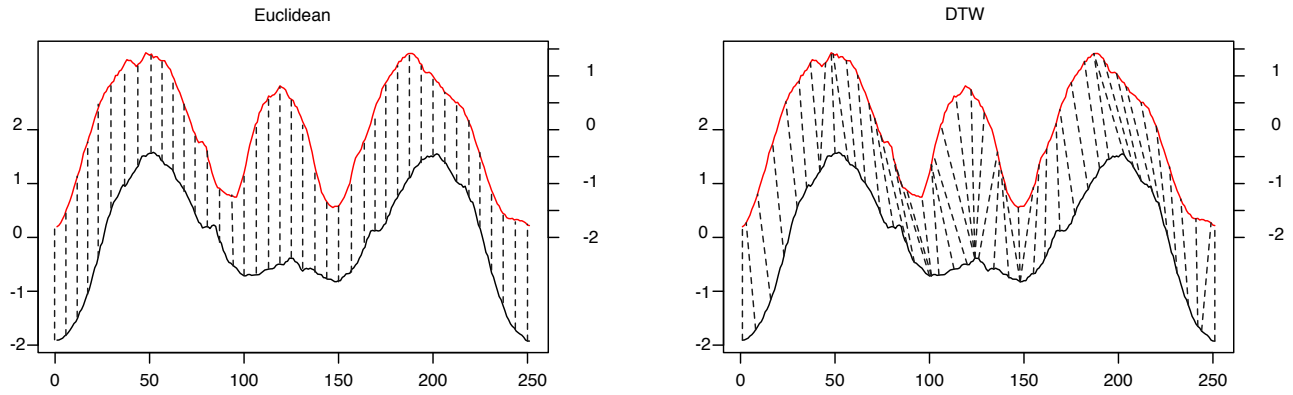

Fig. 2. Illustration of how Euclidean and DTW aligns two time series and calculates their distance.

Relational database management systems (RDBMS) have been widely used for decades and they are proven to be reliable in data persistence as well as easy to retrieve back data. On the other hand, there is one important issue with RDBMS: they often require thorough understanding the relationship between all different data sources and it is a prerequisite to have experts design database schemas before collecting any data in production. Additionally, modifications to database schemas in later phases (e.g., operational phase) are challenging even for expert database administrators. This fact can be a high barrier to users despite good commercial support of RDBMS. Fortunately, recent development of next generation database systems – such as MongoDB, Neo4j, Redis, Elasticsearch, etc. – offer much more flexible schemas, since most of them are based on key-value or document stores. Furthermore, next generation databases offer great reliability, easy maintenance, high security and scalability resulted from distributed computing mechanism and sharding techniques.

From Fig. 1, it is obvious that data persistence sits at the core of the holistic data mining framework from the data flow perspective. As a result, there are some prerequisites for this module. First of all, it must allow flexible database schemas and easy to write and update data from heterogeneous devices. Secondly, it should be able to handle a huge amount of data and allow easy as well as efficient data aggregation. Finally, it should provide flexible access for time series data. After careful consideration, Elasticsearch [22] comes as the most viable option since it supports all of the criteria: 1) as a document store, it provides a RESTful API to conduct data CRUD (Create, Retrieve, Update and Delete) operations; 2) it is essentially a full-text search engine and its internal aggregation mechanism have proven to be extremely efficient; and 3) data values with timestamps can be indexed and thus time series are first-class elements in Elasticsearch.

D. Data preprocessing

Thanks to its intrinsic timestamps, time series can be easily resampled, interpolated and extrapolated [23]. Note that in practice BAS data generally exhibits different statistical characteristics. For instance, different temperature sensors report measurements at different frequencies and the amplitude of values may as well differ. Besides, abnormal and missing values are very common, making the collected data quite noisy and difficult to process for data mining algorithms. To proceed, data cleansing tasks must be conducted as specified below:

1) Resampling. There are several benefits in resampling raw data. First of all, data becomes properly aligned after resampling, which allows most time series mining algorithms to omit timestamps and focus on the values. Besides, resampling can help filtering out noises such as peaky readings and making the data more smooth. Finally, down-sampling techniques can greatly reduce dataset size and improve computation efficiency. Specifically, after down-sampling data records with a uniform frequency (e.g. converting to hourly readings), it becomes easier to proceed with the following steps.
2) Interpolation. Some time series may contain missing values at certain timestamps even after down-sampling. Missing values are common in BAS data, due to failures in sensor hardware, electricity power or communication. There are many missing value imputation techniques [24], however, the most straightforward and helpful one is probably linearly interpolation, since data in BAS are generally redundant and sensor readings within indoor environments do not tend to change drastically.
3) Normalization. Readings from sensors in different locations and of different functionality usually contain values of significantly different amplitude levels. This can result in inaccurate computation especially with distance measures such as the DTW distance. In the time series mining community, it is common practice to Z-normalize data [25], which is done by dividing each value in time series by their corresponding standard deviation for normalization, i.e., $v_i' = v_i / \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(v_i - \mu)^2}$ for $0 \leq i < n$, where $\mu = \frac{1}{n}\sum_{i=0}^{n-1} v_i$. This step is especially important if the overall development of oscillation is more concerned than the absolute terms. However, normalization is not mandatory and it strictly depends on the target application.
4) Smoothing. Research [18] has shown that smoothing time series data can help preserving the most dis-

tinguishable features and improve the performance of time series mining. The most common smoothing technique for time series is moving average. Simple moving average can be defined mathematically as $v'_i = \frac{1}{m}\sum_{j=1}^{m} v_{m*i+j}$, where $2 \leq m < \sqrt{n}$ is the number of data points to average on. Besides moving average, wavelets such as Haar have also proven helpful for time series smoothing [18].

E. Data visualization

Visualization of data from BAS is a key source for understanding the over operational status of a building in a straightforward manner. Some BAS provide some rudimentary visualization means, e.g., heat maps to show realtime information (c.f. Fig. 3 left, however, these means are not sufficient since they only allow operators to compare readings horizontally but not vertically through time. Moreover, building operators may not be always available to monitor realtime situations. When BAS data are modeled as time series, it is then beneficial to take advantage of the temporal information and visualize the evolvement of readings over time. As shown in Fig. 3 (right), historical views of data enables operators not only to navigate back in time to monitor building status when administrators are absent but also to compare the overall development of different sensor readings. Visualization is extremely important to monitor, explore and understand component status within buildings. More examples of visualization techniques will be introduced in the subsequent sections. In fact, all illustrations of data mining results in this paper have been exported from the visualization module. As a result, this visualization module can be of great value for generating reports.

F. Time series mining

Thanks to the vast availability of time series data, there have been many techniques developed for time series mining. Generally, time series mining tasks can be divided into the following categories [26]: query by content (indexing), classification, clustering, motif discovery, anomaly detection, forecasting, segmentation and so on. In this section, the first five categories will be introduced since they are most relevant to BAS mining tasks. Furthermore, this study focuses more on unsupervised algorithms than supervised ones, since the latter requires far more human intervention with data preparation than the former.

1) Query by content: Given a time series database S and a query time series Q, the challenge is to find the most similar time series $X \in S$ that minimizes a given similarity/distance measure $D(Q, X)$ [27]. The most popular distance measure are Euclidean and DTW as illustrated in Fig. 2. Query by content is also known as indexing, which finds its applications mostly in revealing when and where some specific patterns have occurred previously. Indexing time series is one of the mot fundamental and direct usage of similarity measures. Similar tasks to indexing include top-k queries, which tries to find k most similar time series in the database.

2) Classification: One of the major tasks in time series data mining is time series classification (TSC), which involves building a classifier that learns from a training set of labeled data and predicting the class of an unlabeled set of data. Classically, machine learning classification is built by defining two elements: a distance metric to compare samples and a classification algorithm which implements the method of comparison. For example, in k-Nearest Neighbor (kNN) classification, a given sample is directly compared with samples from the training set. The tested sample will be assigned the class label of the sample from the training set which is the closest to it following a distance measure, for instance, the Euclidean distance. This approach can be expensive if the training dataset is large and thus will lead to a high number of pairwise comparisons. More efficient approaches (e.g., Domain Series Corpus [28], [19]) often try to aggregate and abstract features from the training set instead of conducting brute-force pairwise comparison. Time series classification can be helpful in tasks such as profiling electric appliances and conducting non-intrusive load monitoring. However, since time series classification is supervised learning and requires predefined class labels in training set which may require tedious manual work, this study mainly focuses on time series clustering, which is essentially the unsupervised alternative of time series classification.

3) Clustering: Time series clustering is a common type of unsupervised time series mining task that tries to partition time series into homogeneous groups while maximizing within-group similarity and between-group dissimilarity [29]. Clustering algorithms can be categorized into different families based on their underlying models, for instance hierarchical clustering which is based on connectivity and centroid-based clustering (e.g. k-means) where clusters are represented by a representative point. This paper is more interested in the former, since hierarchical clusters can be represented as a dendrogram, which depicts the hierarchy arrangement of clusters that can be merged with another at certain distances. Hierarchical clustering do not attempt to generate an arbitrary number of clusters. Instead, it produces a hierarchy that is easier for users to understand and users can set the break points by themselves. Hierarchical agglomerative clustering has recently received great interests in pattern recognition and become especially popular in financial applications.

Hierarchical clustering can employ either a bottom-up (agglomerative) or top-down (divisive) approach. The former starts with a single instance from the dataset and gradually aggregates instances into clusters until all instances are grouped into a single cluster, while the latter starts with the whole dataset and iteratively divide the dataset into clusters. In general, agglomerative methods are computationally more efficient than divisive ones, thus in this paper the former is more favorable, especially the Ward's method [30], which is a popular algorithm used to minimize the total within-cluster variance. Recall that Ward – as an agglomerative approach – works incrementally, the distance (namely the Ward's Linkage) of clusters $I \cup J$ and K are calculated based on a distance

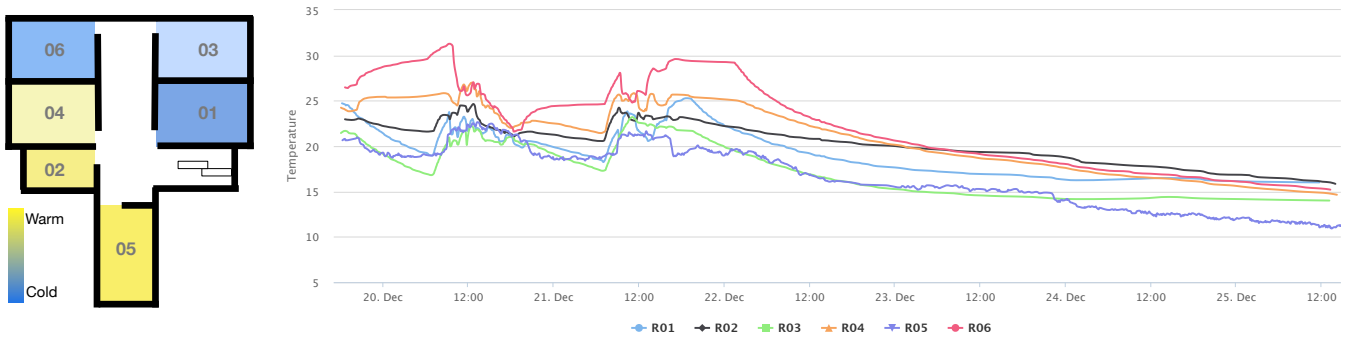

Fig. 3. Heat map of realtime room temperature with simplified floor plan (left) and temperature evolvement of six classrooms during Christmas in 2016 (right).

update formula $D_{Ward}(I \cup J, K)$ as specified below:

$$\sqrt{\frac{(|I|+|K|)D(I,K)^2 + (|J|+|K|)D(J,K)^2 - |K|D(I,J)^2}{|I|+|J|+|K|}}$$

where I and J are two clusters to be joined into a new cluster and K is any other cluster, and $|*|$ denotes the number of instances in one cluster. The computational complexity of Ward is $O(n^2)$, where n is the size of the dataset. Ward is widely available in many software packages, for example in Matlab and Wolfram Mathematica.

4) Motif discovery: Motifs in time series are often referred to as recurring patterns and frequent subsequences [31], [32]. Motif discovery is important since it can summarize and extract helpful information from the time series database and provide useful insights to domain experts [33]. Contrary to query by content that finds user-specified patterns from a time series database, motif discovery aims for finding previously undiscovered but recurrent time series subsequences. In this sense, motif discovery can be considered as the unsupervised version of query by content.

When motif discovery was firstly introduced in 2002, it has a quadratic computation complexity. However, as more and more researchers become interested in it, many algorithms have been proposed to improve both the efficiency as accuracy. For example, recent work by Mueen et al. [34], [35] proposes an exact algorithm to enumerate motifs of all lengths in databases containing millions of time series records.

5) Anomaly detection: In realtime monitoring systems, it is sometimes necessary to detect anomalies or unusual events in order to detect failures or trigger alarms. Contrary to subsequence matching, anomaly detection is the identification of previously unknown and unexpected subsequences. It can be particularly difficult since what constitutes an anomaly can greatly differ depending on the context. There can be two different approaches for anomaly detection in time series: a motif/discord approach that finds abnormal patterns statistically in a large time series database and a comparison-to-prediction approach that tries to identify abnormal points where the actual values deviates from predictions.

Loosely in the motif/discord approach, anomaly detection can be considered the opposite of motif discovery (e.g., discords [36]): subsequences that are not included in any motif discovery results may seem abnormal. That is, if motifs are considered as normal patterns, then the rest can be abnormal ones. Although named as anomalies, they do not necessary constitute an error or harm. Instead, anomalies can also be surprising or interesting patterns.

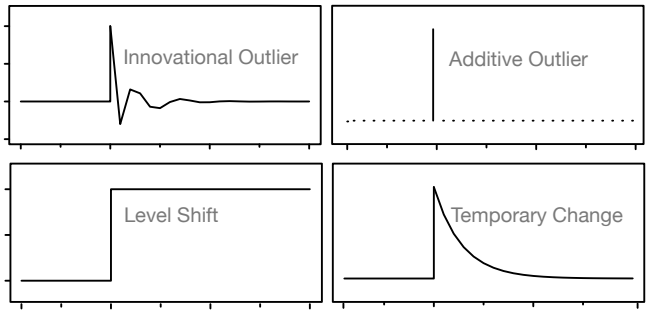

Fig. 4. Illustration of four types of time series outliers.

Anomalies can also be obtained by comparing the forecasted values to actual values. One of the prevailing prediction techniques for time series is probably autoregressive integrated moving average (ARIMA), which fits time series data into models represented as coefficients and tries to provide a better understanding of the data or to predict future data points. Chen and Liu [37] has taken advantage of ARIMA models to detect outliers in time series. Furthermore, they have mathematically defined four different kinds of time series outliers that are illustrated in Fig. 4. Informally, an innovational outlier is a point that affects both the observation itself and subsequent observations, while an additive outlier is a single affected point during all observations. A level shift, as the name suggests, is a shift of observed values during the process. Finally, a temporary change is a special case of innovational outlier, with the exception that subsequent observations eventually returns to normal after the outlier.

In practice, different types of outliers may happen during a short time span. It is thus necessary to have a measure of the outlier effect at a specific point. To this end, an aggregation approach has been proposed in [37] to estimate where outliers in a time series are located. Fig. 5 shows an example of the quantitive process of locating outliers in time series data. After calculation of outlier effects, it is then straightforward and mathematically easy to locate all outliers, as they often happen where there is a level shift in the outlier effect curve.

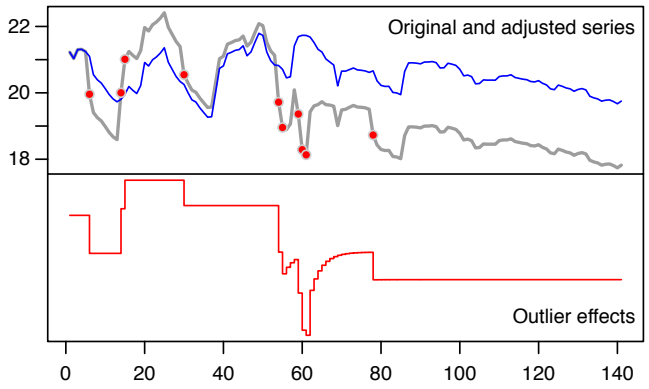

Fig. 5. Example of predicted series (in blue) based on original time series (in gray). Level shifts in the outlier effect curve illustrate where outliers are located (marked in red points).

III. Experiments with real-world data

It is paramount to validate the proposed research methodology in real-world settings. As a result, the architecture proposed in Section II have been implemented and tested in two different scenarios. In the first case, the holistic data mining system started collecting BAS data from a single school building located in western Europe, which has around 100 classrooms, labs and offices located on five different floors. This building was planned and constructed around 2000 and most rooms as well as outside facades are equipped with sensors and actuators to monitor and control temperature and heating, ventilation, illumination, etc. for the ease of building operations. In total, this building has more than a thousand sensors and actuators connected to the BAS. All these sensors and actuators (such as light switches and dimming units) have been connected to a KNX bus, which is a broadcast networking protocol where all communication telegrams passes on the bus and pre-configured source/destination pairs may send and receive only relevant telegrams and react. Over a one-year period (from mid-February 2016 till mid-February 2017), this system has collected 20,196,789 data points to its database. As mentioned previously, KNX is a very popular building automation and control protocol that has been deployed in several millions of installations worldwide.

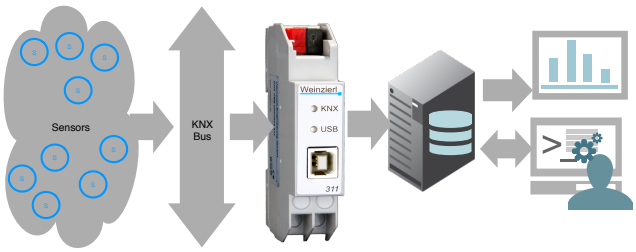

Fig. 6. Overview of data collection and management process.

The system implementation is a simplified version of the proposed architecture depicted in Fig. 1. However, all the core components are present in this system. Fig. 6 illustrates the implementation details. Specifically, thanks to the broadcasting nature of KNX, it is easy to simply attach a KNX-to-USB interface to the KNX bus and listens to every telegram on the bus to a gateway server via the USB interface (in this case, it is a Weinzierl KNX USB Interface 311, which costs around 200 Euros). This gateway then parses and stores all KNX telegrams to the Elasticsearch database hosted on a Linux server. In fact, the gateway is a part of this server – which is an old personal computer and this server is connected to the KNX USB interface. Moreover, this server hosts all the visualization components and data mining modules. When this server is connected to the Internet, remote access can also be granted.

Each record in the database consists of information such as telegram timestamp, source and destination addresses, KNX telegram type and message (generally a numerical value) parsed from this telegram. The commercial BAS system currently being used by the school provides only an interface to monitor real-time readings from each sensor and no history data were store anywhere. As a result, a dashboard (a web application) has been developed for the building operators to not only monitor and view charts about the real-time values of sensors and actuators but also historical records. In the dashboard backend, users and operators may configure a more customized dashboard interface by themselves.

The remaining of this section illustrates and showcases how the holistic mining framework works using less sensitive data such as indoor temperature records and lighting information. More sensitive data – for instance, data from presence sensors – are however not included in this paper in order to protect the privacy of the school where the data has been collected. It is nevertheless easy and straightforward to apply this data mining framework to any time series data regardless of data sources, as all data mining procedures incorporated in PHoliData are generic time series algorithms. A demonstration video of the PHoliData prototype has been recorded and interested audience can view it online[3].

A. Querying interesting patterns

Among the operation tasks, building operators often find certain usage patterns that are interesting. For example, there might be strange oscillation of sensor readings starting from a specific timestamp. When building operators try to understand this type of oscillation, as a first resort they usually seek to find whether similar patterns have occurred in the past. As a result, finding similar patterns from historical records is an important step for understanding the patterns themselves.

To find similar patterns in history, one must first define or describe such patterns. Obviously, for building operators it is infeasible to describe the patterns by abstracting them with mathematical models or natural languages. The easiest way can be just selecting the specific patterns visually and use the selected readings to query the history. To that end, a query-by-content paradigm can be applied. This functionality has been integrated in the holistic data mining framework, where users can select segments on a chart containing the interesting patterns and then the data mining module will try to find in history the corresponding patterns that matches users' selection. Fig. 7 illustrates how one user's query content aligns

[3]https://www.dropbox.com/s/jpvgfznp8j9onav/PHoliData.mp4

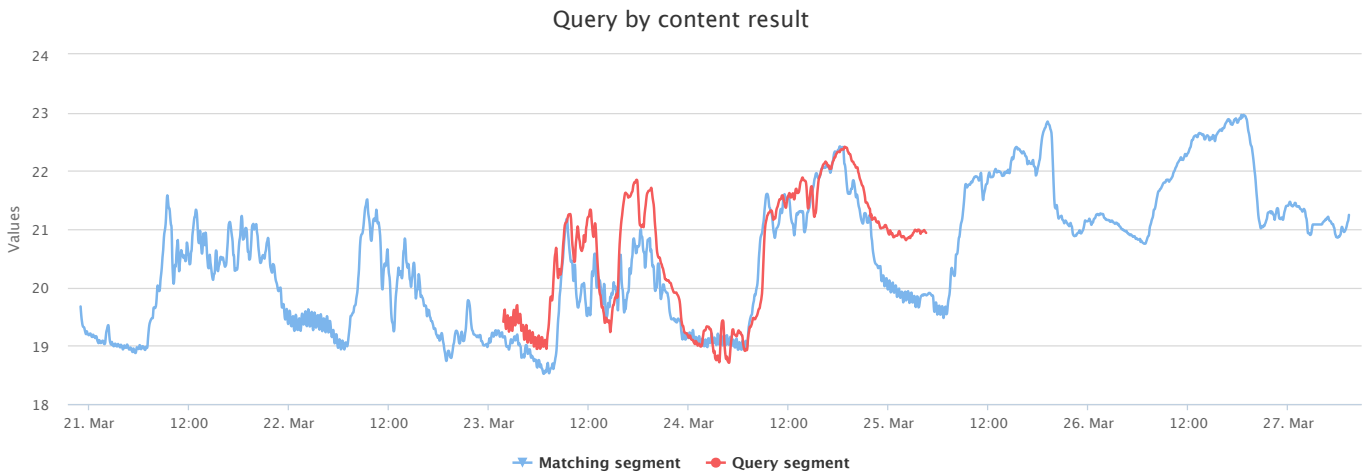

Fig. 7. Example of query-by-content result, where the query content (in red) matches a similar pattern in history.

with historical records. In this figure, user's query concerns indoor temperature movement of a classroom from February 2nd to February 4th in 2017 and the retrieved matching segment spans from March 23rd to March 25th in 2016. As shown, although they are not a perfect point-to-point match, the oscillation of the curve is indeed very similar. DTW distance has been used in this case to evaluate the similarity of two time series, since it allows slight distortions and values overall time series movements more than details. Euclidean distance in this case may not be suitable, because it is more focused on details.

B. Classification

Since time series classification is a supervised learning approach that mandates labelled data as the input of such algorithms, classification tasks often follows unsupervised learning approaches in practice. That is, unsupervised learning methods – e.g., clustering and motif discovery – are exploratory approaches that helps human users finding hidden patterns and understanding them. On the other hand, supervised learning – e.g., classification and regression – are often used to find known patterns defined *apriori*.

An example of applying time series classification to the smart buildings domain is Non-Intrusive Load Monitoring (NILM), which tries to disaggregate households' single-point energy consumption measurements into individual devices' consumption in a non-intrusive manner. Previous work [38] by the authors of this paper has taken advantage of the state-of-the-art time series classification approach to profile the electricity usage patterns of different household appliances in the aim of establishing a pattern dictionary that will be helpful for identifying different type of appliances during NILM tasks. This approach has been evaluated with a dataset obtained from the UK Household Electricity Usage Survey (HEUS) project [39], which contains electric appliance usage readings from 251 households within a one-year span (2010 to 2011). Results have proven that it is indeed feasible to use time series classification techniques for NILM.

Since the dataset of this study does not contain any electricity metering data or any labelled data, it is thus not possible to present the application of time series classification in this paper. However, if labelled time series data are present, the proposed framework is able to conduct classification tasks with ease. In the near future, an on-site monitoring of devices with the help of building operators and residents will be conducted in order to establish ground-truth for appliance profiles, so that time series classification algorithms will produce meaningful NILM results.

C. Clustering

Recall that Fig. 3 (right) shows the temperature readings in six rooms for a few days. When the time span and number of rooms increases, it can be difficult for human eyes to investigate how different or similar reading series are. As a result, there needs to be a more straightforward way to illustrate the similarities of readings from different locations. One solution is to generate a distance matrix diagram of temperature movements from different rooms where darker blocks indicate more differences rather than similarities (cf. Fig. 8 left). However, it may take some time (even for experts) to identify that $R6$ and $R4$ are more different than others. When applying agglomerative clustering techniques on the temperature movements (cf. Fig. 8 right), it has produced two bottom level clusters $\{R1, R3\}$ and $\{R2, R5\}$, and moving up from the latter, $R4$ can be attached to $\{R2, R5\}$ to form a larger cluster, which can then be joined by $R6$. After rearranging the distance matrix by the hierarchical clustering results and generating the clustergram, it is now straightforward to that $R6$ temperate movements are quite different from others. Furthermore, clustergrams give a simplified view of how similar or different each reading series are compared with peers. Consequently, time series clustering can be used both for outlier detection and activity inference, which will be discussed in detail in the remaining of this subsection.

1) Outlier detection with hierarchical clustering: There can be different types of outlier detection mechanisms. For example, a sudden temperature drop may indicate facility failures. In the case, an outlier is only isolated to its own historical values. On the other hand, an outlier can also mean inconsistent behaviors or movements compared to peer readings. For instance, if temperature in one room R is increasing while neighboring rooms are cooling down, it may indicate an

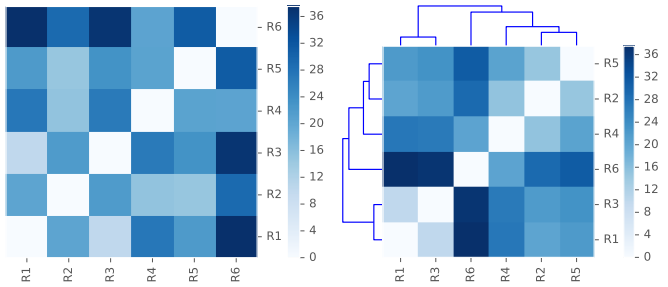

Fig. 8. Distance matrix of temperature movements with Euclidean distance (left) and agglomerative clustering clustergram of temperature readings for six rooms with Ward (right).

outlier. This outlier may come from malfunctioning cooling system in R, or possibly malfunctioning heating systems in all the other rooms. In either case, this outlier indicates inconsistency that requires human examination or intervention.

The plug-and-play data mining framework proposed in this paper has integrated the outlier by clustering mechanism. For example, when information about how long (in percentage) each light is turned on for a specific time period are modeled as time series, and an agglomerative clustering algorithm is applied to the modeled data, it is straightforward to examine the inconsistent lighting behaviors. As the clustergram shown in Fig. 9, where both the clustering hierarchy and the distance heat map are present (darker color indicates more inconsistency), it is clear that four lights behave different than all the rest. An investigation of these time series readings leaves some hint of why these four lights have abnormal behaviors. Fig. 10 shows three series on lighting information. It is clear that while lights in the small library are turned on and off more or less periodically, lights in the large library and especially in the WC are often left on during nights and sometimes even on weekends.

Plausibly, it is relatively easier to use simple rules to find such anomalies in the lighting system. For instance, it may be efficient to just examine the switch status of lights during nights and weekends. However, this kind of rules allows little flexibility and abuse of such rules will likely cause user confusion and bad experience. Furthermore, signals and hints flowing from other systems (e.g. HVAC) may require much more complicated rules, which will be too difficult to maintain in the end. As a result, using clustergrams to find abnormal or interesting patterns and behaviors can be more flexible and reliable.

2) Indoor activity inference: Recent research on BAS promotes the idea of collectively taking advantage of both real-time occupancy information and occupant preferences when designing more efficient building control systems. For instance, Chen et al. [40] propose a control system that keeps track of occupants' real-time indoor location to enable fine-grained control of ambient environment including lighting, cooling, heating, etc. On the other hand, as sensors and actuators are deployed in buildings and these systems are connected to external networks such as the Internet, occupant security and privacy become a more challenging task since sensor data can be leveraged to make unwanted inferences about occupants and their behaviors [41]. As a result, indoor occupant activity inference and detection can be a softer approach when occupants are becoming increasingly concerned about their own privacy. For instance, motion sensors and smart meters can be used for detecting whether a room is occupied and even for analyzing occupant identities [42]. A more recent work [43] explores the resonance effects of rooms and brings up models to infer the number of occupants by observing changes in the ultrasonic spectrum reflected back from a centrally located ultrasonic chirp transmitter. Furthermore, Jin et al. [44] try to infer implicit factors by indirect measurements based on the physical environment. They argue that occupancy can be inferred by indoor CO_2 concentration. In this section, an investigation of inferring indoor activity using room temperature movements is presented.

It is proven that indoor temperature movements are closely correlated with rooms' physical locations, especially when human factors are not taken into account [23]. As a result, it may be possible to infer occupant activities by proxy of temperature movements when human factors are indeed considered. Using agglomerative clustering of indoor temperature movements that are largely influenced by human factors, it is possible to infer which type of activity is prevalent in one room. To validate this point, temperature movements of 20 rooms that serve different functionality are selected and clustered. Out of these rooms, some are offices or labs, while others are normal classrooms or libraries. Note that these rooms are located on different floors of the building and rooms with similar functionality are generally not physically close to each other.

Fig. 11 presents the agglomerative clustering results. It has turned out that rooms with similar activities or functionality are generally clustered together. For instance, offices seem to have similar temperature movements and science labs do not share much similarity with other type of rooms other than slight similarity with offices. Besides, clustering results suggest that temperature movements in cafeteria, auditorium and reception are quite similar, probably due to the fact that all these rooms see bursts of occupants at specific time slots. Moreover, the rooms hosting kindergartens and preschool classes fall into one cluster, which can be joined by a meeting room, probably indicating that such rooms usually have smaller number of occupants. Last but not least, classrooms for training purposes (art, music and culinary) also have similar temperature movements.

Note that although clustering of indoor temperature movements is not able to predict what exactly a specific occupant activity is at a given moment, by comparing with other activity traces it is still possible to infer roughly what such an activity can possibly be. Furthermore, if a database of how different occupant activities can impact indoor temperature movements is curated with higher measuring requirements (e.g., higher sampling frequency, more accurate measurements and larger amount of records), it may be well possible that finer-grained activity inference can be achieved.

D. Motif discovery

Motif discovery in time series mining is an important unsupervised learning task that finds hidden and recurring patterns

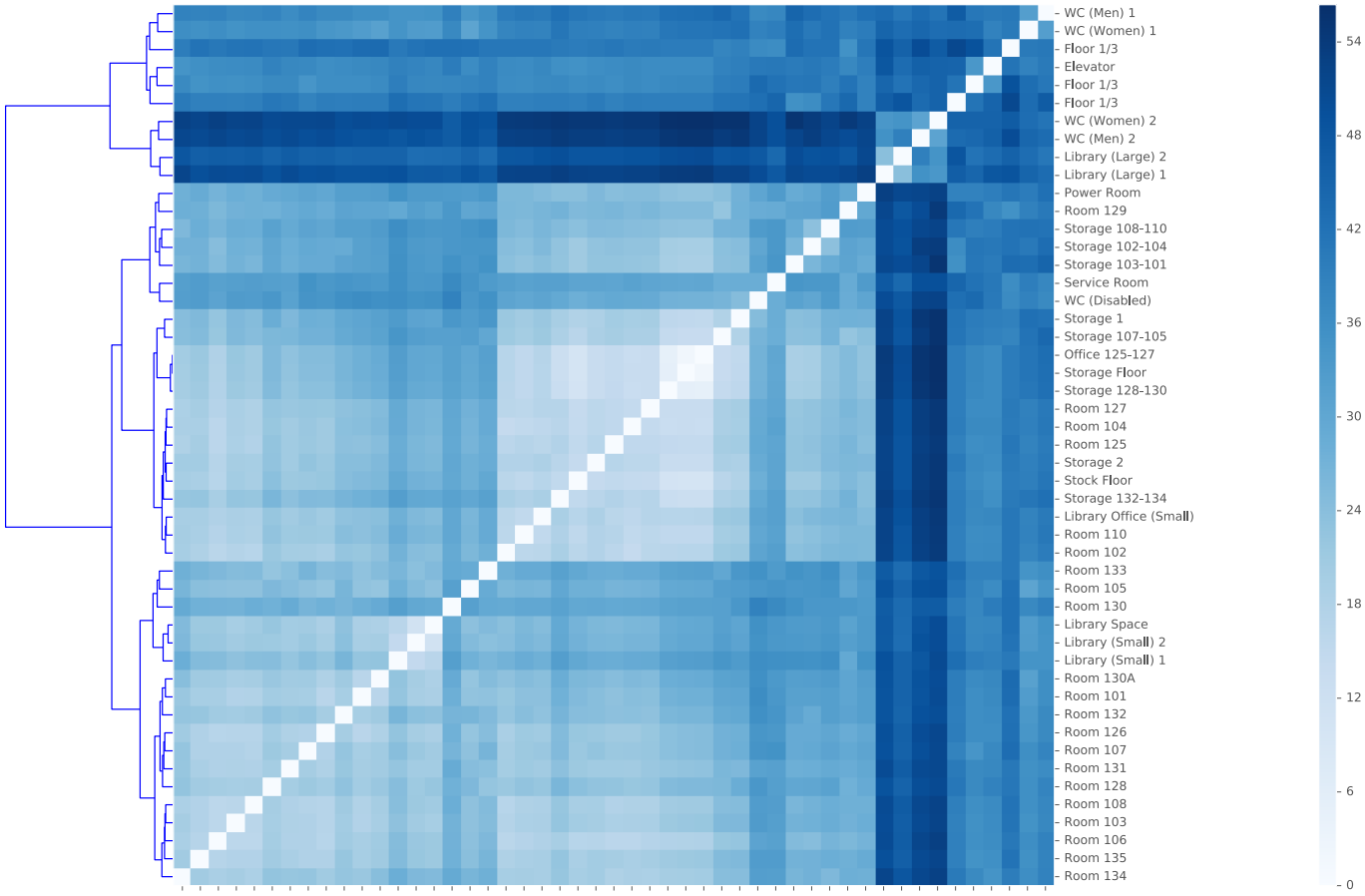

Fig. 9. Agglomerative clustering on lighting information of a single floor in the experiment subject school.

– which are time series subsequences that have occurred more than once in a non-trivial manner, that is, motifs are too similar to happen at random [35]. In building operations, operators may be interested in such recurring patterns. For example, operators may need to investigate the electricity and water usage patterns in order to make predictions and to locate abnormal usages.

Fig. 12 illustrates the motif discovery process with indoor temperature readings. The subfigure on top shows the latest overall temperature movements from the beginning of February to the time of writing this paper. Although humans are keen at finding patterns at times, it can be extremely effort-consuming and error-prone. On the other hand, computer algorithms are able to locate motifs easily. For example, The middle and bottom subfigures in Fig. 12 showcase two motifs found from the temperature movements. With subsequences aligned together by different motifs, it is indeed straightforward and convincing to extract recurring patterns. After these motifs are revealed, it is also possible to use them as query contents and try to find similar subsequences in further history – e.g., using the query-by-content functionality introduced earlier – so as to understand better the motifs themselves.

E. Anomaly detection using trend prediction

As discussed earlier, anomaly detection can be done either considering only a single data source (e.g., temperature readings from one room) or taking advantage of collective data (e.g., temperature readings from rooms next to each other). Since the latter has been demonstrated in Section III-C1, this section presents a case study that tracks single time series and locate anomalies.

Fig. 13 demonstrates how anomaly detection works with temperature readings from a classroom during Christmas in 2016. As shown, the blue curve representing the classroom temperature gradually decreases starting from 22th of December and then increases to normal from 28th of December (Monday). When calculating the forecasted temperatures (in black curve) and outlier effects (shown in red, the outlier types have been omitted for the sake of visualization), it is obvious that there are many anomalies located during the aforementioned two dates. This system successfully detects exactly when the temperature starts to drop and to increase. When consulting the building operator, it then became clear that the heating system has been switch off during during Christmas vacation and then turned on again during the following Monday. These two points matches well with the results from the anomaly detection module.

Although in this specific case it is the building operator that has triggered the anomaly, the anomaly detection can nonetheless be beneficial for detecting real outliers. For instance, it is possible the heating system in certain rooms stops working when no one is monitoring real-time readings (e.g., during nighttime or vacation period). Besides, in some cases it may be easy for human users to spot suspecting segments

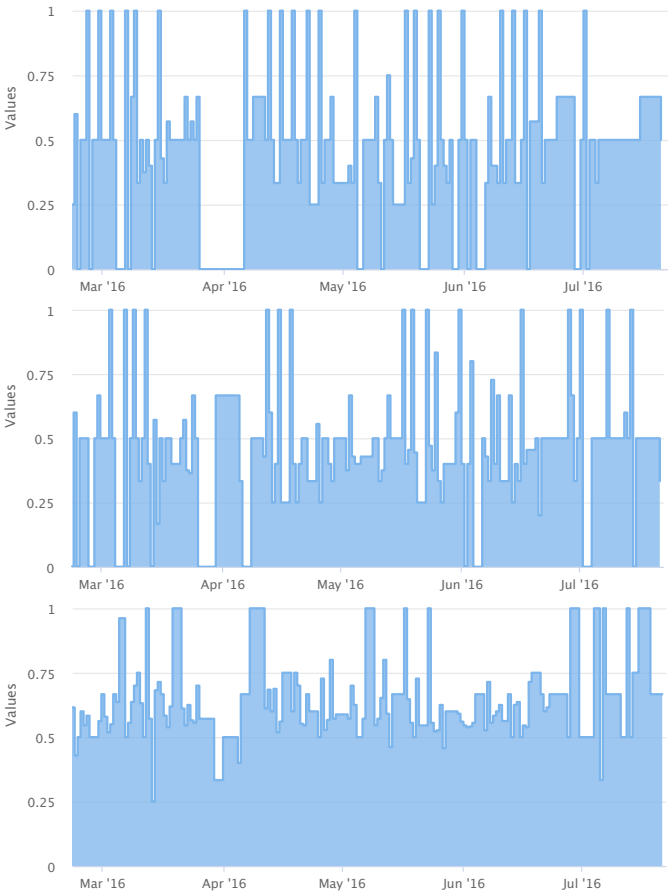

Fig. 10. Lighting information of the small library (top), large library (middle) and WC (bottom), where values indicate the fraction of unit time when lights are turned on.

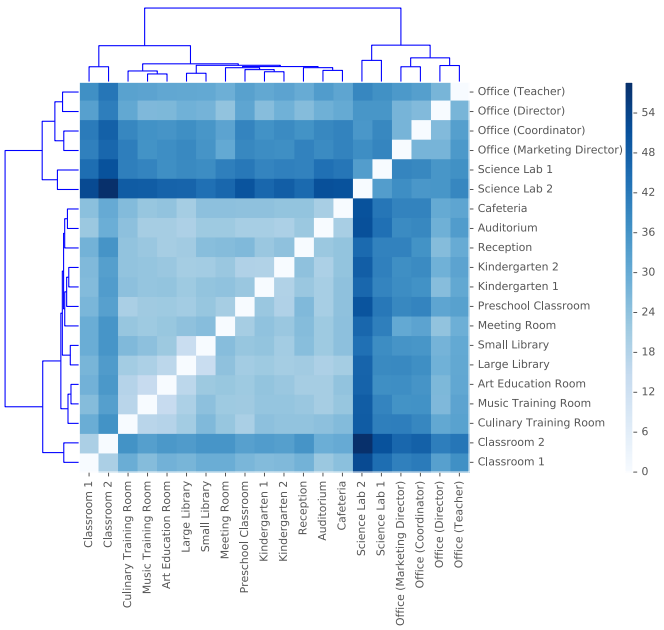

Fig. 11. Agglomerative clustering on temperature movements of 20 rooms with different functionality.

in time series, this process can otherwise be time-consuming and prone to errors. With the help of an anomaly detection

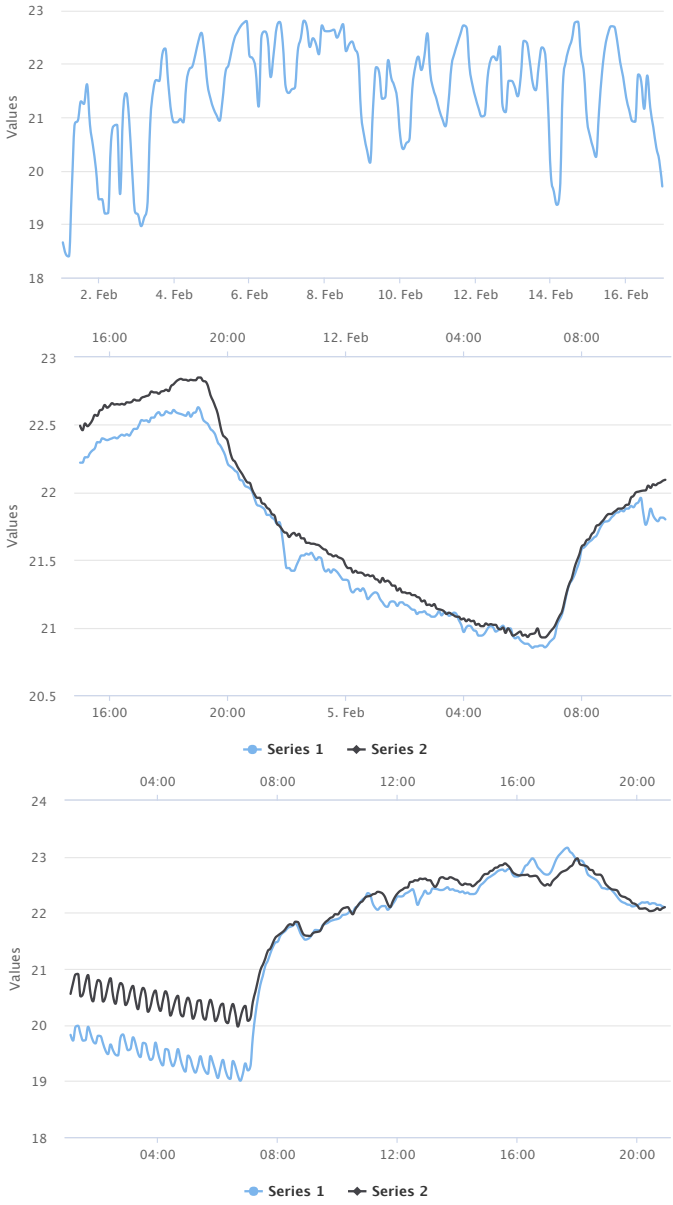

Fig. 12. Example of motif discovery with temperature data. Top: input temperature movements. Middle: one discovered motif with occurrences on February 5th and February 12th. Bottom: another motif with occurrences on February 14th and February 15th.

mechanism, the system can find out the anomalies in time and alert relevant parties before the situation deteriorates.

Finally, this case study demonstrates outlier detection results for a single time series, it is as well feasible to independently detect outliers from multiple series and then aggregate the detected outliers to form collective outliers, so as to eliminate noises in the detection process. For example, there is also an anomaly on 26th of December in Fig. 13, where there is a sudden temperature drop. If it is due to a temporary sensor failure that are not seen in other data readings, then such outliers can be safely removed during the outlier aggregation process.

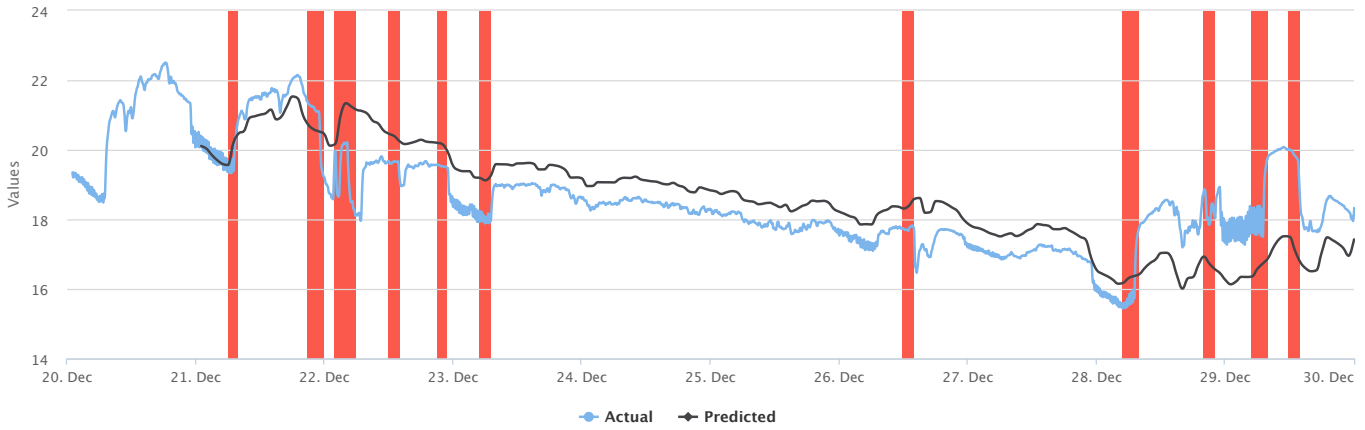

Fig. 13. Anomaly detection results using ARIMA models. Red bands indicate where possible anomalies are located.

IV. Towards plug-and-play data mining in smart buildings

It is known that smart buildings components suffer from many interoperability issues, including hardware, network and application interoperability issues. As a generic data mining framework for smart buildings, it may be of little use if PHoliData can only applied to specific buildings or settings. In case the framework can not be directly used in different settings, the cost of adapting this framework to a different scenario should be minimum.

In order to test the plug-and-play capability of the proposed framework, it must be tested against different real-world settings. To that end, an experiment has been conducted to take advantage of the prototype framework used in the previous section and apply it to a different building with totally different sensors and networking standards. Specifically, this new setting involves mainly data records from electricity meters, heating systems and water valves from an annex of an ancient castle that dates back to the 13th century. A list of changes made to the original data mining framework is enumerated as follows:

1) Adaptor for data collection module. Since operators of the new building are mainly concerned with sensors such as smart meters and heating systems which are not connected to a KNX bus, a different adaptor has to be installed in place to collect data from these sensors. In this case, a LOYTEC LGATE-950 gateway has been used to communicate with energy meters (Modbus), water and heat counters (MBus).
2) Script for feeding data into the database. This script reads the output of the gateway software periodically and stores the data into Elasticsearch server. Since Elasticsearch does not require predefined database schema, it thus allows painless integration regarding data persistence.
3) Sensor list with names. Sensors in buildings are often identified by unified IDs. In KNX, such IDs are network addresses; in this case, the sensor list is already available and can be obtained from the building management team. This list (in Excel, CSV or JSON format) is then fed into the framework for visualization and data mining purposes.

After these steps, the data mining framework is fully functional: data preprocessing is done automatically with Elasticsearch's aggregation functions and building operators can start exploring the realtime as well as historical status of their buildings with the help of visualization and data mining modules. As shown, the efforts for adapting the proposed framework to a totally different scenario are indeed not much. Furthermore, the plug-and-play capability of PHoliData can be further improved by supporting major standards in the building industry, such as BACnet and LonTalk. After major standards are supported in the framework, the first two steps can be discarded when adapting it to new different settings.

V. Conclusions and future work

In order to tackle energy efficiency and user comfort challenges in smart buildings, modern buildings are more and more equipped with various sensors and actuators, which generate a lot of data within buildings. While such data should have great value for understanding the status and occupant behaviors, they are generally underexploited due to technical and cost limitations. This study has proposed to take full advantage of the data flowing inside buildings in order to diagnose issues, explore solutions and improve occupant-building interactions in smart buildings. To that end, this paper unveiled a plug-and-play and holistic data mining framework named PHoliData for smart buildings to collect, store, visualize and mine useful information and domain knowledge from data in smart buildings. PHoliData allows non technical experts to easily explore and understand their buildings with minimum IT support. An architecture of PHoliData has been introduced and a prototype has been implemented and tested against two real-world settings. Discussions with industry experts indicate that the system to be extremely helpful for understanding buildings, since it can provide hints about energy efficiency improvements. Finally, extensive experiments have demonstrated the feasibility of such a framework in practice and its advantage and potential for buildings operators.

Although the current design and implementation of the proposed framework does not ensure full capability of plug-

and-play, it is on its way of evolvement towards this ultimate goal. It is foreseen that in the near future the plug-and-play capability of PHoliData will be further improved by supporting major standards in the building industry, such as BACnet and LonTalk. Besides, more relevant data mining algorithms – e.g., multivariate time series mining and association rule mining – can be included in PHoliData, so that more domain knowledge can be extracted. Moreover, machine learning and data mining techniques can also be used for learning operators' preferences about notifications on events or alerts, so that different operators receive different streams of notifications according to their preferences. Finally, automatic report generation can be included as a component in the framework, in order to provide more straightforward and standardized presentation about buildings' status.

ACKNOWLEDGEMENTS

This study is supported by Paul Wurth S.A. and Luxembourg Ministry of Economy. The authors would like to thank International School of Luxembourg for offering the opportunity of collecting data and giving valuable feedbacks.

REFERENCES

[1] M. Molina-Solana, M. Ros, M. D. Ruiz, J. Gómez-Romero, and M. Martin-Bautista, "Data science for building energy management: A review," *Renewable and Sustainable Energy Reviews*, vol. 70, pp. 598–609, 2017.

[2] T. A. Nguyen and M. Aiello, "Energy intelligent buildings based on user activity: A survey," *Energy and buildings*, vol. 56, pp. 244–257, 2013.

[3] M. Ippolito, E. R. Sanseverino, and G. Zizzo, "Impact of building automation control systems and technical building management systems on the energy performance class of residential buildings: An italian case study," *Energy and Buildings*, vol. 69, pp. 33–40, 2014.

[4] D. T. Delaney, G. M. O'Hare, and A. G. Ruzzelli, "Evaluation of energy-efficiency in lighting systems using sensor networks," in *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. ACM, 2009, pp. 61–66.

[5] V. Garg and N. Bansal, "Smart occupancy sensors to reduce energy consumption," *Energy and Buildings*, vol. 32, no. 1, pp. 81–87, 2000.

[6] K. Padmanabh, A. Malikarjuna V, S. Sen, S. P. Katru, A. Kumar, S. K. Vuppala, S. Paul *et al.*, "isense: a wireless sensor network based conference room management system," in *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. ACM, 2009, pp. 37–42.

[7] A. Kusiak, M. Li, and Z. Zhang, "A data-driven approach for steam load prediction in buildings," *Applied Energy*, vol. 87, no. 3, pp. 925–933, 2010.

[8] Z. Yu, F. Haghighat, B. C. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy and Buildings*, vol. 42, no. 10, pp. 1637–1646, 2010.

[9] Z. Y. Li, "An empirical study of knowledge discovery on daily electrical peak load using decision tree," in *Advanced Materials Research*, vol. 433. Trans Tech Publ, 2012, pp. 4898–4902.

[10] Z. J. Yu, F. Haghighat, B. C. Fung, and L. Zhou, "A novel methodology for knowledge discovery through mining associations between building operational data," *Energy and Buildings*, vol. 47, pp. 430–440, 2012.

[11] F. Xiao and C. Fan, "Data mining in building automation system for improving building operational performance," *Energy and buildings*, vol. 75, pp. 109–118, 2014.

[12] M. Li, L. Miao, and J. Shi, "Analyzing heating equipment's operations based on measured data," *Energy and Buildings*, vol. 82, pp. 47–56, 2014.

[13] C. León, F. Biscarri, I. Monedero, J. I. Guerrero, J. Biscarri, and R. Millán, "Integrated expert system applied to the analysis of non-technical losses in power utilities," *Expert systems with applications*, vol. 38, no. 8, pp. 10 274–10 285, 2011.

[14] P. H. Shaikh, N. B. M. Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim, "A review on optimized control systems for building energy and comfort management of smart sustainable buildings," *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 409–429, 2014.

[15] B. Hu, Y. Chen, and E. Keogh, "Time series classification under more realistic assumptions," in *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 2013, pp. 578–586.

[16] N. Balta-Ozkan, R. Davidson, M. Bicket, and L. Whitmarsh, "The development of smart homes market in the uk," *Energy*, vol. 60, pp. 361–372, 2013.

[17] KNX Association, "KNX is the most popular protocol in European markets," 2012, https://www.knx.org/knx-en/news/New-Press-release-KNX-is-the-most-popular-protocol-in-European-markets/details.php?ref=1666. Accessed on February 1st, 2017.

[18] D. Li, T. F. Bissyande, J. Klein, and Y. Le Traon, "Time series classification with discrete wavelet transformed data: Insights from an empirical study," in *The 28th International Conference on Software Engineering and Knowledge Engineering (SEKE 2016)*, 2016.

[19] D. Li, T. F. Bissyandé, J. Klein, and Y. Le Traon, "DSCo-NG: A Practical Language Modeling Approach for Time Series Classification," in *International Symposium on Intelligent Data Analysis*. Springer International Publishing, 2016, pp. 1–13.

[20] C. Fan, F. Xiao, H. Madsen, and D. Wang, "Temporal knowledge discovery in big bas data for building energy management," *Energy and Buildings*, vol. 109, pp. 75–89, 2015.

[21] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Mining and knowledge discovery*, vol. 15, no. 2, pp. 107–144, 2007.

[22] C. Gormley and Z. Tong, *Elasticsearch: The Definitive Guide*. " O'Reilly Media, Inc.", 2015.

[23] D. Li, T. F. Bissyandé, J. Klein, and Y. Le Traon, "Sensing by Proxy in Buildings with Agglomerative Clustering of Indoor Temperature Movements," in *The 32nd ACM Symposium on Applied Computing (SAC 2017)*, Marrakesh, Morocco, April 2017, pp. 477–484.

[24] P. Royston, "Multiple imputation of missing values," *Stata journal*, vol. 4, no. 3, pp. 227–41, 2004.

[25] D. Q. Goldin and P. C. Kanellakis, "On similarity queries for time-series data: constraint specification and implementation," in *International Conference on Principles and Practice of Constraint Programming*. Springer, 1995, pp. 137–153.

[26] C. A. Ralanamahatana, J. Lin, D. Gunopulos, E. Keogh, M. Vlachos, and G. Das, "Mining time series data," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 1069–1103.

[27] R. A. G. Psaila and E. L. Wimmers Mohamed &It, "Querying shapes of histories," *Very Large Data Bases. Zurich, Switzerland: IEEE*, 1995.

[28] D. Li, L. Li, T. F. Bissyande, J. Klein, and Y. Le Traon, "DSCo: A Language Modeling Approach for Time Series Classification," in *12th International Conference on Machine Learning and Data Mining (MLDM 2016)*, 2016.

[29] T. W. Liao, "Clustering of time series data – a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[30] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[31] J. Lin, E. Keogh, S. Lonardi, and P. Patel, "Finding motifs in time series," in *Proc. of the 2nd Workshop on Temporal Data Mining*, 2002, pp. 53–68.

[32] P. Patel, E. Keogh, J. Lin, and S. Lonardi, "Mining motifs in massive time series databases," in *Proceedings of IEEE International Conference on Data Mining*. IEEE, 2002, pp. 370–377.

[33] N. C. Castro and P. J. Azevedo, "Significant motifs in time series," *Statistical Analysis and Data Mining*, vol. 5, no. 1, pp. 35–53, 2012.

[34] A. Mueen, "Time series motif discovery: dimensions and applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 2, pp. 152–159, 2014.

[35] A. Mueen and N. Chavoshi, "Enumeration of time series motifs of all lengths," *Knowledge and Information Systems*, vol. 45, no. 1, pp. 105–132, 2015.

[36] E. Keogh, J. Lin, and A. Fu, "Hot sax: Efficiently finding the most unusual time series subsequence," in *Data mining, fifth IEEE international conference on*. Ieee, 2005, pp. 8–pp.

[37] C. Chen and L.-M. Liu, "Joint estimation of model parameters and outlier effects in time series," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 284–297, 1993.

[38] D. Li, T. Bissyandé, S. Kubler, J. Klein, and Y. Le Traon, "Profiling household appliance electricity usage with n-gram language modeling,"

in *2016 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2016, pp. 604–609.

[39] J.-P. Zimmermann, M. Evans, J. Griggs, N. King, L. Harding, P. Roberts, and C. Evans, "Household electricity survey: A study of domestic electrical product usage," *Intertek Testing & Certification Ltd*, 2012.

[40] H. Chen, P. Chou, S. Duri, H. Lei, and J. Reason, "The design and implementation of a smart building control system," in *e-Business Engineering, 2009. ICEBE'09. IEEE International Conference on*. IEEE, 2009, pp. 255–262.

[41] T. Zhu, S. Xiao, Q. Zhang, Y. Gu, P. Yi, and Y. Li, "Emergent technologies in big data sensing: a survey," *International Journal of Distributed Sensor Networks*, vol. 2015, p. 8, 2015.

[42] L. Yang, K. Ting, and M. B. Srivastava, "Inferring occupancy from opportunistically available sensor data," in *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*. IEEE, 2014, pp. 60–68.

[43] O. Shih and A. Rowe, "Occupancy estimation using ultrasonic chirps," in *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems*. ACM, 2015, pp. 149–158.

[44] M. Jin, N. Bekiaris-Liberis, K. Weekly, C. Spanos, and A. M. Bayen, "Sensing by proxy: Occupancy detection based on indoor co2 concentration," in *The 9th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM'15)*, 2015, pp. 1–10.