Energy and Buildings 109 (2015) 75–89

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Temporal knowledge discovery in big BAS data for building energy management

Cheng Fan[a], Fu Xiao[a,*], Henrik Madsen[b], Dan Wang[c]

[a] Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong
[b] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[c] Department of Computing, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Received 6 August 2015
Received in revised form
16 September 2015
Accepted 23 September 2015
Available online 26 September 2015

Keywords:
Temporal knowledge discovery
Time series data mining
Big data
Building automation system
Building energy management

ABSTRACT

With the advances of information technologies, today's building automation systems (BASs) are capable of managing building operational performance in an efficient and convenient way. Meanwhile, the amount of real-time monitoring and control data in BASs grows continually in the building lifecycle, which stimulates an intense demand for powerful big data analysis tools in BASs. Existing big data analytics adopted in the building automation industry focus on mining cross-sectional relationships, whereas the temporal relationships, i.e., the relationships over time, are usually overlooked. However, building operations are typically dynamic and BAS data are essentially multivariate time series data. This paper presents a time series data mining methodology for temporal knowledge discovery in big BAS data. A number of time series data mining techniques are explored and carefully assembled, including the Symbolic Aggregate approXimation (SAX), motif discovery, and temporal association rule mining. This study also develops two methods for the efficient post-processing of knowledge discovered. The methodology has been applied to analyze the BAS data retrieved from a real building. The temporal knowledge discovered is valuable to identify dynamics, patterns and anomalies in building operations, derive temporal association rules within and between subsystems, assess building system performance and spot opportunities in energy conservation.

1. Introduction

The building sector is evolving to be the greatest energy consumer around the world, accounting for 40% of the global energy use and one third of the global greenhouse gas emissions [1,2]. As a result, building energy efficiency has become one of the top concerns of a sustainable society and attracted increasing research and development efforts in recent years. Thanks to the advances of information, computing and control technologies, building automation system (BAS) provides a valuable network-based digital platform for automatically managing complex building systems, including heating, air conditioning, ventilation, lighting, vertical transportation, fire safety and security systems. It is estimated that the potential energy savings from the adoption of advanced building automation technologies might reach 22% by 2028 for the European building sector [3]. Besides fulfilling the online monitoring and control functions, BASs record thousands of real-time measurements and control signals, and the amount of data keeps growing in the building life-cycle. Most of the existing building management strategies are developed based on domain expertise or small subsets of the BAS data. There are increasing interests in systematically analyzing the big BAS data and discover knowledge for improving building performance.

Data mining (DM) is a promising technology, which can effectively discover interesting and potentially useful knowledge from big data. Some efforts have been made to investigate the potentials of DM in the building field. DM techniques have been adopted at the building design, construction, and operation stages [4]. The building operation stage draws particular attention, as it accounts for 80–90% of the total building green gas emission and is directly linked to occupant comforts and the realization of building functionality [5]. In general, DM techniques can discover two types of knowledge, i.e., predictive and descriptive. Predictive DM is often used to capture the complex and nonlinear relationships between inputs and outputs. It has been applied at the building operation stage to the prediction of building energy consumption [6–8], thermal load [9,10], indoor environment [11,12], and system performance indices [13–15]. Descriptive DM is used to discover

[*] Corresponding author.
E-mail address: linda.xiao@polyu.edu.hk (F. Xiao).

http://dx.doi.org/10.1016/j.enbuild.2015.09.060
0378-7788/© 2015 Elsevier B.V. All rights reserved.

the associations, correlations, and intrinsic data structure in big data. Compared to predictive DM, descriptive DM is more flexible in applications, as it does not involve a training process and the knowledge discovery process is not guided by pre-defined targets. Descriptive DM has been mainly applied at the building operation stage to fault detection and diagnostics [4,16–18]. Popular techniques include association rule mining, clustering analysis, and anomaly detection.

Despite of the encouraging research outcomes, previous studies of analyzing big BAS data usually considered the BAS data as cross-sectional data, and hence the knowledge discovered mainly includes the concurrent relationships among different variables, such as relationships between the power consumptions of the primary air units and lifts in a building [18]. BAS data are usually recorded in a two-dimensional matrix, with each column representing a measured variable (such as temperature, flow rate, power and control signal), and each row representing an observation/sample at a specific time instant. Each observation is a vector of many measurements and control signals. The time interval between two consecutive observations is usually fixed and may vary from several seconds to tens of minutes. The first two columns usually store the date and the time. Considering that BAS data are in essence multivariate time series data, the cross-sectional knowledge discovered may not be able to fully capture the relationships over time. Building operations are typically dynamic due to the changes in indoor and outdoor operating conditions, such as the outdoor climate conditions, indoor occupant number and utilization of indoor electric appliances. Meanwhile, the changes hardly occur simultaneously which results that the dynamics in building operations are very complicated. For instance, the indoor temperature is influenced by the outdoor air temperature. However, when the infiltration is not significant, these two temperatures rarely change simultaneously due to building thermal mass. Time lags between them often bring challenges to the sequence control of chiller plants. The dynamics are usually complicated and have great influences on control performance, interactions among building components and integrations between buildings and communities (e.g., electricity power grid) [19]. In practice, it is desired to discover such temporal knowledge hidden in BAS data. Advanced tools and methods for temporal knowledge discovery should be developed for this purpose.

Conventional time series analytics, such as the autoregressive moving average models (ARMA), are mainly used for solving predictive tasks in the field of building management, including the prediction of building electricity consumption [20,21], building thermal load [22,23] and indoor environment [24,25]. In recent years, various approaches have been developed to mine temporal knowledge in different formats, such as events, clusters, motifs and temporal association rules [26–30]. However, only limited studies have been performed to explore their potential in analyzing BAS data. The complex event processing (CEP), which is a method well suited for the processing of information flows, has been adopted to utilize time series data in building operations [26–28]. Renners et al. applied CEP to correlate the sensor data and provide real-time reactions to building management [26]. Wen et al. developed a CEP-based method to derive knowledge from building energy data and applied it for building controls [27]. In these studies, domain expertise plays an important role in defining events and rules. Time series data mining enables an approach to discover interesting and previously unknown temporal knowledge. Patnaik et al. adopted the motif discovery technique to mine chiller operation data in data centers [31]. Motifs (i.e., frequent sequential patterns) were successfully discovered to identify energy-efficient operating patterns. Miller, Nagy and Schlueter used a similar method to analyze building energy consumption data [32]. Energy consumption motifs were extracted for building performance characterization. Discords, or infrequent sequential patterns, were identified and used for fault detection. Their work demonstrated the encouraging potentials of time series data mining in the knowledge discovery of BAS data for managing building operations. Currently, the potential and applicability of various time series data mining techniques in mining big BAS data are still uncertain considering unique characteristics of BAS data, such as low quality, nonlinearity, multiple scales or units, and multicollinearity. A generic and systematic methodology for discovering temporal knowledge in big BAS data is needed for developing applicable tools in BAS.

This study proposes a generic methodology for mining temporal knowledge hidden in big BAS data and demonstrates its applications in real cases. We first briefly introduce time series data mining techniques, and specifically highlight the differences between cross-sectional data mining and time series data mining. Then, the generic methodology is presented, which consists of data preprocessing, data partitioning, temporal knowledge discovery, and post-mining. Two methods are developed to improve the efficiency in post-mining. The methodology has been applied to analyze the BAS data retrieved from the tallest building in Hong Kong. Valuable temporal knowledge has been discovered for building operations and performance management.

2. Description of research methodology

Based on a comprehensive exploration of advanced DM techniques, in-depth analysis of BAS data characteristics as well as specific considerations for practical applications, we have developed a generic framework for knowledge discovery in BAS data [4]. The framework consists of four major phases, i.e., data preprocessing, data partitioning, knowledge discovery and post-mining. Each phase was specifically designed considering the BAS data quality and structure, data format requirement of DM techniques, interpretation and selection of knowledge discovered, and application of the knowledge to building performance assessment, diagnosis and optimization. It is a generic framework proposed for analyzing big BAS data using DM techniques. The methodology presented in this study is also developed within this framework, as shown in Fig. 1, and further enriches the framework by integrating time series data mining techniques for temporal knowledge discovery. Three tasks are performed in the first phase, including data cleaning, period estimation and data transformation. Phase 2 adopts the evidence accumulation clustering to partition the SAX subsequences. Phase 3 adopts two techniques, i.e., motif discovery and temporal association rule mining, to discover two different types of knowledge. Two post-mining methods are developed in Phase 4 to improve the efficiency and effectiveness of handling the large amount of knowledge discovered in Phase 3. The details of each phase are introduced in the following subsections.

2.1. Data preprocessing

Data preprocessing fulfills three tasks, i.e., data cleaning, period estimation, and data transformation, with the aims to enhance the data quality, explore the intrinsic characteristics in BAS time series data, and prepare the raw data with suitable format for data mining.

2.1.1. Data cleaning

Data cleaning aims to improve BAS data quality by filling missing values and detecting outliers in raw BAS time series data. Missing values widely exist in BAS data due to sensor malfunctions or signal transmission problems in BAS network. Popular methods to impute missing data include moving window, random imputation, and inference-based methods [33]. Moving window-based methods are easy to implement and have a fairly good performance when the duration of missing values is not very long. Otherwise,

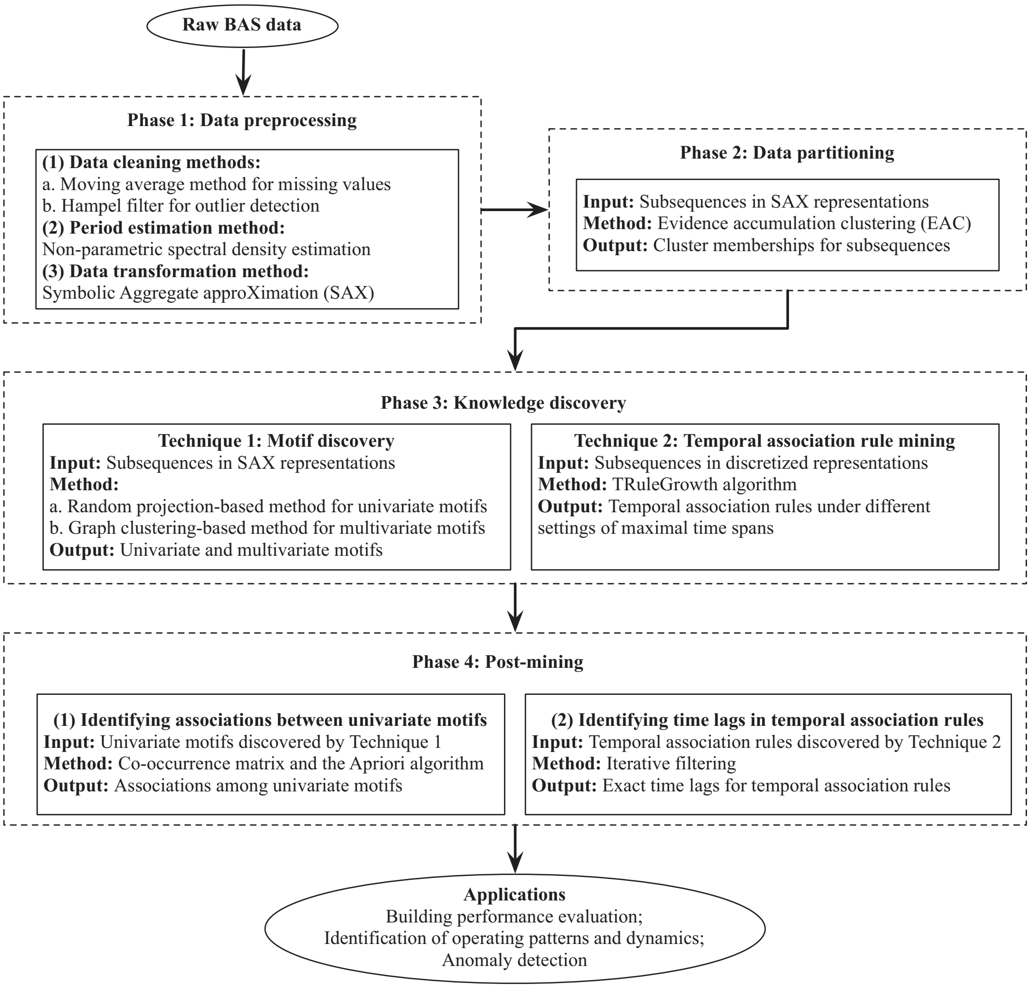

Raw BAS data
Phase 1: Data preprocessing
(1) Data cleaning methods:
a. Moving average method for missing values
b. Hampel filter for outlier detection
(2) Period estimation method:
Non-parametric spectral density estimation
(3) Data transformation method:
Symbolic Aggregate approXimation (SAX)
Phase 2: Data partitioning
Input: Subsequences in SAX representations
Method: Evidence accumulation clustering (EAC)
Output: Cluster memberships for subsequences
Phase 3: Knowledge discovery
Technique 1: Motif discovery
Input: Subsequences in SAX representations
Method:
a. Random projection-based method for univariate motifs
b. Graph clustering-based method for multivariate motifs
Output: Univariate and multivariate motifs
Technique 2: Temporal association rule mining
Input: Subsequences in discretized representations
Method: TRuleGrowth algorithm
Output: Temporal association rules under different settings of maximal time spans
Phase 4: Post-mining
(1) Identifying associations between univariate motifs
Input: Univariate motifs discovered by Technique 1
Method: Co-occurrence matrix and the Apriori algorithm
Output: Associations among univariate motifs
(2) Identifying time lags in temporal association rules
Input: Temporal association rules discovered by Technique 2
Method: Iterative filtering
Output: Exact time lags for temporal association rules
Applications
Building performance evaluation;
Identification of operating patterns and dynamics;
Anomaly detection

Fig. 1. Outline of the research methodology.

it is recommended to use the inference-based methods or simply exclude the observations with long missing values from analysis. Outliers in a time series are observations that are highly unlikely to occur based on the variation seen in the rest of the time series. They can be classified into two types, i.e., points as outliers and subsequence as outliers [34]. It is recommended that the data preprocessing phase only handles the first type of outliers in the raw time series data, as the identification of the second type of outliers may overlap with mining discords (i.e., infrequent sequential patterns) in the later process. The outlier detection methods can be grouped into three categories, i.e., prediction-based, profile-based, and deviant-based methods [34]. The prediction-based methods detect outliers by comparing the actual measurements with their expected or predicted values from statistical analysis or machine learning algorithms. The profile-based methods use historical data to construct a normal profile, which is usually presented in the form of expected means and confidence intervals at different time. Each observation is compared with the normal profile to decide whether it is an outlier or not. The deviant-based methods identify outliers from a perspective of information theory. An observation is an outlier if removing it from the time series leads to a much more succinct representation of the original time series [34].

In this study, the moving window-based method is used to impute the missing values with a short duration, i.e., less than 2-hour. Any missing values with a longer time duration are excluded from analysis. This study adopts the Hampel filter to identify outliers. It is a nonlinear filter which shows high effectiveness in processing time series data [35]. For each observation, the Hampel filter calculates the median and the median absolute deviation (MAD) considering a moving window size of $2k+1$. k is the number of observations before and after the observation concerned. A parameter θ, which usually ranges from 0 to 5, is predefined to generate thresholds for outlierness evaluation, i.e., Median $\pm \theta \times$ MAD. Any observation falls beyond the range of the thresholds is identified as an outlier and is replaced by the median. The smaller the parameter, the more aggressive the detection algorithm is and more observations will be identified as outliers. This study sets θ as 3, which is in accordance with the Ron Pearson's 3-sigma rule.

2.1.2. *Period estimation*

This step is specifically developed for time series data, considering that long time series data usually exhibit periodicity, and consequently motifs and association rules periodically repeat. Finding the period in the time series data and then segment those data into short subsequences can considerably reduce the mining load. It is a common practice in time series data mining, particularly in handling very long time series data like BAS data. The repeating daily working schedule of building users (e.g. office hours and non-office hours) results that the operating schedules of major systems and equipment (such as air conditioning, lighting and lift systems) usually repeat daily. Obviously, the BAS data exhibit daily periodicity. In view of this, Miller et al. segmented the time series of building energy consumption data into daily sequences in their study [32]. This study attempts to adopt a data-driven approach to estimating the intrinsic periods embedded in BAS data. There are two purposes for doing this: firstly to minimize the dependence on domain knowledge in the knowledge discovery process; secondly to maximize the possibility of discovering new knowledge, or new periods in BAS data in our case. Periods in time series data can be detected using the spectral density estimation methods, which can be either parametric or non-parametric. The parametric methods first model the time series using time series modeling techniques, such as autoregressive and moving average (ARMA). The spectral density is then estimated based on the model parameters. By contrast, the non-parametric methods estimate the spectral density by taking the Fourier transformation of the autocorrelation function. Considering that the building data usually present diurnal, weekly and annual seasonality, the resulting parametric models can be very complex. Therefore, this study applies the non-parametric method to period estimation.

2.1.3. Data transformation

Data transformation prepares the time series data with suitable formats to meet the following two needs. Firstly, different mining techniques require different data formats (e.g., numerical or categorical) and BAS data exhibit diversity in units, scales, and data types. Secondly, the computation load is a big concern due to the huge volume of big data, which can be alleviated by effectively reducing the volume of the data without losing valuable information embedded in the data. In this study, the symbolic approximation aggregate (SAX) method is proposed to transform the original time series BAS data into meaningful symbols [32,36]. The SAX method transforms a numeric time series into a symbol stream and the length of the symbol stream is much shorter than the original time series. It can therefore reduce the data size.

To perform SAX, a univariate time series of length *n* is firstly standardized to have a zero mean and a standard deviation of 1 and then segmented into *m* subsequences with a window size of *q*. One of the typical methods to segment the time series is based on the period detected in the previous step. For example, if the period estimated is 24 h, one day BAS data will form one subsequence. Two parameters need to be defined to perform SAX, i.e., the word size *W* and the alphabet size *A*. A set of breakpoints (e.g., $\beta_1, \beta_2, \ldots, \beta_{A-1}$) are determined in such a manner that the area under the *N*(0,1) Gaussian curve from β_i to β_{i+1} is 1/*A*. Each interval will be assigned with an alphabet (e.g., *a*, *b*, and *c*) and the number of alphabets used is the alphabet size, *A*. Given the word size (*W*), each subsequence in the window size of *q* can be divided into *W* equal sections, and the means of each sections are calculated. According to which interval (i.e., β_i to β_{i+1}) the mean lies within, the corresponding alphabet is assigned to the section. In this way, each subsequence can be represented by a SAX word which consists of *W* alphabets. For example, *abca*, *aabc*, *bcca* are SAX words given *W* = *4* and *A* = *3*. In these SAX words, the alphabet size (*A*) is 3, so three alphabets (i.e., *a*, *b* and *c*) are used; the word size (*W*) is 4, so each SAX word consists of four alphabets. The original time series is transformed into a string of alphabets. The larger the alphabet size (*A*) and the word size (*W*), the more detailed information retained in the symbolic stream. However, the reduction of computation load becomes less. Therefore, there is a trade-off, which will be discussed in the later case studies.

The distance between two SAX representations are calculated as $\sqrt{q/w \times \sqrt{\sum_{i=1}^{w} dist(S_i, B_i)^2}}$, where *S* and *B* are two SAX representations, and *dist()* is the distance function for SAX symbols. Table 1 presents an example of distance matrix between symbols considering an alphabet size of 4. The value in *cell*(*x*,*y*) is calculated using Eq. (1). A dissimilarity matrix considering different SAX representations can be computed accordingly. More details can be found in [36].

Table 1
An example distance matrix for SAX symbols.

Distance	*a*	*b*	*c*	*d*
a	0	0	0.67	1.34
b	0	0	0	0.67
c	0.67	0	0	0
d	1.34	0.67	0	0

Besides SAX, difference-based and dictionary-based methods are also capable of transforming time series into symbols [37–39]. The difference-based method transformed the raw time series into symbols based on their first- or higher-order differences. It can be used when the changes between successive time steps are more important than the absolute values [39]. The dictionary-based methods transform the time series into symbols by matching the raw data with predefined patterns in a dictionary. For instance, in the studies performed by Kwac et al. [37] and Gulbinas et al. [38], clustering analysis was applied to generate the representative patterns of daily power consumption, based on which a dictionary was built for symbolization. SAX is selected in this study considering the following two aspects. Firstly, SAX is straightforward to use, as it requires little domain expertise and preprocessing. Secondly, SAX contains an intrinsic distance measure, which provides extra value in the subsequent knowledge discovery [36], as shown in the later part.

$$cell(x, y) = \begin{cases} 0 & \text{if } |x - y| \le 1 \\ \beta_{\max(x,y)-1} & \text{otherwise} \end{cases} \qquad (1)$$

2.2. Data partitioning

Due to the changing operating conditions and complicated system dynamics and interactions, the big BAS data usually scatter in a high-dimensional space. To enhance the reliability and sensitivity of the mining results, data partitioning is carried out to divide the data into several groups or clusters, with the aim of maximizing the intra-group similarities while minimizing the inter-group similarities. Knowledge discovery are then performed on each group separately. Clustering analysis is a suitable DM technique to perform this task. Despite of the large number of clustering algorithms being available, no single algorithm is able to identify all kinds of cluster shapes and data structures in practice [40]. It is usually very difficult to find out the optimal clustering algorithm and the settings of its parameters. Some methods have been developed to facilitate the decision-makings, based on either internal (e.g., Dunn index and Davies-Bouldin index) or external validation indices (e.g., purity and mutual information). However, no validation method can impartially evaluate the results of any clustering algorithm [41]. A common practice is to try out a large number of algorithms with different parameters in order to obtain desired the clustering results. The process can be computationally expensive and time-consuming.

Ensemble learning is capable of enhancing the clustering performance by combining a number of base learners, whose individual performance may be poor [40,41]. The evidence accumulation clustering (EAC) is a method designed to apply ensemble learning on clustering analysis [34]. One advantage of the EAC over other conventional clustering methods is that it has the ability to discover clusters with various sizes and shapes. In addition, the method can automatically determine the optimal cluster number, which provides great flexibility in analyzing data with unknown characteristics. The partition around medoids (PAM) is selected as the base algorithm for EAC. PAM shares a similar partitioning mechanism as the popular *k*-means algorithm. Compared to the *k*-means, PAM is more robust to outliers and noises and can take a dissimilarity matrix as inputs. Therefore, PAM is more compatible with time series data in SAX representations.

Three parameters needs to be defined to perform EAC, i.e., the total iteration number E, the lower and upper limits of the cluster number K_{lower} and K_{upper}. E sets of clustering results are generated by PAM with different cluster numbers (i.e., randomly sampled from K_{lower} to K_{upper} in each iteration) and the dimension of input data. These E sets of clustering results are then transformed into a co-occurrence matrix. Assuming that the data contains n observations, the co-occurrence matrix C has a dimension of $n \times n$. The value of $C_{i,j}$ is the number of times when observations i and j are grouped in the same cluster divided by the total iteration number E. The final clustering result is obtained by using hierarchical agglomerative method to cluster the co-occurrence matrix. More details can be found in [40].

2.3. Temporal knowledge discovery

After the data are preprocessed and partitioned, appropriate DM techniques will be applied for knowledge discovery. The typical descriptive knowledge types in time series data include motifs, discords and temporal association rules [29].

2.3.1. Motif discovery

Motif, or frequent sequential pattern, is a typical knowledge type which can be discovered in time series data. Motifs are valuable to temporal association rule mining, discord (i.e. infrequent sequential pattern) detection, and time series classification [42].

Motif discovery has been mainly applied to analyze univariate time series in previous studies. Conventional motif discovery methods are based on exhaustive search, which results that the computational costs increase dramatically for long time series and is therefore not applicable to big data. In view of this, a more efficient algorithm, which is based on random projection and compatible with SAX representations [42], is selected to discover univariate motifs. Assuming that the time series has a length of n and the sliding window size is q, a matrix containing all the subsequences (denoted as M_1) can be constructed and has a dimension of $(n-q+1)\times q$. Each subsequence is transformed into a SAX representation. Assuming the word size is W, the new matrix containing the SAX representations (denoted as M_2) has a dimension of $(n - q + 1) \times w$. Random projection is performed by randomly picking s columns from M_2, where s ranges from 1 to $W-1$. A collision matrix, which has a dimension of $(n-q+1)\times(n-q+1)$, is constructed to record the times of being identical for two subsequences after a number of random projections. A tentative univariate motif is identified if the two subsequences result in a high value in the collision matrix. Potential members of this tentative univariate motif can then be identified by calculating the Euclidean distance in the original numeric representations.

Several methods have been developed to identify motifs in multivariate time series data, such as PCA-based and density estimation-based methods [43,44]. Those methods can successfully identify synchronous multivariate motifs. However, their practical value in analyzing real-world data is limited, as the motifs in multivariate time series data do not necessarily start at the same time and their duration may vary as well. We can see a lot of such examples in building operations. For example, when the air conditioner or chiller is turned on, the indoor temperature will not change immediately due to the thermal mass. The sudden increase of the lift power consumption in the morning peak hour does not correspond to a large increase in the chiller power consumption due to the pre-cooling strategy. In this study, multivariate motif discovery algorithm proposed in [45] is adopted. The main advantage is that, firstly, both synchronous and non-synchronous multivariate motifs can be discovered, and secondly, the multivariate motifs identified may consist of all univariate motifs or any subset of the univariate motifs. The method first performs univariate motif discovery on the time series of each variable. A graph clustering approach is then applied to identify multivariate motifs. A directed coincidence graph G is constructed. Each motif r_i is represented by a vertex v_i. $e_{i,j}$ represents the edge connecting the vertex v_i and v_j. The weight of $e_{i,j}$ is denoted as $w_{i,j}$ and calculated as $coincident(r_i, r_j)/size_i$, where $coincident(r_i, r_j)$ is the total number of times that a temporal overlap is found between r_i and r_j and the $size_i$ is the number of occurrence of r_i. A parameter, α, ranging from 0 to 1, is user-specified as the minimum correlation between univariate motifs based on which a multivariate motif could be constructed.

2.3.2. Temporal association rule mining

The difference between association rule mining (ARM) and temporal association rule mining (TARM) lies in whether the temporal information is contained in the rule or not. ARM was mainly used to discover cross-sectional associations, where the temporal information is neglected. The typical format of ARM is $A \to B$, where $A \cap B = \emptyset$. It states that if A happens, B will also happen. An association rule is derived if both the rule support and confidence exceed the user-defined thresholds. The support of a rule is the fraction between the number of times when both the antecedent and consequent take place and the total number of records. The confidence of a rule is the conditional probability of the consequent given the antecedent. The interestingness of the association rules can be evaluated using the *lift*, which is the ratio between the rule confidence and the support of consequent. It measures the dependency and correlation between the antecedent and the consequent of a rule. Potentially useful rules usually have a *lift* larger than 1, indicating that the occurrence of the antecedent positively influences the occurrence of consequent.

Temporal association rule mining (TARM) is of particular interest in mining BAS data because of the complicated dynamics in building operations. TARM, or sequential rule mining, discovers associations among variables while providing an insight into the temporal dependency. The general format of temporal association rules is also $A \to B$, where $A \cap B = \emptyset$. However, the temporal dependency is contained, indicating that B will take place after A. Various algorithms have been developed for deriving temporal association rules, such as the SPADE and CMRules [46,47]. In engineering practice, temporal rules that are valid within a limited time span are of special interest. The format of such temporal rules is $A \xrightarrow{t} B$, which means that B will occur within t time units after the occurrence of A. Therefore, the TRuleGrowth algorithm, which can derive temporal association rules under the constraint of maximum time span [48], is selected in this study. To perform this algorithm, three parameters need to be defined, i.e., the minimum support, minimum confidence, and the maximum time span. The other advantage of the TRuleGrowth algorithm is that it can greatly reduce the number of rules generated by controlling the maximum time span. Consequently, the post-mining phase consumes much less time.

2.4. Post-mining

The post-mining phase aims to build a bridge between knowledge discovered in Phase 3 and practical applications, such as building performance assessment, fault diagnosis and optimization. It usually needs domain expertise to select, interpret and apply the knowledge discovered [4,18,32]. The process can be very time-consuming, due to the large amount of knowledge discovered and the diversity of knowledge representations (e.g., rules, clusters, decision trees). Application of the motifs and temporal association

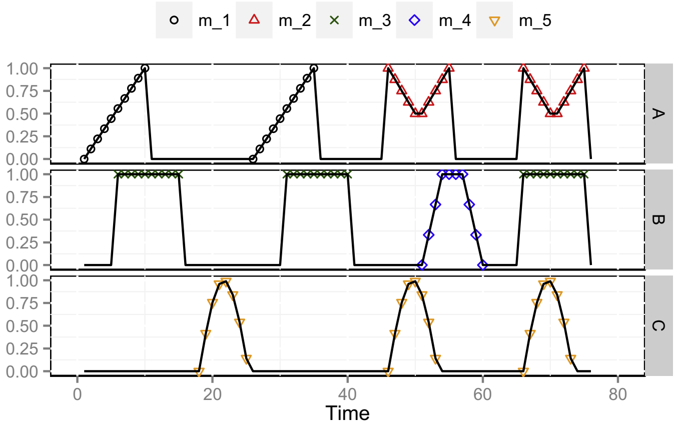


Fig. 2. An example of univariate motifs discovered in three dimensions.

rules is straightforward. They can be used as the references for normal operations and anomalies can be detected if building operation patterns are different from those frequent patterns or violate the association rules. In this study, two methods are specifically developed to enhance the efficiency in post-mining and maximize the practical values of temporal knowledge discovered.

2.4.1. Identify associations between univariate motifs

Building operations involves multiple separate and interactive subsystems, such as air conditioning, mechanical ventilation, lift, lighting and security systems. Univariate motifs usually represent the frequent sequential operation patterns of each system. It is reasonable to link the associations among univariate motifs with the interactions among subsystems. Multivariate motifs can provide general information on which univariate motifs frequently occur together. However, they hardly quantify the relationships among univariate motifs and this limits their practical value. For example, a multivariate motif cannot answer, if one univariate motif occur, whether the other univariate motifs in it will occur or not with certain probability. In this study, a post-mining method is designed to explore the associations among univariate motifs which can directly answer this question. This method is an extension of association rule mining. Given a multivariate time series data, the univariate motif discovery algorithm is applied to each univariate time series separately to find univariate motifs. These univariate motifs are then labeled as $m_1, m_2, \ldots, m_L$, where L is the total number of univariate motifs discovered. Afterward, a co-occurrence matrix is constructed. The matrix has L columns. The values of each row are either 1 or 0, indicating whether an occurrence of a univariate motif is observed or not. Once the matrix is constructed, the Apriori algorithm is used to discover associations between univariate motifs. Two parameters, i.e., the minimum thresholds for support and confidence, are defined for rule induction. Three statistics, including the support, confidence and lift, can be generated with each association rule to facilitate decision making.

An example for construction of a co-occurrence matrix is given here. Fig. 2 illustrates five univariate motifs (i.e., m_1–m_5) in the sequences of three variables, A, B and C. The motifs in A and B, m_1–m_4, have a time duration of 10 while m_5 in C has a time duration of 8. The co-occurrence matrix is constructed as shown in Table 2. The numbers (0 or 1) in each row show the occurrence of the corresponding motifs. For example, the second row shows that only motif 5 occurs during the time period between 18 and 25; the first and third rows show that motifs 1 and 3 occur together twice; the fifth row shows that motifs 2, 4, 5 occur together for once; the sixth row shows that motifs 2, 3 and 5 occur together for once. It should be noted that, although the occurrence of the motifs are related to certain time period, the exact time is not considered in constructing the matrix. The frequency of the co-occurrence of multiple univariate motifs is of interest.

Table 2
An example of co-occurrence matrix for mining association rules between univariate motifs.

m_1	m_2	m_3	m_4	m_5
1	0	1	0	0
0	0	0	0	1
1	0	1	0	0
0	1	0	1	1
0	1	1	0	1

The construction of the co-occurrence matrix can be conveniently implemented by programming with the information of starting and ending time instants of all univariate motifs. Once the co-occurrence matrix is ready, the Apriori algorithm is adopted to mine the associations. Setting the minimum thresholds of support and confidence as 0.3 and 0.8 respectively, two rules are derived, i.e., $m_1 \rightarrow m_3$ and $m_2 \rightarrow m_5$. Both rules have a support of 0.4 and a confidence of 1. It means that when motif 1 occurs, the probability of the occurrence of motif 3 is very high.

2.4.2. Identify time lags in temporal association rules

As introduced in Section 2.3.2, the TRuleGrowth algorithm is adopted to discover the temporal association rules under the constraint of a maximum time span. One limitation is that no information is available about the exact time lag, which is the time interval between the antecedent and the consequent. This type of information is valuable for establishing reliable control and performance optimization in building operations. An iterative filtering method is developed to identify the time lag. The method iteratively runs the TRuleGrowth algorithm by changing the maximum time span from 1 to T and the temporal association rules generated at each iteration and the corresponding time lag are stored in the rule sets. The time lag in a temporal association rule can be discovered by matching the rule with the rule sets.

3. Mining real BAS data

3.1. BAS data description

The methodology is applied to analyze the BAS data retrieved from the tallest building in Hong Kong, i.e., the International Commerce Center (ICC). ICC is a high-rise commercial building with a height around 490 *m* and a total floor area of 321,000 m^2. It contains a 4-storey basement for parking, a 6-storey block for shopping and exhibitions, a 65-storey office tower, an observations deck and a 17-storey hotel.

A complex BAS has been installed to monitor and control the building operations. Energy efficiency is one of the major concerns of ICC. The whole building power consumption can be broken down into five parts for different services systems, i.e., the heating, ventilation, and air-conditioning (HVAC) system, normal power and lighting (NLTG), essential power and lighting (ELTG), vertical transportation system (VTS), and plumbing and drainage system (PD). The HVAC system in ICC consists of six subsystems, i.e., chillers, cooling towers, water pumps, primary air-handling units (PAU), air-handling units (AHU), and mechanical ventilation (MV). Besides power consumption data, there are a large number of measurements of temperature, flow rate, pressure, etc., of the water and the process air in the system as well as various status and control signals. There are approximately 950 measurements in the BAS system, which are sampled every 1 or 15 min. The size of annual BAS data in ICC is around 30 gigabytes. Annual operation data in 2014 with a sampling interval of 15-minute are retrieved for analysis in this study. The data consist of 34,950 observations and 78 variables, including the date and time, building cooling load as well as power consumptions of various subsystems. Approximately 0.52% of the data contain missing values and the maximum lasting period is 45-minute. The moving average method with a window size of 8 (i.e., corresponding to 2-hour) is adopted to fill the missing values. The Hampel filter method is applied to detect point-wise outliers and parameters k and θ are selected as 8 and 3 respectively.

3.2. Identification of daily power consumption patterns in building operation

As introduced in Section 2.1, the intrinsic periods in the time series of building total power consumption are estimated using the non-parametric spectral density estimation method. The top three dominant frequencies are 0.0103, 0.0417 and 0.1121, which correspond to periods of 97 (i.e., 1/0.0103), 24 (i.e., 1/0.0417) and 9 (i.e., 1/0.1121) respectively. Since the BAS data are collected at an interval of 15-minute, these three periods are approximately 1-day, 6-hour and 2-hour respectively.

The dominant period in the sequence of building total power consumption is 1-day. Therefore, the whole BAS data are segmented into daily subsequences and then transformed into SAX representations. Increasing the word size W and alphabet size A will lead to a better SAX representation of the original time series. However, the reduction in computation load is less. Miller et al. recommended W and A as 4 and 3 respectively to identify typical patterns in building power consumption data [32]. Actually, the selection of W and A is influenced by the scale of the building, installation capacities (e.g., cooling, heating, total electricity power) and operation strategies. A large building with high installation capacities tends to require large W and A to adequately describe the variation in the original time series data. In this study, W is chosen as 12, considering that 2-hour was identified as one of the dominant periods. Considering that the chiller plant usually accounts for a large proportion of the total power consumption and the maximum running chiller number is 5 in the BAS data to be analyzed, A is chosen as 5 to reflect there are five major levels of power consumption due to the on-off control of chillers. It should be noted that the standardization is only applied to the total building power consumption time series, but not the daily subsequences. The consideration here is to identify typical daily patterns considering both the shape and magnitude.

The SAX representations of daily subsequences are then partitioned into different groups using the EAC method. K_{lower} and K_{upper} are selected as 2 and 20 respectively. The iteration number E is set as 200. As a result, 8 clusters are identified. Clusters 5, 6, 7 and 8 only consists of 6 daily subsequences out 365 subsequences. Those subsequences are actually subsequence-wise outliers, as their shape and magnitude are dramatically different from the others. They are excluded from further analysis. Fig. 3 presents the profiles of daily subsequences in Clusters 1–4. Further examination of each cluster shows that Clusters 1–4 can be best interpreted using the climate and day type. Cluster 1 includes weekends in cold season and Cluster 4 contains weekdays in hot season. Cluster 2 and Cluster 3 mainly include weekdays in cold season and weekends in hot season respectively. The clustering results are coincident with the results obtained in our previous study [4,18] and domain knowledge. It indicates that the SAX transformation can very well preserve the important information in original time series data.

3.3. Identify frequent operating patterns of subsystems

Univariate and multivariate motif discovery are applied to the 4 clusters separately to identify the frequent operating patterns. Considering that the daily operating conditions (including outdoor weather conditions and indoor occupancy and equipment utilization conditions) varies largely, it is more meaningful to discover motifs in building operations with smaller lengths, compared with the above identification of power consumption pattern. In this study, the length of the univariate motifs to be discovered is set as 6-hour, as it is identified as the second dominant period in the building power consumption data. More specifically, subsequences are segmented using a 6-hour sliding window, which means the subsequences created are overlapping. Standardization is performed for each subsequence in each cluster. SAX representations are created using the setting of $W = 6$ and $A = 5$. In such a case, each SAX symbol represents the hourly mean and has five possible levels. The iteration number for random projection is 100. During each iteration, 4 out of 6 SAX symbols are randomly selected for comparison, which means that subsequences belonging to the same motif can be different at one position at most [42].

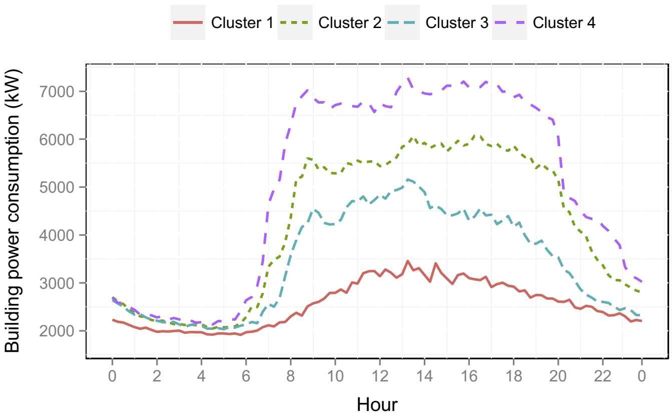


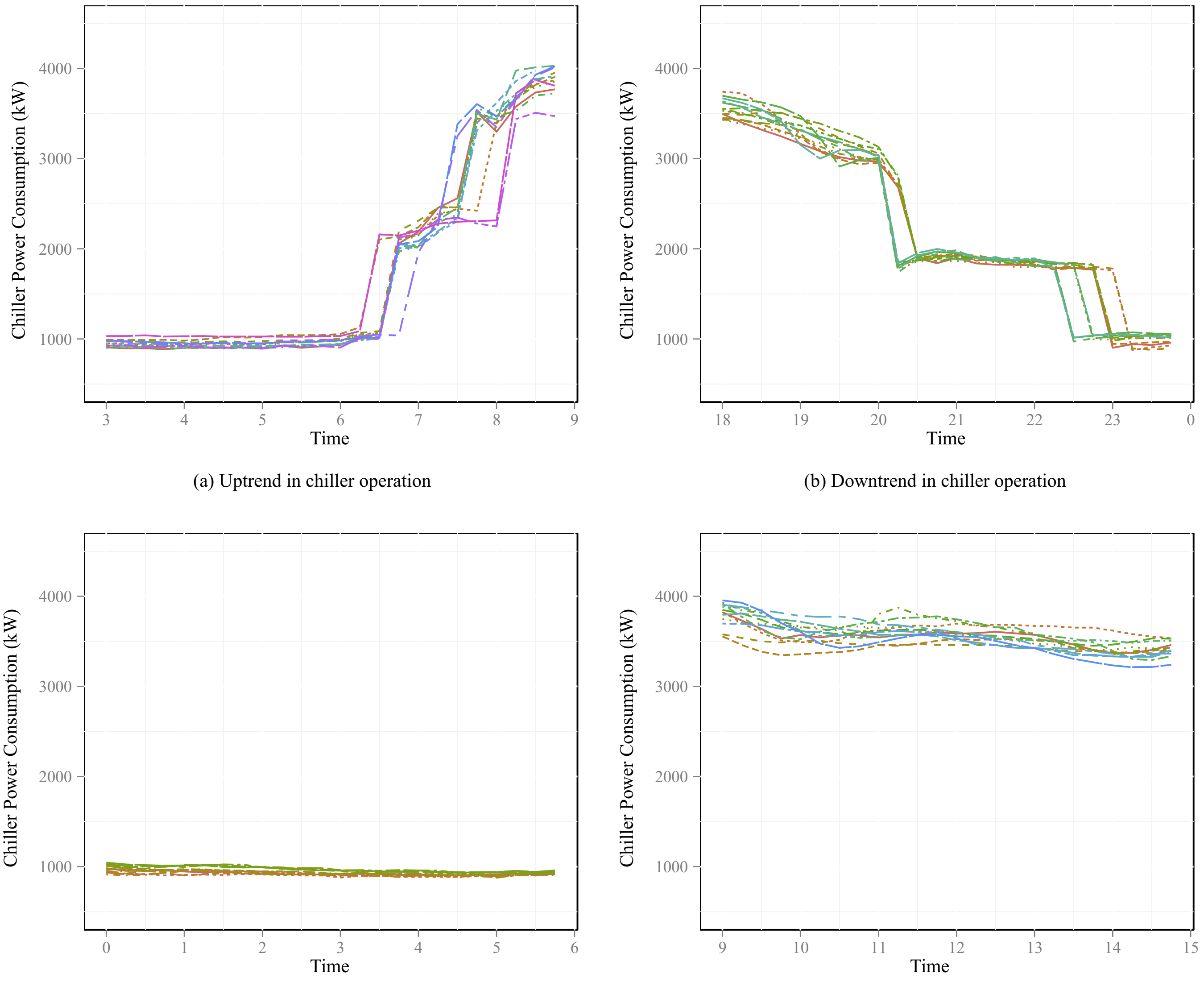
Fig. 3. Four typical prototypes of daily building power consumption.

Table 3
A summary of univariate motifs discovered in Cluster 4.

Subsystems	Chiller	CT	SCHWP	AHU	PAU	MV	VTS	NP	EP	PD
Motif no.	15	9	10	17	14	19	4	15	3	3

Table 3 summarizes the number of univariate motifs discovered for each subsystem in Cluster 4 (i.e., weekdays in hot season). Fig. 4 presents 4 motifs discovered in the time series of the aggregated chiller power consumption in Cluster 4. Each curve represents an occurrence of the corresponding motif. It is apparent that the time series subsequences belonging to the same motif are very similar in their shapes and magnitudes. An uptrend in chiller power consumption is observed in Fig. 4a. It is shown that two chillers are sequentially switched on at the beginning of working hours (i.e., 6:00 a.m. to 9:00 a.m.) to cope with the upcoming morning peak of occupancy and equipment utilization. The chiller switch-off process shares a similar pattern and two chillers are sequentially switched of (Fig. 4b). The other two motifs, as shown in Fig. 4d and c, present relatively steady operating conditions. The chiller operation between 0:00 a.m. and 6:00 a.m. is steadily maintained at a low level due to the absence of occupancy. By contrast, the chiller power consumption is maintained at a much higher level between 9:00 a.m. to 3:00 p.m. A slight decrease can be observed from 1:00 p.m. to 2:00 p.m., which is in accordance with the lunch time for most companies in ICC.

Typical operating behaviors can be obtained by analyzing the univariate motifs identified. For instance, Fig. 5 presents 2 frequent patterns for the AHU operation between 9:00 p.m. and 3:00 a.m. in Cluster 4. The main difference is that a sudden drop in AHU power consumption is observed at 12:00 a.m. in Fig. 5a, while the AHU power consumption gradually decreases in Fig. 5b. After



(a) Uptrend in chiller operation

(b) Downtrend in chiller operation

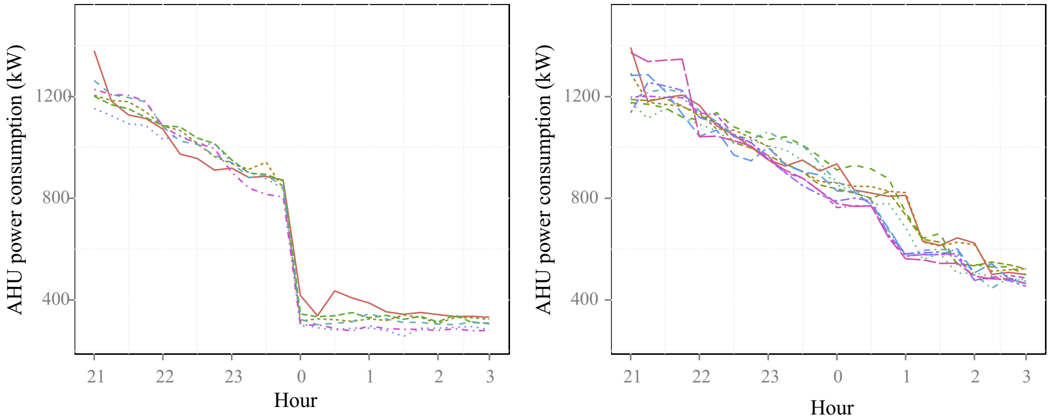
(c) Horizontal trend at low level

(d) Horizontal trend at high level

Fig. 4. Examples of univariate motifs in chiller operation in Cluster 4.

carefully examined the original data, it is found that the AHU power consumption measured at three mechanical floors (i.e., 6/F, 42/F and 78/F) simultaneously drop at 12:00 a.m. in pattern 1. By contrast, the drops are observed at 10:00 p.m., 1:00 a.m. and 2:00 a.m. for the AHUs at 42/F, 78/F and 6/F slightly and gradually in pattern 2.

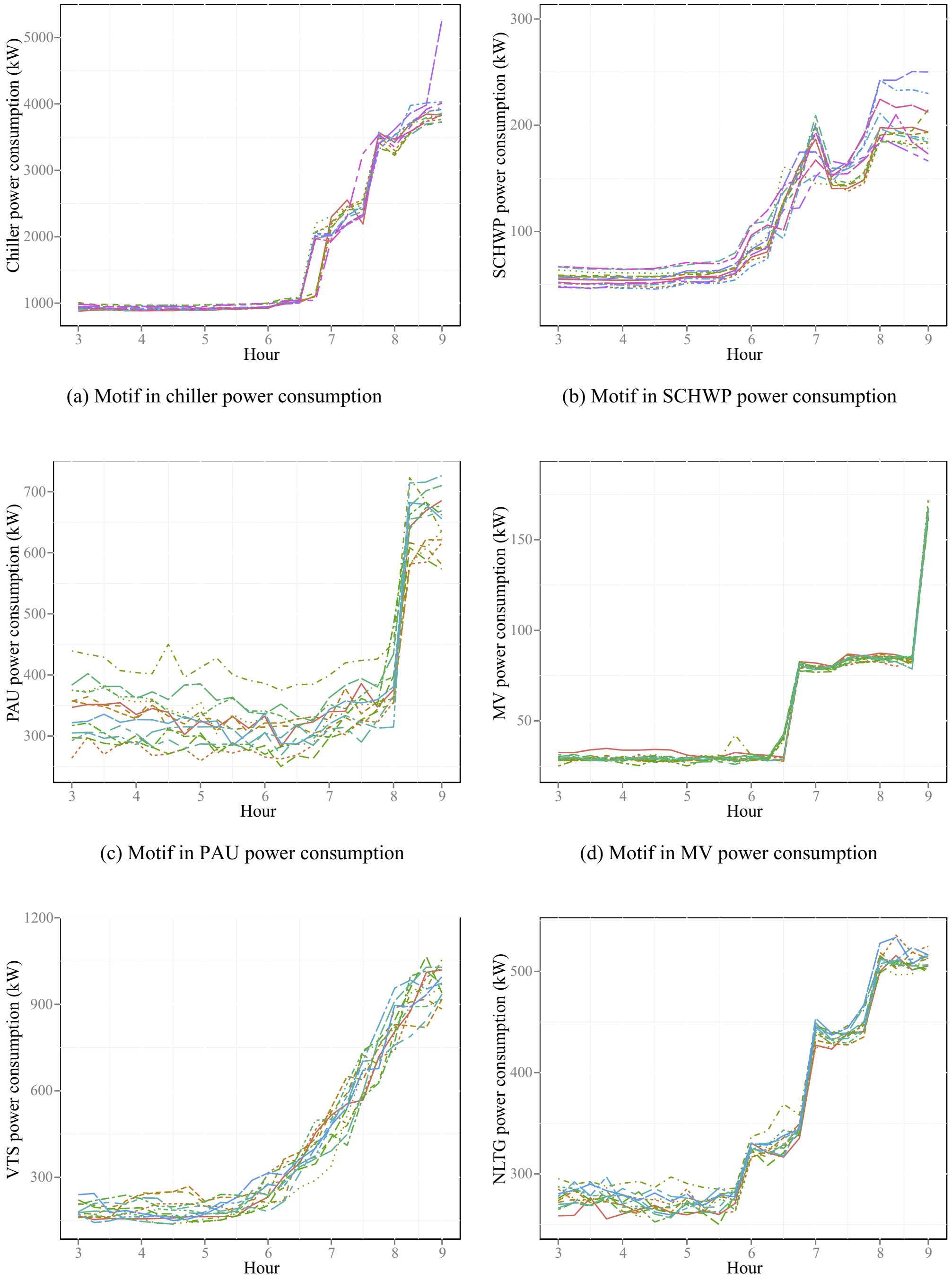
As introduced in Section 2.3.1, univariate motifs discovered are used to identify multivariate motifs. The algorithm can discover



(a) Pattern 1

(b) Pattern 2

Fig. 5. Typical AHU operating patterns between 9:00 p.m. and 3:00 a.m. in Cluster 4.



(a) Motif in chiller power consumption

(b) Motif in SCHWP power consumption

(c) Motif in PAU power consumption

(d) Motif in MV power consumption

(e) Motif in VTS power consumption

(f) Motif in NLTG power consumption

Fig. 6. An example of multivariate motif in Cluster 4.

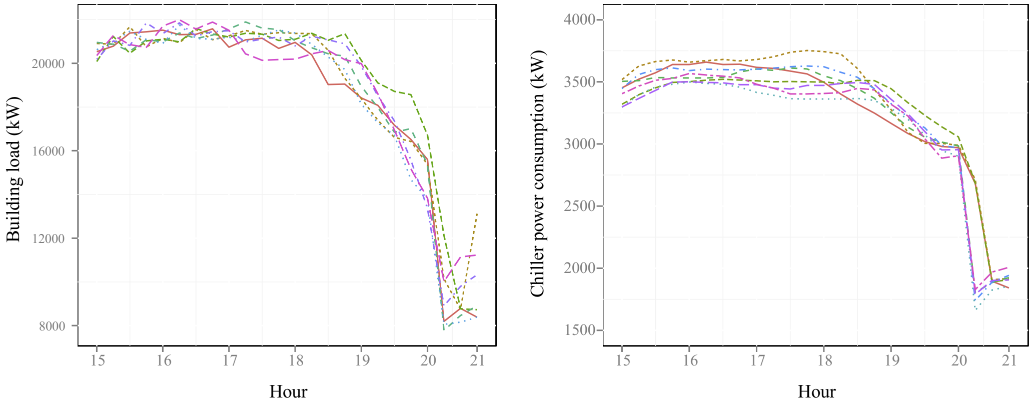
both synchronous and non-synchronous multivariate motifs. The parameter α is set as 0.8. Fig. 6 presents an example of simultaneous multivariate motif. It depicts the building dynamic operations for different subsystems during 3:00 a.m. to 9:00 a.m. The chiller power consumption starts to rise from 6:30 a.m. and two chillers are sequentially switched on. A rise in SCHWP power consumption can be observed accordingly to circulate the chilled water. The PAU power consumption stays steady at low-level until 8:00 a.m. This is because ICC adopts demand-controlled ventilation to control the PAU and the occupancy increases from 8:00 a.m. because people start to work. A rise in MV power consumption can be observed at 8:00 a.m., which is also due to the occupancy change. In addition, the MV power consumption also undergoes an increase at around 6:30 a.m., which relates to the activation of the precooling strategy. Similarly, uptrends in VTS and NLTG power consumptions can be observed in Fig. 6e and f to cope with the increase in occupancy. These motifs show that the HVAC system in ICC is under reliable control and operations well meet the expectations. ICC was awarded as an Intelligent Building of 2011 by the Asian Institute of Intelligent Buildings, partly owing to the advanced BAS installed in ICC.

3.4. Identify temporal association rules between subsystem operations

This section focuses on discovering the temporal associations between the operations of different subsystems. The operation of each subsystem at certain time instant is represented by two

Table 4
Examples of temporal associations discovered.

Rule	Antecedent	Consequent	Time lag (15-min)	Support	Confidence
1	AHU = *Low*, 5	Chiller = *Low*, 4	1	1.00	1.00
2	AHU = *Low*, 5	Chiller = *Low*, 5	2	0.89	0.89
3	NLTG = *Low*, 7	PAU = *Low*, 7	9	0.78	0.82



(a) Antecedent motif

(b) Consequent motif

Fig. 7. Association between building cooling load and chiller motifs in Cluster 4.

features, i.e., level and trend. The power consumption data of each subsystem are categorized into three levels, i.e., *Low*, *Medium* and *High*. The trend is defined based on the changes between successive time step and categorized into 1–7, indicating large decrease, moderate decrease, slight decrease, steady, slight increase, moderate increase and large increase. The categorization thresholds are determined using *k*-means clustering algorithm. The TRuleGrowth algorithm is applied with the minimum support and confidence being set as 0.2 and 0.8 respectively. The maximum time span changes from 1 (i.e., 15-minute) to 12 (i.e., 3-hour). The post-mining method described in Section 2.4.2 is applied to find the exact time lag in temporal association rules.

Table 4 presents three example rules describing the inter-subsystem temporal associations in the multivariate motif shown in Fig. 6. The first rule shows that when the AHU power consumption is *Low* and experiencing a slight increase at time *T*, the chiller power consumption will be *Low* and stay steady at time *T* + 1. The second rule shows that given the same antecedent, a slight increase in the chiller power consumption will be observed at *T* + 2. These two rules demonstrate that the change in AHU and chiller operation is not synchronous and the time lag is around 15 min. The last example rule describes the temporal association between the NLTG and the PAU power consumptions. It states that when the NLTG consumption is *Low* and experiencing a significant increase at time *T*, a significant increase in the PAU power consumption will be observed at *T* + 9. The result's validity can be verified by manually inspecting Fig. 6. For instance, the first significant increase in NLTG and PAU power consumptions take place at around 5:45 a.m. and 8:00 a.m. respectively and therefore, the time lag for the third rule should be 9 unites of time (i.e., 135 min).

4. Applications of temporal knowledge discovered

A straightforward approach to applying the temporal knowledge discovered to building management is to build a database of motifs and temporal association rules as the benchmark of building operations. Then, the real-time BAS time series data are compared with the benchmarked operations to identify any possible anomalies. The post-mining methods developed in this study provide two more approaches to such applications. The following parts demonstrate these applications.

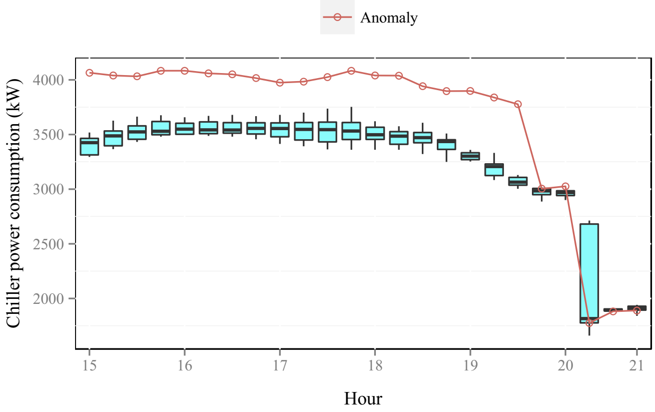
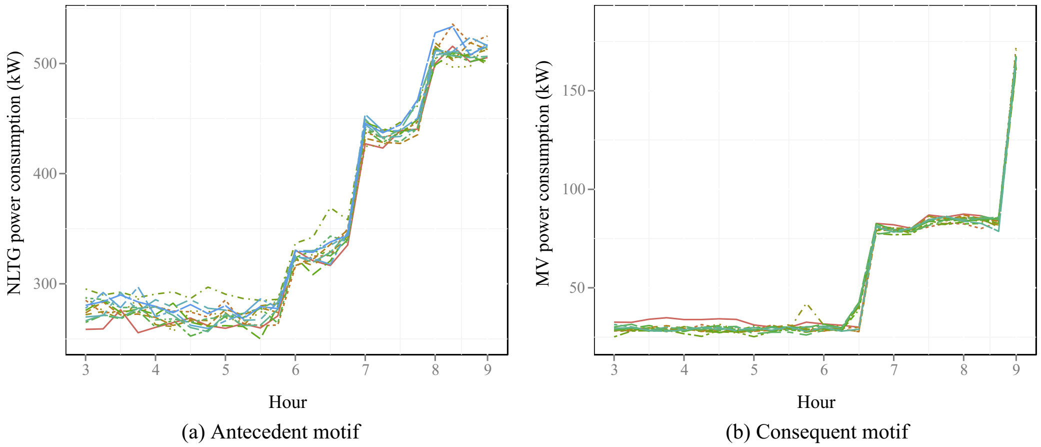


Fig. 8. Comparison of chiller operations.

4.1. Applications of associations between univariate motifs

The post-mining method introduced in Section 2.4.1 is applied to discover associations between univariate motifs. To illustrate, 103 univariate motifs which are discovered in Cluster 4 are used for analysis. The Apriori algorithm is applied with the minimum support and confidence set as 0.1 and 0.8 respectively. These thresholds are set in such a way to ensure the discovery of strong but not necessarily frequent associations. 144 association rules are discovered. The association rules obtained can be applied to find anomalies in operation, such as less energy-efficient operations, faulty operations, as well as normal but rare operations.

As shown in Fig. 7, one rule is Cooling Load = Motif 3 → Chiller = Motif 11. It describes the association between Motif 3 in the building cooling load and Motif 11 in the chiller power consumption, which both take place between 3:00 p.m. and 9:00 p.m. Atypical patterns are identified by finding the time series data which meet the antecedent but not the consequent. An example



(a) Antecedent motif (b) Consequent motif

Fig. 9. Association between NLTG and MV motifs in Cluster 4.

is presented in Fig. 8. Motif 11 in the chiller power consumption is shown using blue boxplots and the atypical chiller operation is shown using the red solid line. Given the same building load demand, the atypical operation results in much higher chiller power consumption during the period from 3:00 p.m. to 7:30 p.m. The mean chiller coefficient of performance (COP) decreases from 5.82 to 5.12 (i.e. 12% drop in energy efficiency) when the atypical operation takes place. It is found out by examining original data that during chiller Motif 11, three chillers are running at a nearly full-load condition. By contrast, 4 chillers are switched-on during the atypical operation with a lower part-load ratio. In such a case, the identified atypical operation resulted in a less energy efficient operation.

Another example rule is NLTG = Motif 6 → MV = Motif 9. It describes the association in the operating patterns of the normal power and lighting (NLTG) and mechanical ventilation (MV). These two motifs both take place between 3:00 a.m. and 9:00 a.m. and are shown in Fig. 9. Fig. 10 compares an atypical MV operation with the MV Motif 9. Starting from 4:30 a.m., the atypical operation has higher MV consumption than that in MV Motif 9. Further investigation shows that the difference is caused by the MV at the third mechanical floor (i.e., 78/F). Normally, the MV consumption at the third mechanical floor is maintained at around 20 kW between 3:00 a.m. and 9:00 a.m. During the atypical operation, it experiences a sudden increase from 20 kW to 45 kW at 4:30 a.m. and is maintained at that level afterwards. One possible reason for this increase is due to the occupancy change in the corresponding office zone. However, the occupancy in office zone is unlikely to change at 4:30 a.m. In addition, the NLTG consumption is also subject to the influence of occupancy and no significant difference is observed during atypical operation. Such atypical operation may be due to the interference of manual control.

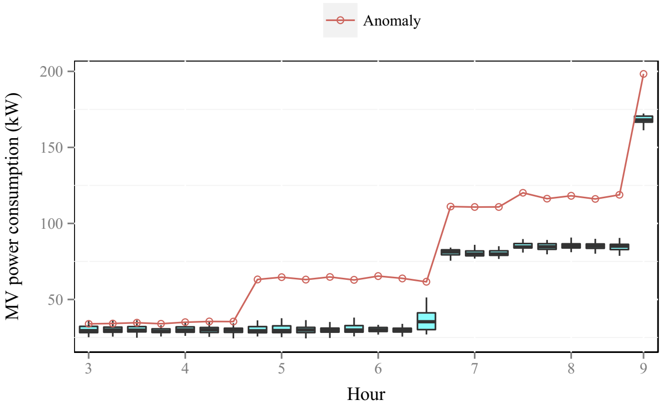
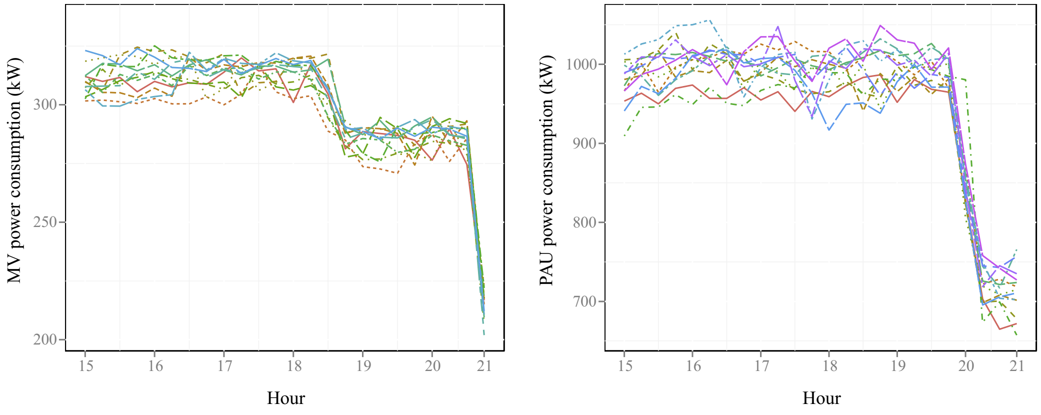


Fig. 10. Comparison of MV operations.

Another example rule describes the association between MV Motif 18 and PAU Motif 14. As shown in Fig. 11, both motifs take place between 3:00 p.m. and 9:00 p.m. The two drops in MV consumption at around 6:30 p.m. and 8:45 p.m. are due to the decrease in MV consumption at the second and the first mechanical floors



(a) Antecedent motif (b) Consequent motif

Fig. 11. Association between MV and PAU motifs in Cluster 4.

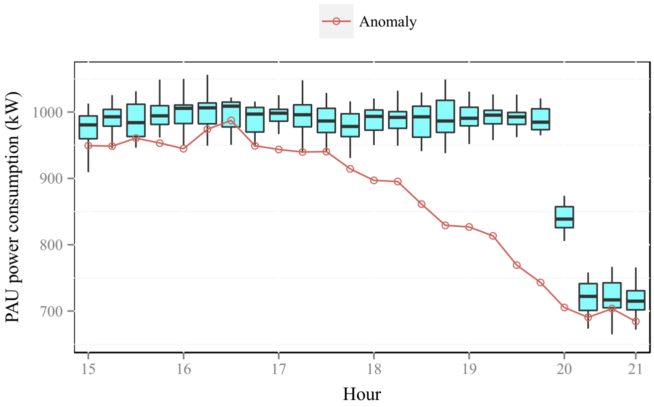


Fig. 12. Comparison of PAU operations.

respectively. By contrast, one significant drop in the PAU consumption is observed at around 8:00 p.m., which is due to the huge decrease in office occupancy. An atypical operation is identified and its PAU consumption is compared with the PAU Motif 14 in Fig. 12. Compared with PAU Motif 14, the PAU consumption in atypical operation is much smaller from 5:30 p.m. to 8:00 p.m. The reason behind is that the next day is a public holiday in Hong Kong and many offices have their employees released at around 5:00 p.m. Consequently, a power reduction in PAU consumption is observed. In such a case, the atypical operation identified is a normal but rare operation.

4.2. *Application of temporal association rules*

4.2.1. *Temporal anomaly detection*

Temporal anomaly can be detected using the temporal association rules. Two approaches are possible. If the anomaly widely exists in the time series data, a temporal association rule specifying such atypical association will be derived. In such a case, temporal anomaly can be detected by finding those observations which are in accordance with these temporal association rules. However, those anomaly data are seldom available. The second approach is more practically feasible. A knowledge database of normal temporal association rules can be constructed. Temporal anomaly can be detected by finding those observations which fail to meet the rules in the database.

An example is given here. Two rules with a time lag of 15-minute are derived to describe temporal associations in the chiller operation between 6:00 a.m. and 12:00p.m.: Chiller = *High*, 5 $\xrightarrow{T=1}$ Chiller = *High*, 4 and Chiller = *High*, 5 $\xrightarrow{T=1}$ Chiller = *High*, 3. These two rules specify that two possible operating modes are possible at time $T+1$ given the chiller power consumption at time T is *High* and has a slightly increasing trend. Fig. 13 presents the subsequences which fulfill these two rules. The chiller power consumption at $T+1$ will remain at *High* level, with either a steady or a slightly decreasing trend. Temporal anomalies can be detected by finding subsequences which fail to meet the rule consequent given the same antecedent. Fig. 14 presents an example of such anomalies. The anomaly is shown in red solid line. It meets the rule antecedent at time T; however, the operating mode at time $T+1$ becomes *Medium* and has a significant decreasing trend. Further investigation shows that at 8:30 a.m., Chiller 4 was switched off while two other chillers were switched on as replacement. After consulting with the operation staff, it is found that Chiller 4 was manually switched off due to its high operating current.

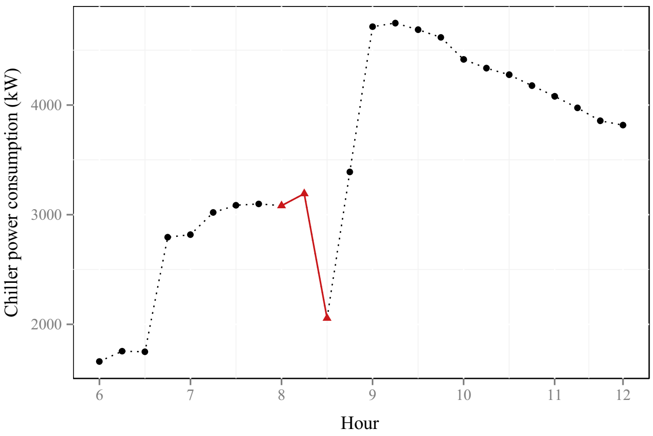
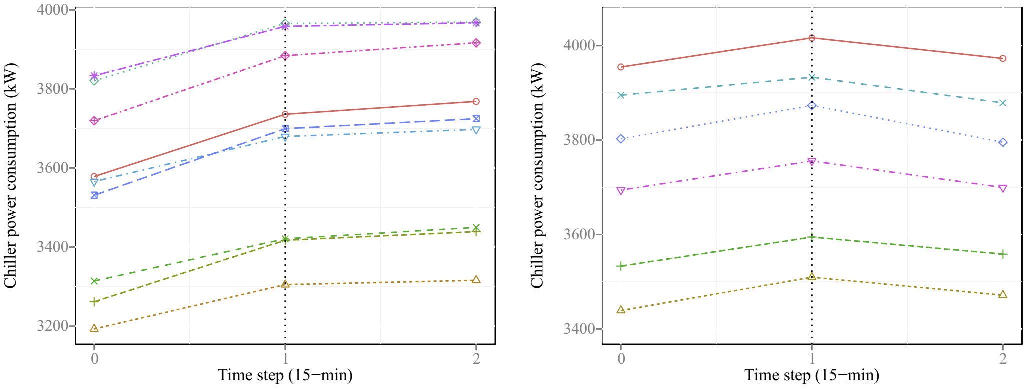


Fig. 14. An example of temporal anomalies.

4.2.2. *Characterization of building dynamics*

The extraction of time lag between the antecedent and the consequent of temporal association rules helps to characterize the building dynamics. Table 5 presents six example rules describing temporal associations in chiller operation. These rules are extracted from two chiller motifs, which are shown in Fig. 15. The first three rules are derived from the chiller Motif A, which takes place between 6:00 a.m. and 12:00 p.m. Rule 1 indicates that if the chiller power consumption is *Low* and experiencing a significant increase at time T, the operating mode will be at *Medium* level and under slight increase at $T+3$. The second rule shows that once the chiller



(a) Chiller=High, 5 $\xrightarrow{T=1}$ Chiller=High, 4

(b) Chiller=High, 5 $\xrightarrow{T=1}$ Chiller=High, 3

Fig. 13. Examples of temporal associations in chiller operation.

Table 5
Temporal associations in chiller operations.

Rule	Motif	Antecedent	Consequent	Time lag (15-minute per unit)	Support	Confidence
1	A	Chiller = *Low*, 7	Chiller = *Medium*, 5	3	0.92	0.97
2	A	Chiller = *Low*, 7	Chiller = *High*, 4	10	0.81	0.92
3	A	Chiller = *Low*, 7	Chiller = *Medium*, 7	4	0.83	0.83
4	B	Chiller = *High*, 1	Chiller = *Medium*, 4	3	0.36	0.87
5	B	Chiller = *Medium*, 1	Chiller = *Low*, 4	3	0.78	0.85
6	B	Chiller = *High*, 1	Chiller = *Medium*, 1	12	0.44	0.84

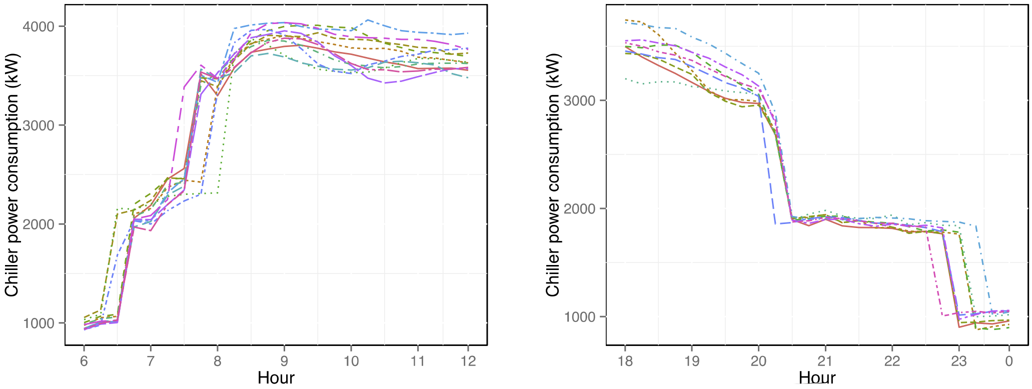


Fig. 15. Two examples of chiller operation motifs.

power consumption starts to increase significantly at *Low* level, it will reach its steady state at *High* level at $T+10$. Rule 3 shows that the time lag between two significant increases at *Low* and *Medium* levels is around 1 h. The latter three rules are derived from the chiller Motif B, which occurs between 6:00 p.m. and 12:00 a.m. Rule 4 states that if the chiller consumption is *High* and experiencing a significant decrease at time *T*, its steady state at *Medium* level will be reached at $T+3$, i.e., 45 min later. Similarly, Rule 5 describes that the time needed for the chiller power consumption to reach its steady state from the *Medium* to *Low* level is also 45 min. The last rule quantifies that the time lag between the huge decrease at *High* and *Medium* levels is around 3 h. The result is verified by checking Fig. 15. The knowledge discovered in this subsection helps to quantify the building dynamics from two perspectives, i.e., the power consumption level and relative changes between successive time steps (i.e., trend). The temporal interactions and dynamics can be automatically extracted. Useful insights can be gained into how building subsystems react to a certain change in operation over time. The temporal associations discovered can be used to facilitate the optimal control and decision-makings in building operation, e.g., chiller sequence control and integration between individual buildings and large power grid systems.

5. Conclusions and discussions

BAS data are in essence multivariate time series data. Currently, few studies have addressed temporal knowledge discovery and applications in big BAS data. This study proposes a generic temporal knowledge discovery methodology for mining big BAS data. A diversity of time series data mining techniques and their practical potentials in analyzing big BAS data for building operations and performance management are explored in this study. Rather than addressing pre-defined specific problems, the methodology developed mainly aims to discover unknown temporal knowledge by adopting unsupervised DM techniques to mine the big BAS data. The intention is to let the data tell the story and then, using domain knowledge to interpret, select and apply the knowledge discovered. The methodology proposed serves as a prototype of big data analysis tools which can be integrated with modern building automation systems to realize automatic knowledge discovery and applications.

Enabling the building automation industry to benefit from advances in big data analysis is a non-trivial task. It requires building professionals to thoroughly understand the mechanisms of both DM algorithms and building operations. A lot more work needs to be done other than simply applying DM algorithms to analyze BAS data. One major concern in mining BAS data is the computational challenge brought by the massive data amount. From a technological perspective, this challenge can be tackled by using high-performance computing machines or cloud-based computing. The adoption of suitable data transformation methods and more computationally efficient DM algorithms can provide an alternative solution. This study shows that the SAX method is capable of reducing the data numerocity while preserving the majority of the information contained in the BAS power consumption data. The univariate motif discovery algorithm adopted in this study is based on the concept of combinatorial search rather than exhaustive search and thereby the required computational costs can be largely reduced. Another essential element in the knowledge discovery process is feature engineering, which refers to the development of new features based on the original data. It can greatly enhance the mining result quality. Besides the power consumption level, this study includes the changing trend to describe the mode of each subsystem at each time step. The temporal association rules discovered are more meaningful and straightforward for knowledge interpretation and application, compared with those obtained using other features as inputs (e.g., the power consumption level alone).

Time series data mining can discover large amounts of knowledge with different types, such as clusters, univariate and multivariate motifs, and temporal association rules. It is challenging and time-consuming to interpret and apply the knowledge discovered. This study develops two methods for the efficient post-processing of knowledge discovered. The first method uses a co-occurrence matrix to map the relationship between univariate motifs. Reliable associations between univariate motifs are derived which provides a novel and convenient approach to utilizing univariate motifs. The second method utilizes a filtering method to improve the temporal association rules mining algorithms with the accurate estimation of time interval between the antecedent and the consequent. The time interval or lag provides valuable insights into building dynamics and HVAC performance characteristics. The methodology has been applied to analyze the BAS data retrieved from the tallest building in Hong Kong. The knowledge discovered has been successfully used to identify anomalies in building operations and characterize the building dynamics. The open-source software *R* and *SPMF* were used to perform the mining.

The main purpose of this study is to bridge the knowledge gap between building professionals and advanced data analytics. One limitation of this research is that it only considers power consumption data. The data transformation methods considering the physical variables in BAS data, such as the temperature, relative humidity, water flow rate and pressure, are more challenging. In addition, even though two methods have been developed to enhance the post-mining efficiency, the amount of knowledge to be examined by domain expertise is still large. For instance, 103 association rules are obtained from mining the univariate motifs in one data cluster. Future research will focus on developing transformation methods for various types of physical variables in BAS data and propose solutions to further enhance the post-mining efficiency.

Acknowledgements

The authors gratefully acknowledge the support of this research by the Research Grant Council (RGC) of the Hong Kong SAR (152181/14E). Henrik Madsen is partly funded by the Danish CITIES project (DSF-1305-00027B) which is hence also acknowledged.

References

[1] L. Perez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy Build. 40 (2008) 394–398.

[2] United Nations Environment Programme (UNEP), Common Carbon Metric for Measuring Energy use and Reporting Greenhouse Gas Emissions from Building Operations, 2009, http://www.unep.org/sbci/pdfs/UNEPSBCICarbonMetric.pdf (accessed on 09.04.15).

[3] P. Waide, J. Ure, N. Karagianni, G. Smith, B. Bordass, The Scope for Energy and CO_2 Savings in the EU Through the Use of Building Automation Technology, Final Report for the European Copper Institute, 2013, August.

[4] C. Fan, F. Xiao, C.C. Yan, A framework for knowledge discovery in massive building automation data and its applications in building diagnostics, Autom. Constr. 50 (2014) 81–90.

[5] F. Dalene, Technology and information management for low-carbon building, J. Renew. Sustain. Energy 4 (2012) 041402, http://dx.doi.org/10.1063/1.3694120.

[6] B. Dong, C. Cao, S.E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy Build. 37 (2005) 545–553.

[7] A.S. Ahmad, M.Y. Hassan, M.P. Abdullah, H.A. Rahman, F. Hussin, H. Abdullah, R. Saidur, A review on applications of ANN and SVM for building electrical energy consumption forecasting, Renew. Sustain. Energy Rev. 33 (2014) 102–109.

[8] C. Fan, F. Xiao, S.W. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Appl. Energy 127 (2014) 1–10.

[9] A. Kusiak, G.L. Xu, Z.J. Zhang, Minimization of energy consumption in HVAC systems with data-driven models and an interior-point method, Energy Convers. Manag. 85 (2014) 146–153.

[10] S.K. Kwok, K.K. Yuen, W.M. Lee, An intelligent approach to assessing the effect of building occupancy on building cooling load prediction, Build. Environ. 46 (2011) 1681–1690.

[11] S.M. Wu, D. Clements-Croome, Understanding the indoor environment through mining sensory data—a case study, Energy Build. 39 (2007) 1183–1191.

[12] G. Kim, L. Schaefer, T.S. Lim, J.T. Kim, Thermal comfort prediction of an underfloor air distribution system in a large indoor environment, Energy Build. 64 (2013) 323–331.

[13] Z. Yu, F. Haghighat, C.M. Fung, H. Yoshino, A decision tree method for building energy demand modeling, Energy Build. 42 (2010) 1637–1646.

[14] J.S. Chou, Y.C. Hsu, L.T. Lin, Smart meter monitoring and data mining techniques for predicting refrigeration system performance, Expert Syst. Appl. 41 (2014) 2144–2156.

[15] E.U. Kucuksille, R. Selbas, A. Sencan, Prediction of thermodynamic properties of refrigerants using data mining, Energy Convers. Manag. 52 (2011) 836–848.

[16] D.F.M. Cabrera, H. Zareipour, Data association mining for identifying lighting energy waste patterns in educational institutes, Energy Build. 62 (2013) 210–216.

[17] A. Capozzoli, F. Lauro, I. Khan, Fault detection analysis using data mining techniques for a cluster of smart office buildings, Expert Syst. Appl. 42 (2015) 4324–4338.

[18] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy Build. 75 (2014) 109–118.

[19] X. Xue, S.W. Wang, Y.J. Sun, F. Xiao, An interactive building power demand management strategy for facilitating smart grid optimization, Appl. Energy 116 (2014) 297–310.

[20] A. Azadeh, M. Saberi, S.F. Ghaderi, A. Gitiforouz, V. Ebrahimipour, Improved estimation of electricity demand function by integration of fuzzy system and data mining approach, Energy Convers. Manag. 49 (2008) 2165–2177.

[21] I. Fernandez, C. Borges, Y. Penya, Efficient building load forecasting, in: Proceedings of the 16th IEEE International Conference of Emerging Technologies and Factory Automation, 5–9 September, Toulouse, France, 2011.

[22] Y. Yao, Z.W. Lian, S.Q. Liu, Z.J. Hou, Hourly cooling load prediction by a combined forecasting model based on analytic hierarchy process, Int. J. Therm. Sci. 43 (2004) 1107–1118.

[23] M. Kawashima, C.E. Dorgan, J.W. Mitchell, Hourly thermal load prediction for the next 24 h by ARIMA, EWMA, LR, and an artificial neural network, ASHRAE Trans. 101 (1995) 186–200.

[24] J.C.M. Yiu, S.W. Wang, A multiple ARMAX modeling scheme for forecasting of air conditioning system performance, Energy Convers. Manag. 48 (2007) 2276–2285.

[25] F. Zamora-Martinez, P. Romeu, P. Botella-Rocamora, J. Pardo, Towards energy efficiency: forecasting indoor temperature via multivariate analysis, Energies 6 (2013) 4639–4659.

[26] L. Renners, R. Bruns, J. Dunkel, Situation-aware energy control by combining simple sensors and complex event processing, in: Workshop on AI Problems and Approaches for Intelligent Environment, August, Montpellier, France, 2012.

[27] Y.C. Wen, G.Y. Lin, T. Sung, M. Liang, G. Tsai, M.W. Feng, A complex event processing architecture for energy and operation management, in: The 5th ACM International Conference on Distributed Event-Based Systems, July 11–15, New York, USA, 2011.

[28] J. O'Donnell, E. Corry, S. Hasan, M. Keane, E. Curry, Building performance optimization using cross-domain scenario modeling, linked data, and complex event processing, Build. Environ. 62 (2013) 102–111.

[29] T.C. Fu, A review on time series data mining, Eng. Appl. Artif. Intell. 17 (2011) 164–181.

[30] H. Madsen, Time Series Analysis, 1st ed., Chapman & Hall/CRC Texts in Statistical Science, 2007.

[31] D. Patnaik, M. Marwah, R.K. Sharma, N. Ramakrishnan, Temporal data mining approaches for sustainable chiller management in data centers, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–29.

[32] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Autom. Constr. 49 (2015) 1–17.

[33] A. Gelman, J. Hill, Data analysis using regression and multi-level/hierarchical models, in: Analytical Methods for Social Research, 1st ed., Cambridge University Press, 2006.

[34] M. Gupta, J. Gao, C.C. Aggarwal, J.W. Han, Outlier detection for temporal data: a survey, IEEE Trans. Knowl. Data Eng. 15 (2014) 1–20.

[35] R.K. Pearson, Outliers in process modeling and identification, IEEE Trans. Control Syst. Technol. 10 (2002) 55–63.

[36] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: a novel symbolic representation of time series, Data Min. Knowl. Discov. 15 (2007) 107–144.

[37] J. Kwac, J. Flora, R. Rajagopal, Household energy consumption segmentation using hourly data, IEEE Trans. Smart Grid 5 (2014) 420–430.

[38] R. Gulbinas, A. Khosrowpour, J. Taylor, Segmentation and classification of commercial building occupants by energy-use efficiency and predictability, IEEE Trans. Smart Grid 6 (2015) 1414–1424.

[39] C.S. Daw, C.E.A. Finney, E.R. Tracy, A review of symbolic analysis of experimental data, Rev. Sci. Instrum. 74 (2003) 915–930.

[40] A.L.N. Fred, A.K. Jain, Combining multiple clustering using evidence accumulation, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 835–850.

[41] S. Vega-Pons, J. Ruiz-Schulcoper, A survey of clustering ensemble algorithms, Int. J. Pattern Recognit. Artif. Intell. 25 (2011) 337–372.

[42] B. Chiu, E. Keogh, S. Lonardi, Probabilistic Discovery of Time Series Motifs, ACM SIGKDD, Washington, DC, USA, 2003, pp. 493–498.

[43] Y. Tanaka, K. Iwamoto, K. Uehara, Discovery of time-series motif from multi-dimensional data based on MDL principle, Mach. Learn. 58 (2005) 269–300.

[44] D. Minnen, C.L. Isbell, I. Essa, T. Starner, Discovering multivariate motifs using subsequence density estimation and greedy mixture learning, in: The 22nd National Conference on Artificial Intelligence, vol. 1, 2007, pp. 615–620.
[45] A. Vahdatpour, N. Amini, M. Sarrafzadeh, Towards unsupervised activity discovery using multi-dimensional motif detection in time series, in: 21st International Joint Conference on Artificial Intelligence, IJCAI, 2009.
[46] M.J. Zaki, SPADE: an efficient algorithm for mining frequent sequences, Mach. Learn. 42 (2001) 30–60.
[47] P. Fournier-Viger, U. Faghihi, R. Nkambou, E.M. Nguifo, CMRules: mining sequential rules common to several sequences, Knowl. Based Syst. 25 (2012) 63–76.
[48] P. Fournier-Viger, C.W. Wu, V.S. Tseng, R. Nkambou, Mining sequential rules common to several sequences with the window size constraint, in: Advances in Artificial-25th Canadian Conference on Artificial Intelligence, Toronto, Canada, 2012, pp. 299–304.