Higher Order Mining for Monitoring District Heating Substations

Shahrooz Abghari*, Veselka Boeva*, Jens Brage†, Christian Johansson†, Håkan Grahn*, Niklas Lavesson‡
*Department of Computer Science, Blekinge Institute of Technology, Sweden, firstname.lastname@bth.se
†NODA Intelligent Systems AB, Sweden, firstname.lastname@noda.se
‡Department of Computer Science and Informatics, Jönköping University, Sweden, firstname.lastname@ju.se

***Abstract*—We propose a higher order mining (HOM) approach for modelling, monitoring and analyzing district heating (DH) substations' operational behaviour and performance. HOM is concerned with mining over patterns rather than *primary* or *raw* data. The proposed approach uses a combination of different data analysis techniques such as sequential pattern mining, clustering analysis, consensus clustering and minimum spanning tree (MST). Initially, a substation's operational behaviour is modeled by extracting weekly patterns and performing clustering analysis. The substation's performance is monitored by assessing its modeled behaviour for every two consecutive weeks. In case some significant difference is observed, further analysis is performed by integrating the built models into a consensus clustering and applying an MST for identifying deviating behaviours. The results of the study show that our method is robust for detecting deviating and sub-optimal behaviours of DH substations. In addition, the proposed method can facilitate domain experts in the interpretation and understanding of the substations' behaviour and performance by providing different data analysis and visualization techniques.**

***Index Terms*—District Heating Substations; Clustering Analysis; Minimum Spanning Tree; Data Mining; Higher Order Mining; Outlier Detection; Fault Detection**

I. Introduction

A district heating (DH) system provides a number of *buildings* with heat and domestic hot water from a *central boiler plant* through a *distribution network* for a limited geographical area. Different components of the DH system at the primary side are shown in Figure 1. The provided heat transfers through substations from the distribution network into consumers' buildings (the secondary side of the DH system) to get heat and domestic hot water on demand. The DH substations consist of different components and each can be a potential source of faults. Faults in substations can arise from stuck valves, fouled heat exchangers, malfunctions in temperature transmitters, control systems and many more [1], [2].

Gadd and Werner [3] divide faults in DH substations and secondary systems into three categories as follows: 1) faults resulting in comfort problems such as lack of enough heat or physical issues such as water leakage, 2) faults with known cause but unsolved since their identification are time demanding and costly, and 3) faults that require advanced fault detection systems. Faults in substations do not necessarily result in comfort problems for the consumers. Instead, in most cases they cause sub-optimal behaviour for a long time before they are noticed. Therefore, early detection of faults and deviations can reduce the maintenance cost and help avoid abnormal event progression. This in return makes it possible to lower the system's temperatures and provides an opportunity to use renewable and other low-value energy sources such as excess heat.

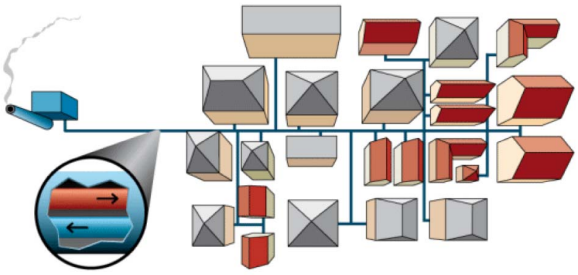


Fig. 1: District heating system (picture borrowed from the Swedish Energy Agency webpage[1])

In this study, we use a combination of data analysis techniques for modelling, monitoring, and analyzing the DH substations' operational behaviours. We propose a higher order mining (HOM) approach to facilitate domain experts in understanding faulty and deviating (sub-optimal) DH substations' behaviours. HOM is a sub-field of knowledge discovery that is applied on non-primary, derived data, or patterns to provide human-consumable results [4]. In our approach we apply sequential pattern mining on raw data, perform clustering analysis, consensus clustering, and minimum spanning tree (MST) construction on the extracted patterns.

We initially discretize the data and extract weekly frequent patterns. The patterns extracted for each week are grouped into clusters which model a DH substation's operational behaviour. Next, we analyze and assess the similarity between substation behaviours for every two consecutive weeks. The assessed similarity can be used to measure the discrepancy between the substation performance within the studied time period. When the discrepancy is significant (above a given threshold) we perform further analysis by integrating the produced clustering solutions into a consensus clustering. We further apply the MST algorithm, which builds an MST, by considering the exemplars of the built consensus clustering solution as nodes

[1] http://www.energimyndigheten.se/en/sustainability/households/heating-your-home/district-heating/

978-1-7281-4493-1/19/$31.00 ©2019 IEEE
DOI 10.1109/DSAA.2019.00053

and the distance between them as edges. An MST is a tree with a minimum traversing cost, and in order to identify deviating behaviours, we cut the longest edge(s) of the MST. Small and distant sub-trees can be interpreted as outliers. In addition, we calculate the similarity between the clustering solutions generated for each two consecutive weeks for the whole heating season. The calculated similarities can be used to create a performance signature profile of the substation and further be applied for comparing the performance of substations that belong to the same heat load category.

The main contribution of this study is a data-driven approach based on the HOM paradigm that combines sequential pattern mining, clustering analysis, and MST. We show the applicability of the approach in district heating domain to:

1. Model and analyze weekly DH substation's operational behaviour.
2. Monitor DH substation's performance by assessing the similarity among the weekly built models and creating a substation's performance signature profile for the whole studied period.
3. Identify deviating/faulty and sub-optimal substation's behaviours by integration analysis of the weekly built models.

II. Related Work

Fault is an abnormal state within the system that may cause a failure or a malfunction. Fault detection is the identification of an unacceptable deviation of at least one feature of the system from the expected or usual behaviour [5]. Fault detection has been researched and studied in different domains. There are several factors such as the nature of data, availability of labeled data, constraints and requirements of a fault detection problem that makes it domain specific [6]. In other words, most fault detection techniques are developed to address specific problems.

Katipamula and Brambley [7], [8] conducted an extensive review in two parts on fault detection and diagnosis (FDD) for building systems. They classified FDD methods based on the availability of a priori knowledge for formulating the diagnostics and highlighted their advantages and disadvantages.

In a recent review, Djenouri et al. [9] focused on the usage of machine learning in smart building applications. The authors classified the existing solutions in two main categories; occupancy monitoring such as user preferences, and energy/device-centric solutions. In each group the existing solutions were presented, discussed, and compared.

Gadd and Werner [3] showed that hourly meter readings can be used for detecting faults at DH substations. The authors identified three fault groups: 1) low average annual temperature difference, 2) poor substation control, and 3) unsuitable heat load pattern. The results of the study showed that low average annual temperature differences are the most important issues, and that addressing them can improve the efficiency of the DH systems. However, solving unsuitable heat load patterns is probably the easiest and the most cost-effective fault category to be considered.

Xue et al. [10] applied clustering analysis and association rule mining to detect faults in substations *with* and *without* return-water pressure pumps. Cluster analysis was applied in two steps 1) to partition the substations based on monthly historical heat load variations and 2) to identify daily heat variation using hourly data. The result of the clustering analysis was used for feature discretization and preparation for association rule mining. The results of the study showed the method can discover useful knowledge to improve the energy performance of the substations. However, for temporal knowledge discovery, advanced data mining techniques are required.

Månsson et al. [2] proposed a method based on gradient boosting regression to predict an hourly mass flow of a well performing substation using only a few number of features. The built model is tested by manipulating the well performing substation data to simulate two scenarios: communication problems and a drifting meter fault. The model prediction performance is evaluated by calculating the hourly residual of the actual and the predicted values on original and faulty datasets. Additionally, cumulative sums of residuals using a rolling window that contains residuals from the last 24 hours were calculated. The results of the study showed that the proposed model can be used for continued fault detection.

Ece et al. [11] proposed an approach for automatically discovering heat load patterns in DH systems. Heat load patterns reflect yearly heat usage in an individual building and their discovery is crucial for effective DH operations and managements. The authors applied k-shape clustering [12] on smart meter data to group buildings with similar heat load profiles. Additionally, the proposed method was shown to be capable of identifying buildings with abnormal heat profiles and unsuitable control strategies.

Sandin et al. [13] used probabilistic methods and heuristics for automated detection and ranking of faults in large-scale district energy systems. The authors studied a set of methods ranging from limit-checking and basic model to applying more sophisticated approaches such as regression modelling, clustering analysis on hourly energy metering.

Our current work is devoted to modelling, monitoring and analyzing the DH substations' operational behaviours by following the HOM paradigm. In this study, hourly data is transformed into categorical data and sequential pattern mining is used to extract weekly frequent patterns. After this step, we only focus on non-primary data (the extracted patterns) to perform different levels of clustering analysis and knowledge discovery to facilitate the domain experts in understanding the substations' operational behaviours. In contrast to the studies discussed above, by extracting weekly patterns we are able to monitor and assess the operational behaviours of a DH substation with respect to all selected features. This can support domain experts in better understanding the underlying specifics of the detected deviations by supplying them with a more complete view of the monitored phenomenon.

III. Methods and Techniques

A. Sequential Pattern Mining

Sequential pattern mining is the process of finding frequently occurring patterns in a sequence dataset. The records of the sequence dataset contain sequences of events that often have chronological order. In this study, we apply the PrefixSpan algorithm [14] to extract frequent sequential patterns. PrefixSpan applies a prefix-projection method to find sequential patterns. Given a sequence dataset and a user-specified threshold, the dataset is first scanned in order to identify all frequent items with the length one in sequences. Using a divide and conquer fashion the search space is divided into a number of subsets based on the extracted prefixes. Finally, for each subset a corresponding projected dataset is created and mined recursively.

B. Clustering Analysis

1) Affinity Propagation: We use the affinity propagation (AP) algorithm [15] for clustering the extracted patterns. AP is based on the concept of exchanging messages between data points. The exchanged messages at each step assist AP to choose the best samples as exemplars and which data points should choose those samples to be their exemplars. Unlike most clustering algorithms, such as k-means [16] which requires the number of clusters as an input, AP estimates the optimal number of clusters based on the data provided and the chosen exemplars are real data points. These characteristics make AP a suitable clustering algorithm for this study.

2) Consensus Clustering: Gionis et al. [17] proposed an approach for clustering that is based on the concept of aggregation. They are interested in a problem in which a number of different clustering solutions are given on some datasets of elements. The objective is to produce a single clustering of the elements that agrees as much as possible with the given clustering solutions. Consensus clustering algorithms deal with similar problems to those treated by clustering aggregation techniques. Namely, such algorithms try to reconcile clustering information about the same data phenomenon coming from different sources [18] or from different runs of the same algorithm [19]. In this study, we use the consensus clustering algorithm proposed in [18] in order to integrate the clustering solutions produced on the datasets collected for two consecutive weeks. We consider the exemplars (the representative patterns) of the produced clustering solutions. These exemplars are then divided into k groups (clusters) according to the degree of their similarity by applying the AP algorithm. Subsequently, the clusters whose exemplars belong to the same partition are merged in order to obtain the final consensus clustering.

C. Distance Measure

The similarity between the extracted patterns are assessed with Levenshtein distance (LD) metric [20]. The LD, also known as edit distance, is a string similarity metric that measures the minimum number of editing operations (insertion, deletion and substitution) required to transform one string into the other. We have used the normalized LD where score *zero* implies 100% similarity between the two patterns and *one* represents no similarity. LD is a simple algorithm capable of measuring the similarity between patterns with different lengths. Although in this study the extracted patterns have similar lengths, using LD can provide more flexibility when patterns with different lengths are required to be studied. Therefore, we choose LD as the similarity measure.

Given two clustering solutions $C = \{C_1, C_2, \ldots, C_n\}$ and $C' = \{C'_1, C'_2, \ldots, C'_m\}$ of datasets X and X', respectively the similarity, S_w, between C and C' can be assessed as follows:

$$S_w(C, C') = \frac{\sum_{i=1}^{n}(min_{j=1}^{m} w_i.d(c_i, c'_j))}{2} + \frac{\sum_{j=1}^{m}(min_{i=1}^{n} w'_j.d(c_i, c'_j))}{2}, \quad (1)$$

where c_i and c'_j are exemplars of the clustering solutions C_i and C'_j, respectively. The weights w_i and w'_j indicate the relative importance of clusters C_i and C'_j compared to other clusters in the clustering solutions C and C', respectively. For example, a weight w_i of a cluster C_i can be calculated as the ratio of its cardinality to the cardinality of the dataset X, i.e., $w_i = |C_i|/|X|$. The S_w has values in a range of [0,1]. Scores of *zero* imply identical performance while scores close to *one* show significant dissimilarities.

D. Minimum Spanning Tree

Given an undirected and connected graph $G = (V, E)$, a spanning tree of the graph G is a connected sub-graph with no cycles that include all vertices. A minimum spanning tree (MST) of an edge-weighted graph is a spanning tree where the sum of the weights of its edges is minimum among all the spanning trees. MSTs have been studied and applied in different fields including cluster analysis and outlier detection [21], [22], [23], [24], [25]. In this study we apply an MST on top of the created consensus clustering solution to further analyse the deviating substations' behaviours. We use Kruskal's algorithm [26] for building the MST. Kruskal's algorithm follows a greedy approach, i.e., at each iteration it chooses an edge which has least weight and adds it to the growing spanning tree. The algorithm first sorts the edges of G in an increasing order with respect to their weights. Then, it starts adding edges in sorted order and only those that do not form a cycle in the MST.

IV. Proposed Method

We propose a higher order mining approach for modelling, monitoring, and analyzing the DH substations' operational behavior and performance. The proposed approach uses a combination of different data analysis techniques such as sequential pattern mining, clustering analysis, consensus clustering, and the MST algorithm. Note that the last three data mining techniques are not applied on primary data, but on derived patterns and built models, i.e., they fall into the

HOM paradigm. The latter brings new potential and perspective for knowledge discovery by generating more human-understandable results and additionally facilitating the comparative analysis among substations. Thus, in the proposed approach the available data are initially partitioned across the time axis on a weekly basis, allowing for conventional mining within each week and for higher order mining over the patterns extracted from the weeks. This facilitates not only revealing similarities and interesting differences among the substation's weekly operational behaviors but also contributes to a more tractable process, e.g., modelling the substation performance for the whole heating season. Some of the above mentioned data analysis techniques have also been used in [27] where the authors proposed a method for identification of sequences of unexpected events in data streams. In this study, these techniques are reproduced, evaluated and generalized for solving different applied problems.

The main steps of the proposed method are as follows:

A. Data Preprocessing

In this step, we first remove all the duplicates and impute missing values. Missing values can occur due to connection problems of measuring instruments such as energy meters. There are different imputation methods such as mean substitution, hot-deck imputation [28], regression analysis, and multiple imputation [29]. In this study, we apply a simple approach for imputation of missing values, i.e., each missing value is replaced by averaging its neighbours. The first and the last missing values are replaced with the next and the previous available values, respectively. In our future work, we plan to study and compare different methods for imputation of missing values, e.g., such that consider the correlation structure of the data in order to select the imputation method that is best suited for DH substation related data.

Faults in measurement tools can appear as extreme values or sudden jumps in the measured data. We use a Hampel filter [30] which is a median absolute deviation (MAD) based estimation to detect and smooth out such extreme values. The filter computes the median, MAD, and the standard deviation (SD) over the data in a local window. In this study, the size of the window is considered to be seven, i.e., 3-neighbours on each side of a sample and the threshold for extreme value detection is set to be three. Therefore, in each window a sample with the distance three times the SD from its local median is considered as an extreme value and is replaced by the local median.

Since we are monitoring the operational behaviour of the substations based on outdoor temperature, 5 out of 10 features that have a strong negative correlation with the outdoor temperature are selected. These features are as follows: 1) Primary temperature difference ($\Delta T_{1^{st}}$), 2) Secondary temperature difference ($\Delta T_{2^{nd}}$), 3) Primary mass flow rate ($G_{1^{st}}$), 4) Primary heat ($Q_{1^{st}}$), and 5) Substation efficiency (E_s^T). The substation efficiency is calculated by considering features from both primary and secondary sides as follows:

$$E_s^T = \frac{\Delta T_{1^{st}}}{T_{s,1^{st}} - T_{r,2^{nd}}} \tag{2}$$

where, $\Delta T_{1^{st}}$ is the difference between primary supply and return temperatures, $T_{s,1^{st}}$ is the primary supply temperature, and $T_{r,2^{nd}}$ is the return temperature at the secondary side. Notice that the efficiency of a well-performed substation should be close to 1 in a normal setting. However, due to affect of the domestic hot water generation on the primary return temperature, the E_s^T can be higher than 1.

Table I shows all features included in the dataset. The features, 4-6, 9, and 10 in bold font are selected in this study due to their strong correlation with outdoor temperature. The selected features have a linear correlation with outdoor temperature, therefore, each one can be regressed on outdoor temperature. The created regression model for each feature gives us an approximation of that feature. We use this to compute the residuals $r = measured_{value} - predicted_{value}$. Note that the aim here is twofold. Firstly, to reduce the effect of outdoor temperature on the remaining features to detect deviating behaviours of the DH substations while the outdoor temperature is below 10 °C. Secondly, to de-trend the seasonality of the data.

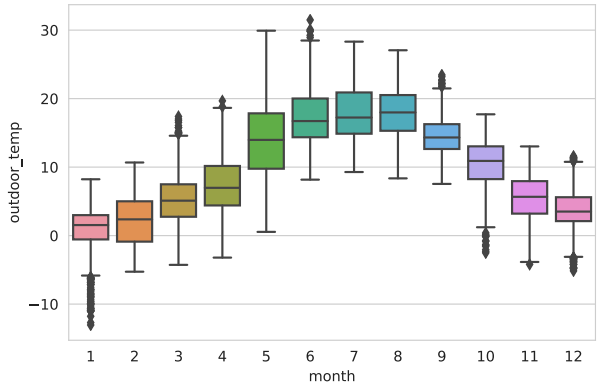
TABLE I: Features included in the dataset

No.	Feature	Notation	Unit/Format
1	T_o	Outdoor temperature	°C
2	$T_{s,1^{st}}$	Primary supply temperature	°C
3	$T_{r,1^{st}}$	Primary return temperature	°C
4	$\mathbf{\Delta T_{1^{st}}}$	**Primary temperature difference**	**°C**
5	$\mathbf{G_{1^{st}}}$	**Primary mass flow rate**	$\mathbf{m^3/h}$
6	$\mathbf{Q_{1^{st}}}$	**Primary heat**	**kW**
7	$T_{s,2^{nd}}$	Secondary supply temperature	°C
8	$T_{r,2^{nd}}$	Secondary return temperature	°C
9	$\mathbf{\Delta T_{2^{nd}}}$	**Secondary temperature difference**	**°C**
10	$\mathbf{E_s^T}$	**Substation efficiency**	**%**

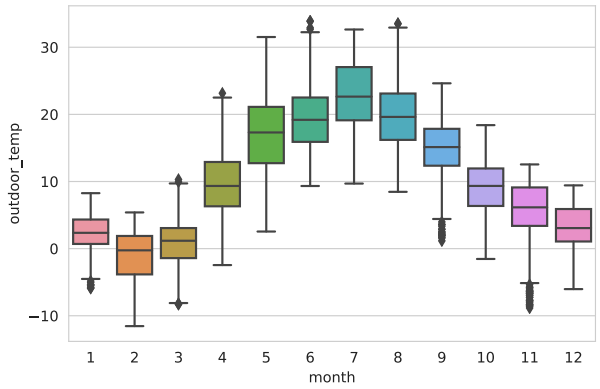
Note. Features in bold font are selected in this study due to their strong correlation with outdoor temperature.

In this study we only assess the substations' behaviour while space heating is needed. Figure 2 shows the yearly seasonality of outdoor temperature measured by building F. As one can see the average outdoor temperature in January - April and November - December is below 10 °C.

In order to prepare the data for the next step, we apply z-score normalization on each feature and for every 24-hour period: $z = \frac{x-\mu}{\sigma}$, where x is a feature's value, μ and σ are the mean and the standard deviation of the feature within the 24-hour period, respectively. We perform the normalization to scale the features to have a mean of zero and a standard deviation of one. This makes it possible to assess a DH substation's operational behaviours on a weekly basis and in comparison with other substations with the same heat load profile. In other words, the z-score normalization is relevant when the general shape of a feature, rather than its amplitude, is important.



(a) 2017



(b) 2018

Fig. 2: Yearly seasonality of outdoor temperature for building F in **a)** 2017 and **b)** 2018.

As it was mentioned earlier, the proposed approach partitions the available data across the time axis on a weekly basis in order to extract patterns within each week. Therefore, it is necessary to convert the continuous features to categorized or nominal features, i.e., *data discretization* must be conducted. This process can be performed in a supervised (class information is taken into account to find proper intervals) or an unsupervised fashion [31]. In this study, due to unavailability of labelled data, k-means-based discretization is used. Note that the size of k is set to be four, the same as the number of season periods in Sweden. As a result of the discretization process feature categories are defined as *low*, *low_medium*, *medium_high*, and *high*.

B. Data Segmentation And Pattern Extraction

The size of the time window (partition) for pattern extraction is important for further analysis. The proper partition length leads us to monitor operational behaviour of the substations rather than the residents' behaviour. Therefore, after performing some preliminary tests and having discussions with domain experts, the time window is set to be a week. The PrefixSpan algorithm is used to find frequent sequential patterns with the length of five in each week. Those sequential patterns that satisfy the user-specified support are considered as frequent ones. In this study, the user-specified support threshold is set to be 1, i.e., any patterns that appear at least once will be considered.

C. Data Analysis

The data analysis step can be further broken down into three sub-steps: a) clustering of the extracted patterns, b) assessing a substation's behaviour by comparing the clustering solutions produced for every two consecutive weeks, and c) conducting further analysis and evaluation of the observed behaviour by building a minimum spanning tree and detecting the potential outliers. Sub-steps b) and c) can facilitate the domain experts in further analysis and better understanding of the DH individual substations' behaviour and also in comparison among the substations belonging to the same heat load category.

a) *Clustering frequent sequential patterns:* At this step, the substation's weekly operational behaviour is modeled. This is preformed by clustering the extracted patterns based on their similarities into groups. The similarity between the patterns are calculated using LD.

b) *Assessing a substation's behaviours:* The similarity between substation behaviours is analyzed and assessed for every two consecutive weeks. This is done through pairwise comparison of the exemplars of the clustering solutions using equation 1. The assessed similarity can be used to measure the discrepancy between the substation performance within every two weeks period. When the discrepancy is significant (above a given threshold, e.g., more than 25%) and the weekly average temperature is below 10 °C, further analysis is preformed by integrating the produced clustering solutions into a *consensus clustering*. Moreover, the assessed similarities for the whole period can be used to build a signature profile of the substation's performance. In addition, such performance profiles can be applied for comparing the substations belonging to the same heat load category.

c) *Building a minimum spanning tree and detecting outliers:* The consensus clustering solution is used for building an MST, where the exemplars are tree nodes and the distances between them represent the tree edges. Notice that an MST is a tree with a minimum traversing cost. In order to identify unusual behaviours, the longest edge(s) of the MST is removed. Smallest and distant sub-trees created by the cut can be interpreted as outliers.

V. Experimental Design

A. Dataset

The data used in this study is provided by an energy company located in Southern Sweden. The dataset consists of hourly average measurements from 82 buildings equipped with the company's smart system. The collected data was obtained during February 2014 until December 2018. This means 43,800 instances per building (24 instances per day).

TABLE II: Detected deviations in different months during 2017.

No.	Building	January	February	March	April	May	June - September	October	November	December	Total count
1	C	107	18	162	45	–	–	–	17	–	349
2	E	74	42	143	53	–	–	28	99	83	522
3	F	–	33	51	60	–	–	80	87	159	470
4	L	114	42	186	193	63	–	126	111	62	897
5	**M-S**	35	166	60	–	–	–	80	33	42	416
6	**P-S**	40	23	99	323	–	–	57	120	40	702
7	S-1	21	100	28	34	–	–	–	122	66	371
8	S-2	144	59	–	–	–	–	37	45	31	316
9	**S-S**	112	87	9	9	–	–	51	118	218	604
10	**O-S**	139	26	77	–	–	–	22	85	25	374

Note. Buildings in bold font are schools. Highlighted numbers shows the least number of detected deviations per month and in total. '–' means the biweekly average outdoor temperature was above 10 °C, therefore, no further analysis has been done.

However, since most of the buildings have a high proportion of missing values and rows in the time span of 2014 to 2016, we have focused our analysis on data collected for 10 randomly selected buildings for the period covering the recent two years (2017 and 2018). The selected buildings are considered as representatives for the whole set of available buildings. We discuss and interpret the results obtained on their data for the rest of our study.

B. Implementation And Availability

The proposed approach is implemented in Python version 3.6. The Python implementations of the PrefixSpan algorithm and the edit distance are fetched from [14] and [20], respectively. The affinity propagation algorithm and the k-means-based discretization are adopted from the scikit-learn module [32]. For constructing and manipulating a minimum spanning tree the NetworkX package is used [33]. The NetworkX package uses Kruskal's algorithm for constructing the MST. The implemented code and the experimental results are available at GitHub[2].

VI. Results and Discussion

We have studied substations' operational behaviour of 10 buildings during a period of two years (2017 and 2018). For each building, we first model the substation's weekly operational behaviour. This is performed by grouping the extracted frequent patterns into clusters of similar patterns. In order to monitor the substation's performance, we analyze and assess the similarity between substation's behaviours for every two consecutive weeks. The assessed similarity is used to measure the discrepancy between the substation's performance within every two week period. When the discrepancy becomes more than 25% (a user-specified threshold) and if the biweekly average temperature is below 10 °C, further analysis is conducted by integrating the produced clustering solutions into a consensus clustering. The obtained consensus clustering solution is used for building an MST, where the exemplars are tree nodes and the distances between them represent the tree edges. In order to identify unusual behaviours, the longest edge(s) of the MST is removed. The smallest sub-trees created by the cut are interpreted as faults or deviations. Table II shows the total number of detected deviations per hour for each building. Notice that *four* out of the *ten* studied buildings are schools and the rest are residential buildings.

[2] https://github.com/shahrooz-abghari/HOM-DH-Monitoring

A. Building Substation Performance Signature Profile

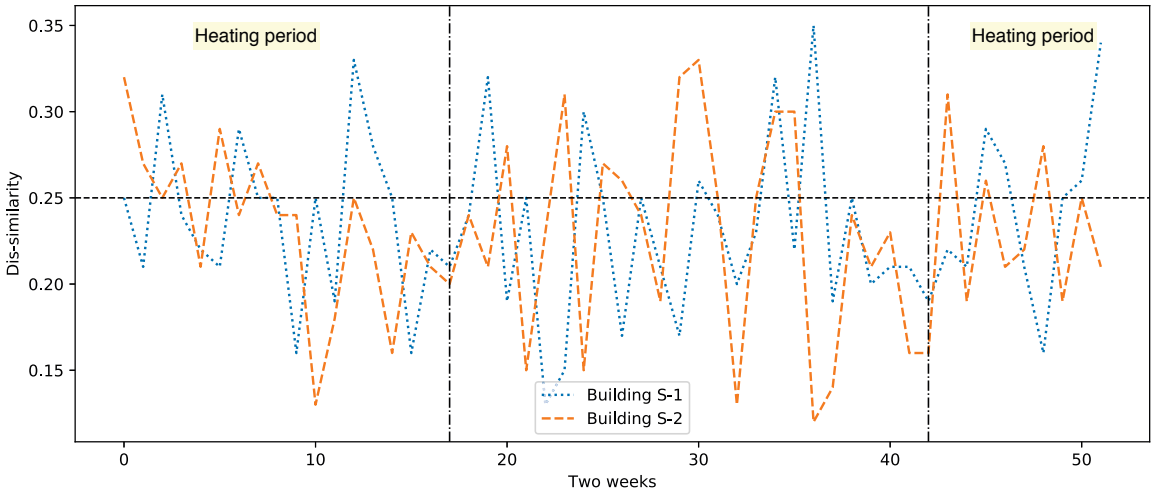
As we mentioned earlier in Section IV-C, the assessed similarities of a substation's operational behaviour can be used to build the substation performance signature profile for the entire studied period. Additionally, such profiles can be used for comparing the substations belonging to the same heat load category. Figure 3 (a) shows the signature profiles of two residential buildings, *S-1* and *S-2* in 2017. The two buildings are quite similar and located in the same block. Figure 3 (b) depicts the substations' performance signature profiles of four school buildings for the same year. In this plot, the substations' performance of buildings *S-S* and *O-S* among others have some similarities.

Although, the expectation was to observe similar performance signatures from the buildings that are in the same category, in most cases the substations show quite different behaviours. The main reasons can be related to the difference between average outdoor temperature within two weeks, social behaviour of people, special holidays, and/or faulty substations and equipment. It is also the case that buildings of same build behave differently mostly due to installation issues or unsuitable configurations. Nevertheless, this requires further analysis by domain experts.

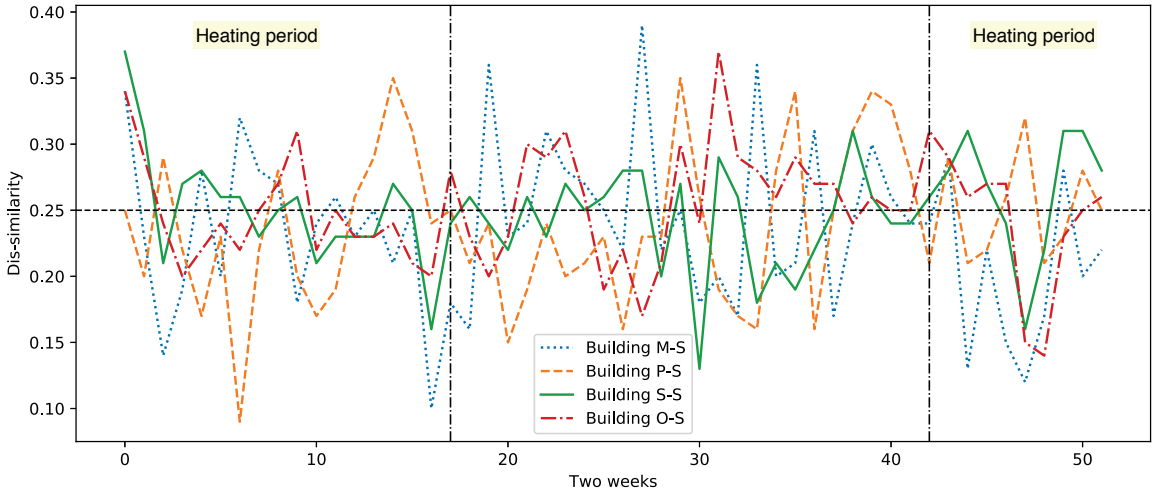
For the rest of this section we focus on building F and discuss the corresponding results generated on its data.

B. Modelling Substation Operational Behaviour

Weekly operational behaviour of a substation can be modeled by clustering the extracted patterns based on their similarities into groups. Using the AP algorithm each cluster can be recognized by its exemplar, a representative pattern of the whole group. Figure 4 (a) shows the substation's operational model for week 16, 2017. As one can see, the clustering solution contains 16 clusters which represents 168 (*24 hours* $\times$ *7 days*) different patterns. Each cluster models the substation's



(a) Residential buildings



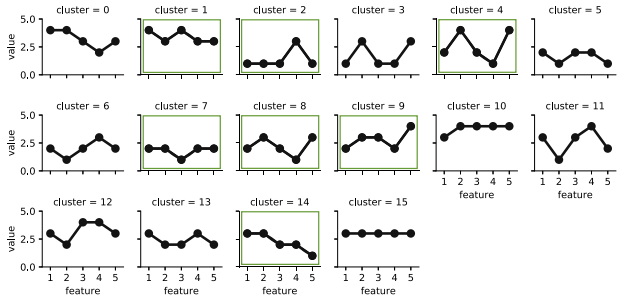
(b) School buildings

Fig. 3: Operational performance signature profiles of **a)** two residential buildings (substations) located in the same block and **b)** four schools. The signatures show the biweekly performance of the substations in 2017.

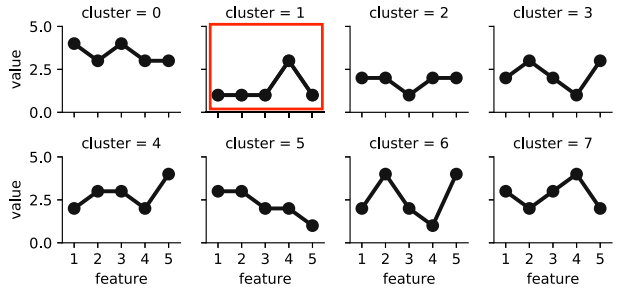
operational behaviour for some hours up to a couple of days, based on its frequency. The number of clusters in each clustering solution can be interpreted as different operational modes of the substation for the studied week. When the number of clusters is high this means that the substation has operated in more different modes. The latter can be related to the difference between outdoor temperature during days and nights, social behaviour of the tenants, special holidays, and/or faulty substations and equipment. The extracted patterns in this study contain five features. Each feature can have a value within a range *low*, *low_medium*, *medium_high*, and *high*, where *low* is represented by one and *high* by four, respectively. Notice that these values show the residuals between measured and predicted values by the created regression model for each feature.

We further analyze the operational behaviour models of weeks 16 and 17 by calculating the similarity between the exemplars of the corresponding clustering solutions. The calculated dissimilarity is above 25% and the average weekly outdoor temperature below 10 °C. Therefore, the proposed method integrates the clustering solutions into consensus clustering. Figure 4 (b) represents the substation's operational behaviour model for the studied two weeks. The model contains eight clusters and majority of the patterns, *seven* out of *eight* (the green framed clusters in Figure 4 (a)) are shared between the two weeks. Only cluster 7 belongs to week 17. In order to detect deviating behaviour, first an MST is built on top of the consensus clustering solution. Next, the longest edge(s) of the tree is removed. This transforms the MST into a forest. Sub-tree(s) with the smallest size can be marked as deviating

behaviour. As one can notice cluster 1 (framed in red in Figure 4 (b)) is detected as an outlier. This cluster appears in 11 days and in total 30 times out of 336 (*24 hours* $\times$ *14 days*). The data collected for these particular days can be further analyzed by domain experts to get better insight and understanding of the identified deviating behaviour.



(a) The operational behavior of building F's substation for week 16 of 2017. The substation's behaviour is modeled by 16 clusters. Each cluster is shown by its exemplar. Each feature can have a value within a range *low*, *low_medium*, *medium_high*, and *high*, where *low* is represented by one and *high* by four, respectively. The green framed clusters are the exemplars presented also in the consensus clustering solution (see Fig. 4 (b)).



(b) The consensus clustering integrating the clustering solutions for weeks 16 and 17 of 2017. The majority of exemplars (*seven* out of *eight*) are similar in both weeks while only one exemplar is chosen from week 17. After building an MST on top of the consensus clustering solution, cluster 1 (the red framed plot) is detected as deviating behaviour of the substation for week 16 and 17.

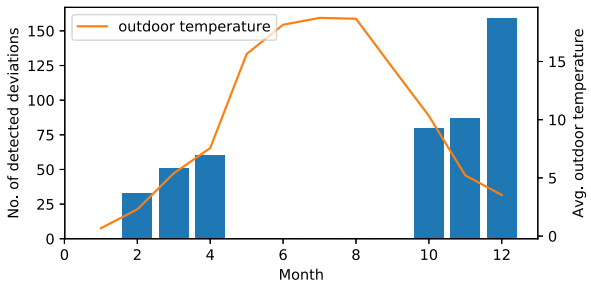
Fig. 4: The operational behaviour of building F's substation during weeks 16 and 17 of 2017.

C. Substation Performance

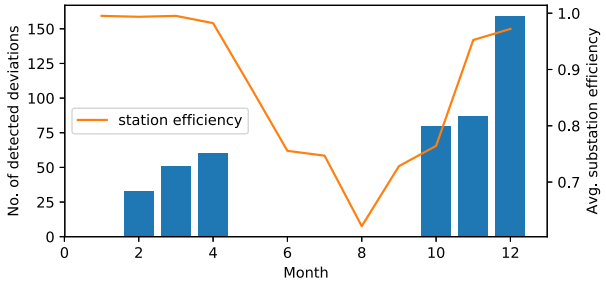
Substation efficiency, E_s^T, can be used as an indicator to assess a substation's operational behaviour throughout the entire year. Figure 5 depicts the detected deviations for building F's substation using its average efficiency and outdoor temperature for year 2017. Notice that in this study we only consider the smallest sub-tree(s) after cutting the longest edge(s) of an MST as outlier(s). Nevertheless, one can consider sorting the sub-trees based on their size from smallest to the largest for further analysis. Alternatively, one can define a threshold and cut those edges with a distance greater than the threshold.

The deviations that are marked in Figure 5 represent the less frequent patterns, which here are considered as abnormal behaviours. For instance, in weeks 16 and 17, cluster 1 is marked as deviating behaviour of the substation due to its proportion which is $\frac{30}{336} = 9\%$. Possibly, this can be interpreted that the following cluster is evolving, i.e., it might disappear or expand.

We can analyze and assess the substation's operational behaviours for the whole year by counting the number of detected deviations in each month. In Figure 6, one can see that the number of detected outliers in January to April compared to October to December is lower. In addition, substation efficiency on average is closer to 100% for the first four months. This means that when the average outdoor temperature is approximately 3 °C the substation performs better. During October to December, on the other hand the efficiency by decreasing the average temperature to below 10 °C gets closer to 98%. This might be partly related to the fact that the outdoor temperature during this period of time is frequently fluctuating between above and below 10 °C. However, since the number of detected deviations are doubled by December this can be related to some kind of fault in the substation.



(a) No. of detected deviations against average outdoor temperature



(b) No. of detected deviations against average substation efficiency

Fig. 6: Monthly number of detected deviations against **a)** average outdoor temperature, **b)** average substation efficiency for building F in 2017.

Figure 7 provides more detailed visualization showing detected deviations aggregated each day of the week per month for both average outdoor temperature and substation efficiency. As one can notice the substation efficiency shows a sudden drop in October. Further analysis of the data reveals that the

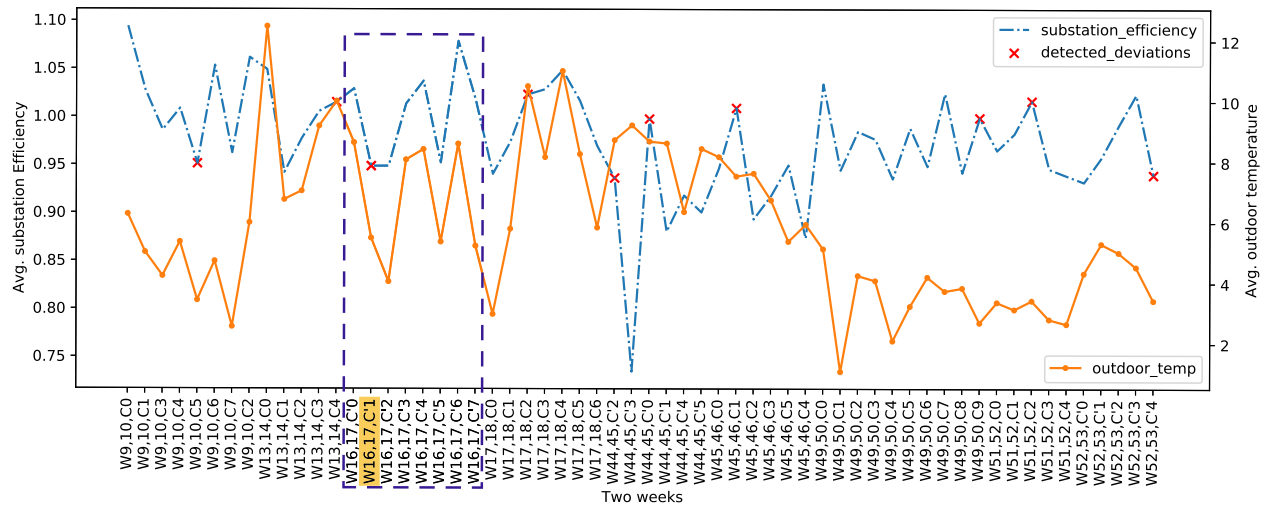
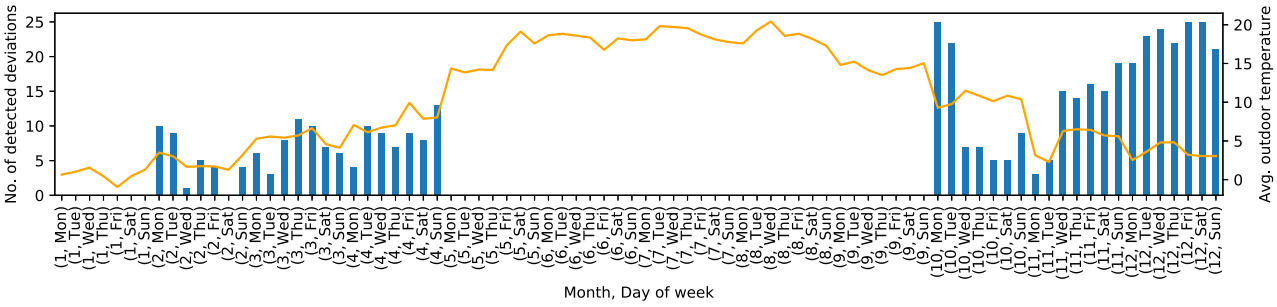
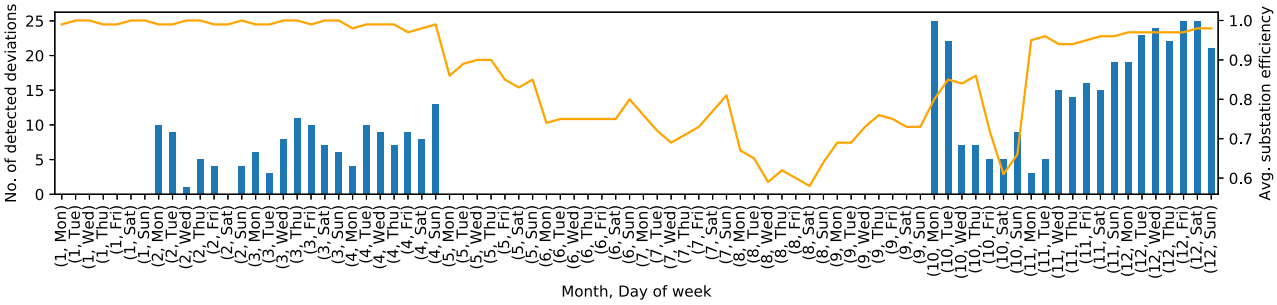


Fig. 5: Detected deviations for building F in 2017. Each deviation is detected by building an MST on top a consensus clustering solution for two weeks and cutting the longest edge(s) of the MST. The detected deviation are shown using average substation efficiency and outdoor temperature for every two weeks.



(a) No. of detected deviations against average outdoor temperature



(b) No. of detected deviations against average substation efficiency

Fig. 7: Number of detected deviations aggregated per weekday for each month against **a)** average outdoor temperature, **b)** average substation efficiency for building F in 2017.

substation was turned off from October 20th for almost four days.

VII. Conclusion And Future Work

We have proposed a higher order data mining approach for modelling, analyzing and monitoring the operational behaviour of DH substations. The proposed approach initially partitions the available data across the time axis on a weekly basis, allowing for conventional mining within each week and for higher order mining over the extracted patterns from the weeks. As demonstrated by the conducting experiments, this

facilitates not only revealing similarities and interesting differences among the substation's weekly operational behaviors, but also contributes to generating human-understandable and tractable results.

The approach has been applied to and evaluated on over two years of data for 10 buildings that are chosen at random. The results have shown that the method is robust in identifying and analyzing deviating and sub-optimal behaviours of the DH substations. In addition, the proposed approach provides different techniques for monitoring and data analysis, which can facilitate domain experts in the interpretation and better understanding of the DH substations' operational behaviour and performance.

For future work we aim to pursue further analysis and evaluation of the proposed approach on richer data sets coming from different sources (e.g, different types of buildings) by cooperating more closely with the domain experts.

In the long-term perspective, we are interested in deriving weekly patterns, discriminative in terms of the substation behavior, and further linking the derived patterns (a substation's operational modes) to performance indicators such as efficiency. In addition, we have the ambition to extend the proposed approach with means for root-cause analysis and diagnosis of detected deviations by considering secondary side data.

ACKNOWLEDGMENT

This work is part of the research project "*Scalable resource-efficient systems for big data analytics*" funded by the Knowledge Foundation (grant: 20140032) in Sweden.

REFERENCES

[1] S. Frederiksen and S. Werner, *District heating and cooling*. Studentlitteratur Lund, 2013, vol. 579.

[2] S. Månsson, P.-O. J. Kallioniemi, K. Sernhed, and M. Thern, "A machine learning approach to fault detection in district heating substations," *Energy Procedia*, vol. 149, pp. 226–235, 2018.

[3] H. Gadd and S. Werner, "Fault detection in district heating substations," *Applied Energy*, vol. 157, pp. 51–59, 2015.

[4] J. F. Roddick, M. Spiliopoulou, D. Lister, and A. Ceglar, "Higher order mining," *ACM SIGKDD Explorations Newsletter*, vol. 10, no. 1, pp. 5–17, 2008.

[5] R. Isermann, *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Science & Business Media, 2006.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, p. 15, 2009.

[7] S. Katipamula and M. R. Brambley, "Methods for fault detection, diagnostics, and prognostics for building systems-A review, part I," *Hvac&R Research*, vol. 11, no. 1, pp. 3–25, 2005.

[8] ——, "Methods for fault detection, diagnostics, and prognostics for building systems-A review, part II," *Hvac&R Research*, vol. 11, no. 2, pp. 169–187, 2005.

[9] D. Djenouri, R. Laidi, Y. Djenouri, and I. Balasingham, "Machine learning for smart building applications: Review and taxonomy," *ACM Computing Surveys*, vol. 52, no. 2, p. 24, 2019.

[10] P. Xue, Z. Zhou, X. Fang, X. Chen, L. Liu, Y. Liu, and J. Liu, "Fault detection and operation optimization in district heating substations based on data mining techniques," *Applied Energy*, vol. 205, pp. 926–940, 2017.

[11] E. Calikus, S. Nowaczyk, A. Sant'Anna, H. Gadd, and S. Werner, "A data-driven approach for discovery of heat load patterns in district heating," *arXiv preprint arXiv:1901.04863*, 2019.

[12] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 2015, pp. 1855–1870.

[13] F. Sandin, J. Gustafsson, and J. Delsing, *Fault detection with hourly district energy data: Probabilistic methods and heuristics for automated detection and ranking of anomalies*. Svensk Fjärrvärme, 2013.

[14] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth," in *Proc. of the 17th Int'l Conf. on Data Engineering*, 2001, pp. 215–224.

[15] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[16] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proc. of the Fifth Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[17] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Transaction of Knowledge Discovery Data*, vol. 1, no. 1, 2007.

[18] V. Boeva, E. Tsiporkova, and E. Kostadinova, *Analysis of Multiple DNA Microarray Datasets*. Springer Berlin Heidelberg, 2014, pp. 223–234.

[19] A. Goder and V. Filkov, "Consensus clustering algorithms: Comparison and refinement," in *ALENEX*, 2008, pp. 109–234.

[20] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[21] M.-F. Jiang, S.-S. Tseng, and C.-M. Su, "Two-phase clustering process for outliers detection," *Pattern Recognition Letters*, vol. 22, no. 6, pp. 691–700, 2001.

[22] X. Wang, X. L. Wang, and D. M. Wilkes, "A minimum spanning tree-inspired clustering-based outlier detection technique," in *Ind. Conf. on Data Mining*. Springer, 2012, pp. 209–223.

[23] G.-W. Wang, C.-X. Zhang, and J. Zhuang, "Clustering with prims sequential representation of minimum spanning tree," *Applied Mathematics and Computation*, vol. 247, pp. 521–534, 2014.

[24] A. C. Müller, S. Nowozin, and C. H. Lampert, "Information theoretic clustering using minimum spanning trees," in *Joint DAGM (German Association for Pattern Recognition) and OAGM Symp.* Springer, 2012, pp. 205–215.

[25] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *ACM Sigmod Record*, vol. 30, no. 2. ACM, 2001, pp. 37–46.

[26] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[27] S. Abghari, V. Boeva, N. Lavesson, H. Grahn, S. Ickin, and J. Gustafsson, "A minimum spanning tree clustering approach for outlier detection in event sequences," in *17th IEEE Int'l Conf. on Machine Learning and Applications*, 2018, pp. 1123–1130.

[28] B. L. Ford, "An overview of hot-deck procedures," *Incomplete data in sample surveys*, vol. 2, no. Part IV, pp. 185–207, 1983.

[29] D. B. Rubin, "Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse," in *Proceedings of the survey research methods section of the American Statistical Association*, vol. 1. American Statistical Association, 1978, pp. 20–34.

[30] F. R. Hampel, "A general qualitative definition of robustness," *The Annals of Mathematical Statistics*, pp. 1887–1896, 1971.

[31] S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," *GESTS Int'l Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 47–58, 2006.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2008.