Energy & Buildings 248 (2021) 111195

Contents lists available at ScienceDirect

Energy & Buildings

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

Data mining cubes for buildings, a generic framework for multidimensional analytics of building performance data

Julien Leprince [a,*], Clayton Miller [b], Wim Zeiler [a]

[a] Building Services, Technical University Eindhoven, Eindhoven, The Netherlands
[b] BUDS Lab, National University of Singapore, Singapore, Singapore

ARTICLE INFO

Article history:
Received 27 January 2021
Revised 12 May 2021
Accepted 11 June 2021
Available online 17 June 2021

Keywords:
Data mining
Data cube
Generic method
Multidimensional analytics
Machine learning
Building data

ABSTRACT

Over the last decade, collecting massive volumes of data has been made all the more accessible, pushing the building sector to embrace data mining as a powerful tool for harvesting the potential of big data analytics. However repetitive challenges still persist emerging from the need for a common analytical frame, effective application- and insight-driven targeted data selection, as well as benchmarked-supported claims. This study addresses these concerns by putting forward a generic stepwise multidimensional data mining framework tailored to building data, leveraging the dimensional-structures of data cubes. Using the open Building Data Genome Project 2 set, composed of 3053 energy meters from 1636 buildings, we provide an online, open access, implementation illustration of our method applied to automated pattern identification. We define a 3-dimensional building cube echoing typical analytical frames of interest, namely, bottom-up, top-down and temporal drill-in approaches. Our results highlight the importance of application and insight driven mining for effective dimensional-frame targeting. Impactful visualizations were developed allowing practical human inspection, paving the path towards more interpretable analytics.

1. Introduction

With Europe becoming climate neutral by 2050, meeting carbon dioxide emission targets is being promoted as a major step towards curtailing the rise of global heating. Increased renewable energy shares in the existing energy power systems combined with reduced energy consumption constitute a common adopted long term energy strategy by most countries, including the Netherlands. The International Energy Agency (IEA) indicates buildings to be the largest energy consumer in the world, accounting for over one-third of overall final energy consumption [1]. As a result, building energy efficiency has become one of the main concerns for reaching environmental and sustainability targets. The issue has drawn increasing research and development efforts in recent years. Advances in information systems, computing power and control technologies for optimal resource management, have endowed building automation systems (BAS) with enhanced energy savings ranging from 20 to 35% [2], while generating a huge amount of data from a wide range of appliances every day. These include essential buildings indoor environment quality condition processes such as ventilation, lighting, air conditioning and heating but also home appliances such as dish washers, laundry, kitchen devices and home entertainments.

Yet, BAS data are rarely fully exploited and interpreted. Improving building energy efficiencies with such pools of data remains a challenge; how does one approach the analysis of various associations and correlations amongst multi-temporal, i.e., seconds to hourly resolutions, with daily to decades horizons, and high dimensional data? What methods to follow to acquire useful, interpretable insights on building energy performance and reduce its consumption? Such questions root upon causes usually involving poor data quality, resulting from a large share of missing values and outliers, coupled to lack of efficient and convenient analytical tools & methods for large data sets. Additionally, most BASs only perform basic data analytics and visualizations, such as historical tracking, moving averages and threshold-based anomaly detections which have pushed the building automation industry to new data driven methods and tools to harvest these data pools, namely, data mining.

[*] Corresponding author.
E-mail addresses: j.j.leprince@tue.nl (J. Leprince), clayton@nus.edu.sg (C. Miller), w.zeiler@tue.nl (W. Zeiler).

https://doi.org/10.1016/j.enbuild.2021.111195
0378-7788/© 2021 The Author(s). Published by Elsevier B.V.

1.1. Data mining

Data mining (DM) has grown from a promising technology to an established powerful and effective analytical tool to interpret massive and complex data. In 2001, MIT reviewed DM as one of the top 10 emerging technologies that will change the world [3], while it has now accumulated over a hundred thousand publication records over the last 20 years in a wide variety of fields [4], including medicine, retails telecommunication, financial services and target marketing [5]. In the building sector, the effervescence surrounding the technology was such that recent reviews employed text mining tools to fully uncover the extent of developments in the field [6]. DM is a multi-disciplinary subject, integrating combined techniques from statistics, machine learning and artificial intelligence thanks to high performance computing. It is the core process of identifying valid, useful and understandable patterns from large and complex datasets, known as Knowledge Discovery in Databases (KDD). The DM taxonomy established by Oded and Lior [5] distinguishes two preeminent types of DM: verification oriented, where the system verifies a proposed hypothesis, and discovery-oriented, where the system identifies new rules and patterns autonomously. Verification methods include traditional statistical tests such as goodness of fit test, test of hypotheses (i.e. t-test of means, one sample Z-test), and analysis of variance (ANOVA). Discovery methods, on the other hand, are based on inductive learning, where a model is constructed from generalized sufficient numbers of training examples, assuming its applicability to future unseen data. Another terminology, widely used within the machine learning community, preferably considers discovery-oriented DM and separates the techniques into supervised and unsupervised learning. Supervised methods attempt to discover complex and non-linear relationships between input and output target attributes (referred to as independent and dependent variables respectively) by learning from historical data. This type of process largely composes the predictive learning component of DM discovery methods. It has been applied to the building operational stage [7] as it is directly linked to occupant comfort and responsible for 80–90% of the building's total green gas emissions [8]. Applications of supervised learning notably include predictions of building energy consumption [9–12], thermal load [13,14], indoor environment [15,16], and system performance indices [17–19]. Popular supervised methods comprise two predominant groups: classification [20] and regression [21]. Unsupervised learning, also recognized as descriptive learning, groups techniques used to organize instances without pre-specified attributes. It aims at finding underlying associations or data structures between variables. The prominent advantage of unsupervised analytics is its ability to discover formerly unknown knowledge [22,23]. Well established techniques involve clustering, association rule mining (ARM) and anomaly detection. Visualization and summarization techniques, for instance, are DM descriptive methods that are not regarded as unsupervised learning. In opposition to predictive learning, descriptive learning can be viewed as a more flexible application that does not require model training or predefined targets during knowledge discovery. Its main applications encompass fault/anomaly detection and building performance diagnostics [24–26].

1.2. Cube multidimensional analytics

Dealing with the large volumes, velocities and varieties characterizing high-dimensional big building data is a complex task. Common analytical tools developed to tackle multidimensional data rely on exploring different dimensional associations at different levels of aggregation leveraging the structures of a data cube [22]. A data cube is defined as a multidimensional data model allowing data exploration from its structured dimensions, i.e. *dimension table* and *facts*. A data cube is commonly organized around a central theme, represented by a *fact table*, which contains names of the different *facts*, or numeric values, and relational attribute keys. For example, a building fact table could include *time*, *location* or *energy flow* attribute keys linking them to their dimension table. Given a fixed set of dimensions, a *cuboid* can be generated for each subset of the given dimensions. Their combinations result in a *lattice* of cuboids, presenting the data at specific levels of summarization from which a multi-dimension analytical map can be defined. Cuboids forming the lowest level of summarization are denoted *base cuboid*, while the 0-D cuboid, holding the highest level of summarization, is designated the *apex cuboid* (typically referred to by *all*) [22,27]. This lattice of cuboids defines the data cube. While data cubes are commonly represented as 3-D geometrical structures, they are naturally n-dimensional, where each dimension represents objects intended to keep record off. In BAS data, hierarchical relationships are often found within dimension tables, e.g., the time dimension includes a natural tree structure rooted on the *year* attribute, and progressively branching out to *months*, *weeks*, *days*, *hours*. Other hierarchically structured dimensions typically include geographical location and site measurements.

Multidimensional cube space analytics rely on the high-dimensional structure of the data to explore multi-lattice and abstraction levels of the cube. Common dimensional exploration methods rely on bottom-up, top-down approaches, namely *rollup*, where fine granularities are gradually aggregated in coarser ones, and *drilldown*, starting from coarser dimensional granularities down to finer ones. This navigation across multiple cube spaces of interest is called OnLine Analytical Processing (OLAP) [27]. By summarizing & aggregating data subsets at different abstraction levels, this tool has greatly assisted multidimensional analytics. Leveraging this approach, R. Ramakrishnan and B. Chen [28] have put forward a cube-space mining method, taking advantage of the data-cube structure to define and select cuboids of interest to mine over. This way, data mining can be used as a building block within the OLAP analysis to exploit multi-scale knowledge discovery in a defined dimensional frame. Characteristics of the cube-space data mining scheme involve the following three steps: (1) relying on cube space to determine the space of candidates for mining, (2) employing OLAP queries to explore features and targets for mining and (3) adopting data-mining models as building blocks within a multi-step mining process. This exploratory multidimensional DM approach, also known as OnLine Analytical Mining (OLAM) [22], allows the user to effectively select and analyze a relevant subset of data at different granularities and present discovered knowledge at different abstraction levels.

1.3. Motivation

This being said, OLAM has, to the best of the authors knowledge, little to none been practiced in the built environment sector, and while DM extensively demonstrated strength and performance in this domain, barriers still persist avoiding professionals from exploiting the full potential of DM analytics. Previous studies usually relied on predefined problems using only a small subset of building data with few established benchmarks to compare results from one investigation to the next [29]. Additionally, developed research methods commonly follow disparate steps, increasing complexity with unsystematic mining analytical procedures. With the variety and complexity of the most recently developed DM techniques as well as the highly dimensional building data, it has become increasingly challenging for building professionals to (i) effectively target which data dimensions to explore and consider in their analytics, (ii) determine what analytical steps to follow for targeted building data mining and (iii) select the most suitable

DM technique for a particular case study from established references. Realizing the prevalent demand for a common DM framework, noticeable studies have proposed methods applied to BAS data [7,30,31]. However, developed frameworks were usually tailored to DM application-specific cases and failed to address multi-dimensional analytical approaches from orderly steps required for systematic and benchmarked building data analytics. Detailed stepwise generic approaches with established good practices for preprocessing, application-specific and benchmarking procedures are frequently overlooked yet desperately needed. In order to adopt systematic analytical steps from a common framework within the building analysts and research community, several steps still need to be undertook; (*i*) establishing and following a common DM framework and (*ii*) developing and employing open building data toy sets to serve both as benchmarks to case-specific studies while allowing (*iii*) the development of replicable implementations of typical building energy management applications for valuable knowledge transfer. Ensuing these steps would cultivate more generalizable findings and insights while vastly contributing to the practical adoption of a common analytical frame.

This study proposes a response to this appeal and puts forward a multi-dimensional analytical method grounded on a generic data mining framework for building data analysis. It puts together reviewed analytics best practices in a step-wise method tailored to DM application for systematic knowledge discovery in big building data. Contributions of this work can be summarized as threefold;

(i) Putting forward a generic building-tailored DM framework for unified and systematic analytics,

(ii) Framing a multi-dimensional analytical approach to big building data, cutting down the complexity endowed from high-dimensionality, and

(iii) Providing an open access implementation of the presented method relying on a large and open building data set, serving as benchmarks to similar studies and appealing to more reproducible, comparable & empirically validated analytics.

The rest of our paper is divided into two main sections covering the presentation of our proposed multidimensional generic DM framework and an illustrative implementation of the method with an automated pattern identification application. Our method is divided in five main phases, namely building data preprocessing, where we define the building data cube, an exploratory analysis to frame our multi-dimensional analytics, followed by pre-mining, mining, confirmatory analysis and post-mining. We apply the method to a typical descriptive knowledge discovery case on an open building data set, assuring overall reproducibility of the exposed findings.

2. Method

Resulting from an in-depth analysis of DM methods and comprehensive review of domain application driven techniques we propose a generic DM framework tailored to multidimensional building data. The developed method is founded on established methodology from the literature. Notable existing frameworks typically involve four major phases, i.e. data preprocessing, data partitioning, knowledge discovery (data mining) and post-mining. In particular, the generic framework developed by Fan, Xiao and Yan [10] designed for BAS data knowledge discovery englobed building performance assessment, diagnosis and optimization as possible applications. Our method follows similar steps yet extends it from a multidimensional view point leveraging both descriptive and predictive mining techniques while importantly stressing prerequisites for reproducible and generalizable results from benchmarks. It puts forward a generic feed-forward and back process to follow while attempting any building mining process and differentiates mining application-dependent steps from generic DM ones from a unified and interpretable method. Our method is illustrated in Fig. 1 where two tasks are performed in the data preprocessing phase, including data integration and data cleaning. Multidimensional data exploration then follows, incorporating benchmark reports and cube lattice selection with OLAP exploration. After, a pre-mining phase incorporates data transformation and mining-specific steps. Next, the mining stage takes place and an important confirmatory analysis phase is there after carried out with validation methods leading to algorithm selection. A feedback loop linking confirmatory analysis to the mining and pre-mining blocks is included in the framework to indicate potential iterative sequence allowing pre-mining steps and mining to be repeated to converge to the desired results for algorithm optimal selection. And the OLAM feedback loop illustrates the repeated mining process over different cube lattices for multidimensional mining. Knowledge interpretation and extraction is then proceeded within the post-mining phase, supported by visualization tools. Finally, discovered knowledge can be used for a defined application, or serve as a preliminary step to another mining phase, as illustrated with the last feedback arrow. This is often the case when mining for association rules or undertaking predictive learning with prior profile clustering for example [31,32]. Details of the evoked phases are developed in the following subsections.

2.1. Preprocessing

Data preprocessing completes two main tasks, i.e. data integration and data cleaning (outlier identification, missing value handling). Data integration refers to the selection of a suitable structure format and data model for the later analysis. Cleaning aims at enhancing data quality to obtain suitable results out of the intended analytics. It has been reported in DM literature that data cleaning should be performed prior to data integration, allowing information industry to benefit from clean ‘usable’ data stored in data warehouses [22]. This work considers the analytical process from a scientific point of view, where data may be cleaned in different ways consequently impacting the later analysis, which is why we recommend the data be stored raw rather than preprocessed[1].

2.1.1. Data integration

Data integration is composed of a first data model definition phase, from which the later integration process can be undertaken. Data model definition constitutes a fundamental first step to structure the multidimensional BAS data under a given schema. Data integration techniques can later be applied consequently providing consistency in naming conventions, encoding structures and attribute measures [22].

2.1.1.1. Data cube map. Establishing the building data mapping serves as an imperative step to the framing and structuring of its various dimensions. Additionally, obtaining clear delineated dimensions allows leveraging the design of a data cube into decomposed lattices that will serve in shaping the later OLAM analytics. A common approach to the data cube model definition originates from the formulation of the analysts’ interrogations. Specifying questions such as “what is the energy consumption

[1] Ideally, if data storage space permits it, both raw and preprocessed versions of the data should be stored, to allow preprocessing reproducibility evaluation as well as alternative variants that could be considered in other studies.

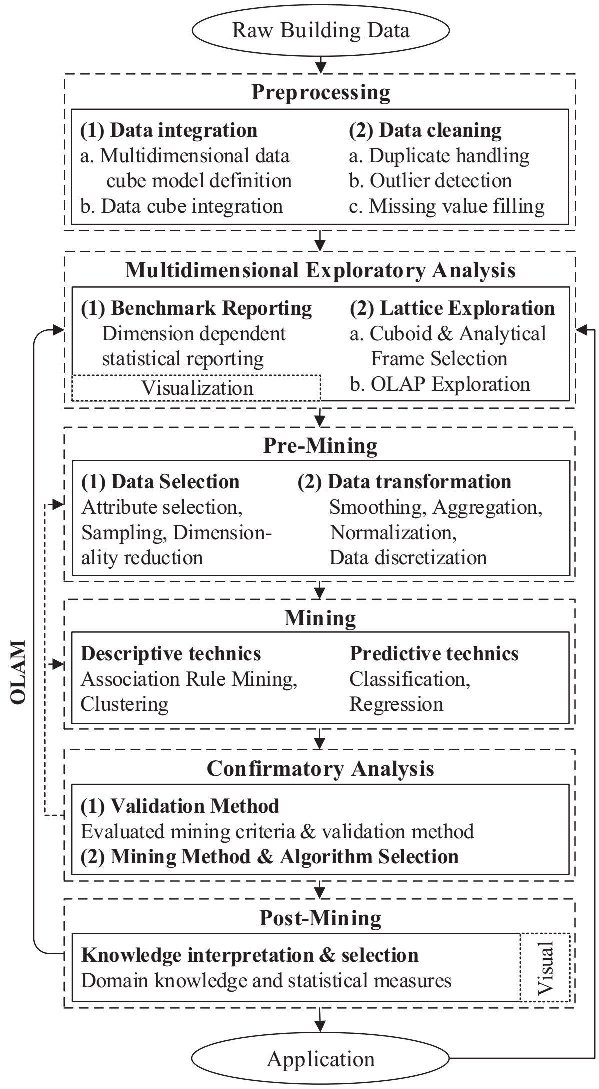
Raw Building Data
Preprocessing
(1) Data integration
a. Multidimensional data cube model definition
b. Data cube integration
(2) Data cleaning
a. Duplicate handling
b. Outlier detection
c. Missing value filling
Multidimensional Exploratory Analysis
(1) Benchmark Reporting
Dimension dependent statistical reporting
Visualization
(2) Lattice Exploration
a. Cuboid & Analytical Frame Selection
b. OLAP Exploration
Pre-Mining
(1) Data Selection
Attribute selection, Sampling, Dimensionality reduction
(2) Data transformation
Smoothing, Aggregation, Normalization, Data discretization
Mining
Descriptive technics
Association Rule Mining, Clustering
Predictive technics
Classification, Regression
Confirmatory Analysis
(1) Validation Method
Evaluated mining criteria & validation method
(2) Mining Method & Algorithm Selection
Post-Mining
Knowledge interpretation & selection
Domain knowledge and statistical measures
Visual
OLAM
Application

Fig. 1. A generic multidimensional data mining framework for building data.

relationship to time?", framing the analysis under the energy and time dimension, or "what was the total energy consumption of a building in a certain location during a specific time interval?", here querying along three dimensions, serves in the conceptualization of the state space to explore and, thus, in the definition of the data cube dimensions.

Building data gathers six types of recorded data, from building operations to metadata combined [33], echoing quite conveniently the 6 facets of a cube; i.e., time, location, building data encapsulating operational and *meta*-data, climate conditions, occupant related information and equipment data. Time data serve as a reference index to the other measured attributes, and indicates *Year*, *Month*, *Day*, *Hour*, *Minute*, *Second*, *Day Type*, generally formatted under the ISO 8601 [34] recognized complete format "YYYY-MM-DD**T**HH:MM:SS**Z**", e.g., 2019-07-16 T19:20:30+01:00. Location straight forwardly regroups spatial delineations such as geographic coordinates or address, which can be divided into numerous granularities, i.e., device, room, zone, system, building, street, district, city, state and country. Building data regroups building characteristics and operational data. Operations cover energy demands originating from building comfort maintenance with heating and cooling loads, lighting and ventilation systems through electric power loads, heat flows or natural gas consumption, but can also cover also water supply. Metadata evokes the building's physical characteristics commonly encasing floor area, number of floors, global insulation coefficient, window-to-wall ratio, date of construction and building type (school, dwelling, office building, hospital, education…). Climate conditions assembles indoor or external environmental conditions with attributes such as dry-bulb temperature, relative humidity, irradiance, wind-speed, precipitations, pressure and air quality but also non-temporal characteristics such as the Koppen climate classification [31]. Occupant information can deal with both occupant characteristics and comfort data. Occupant characteristics are seldom collected as a result of their privacy sensitive nature as well as tediousness to gather through costly surveys. They cover attributes of age, gender, education, lifestyle, annual income and other socio-economic parameters [35,36]. Occupant comfort data relate to physiological, psychological and environmental factors influencing human comfort perception, i.e. thermal, visual and aural comfort [37]. Equipment data possess a non-temporal and operational entity, namely equipment characteristics, i.e. equipment type, efficiency, capacity, and operational system settings, i.e. set-point temperature, inlet and outlet equipment temperatures or pressures, control parameters and events accompanied with eventual respective causes (human or agent initiators) [38].

Given these dimensions, one can group study-specific available data to form a dimensional mapping of the cube. Given an n-dimensional cube, each dimensional-element d_i, with $i \in [1, n]$, can thus be associated into groups of increasing size, i.e. cuboids. Cuboids of a given size compose a lattice, where each lattice $l \in [0, n]$ will thus be composed of $P_{n,l}$ possible partition cuboids from the below equation.

$$P_{n,l} = \binom{n}{l} = \frac{n!}{l!(n-l)!}$$

Fig. 2 illustrates a 4D cube mapping example given the building cube dimensions: time (T), resource consumption (R), External conditions (E) and location (L). The cube can then be reduced by eliminating non-relevant dimensional associations from analyst inspection. Here the 2D cuboid association $\{T, L\}$ can be eliminated as location is, by essence, non-temporal. Consequently, all emerging cuboids can also be eliminated from the cube space, resulting in a reduced state space mapping.

Establishing the cube data mapping provides a conceptual and structured model, dividing data into clear separate dimensions, on which the later analytical processing can be founded on. It may also be noted that the hierarchical structure inherent to some dimensions, e.g., time or location, possess abstraction levels called footprints which represent granularities accessible for later OLAP exploration of the cube space [22].

2.1.1.2. Data cube integration. Integrating the data cube to a suitable format for mining processing then follows. Building data are typically recorded in two dimensional tables where a set of attributes (columns) representing a variable are stored across different instances (rows). Within the defined dimensions stated earlier, different levels of measurements are often required for in depth building energy performance analytics, i.e. from building site scale to room, equipment or component-point measurements, increasing data dimensionality and complexity. For instance, HVAC systems often require multiple outlet temperature, pressure and airflow point measurements, with one aggregated component energy consumption. Differentiating these relationships in an ergonomic and analytically efficient way becomes crucial for effective DM. A prevalent adopted solution proposes common markup language and data structure to organize the collected information: Project Haystack [39]. Data are organized hierarchically from three entities, i.e. *site*, a single building with a unique street address, *Equip*,

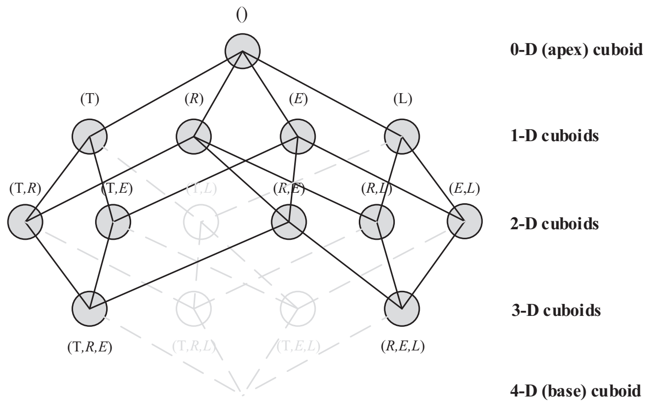

Fig. 2. Building 4D data cube mapping and space reduction example, where (T) relates to time, (R) resource consumption, (E) External conditions and (L) location.

physical or logical pieces of equipment within a site, and *Point*, referring to sensors, actuators or set point values of an equipment. Following this reference, a multi-column format is proposed where sets of attributes are grouped hierarchically under common sites, consequently structuring attributes from similar buildings under a common table.

2.1.2. Data cleaning

Data cleaning is a crucial step to efficient DM analytics aiming to improve data quality by dealing with duplicates, clearing outliers and filling missing values from raw BAS data. It is unfortunately still common to find little to no information on the data cleaning phase of many studies [31,40,41]. This first major phase of DM legitimately effects the outcome of the later analysis and should always be clearly reported to assure proper result benchmarking. To the best of the author's knowledge, every existing BAS analytics from the literature perform missing values filling prior to outlier detection, as a consequence of the few existing methods robust to missing values. This work introduces a shift in this established order to avoid using tampered sets for missing value filling which can result in a greater share of produced outliers, consequently making them harder to identify in the later step.

2.1.2.1. Duplicate data handling. Duplicates in data sets consist of data objects that are corresponding or identical to one another to some extent. In BAS data, these can consist of redundant attributes within a data set, or multiple attributes stored in a common instance (timestamp), sometimes with different values, also referred to as inconsistencies. Their sources cover use of denormalized tables, inaccurate data entry or updating some but not all data occurrences [22]. They can create major issues when merging data from heterogeneous sources and should be handled first within the cleaning phase. Duplicates handling is seldom depicted in BAS literature and usually consists of candid duplicate attribute, or instance, lookup functions coupled to targeted removals if the duplicates are identical. Handling inconsistencies however yields different alternatives, i.e., keep only one duplicate over the others, average the inconsistencies out or remove them from the data. Knowledge on the origin of a set of duplicates can help identify erroneous data and chose an appropriate strategy for duplicate handling.

2.1.2.2. Outlier detection. An outlier can be defined as data point that is significantly dissimilar to other data points or that does not imitate the expected behavior of others [42,43]. In BAS data, outliers can come from measurement faults (sensor), transmission or transcription anomalies due system changes or human errors. Natural outliers reveal unusual but occasional behaviors of the monitored phenomenon. Outliers can be grouped in two main groups, i.e., point and subsequences outliers [44]. This phase of DM should only consider point outliers identification as recommended by the work of Fan et al. [45], not to later overlap with mining typical/atypical patterns [46]. Outlier detection methods include prediction models, profile similarity approaches and deviants identification [44]. Prediction models spot-out outliers by comparing measured values from predicted ones with an outlier score threshold based comparison. The primary variation across models concerns the particular prediction model considered (supervised, unsupervised). The profile similarity approach is based on a reference normal profile built upon historical data to which new time points are compared to. Outliers are then identified from time dependent normal profiles and variance vector comparison with anomaly score. The deviant based method estimates outliers from a minimum description length (MDL) standpoint originating from information theory. If the removal of a point in a time sequence results in a significantly simpler sequence to describe, then it is considered an outlier.

2.1.2.3. Missing value filling. Missing data in BAS are commonplace, with multiple processes being monitored from seconds to hourly frequencies on a yearly basis, gaps are typical within raw BAS data. Missing data originate from error or omissions when data is recorded or transferred [47], imperfect procedures of manual data entry, incorrect measurements, and equipment error [48]. Discontinuities may lead to serious obstacles when analyzing findings [49], e.g., loss of efficiency, complications in data handling and analysis, bias estimates from dissimilar lengths of data and reduction of statistical power (inefficient estimates) [50]. Selecting an appropriate method for missing data handling depends on the time series pattern and the missing data mechanism [51]. Challenges related to these techniques involve, maximizing available data use to preserve covariance structure in multivariate data [52], and incorporating variance estimates of the uncertainty rooted on imputed data [53]. If the gaps represent more than 60 percent of the set however, then no method is judged suitable to cure the set [48]. Missing value filling methods cover either deterministic approaches, known as single imputation, or stochastic ones, also referred to as multiple imputations, where several values are generated for each missing observation to reflect the uncertainty of the missing data [49]. This work proposes a single imputation approach dependent on the length of the missing sections. Explicit modeling with regression can be chosen for missing data sections smaller than 3 consecutive hours, i.e., moving average [45], while longer sections call for implicit modeling using the hot deck method, i.e., where missing values are averaged from identical time intervals and day of the week using sections of 2 weeks. Interested readers are invited to refer to the work of M. Norazian Ramli et. al. [54] for in depth review of imputation methods.

2.1.3. Multidimensional exploratory analysis

This section intends on framing multidimensional data exploration leveraging the BAS data cube representation. It holds the essential role of identifying data structures, distributions and trends, needed for benchmarking purposes and defining appropriate mining approaches for the investigated set. Additionally, it supports more generalizable, interpretable and framed analytics by (*i*) cutting down the complexity of big data from cube lattice selection with OLAP exploration and (*ii*) putting forward important benchmark reporting characteristics. Exploratory Data Analysis (EDA) was originally defined by John W. Turkey as the act of "looking at data to see what is seems to say" [55,56]. It aims at collecting insights into data characteristics to help with the following analysis by answering questions such as; what does the data look like? How can one visualize the data to get a better sense of it all? How are the values distributed and can similarities between attributes be measured [22]? Existing explored characteristics comprise attribute types, i.e., nominal, binary, ordinal, numeric, discrete or continuous, and statistical descriptions, i.e., central tendency, dispersion, variance and correlations. Attribute type exploration is carried out during the first data integration phase, however statistical feature inspection can be performed a priori or posteriori, to data cleaning. EDA is here presented a posteriori to data preprocessing in the DM framework as a necessary step to encase multidimensional mining. First benchmark reporting presents the dimension-specific data structures, providing necessary insights to the later pre-mining phase to which follows, lattice/cuboid selection and OLAP exploration.

2.1.3.1. Benchmark reporting. EDA serves as a necessary data structure reporting appliance to any scientific study. As the work of B. Yildiz et. al. demonstrates, BAS data characteristics should systematically be described to allow validation of a study's true success thanks to a defined analytical framework [57]. Yet, too many studies fail to report these features. Description of household characteristics such as dwelling types, age and physical condition, household loads statistical components and climatic conditions using established classifications, e.g. Koppen Climate Classification [58], should henceforth systematically be reported [57]. While undertaking EDA, it becomes necessary to define what the authors propose to call the *analytical window frame* which encompasses three elements, i.e., data granularity, horizon and frame. Granularity refers to the sampling rate of the data set over the selected dimensions. The horizon entails the largest dimensional attribute considered and the frame defines the dimensional region of interest within the analysis. For example, typical building energy pattern analytics tend to use temporal analytical windows with 15-min to hourly granularity, yearly horizon and daily frame [59].

Statistical features examination is often performed through data visualization. It communicates data structures and tendencies clearly and effectively from graphical representation endowing users with a straightforward understanding of the data [22]. A series of three data visualization techniques are hereby presented capturing the data dimensions' inter- and/or intra-attribute structures.

(i) Combined *half-violin* and *boxplots* allow appreciation of central tendency, distribution and variance with an assessment of statistical inference at a glance via overlaid boxplots [60], while avoiding the redundant mirrored probability density functions of violin plots.
(ii) *Scatterplot matrix* coupled to *correlation matrix* display the marginal dependence structure of the data [61], granting examination of intra-attribute correlations and attribute distributions from bivariate relationships. These plots are favored as particularly effective for feature engineering and visualization [62].
(iii) Weekly framed *heat maps* are suggested as a substitute to *run charts*, enabling inspection of per attribute patterns leveraging a weekly to daily analytical frame of interest using hourly resolution and yearly horizon.

2.1.3.2. Lattice exploration. Lattice exploration embodies the starting step of the iterative cube-space mining, i.e., OLAM loop [28]. A cuboid is firstly chosen from the established data cube dimensions for OLAP exploration of the multidimensional data. For instance, given a 3-D building data cube covering *time*, *site* and *attributes* dimensions, see Fig. 3, three sets of 2-D cuboids can be iteratively selected and explored, i.e. {*time, site*}, {*time, attribute*} and {*site, attribute*}. Typically, analytical frames explored in building performance mining encompass only one of the three presented cuboids, i.e., *top-down*, *bottom-up* and the less common *temporal drill-in* approach, respectively corresponding to cuboids A {*time*, *site*}, B {*time*, *attribute*} and C {*site*, *attribute*}. By examining varying levels of abstractions through lattice exploration, information and insights sharing between them can be exploited; a concept also employed in transfer learning [63]. C. Fan et. al. [64] recently demonstrated its value across buildings for short-term building energy predictions, particularly when measured data are limited. Multidimensional mining can exploit these varying levels of data abstractions from drilling, pivoting, filtering, dicing and slicing of the data cube. Leveraging data visualization to these ends notably expands the power and flexibility of data mining [22].

2.2. Pre-mining

Pre-mining is by nomenclature the phase completed prior to mining. Customarily, this process is treated within preprocessing as it shares the objective of preparing the data for mining [7,22,30,33,45,65]. This study suggests differentiating application-independent steps from the dependent ones and introduces pre-mining in the DM framework as a mining-specific preprocessing phase which can be iterated over in response to confirmatory analysis results. Pre-mining englobes two principal functions, i.e. data selection, for targeted and computationally efficient mining, and data transformation, to prepare the data to a suitable type and range for mining.

2.2.1. Data selection

Data selection, also referred to as data reduction, answers to a necessary step in big-data mining originating from the sheer volume and high dimensionality of the data. Indeed, addition of data volumes from keeping irrelevant attributes or loss of decisive information from withdrawing relevant ones will likely be detrimental to the mining process; it may slow the mining algorithm employed, while leading to discovered patterns of poor quality [22] and has been recognized to play an equally important role as ML model development throughout the pipeline of DM [62]. To that end, data selection encompasses measures that are attribute selection, sampling, and dimensionality reduction techniques.

Attribute selection proposes to straightforwardly reduce the data set dimension by removing irrelevant or redundant attributes (or features). Note that this process can also involve the creation of new attributes, from combined information of removed ones. Its aims at finding a minimum set of attributes while keeping the original probability distribution of the classes as unaffected as possible [22]. Feature selection is commonly conducted by sequential backward selection (SBS), where attributes are sequentially removed till the reduced space contains the desired number of features [62]. Existing techniques commonly evaluate and rank individual or subsets of data attributes, e.g., information gain attribute ranking, relief, principal component, and correlation-based feature selection [66].

Data sampling involves using statistical techniques to select, manipulate and analyze a representative (sub) set of data, usually resulting from the need to reduce the size (dimension) of the enormous data set considered, i.e. under-sampling. Over-sampling on the other hand is less frequent within the big data era, as a result of the overabundance of already collected data. Yet, over-sampling can be used to test the robustness of mining results and highlight sensitivity of the approach to sampled realizations.

Dimensionality reduction techniques, or data reconstruction methods [7], serve as a means to reach reduced representations of the data while minimizing information loss [22]. Main tech-

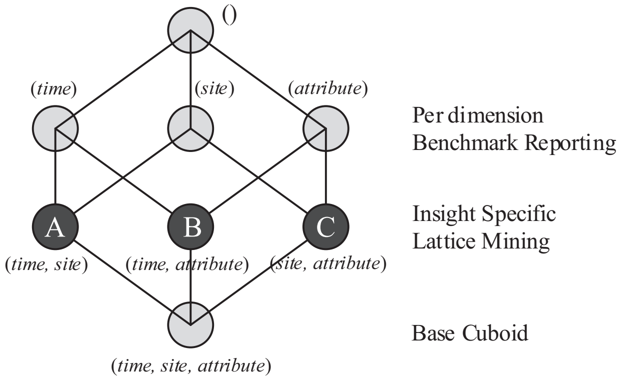

Fig. 3. Building 3D data cube mapping with Benchmark reporting and lattice exploration.

niques include wavelet transform, which provides high and low frequency decompositions of signals based on wavelet approximation coefficients [67], and principal component analysis, where low-dimensional attributes are created from orthogonal linear transformations of the original high-dimensional ones.

2.2.2. Data transformation

Data transformation addresses data conversion to suitable types, ranges and noisiness to serve as DM algorithms input. Indeed, depending on the mining technique considered different data formats are required, e.g. categorical or numerical, while BAS data can exhibit varying units, scales and data type [45]. To this end, this phase covers data normalization, aggregation, smoothing and discretization.

Normalizing a time series consists in scaling its attributes within, or around a smaller range or value, typically [-1, 1] or [0,1]. This step is commonly performed to allow scaled comparisons between dissimilar ranges of attributes, e.g. normalizing features allows balanced contributions in the update of model weights during the training phase of predictive learning. Typical normalization methods cover min–max, z-score and decimal point normalizations [22].

Aggregation similar data groups together, also known as binning or bucketing, consists in applying summary operations to the data. For instance, sample-rate conversions resample the data by aggregating values together at regular instances, i.e., down-sampling, where daily intervals are reduced to weekly or monthly ones. The reverse operation, interpolates data across larger resolutions, e.g. up-sampling to convert hourly instances to 15 min interval ones.

Smoothing serves to remove noise from data which is frequently used to uncover trends in noisy time series and can alleviate overfitting pitfalls of regression models. Usual techniques include binning, with either equal width or frequency, regression or clustering [22]. It can be interestingly noted that the previously presented down-sampling rate conversion method, can also achieve smoothing effects as a data binning technique.

Discretization involves data type transformations such as converting numeric features to interval or conceptual labels [22], e.g., 30–50, 50–70 or *adult*, *senior*, respectively. This step is consistently required for mining algorithms such as Association Rule Mining (ARM), i.e., the frequent-pattern growth and Apriori algorithms that can only handle categorical data [7].

It should be noted that feature construction relates more to data transformation, as the work of J. Han et al. reports [22]. Yet, because this framework treats the ordering in which these steps should be taken, it was chosen to include it within feature selection, to apply data transformation techniques a posteriori to feature engineering.

2.3. Mining

Mining, or knowledge discovery, encapsulates the algorithmic mining of the data which entails a large number of varying techniques. Selecting the appropriate one for a given application is part of the difficulties most data practitioners are faced with and is, naturally, function of the nature of the problem and the given data set, or case study. Going towards more interpretable DM analytics, we propose to group these techniques in two application-oriented groups, i.e., descriptive and predictive techniques. These groups echo the, well established machine-learning families that constitute unsupervised and supervised learning respectively, while clearly distinguishing the typical end goals one can expect from such methods. It is beyond the scope of this study to give a complete review of all existing DM methods, however principal mining groups will here be revised.

2.3.1. Descriptive techniques

By definition, descriptive techniques are diagnostic application-oriented analytics and intend on achieving a better understanding of the causes of a given process, i.e., identifying patterns or abnormal behaviors. Descriptive DM techniques, as opposed to predictive ones, have been judged more capable at discovering previously unknown knowledge from BAS data [7]. Descriptive techniques cover the important DM groups of clustering, and Association Rule Mining (ARM). Clustering is the process of grouping a set of data objects (or observations) into subsets or clusters. Each object within a cluster is similar to one another, yet dissimilar to objects in other clusters. Similarities and dissimilarities are assessed based on attribute values describing the objects and often involve distance measures, or metrics [22]. Clustering algorithms have been broadly applied to identify typical building operation patterns, e.g., building energy demand patterns, indoor environment distribution and building energy system operation patterns. Main clustering algorithms involve k-means clustering with many variants including adaptive k-means and k-shape clustering, Fuzzy C-Means (FMC), support vector clustering, hierarchical clustering or decision tree-based clustering and Self-Organizing Maps (SOM) [59]. ARM is a powerful tool designed to extract association rules amongst attributes from large amounts of operation data. Association rules are commonly an implication of the form “A → B”, where A is defined as the antecedent and B the consequent. In general, ARM can be viewed as a two-step process where all frequent item sets are firstly identified, from which strong association rules from the frequent item sets satisfying minim support and confidence can then be generated [22]. Variations of ARM recently applied in BAS data include Temporal Association Rule Mining (TARM), or sequential rule mining, to encapsulate the temporal dimension within the discovered rule. Common ARM algorithms encompass TRuleGrowth, Weighted ARM, QuantMiner, Apriori, ParaMiner and CloseGraph. Some notable TARM algorithms include TRuleGrowth, SPADE and CMRules [45].

2.3.2. Predictive techniques

Predictive mining intends on determining the likelihood of future events from historical data. It constructs a model, or function from the analysis of sufficient numbers of training sets, i.e., data objects for which the desired output is known. Predictive DM is often employed to capture complex and nonlinear relationships between inputs (independent) and outputs (dependent variable) of an observable phenomenon [45]. It is then employed to predict the discrete or continuous value of observations yet unforeseen. Familiar DM predictive techniques comprise regression and classification based methods. Regression analysis is a statistical methodology most often applied for numeric prediction of missing or unavailable data values. It also covers the identification of distribution trends from available data [22]. Methods include Artificial Neural Networks (ANN), deep neural networks, Support Vector Machines (SVM), Decision-Trees (DT), Genetic Algorithms (GA) and ensemble learning [68]. On the other hand, classification predicts categorical labels (unordered, discrete). Classification forms an analysis that identifies a model describing the data into distinguishable classes or concepts. The models are built on targeted attributes fitted to the value of predictor attributes. Data classification aims to classify data into distinct predefined classes, providing the description categorization and generalization of a given database [31]. It includes algorithms such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Decision-Trees (DT), Bayesian Network (BN) and ensemble models, i.e., random forest [69].

2.4. Confirmatory analysis

Confirmatory analysis provides answers to the questions of model accuracy estimation; "what are appropriate measures of a model's *goodness*?" and, if there are multiple models to choose from, "how to selection the best model?" from them. These enquires relate to method validation, and model selection respectively. This phase embottles the two earlier pre-mining and mining phases and constitutes the keystone of this iterative process. It defines the method around which the mining will be performed, and consequently arises as a founding phase of the analytical approach. Confirmatory analysis in DM echoes the statistical process of evidence evaluation from significance, inference and confidence tests; it is the phase where findings and arguments are put to trial. This phase is usually implicitly included to the earlier mining step and explicitly framing such a key step of the mining process is becoming imperative to approach more interpretable mining analytics. The confirmatory analysis step ergo includes validation method and model selection.

2.4.1. Validation method

Determining *what* decides the goodness of a mining's process and *how* to assess it is by all means what the validation method deals with. The *what* serves to quantify the evaluated characteristics, commonly employing performance metrics. These characteristics can cover speed, robustness, scalability, interpretability and mining-dependent validity indicators, e.g., purity, similarity index, or accuracy [22]. The *how* works towards obtaining representative evaluated characteristics and assures reliable results are obtained. Common methods employed for model assessment contain cross validation, bootstrap, sensitivity analysis and hyper-parameter tuning. Defining how a mining model is validated and under what criteria is the fundamental foundation of any mining process.

2.4.2. Algorithm selection

With evaluation characteristics, metrics and validation method defined and undertook, model and associated parameters are selected from multi- or single-criteria assessments. Most DM work evaluate model quality from one criterion at a time such as accuracy or interestingness with a single-criterion assessment [70]. Some works have proposed multi-criteria evaluation methods to combine multiple measures in their model selection. The review of Aruldoss et al. cover a few of them, namely fuzzy and non-fuzzy analytics hierarchy process, TOPSIS, grey theory, data envelopment analysis, weigthed sum models [71]. Panapakidis and Christoforidis have notably developed a multi-criteria decision method for optimal selection of clustering algorithms applied to load profiling applications [72].

2.5. Post-Mining

Post-Mining intends on bridging practical applications with mined discovered knowledge. This step requires domain expertise for knowledge selection and interpretation which can become particularly time consuming [7,23]. Knowledge selection can refer to varying application-dependent processes, e.g., characterizing identified load profiles from clusters [73,74], or selecting relevant rules for interpretation amongst massive ARM outputs [45]. Typical end-use applications of building data mining englobe building energy load prediction, predictive maintenance, fault detection and diagnosis, building performance analysis and energy management optimization.

3. Implementation

We implement the given method on an established automated pattern filtering application for building performance analysis proposed by Miller et. al, namely *DayFilter* [23]. Using time series Symbolic Aggregate approXimation (SAX) daily profiles are first segmented into W equal sized segments, piece-wise approximated across each of these segments and finally transformed to alphabetic characters according to a chosen alphabet size A, creating breakpoints of equiprobable regions from Gaussian distribution. For example, fixing $W = 4$ and $A = 3$ could produce the sequence '*abca*' where each alphabet character would correspond respectively to a '*low-medium–high-low*' segment values. SAX transformation results in reduced representations of daily profiles allowing computationally efficient differentiation of daily motifs from discords. The method steps are visually presented under Fig. 4, where the arrows in the diagram designate the sequence execution flow from steps 1 to 6. The iterative cube space OLAM process is repeated within steps 3 to 6, where each cuboid selection orients the analysis to either prevalent top-down, bottom-up or temporal drill-in approaches. The complete code implementation of the reported study can be found under the open github repository *https://github.com/JulienLeprince/multidimensional-building-data-cube-pattern-identification*, for more interpretable as well as transparent knowledge transfer.

We use the open data set from the Building Data Genome project 2 (BDG2) [75]. This open set was chosen to allow reproducibility of the given analytics while illustrating how large open source reference sets can be beneficial for DM analytics benchmarking. The BDG2 includes 3053 energy meters from 1636 non-residential buildings located in Europe and, principally, North America. The set covers two full years (2016–2017) at an hourly resolution with multi-meter building measurements as well as weather and building *meta*-data.

We consider a simple 3D data cube regrouping dimensions of {*time*, *site*, *attribute*}, recall Fig. 3, to illustrate the given pattern identification. The {*attribute*} dimension encapsulates weather and building meter-data combined to allow a simplified space cube mapping where its 2D lattice echoes typical study frames of inter- and intra-building analytics. The inter-building analytical frame, i.e. A={*time*, *site*} cuboid, is typically relevant for building stock diagnosis or benchmarking given a fix attribute, while the intra-building frame, i.e. B={*time*, *attribute*} cuboid, serves for within-site diagnosis for the selected building across time. The rather unfamiliar C cuboid regrouping {*site*, *attribute*} dimensions, allows diurnal drill-in exploration of cross-building/attributes combined information from a certain time slice of interest. To grasp the complexity endowed with high dimensionality while keeping the use case relatively simple, we select temporal attributes of electricity, gas, hot water and chilled water of site meter energy consumptions

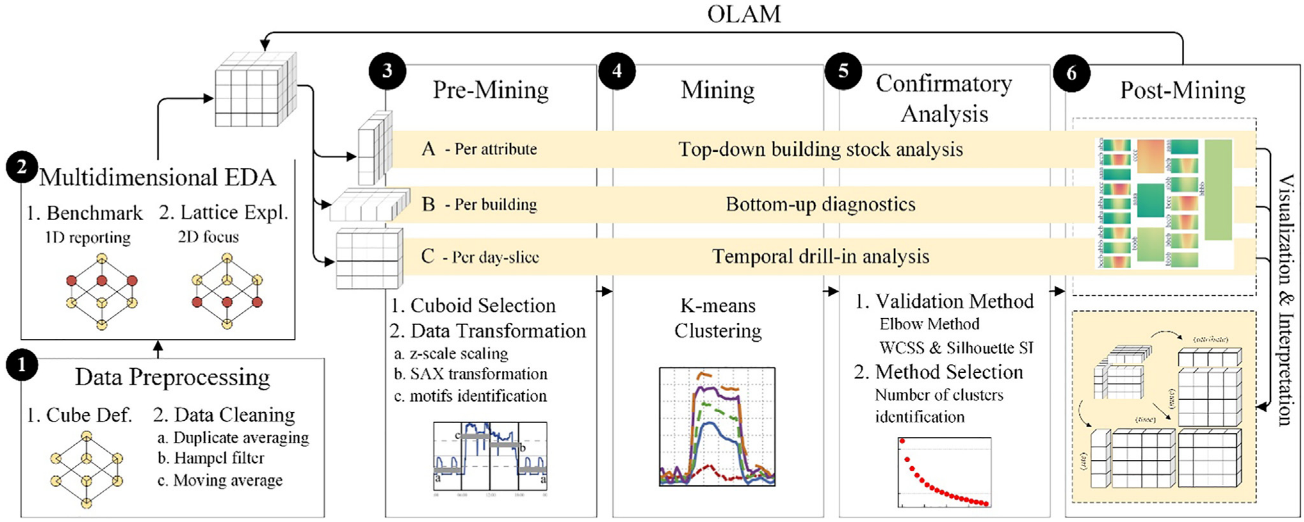

Fig. 4. Diagram of three-dimensional cube space SAX pattern mining steps.

and pair them with external condition attributes of air temperature and sea level pressure.

Meter data contained approximately 4.74%, 6.20% and 6.94% of missing values for electricity, hot and chilled water respectively with maximum lasting periods ranging between 0.25 and 1.5 days, lower and upper quantiles for electricity respectively. First the Hampel filter, an outlier robust rolling window method, was applied to detect point-wise outliers with a window size of 6 time-steps (hours) and a standard-deviation threshold of 3 [76]. A moving average was then used to fill in missing data points for consecutive gaps smaller than 4 h. Greater gaps were averaged from identical time intervals and days of the week using sections of 2 weeks. The dimension dependent EDA of the building stock meta-, weather, and meter-data can be found under the publication's GitHub repository [75] and will thus not be repeated here. The 2D lattice is then selected for cube-space exploration as it encompasses typical study frames while paving the way to the dimensionally more complex 3D base cuboid.

The following mining steps treat the OLAM iterative process of data selection, transformation, clustering, validation and knowledge interpretation, over the selected 2D lattice of the cube. The time series are first normalized through a z-scale transform to obtain an approximate mean of 0 with standard deviations approaching 1 [77]. Echoing the work of Miller et al. [23] we do not normalize the series based on individual sub-sequences and rather take the full temporal scope of the time series into consideration. This allows us to discover patterns leveraging both the magnitude and shape of the original profiles revealing the seasonality within the series. SAX transformation then serves as a blended data dimensionality reduction, aggregation and smoothing technic. It is performed over the time series considering segments W = 4 and alphabet size A = 3. This selection of SAX parameters is driven by the desired signal granularity and coarseness of the reduced time series approximation. More detailed patterns could be generated with increasing segment and alphabet size, however ensuing the findings of Miller et al. these parameters have been found to provide the best balance between the number of patterns generated and detailed resolution required to filter discords in a diurnal frame context. Heuristically, we set an absolute count threshold at 10 to filter motifs from discords, to which succeeds Euclidean distance-based K-means clustering, further reducing the pattern groups. We validate the optimal number of clusters through an elbow method assessment using two similarity indexes, i.e. within-cluster sum of squares (WCSS) and silhouette score. Finally, we present results using expressive visualization tools, allowing human result inspection for efficient and interpretable knowledge extraction. Diurnal heatmaps were retained as a particularly impacting visualization plot allowing 3 dimensional inspections within a 2D domain.

The below sub-sections present the cuboid-specific analytical particularities & diagnostic focus of the evoked pre-mining, mining, confirmatory analysis and post-mining steps, i.e., building benchmarking, in-site view and temporal drill-in analysis, with a closing multi-cube-space visualization interpretation step.

3.1. Building benchmarking

A high level top-down building stock diagnosis is first undertaken to gather insights from building energy consumption profile ranges and orient the subsequent lattice exploration. The {*time*, *site*} dimensions are sliced from the building data cube, and the site electric meter consumption attribute is chosen as both a conventional and representative energy resource consumption metric. Time series normalization is performed per *site* and across *time*, capturing each buildings' energy consumption profile shape and seasonal range. SAX sequences are obtained from the z-scaled time series and discords are filtered out from the settled count threshold. Fig. 5 presents the motif counts obtained across the stock, from which the most frequently observed motif count is *aaaa*, a constant low to null consumption steady-state sequence accounting for 12.58% of overall building stock SAX sequence counts. To group buildings into similarly operating clusters we leverage the dimensionality reduction brought by the SAX transformation and alleviate the computational burden that would follow undertaking clustering on the original daily profiles. We consequently perform clustering on the motif sequence cumulated *counts* across the building pool. This allows to group buildings together based on their motifs distribution across the entire time horizon considered, while being very computationally light. A limitation of considering solely motif counts in the clustering process is however that the similarity between sequences is not accounted for, e.g., *aaba* will be considered as different from *ccbc* as *aaaa*, although it is naturally much closer to the later. While the authors are conscious of this limitation, it is beyond the scope of this article to develop a clustering method accounting for SAX sequences similarities. From the confirmatory analysis results presented in Fig. 6, we fix the optimal number of clusters as 6; a value showing a peak in silhouette score, indicating a slightly higher cluster cohesions, while dis-

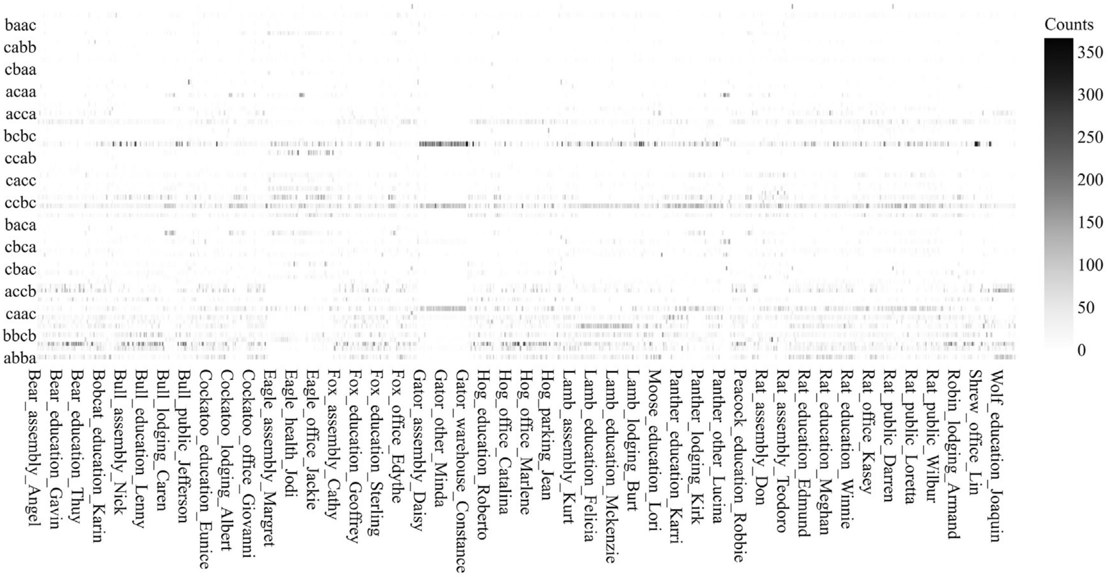

Fig. 5. Heatmap of electric meter consumption SAX counts accross building stock. An interactive visualization version of this figure is available via the public github repository of this study.

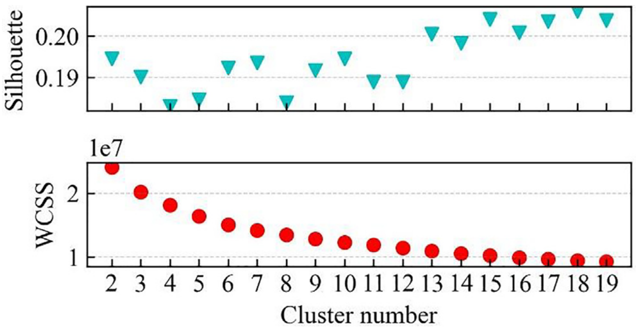

Fig. 6. Cluster similarity index assessement of cross-building stock from electric-meter SAX motif counts.

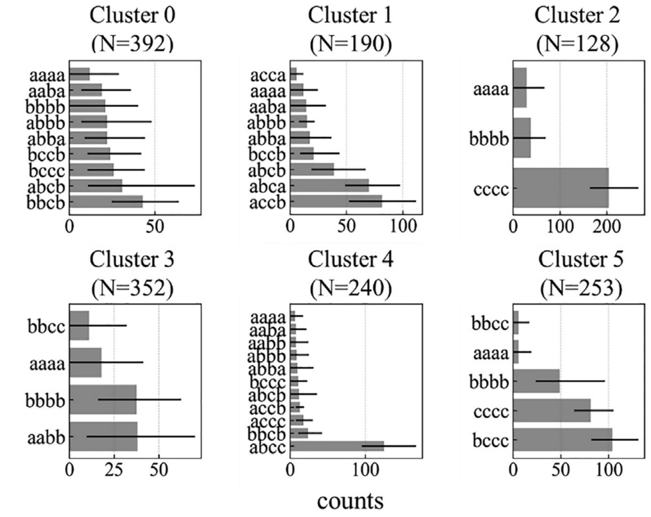

Fig. 7. Building stock electric-meter SAX motifs distribution across clusters. Bar plots represent the sequence count median value while the error bars indicate the lower and upper quantiles.

playing a sufficiently lowered WCSS and acceptably large number of clusters. The clustering results present the distribution of motifs across the obtained clusters under Fig. 7. Cluster 2 stands out as being composed solely of flat daily profiles from either *aaaa*, *bbbb* or *cccc* sequences. These electrical yearly consumption typically point to constant daily rule-based operationally controlled buildings, here mainly present in education, office, assembly and public type buildings, as Fig. 8 shows. Clusters 3 and 5 behave quite similarly, with predominant flat profiles and low numbers of different sequences across the time horizon. Cluster 4 presents a variety of patterns yet with a clear predominant *abcc* sequence across the temporal horizon. Clusters 0 and 1 both show a diversity of profiles, although cluster 1 presents less variability in SAX counts, a likely consequence of it being less populated than cluster 0, the most populated of all 6 groups, collecting close to 400 buildings.

We turn our attention to cluster 0, the most populated cluster of the six, also presenting an interesting variety of motifs across both evaluated years. In the following sub-section, we explore a within-building analytical frame and switch to cuboid B for an in-site, bottom-up, analytical approach.

3.2. In-site view

This analytical frame follows more closely the presented *DayFilter* process of Miller et al., yet extends it with a multi-attribute temporal exploration through the {*time*, *attribute*} cuboid, further bridging the gap between top-down and bottom-up approaches. We explore buildings grouped within the afore-determined cluster 0, thus substantially reducing the initially considered cube-space, and iteratively slice the reduced cube by selecting individual buildings on which to perform automated pattern filtering. Time series are scaled within each {*attribute*} dimensional-frame, transformed to SAX sequences and filtered for motif identification using the same formerly presented process. Fig. 9 presents the sax-grouped daily profiles of building *Fow_education_Melinda* in the form of a cascade heatmap, where motif sequences are explicitly tagged. This allows impactful visual representation of the profiles, as well as their per-sequence group size, in a cross attribute context. Links between distributions of patterns from one dimensional attribute to the other can thus be visually explored, e.g. two largest presented sequences within air temperature across the time horizon are *cccc* and *aaaa* (can also include *aaba*), referring quite straight-

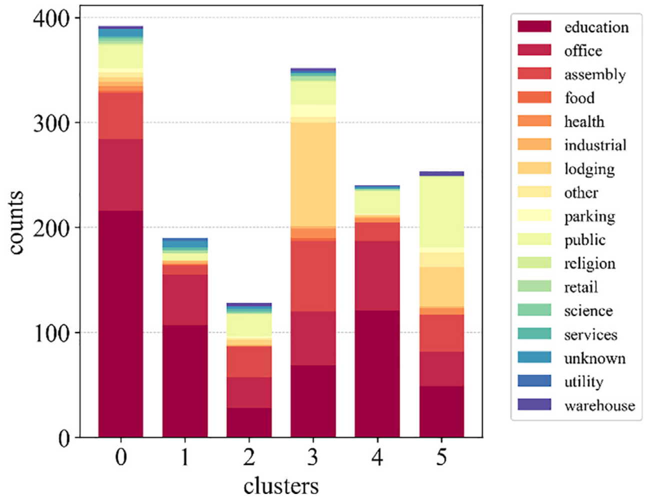

Fig. 8. Building type distribution across identified clusters.

forwardly to typical winter and summer periods, while chilled water possess two similar principal groups, i.e., *cccc* and *aaaa*, hinting to these identical seasonal periods. While this visualization display is powerful, the complexity endowed from exponentially increasing association possibilities between cross-attribute motifs can be limiting for human inspection. The dimensionality reduction provided by the ensuing clustering step takes up this problem, in an {*attribute*} dimensional-frame. After visual inspection of the confirmatory analysis' similarity indexes presented in Fig. 10, we fix the cluster number across attributes to be 4. This value shows low WCSS range and norm while serving as an acceptable trade-off between a reduced number of pattern groups and sufficiently high group variety for detailed attribute pattern characterization. The clustering results depicted under Fig. 11 exhibit close to homogeneous cluster sizes for air temperature alluding to the four seasons of temperate climate zones. Hot and chilled water meter patterns seem to behave in a mirrored fashion with consumption peaks and drops located in either mornings and evenings or evenings and mornings respectively. The electricity meter group size repartition seems closer to hot water consumption for this education building which both seems to testify on the building's operational activity; with three clusters presenting strong daily trends and one close to null consumption, hinting at weekend and holiday-type profiles.

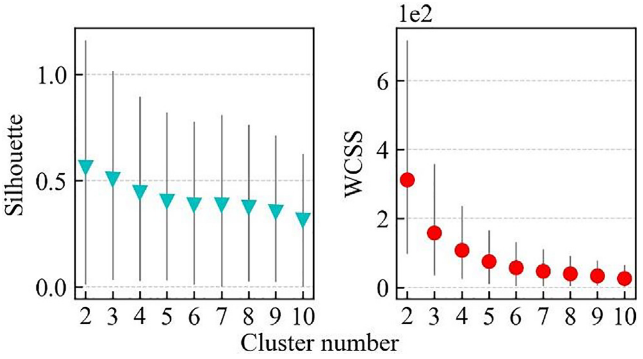

Fig. 10. Cluster similarity index assessement of cross-attributes from Fow_education_Melinda building diurnal motifs. Scatter points illustrate median values of the evaluated similarity index across attributes, while the error bars cover the upper and lower quantiles, representing value variance.

As we inspect the temporal depth of the cuboid and how attribute patterns are cross-distributed, our final cube-space exploration step approaches temporal drill-in analysis, where the complexity of multi-site, multi-attribute dimensions in a set time–space frame is examined.

3.3. Temporal drill-in analysis

We investigate the {*site*, *attribute*} dimensions of cuboid C from iterative daily slices of the building data cube. Supported by the temporal cross-attribute exploration of the former cuboid B, we target the most represented cluster group within the temporal dimension, i.e. a typical day within the summer season, below illustrated by the selection of the day 2016–06–07. Selected daily

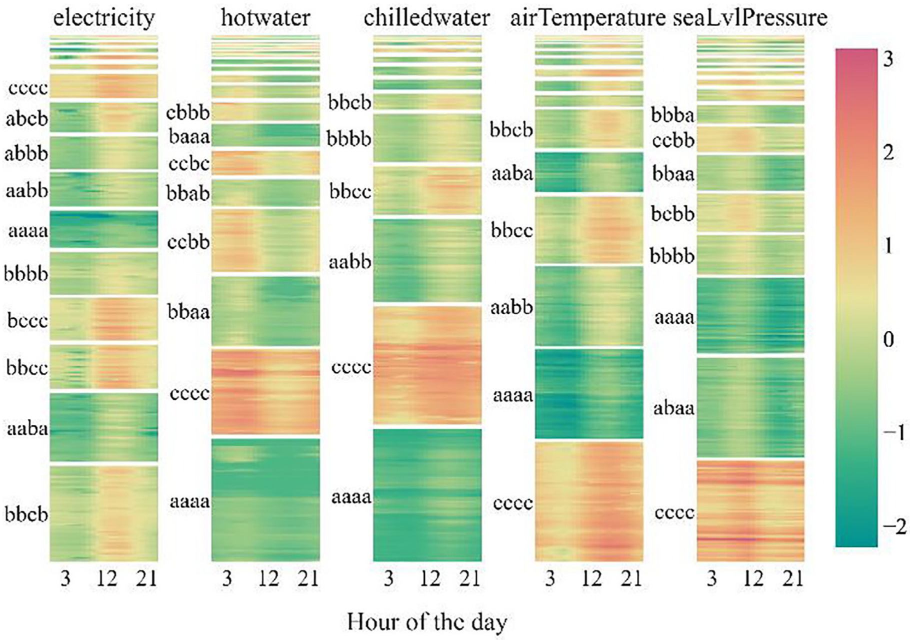

Fig. 9. SAX sequences across Fow_education_Melinda attributes illustrated by daily heatmaps normalized per attribute. SAX motifs are explicitly referenced while discords are ploted but not tagged.

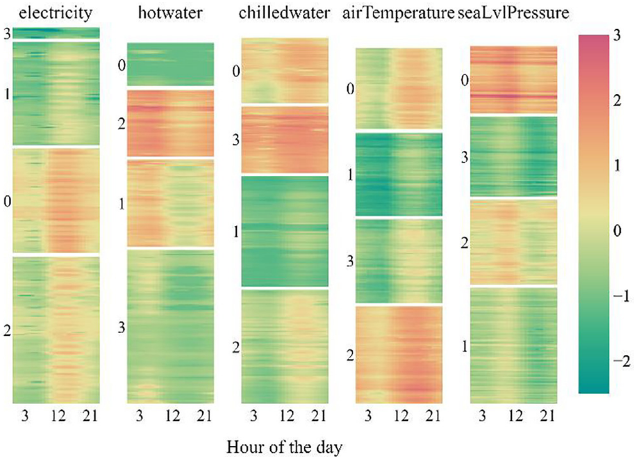

Fig. 11. Dirunal pattern clusters across Fow_education_Melinda attributes illustrated by daily heatmaps normalized per attribute.

profiles are z-scaled per attribute, across the building stock then SAX transformed, resulting in a singular daily sequence per cuboid dimensions. Given the lack of temporal depth of the sequences, notions of patterns and discords become meaningless along the {*time*} dimension. We therefore divert these notions to the {*site*} dimension, where buildings would be examined across their attributes as either behaving similarly to other buildings or not, i.e. motifs and discords respectively, given a certain threshold. We enumerate buildings displaying similar cross-attribute sequences and consider motifs for groups larger than 5 members. Fig. 12 presents aggregated daily attribute values annotated with SAX sequences and building group motifs member counts. From this cross-sectional view, it can be seen that the three most important aggregates possess only electrical meter data with SAX sequences of the three constant *aaaa*, *bbbb* and *cccc* profiles. Discord buildings are filtered out following which clustering can be performed from a weighted average of the daily multi-attribute time series.

Attribute averaging weights

Attributes				
Electric	Hot Water	Chilled Water	Air Temperature	Sea Level Pressure
0.7	0.1	0.1	0.05	0.05

Weights were designed to favor resource energy consumption data from weather conditions. In particular, electric meter was weighted as the preponderant attribute accounting for 70% of the time series weighted average, as reported under Table I. The reduced one-dimensional daily-series are then clustered across sites. Selection of the optimal number of cluster is performed from visual inspection of the confirmatory analysis results presented under Fig. 13. We select this number to be 4, given an over average silhouette score of 0.6 and a flattening WCSS trend. Clustering results are delineated under Fig. 14 in the form of quantile-profile heatmaps for each attribute across the four obtained building clusters, granting a cross-{*site*, *attribute*} dimensional inspection of patterns. The larger building cluster aggregates a total of 813 buildings together while the smaller one only 19. Electrical patterns across the stock seem to follow overall comparable trends, with consumption increases and drops ranging between 6 and 10am and 7–9 pm for their lower and upper quantiles respectively. This comes as a surprising read given the prevalence of the constant SAX sequences previously mentioned and could appear as a notable pitfall of the quantile heatmap visualizations, which solely show hourly quantiles across the cluster instead of original daily profiles. Quite similarly to the previously observed finding within cuboid B, chilled water appears to be positively correlated to outside air temperature across the building stock, with similar daily-temporal tendencies, i.e., lower morning values increasing from 10am, peaking around 2 pm and decreasing in the evening with ranges from 6 to 9 pm. Hot water, for the two smaller clusters, behaves in reverse to chilled water, with a prominent daily peak in the early morning, suggesting bathroom hot water consumption, while the larger N = 813 building group possess a very flat to null consumption over the day, with faint lower and higher demands in the morning and evening respectively. Finally, the temperatures patterns across this selected day are quite typical of warm summer seasons with steep morning increases and smoother afternoon decays.

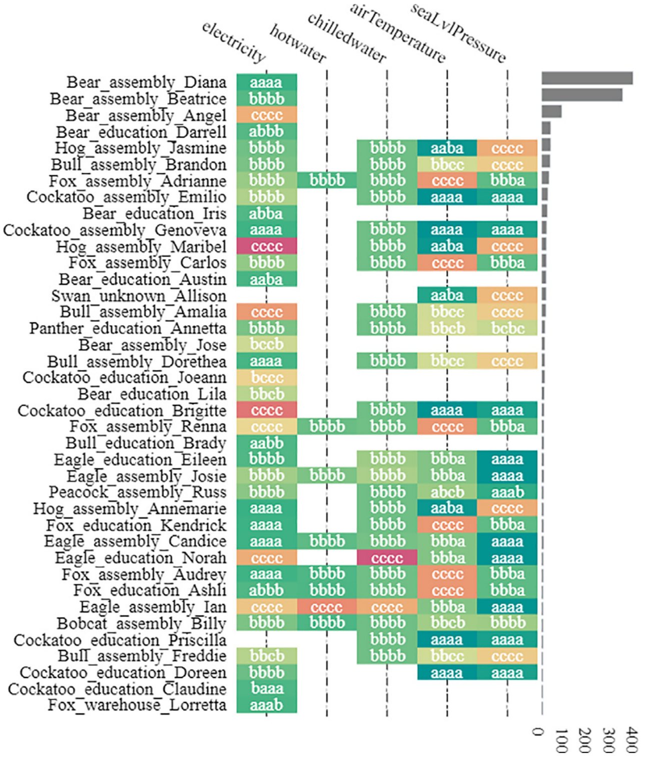

Fig. 12. Grouped building motifs SAX sequences across building stock and attributes on 2016–06–07. Group member counts are presented on the right hand side by a bar chart, and daily aggregate attribute-specific consumptions are illustrated as heatmaps, echoing the classic OLAP approach.

3.4. Towards multi-cube-space visualization

From the examined 2 dimensional lattice of the cube, we reach for a multi-dimensional visual exploration of the highly-dimensional 3D base-cuboid. Daily heatmaps have proven to be powerful visualization tools for 3 dimensional plots, yet the complexity endowed from base-cube visualization needs to be cut-down. To this end we propose combining visual insights from the three afore-examined cuboids to a recomposed, flattened, dimensional visualization of the cube; as if one were studying the cube's pattern rather than the assembled 3-dimensional structure. Fig. 15 presents this multi-cube space visualization, where cuboid A, grouping {*site*, *time*} dimensions, is presented on the lower left corner, cuboid B with {*attribute*, *time*} in the top right corner and cuboid C, gathering {*site*, *attribute*} links both visuals from aligned dimensional sections. Additional rehashed insights from cuboids including {*time*} dimensions were supplemented with aggregated

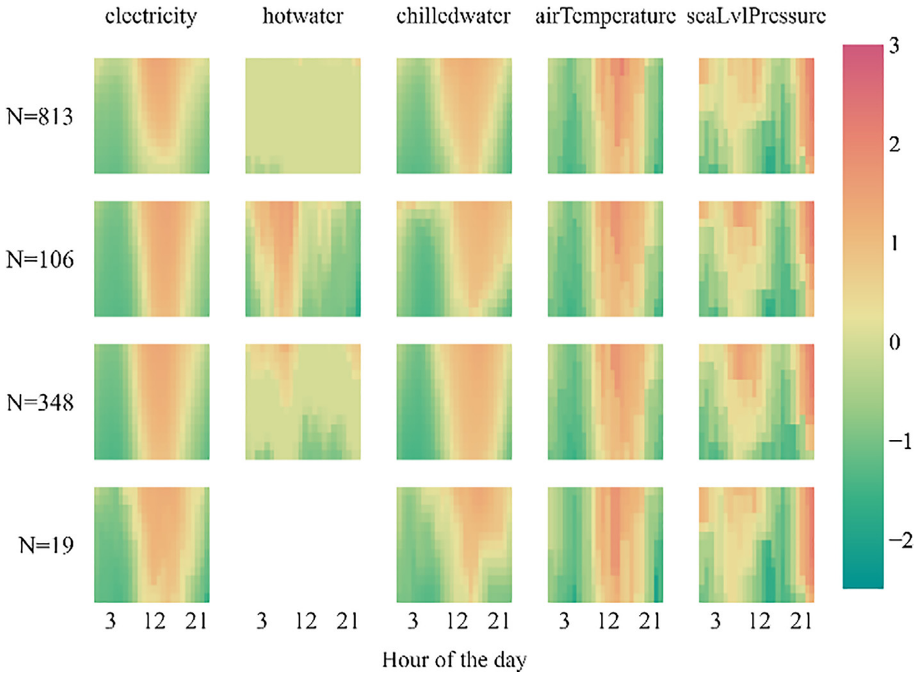

Fig. 13. Cluster similarity index assessement of cross-building motifs stock from weighted averaged one-dimensional time-series from 2016 to 06–07.

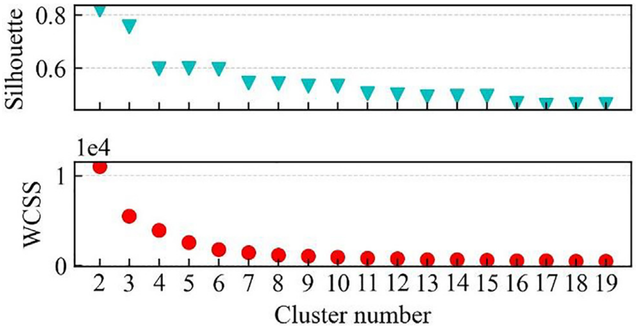

Fig. 14. Motif-building clusters across attributes from day 2016–06–07 illustrated by quantile heatmaps normalized per attribute. Top and lower horizontal heatmap lines report the upper 75% and lower 25% quantiles per hour respectively, with a range of 10 gradually decreasing quantiles in between.

temporal outlooks, here illustrated with barplots resuming the SAX sequences across the temporal study frame. This multi-dimensional view allows distinct knowledge transfer and analytical examination from one dimensional and diagnostic-specific study frame to the next. For instance, while the electricity SAX sequence distribution of cluster 0 within cuboid A should echo that of cuboid B, a building-element subset of cluster 0, the sequence distributions presented are quite different from one another. This stems from the differences in pre-mining normalization frames, where cuboid A scaled electricity consumption over the entire building stock, while cuboid B considered a fixed {*site*} selected subset, consequently resulting in different alphabetical ranges and breakpoints during the SAX transformation process. Yet similar heatmap tendencies may be observed from one cuboid view to the next, i.e., both possess clear summer and winter typical consumption trends with a flat weekend-like *aaaa* consumption profile.

This highlights the importance of per cuboid diagnostic focus; as data analytic choices might be relevantly made for isolated cuboids, mining result comparisons from one cube sub-space to the next ought to be treated cautiously, as a result of different cuboid-specific mining steps. For a common and global multi-dimensional analytical diagnostic, it becomes necessary to follow identical mining tasks at every step of the process. For this work, the importance of framing insight specific steps was chosen to further highlight the significant role of dimensional-frame determination within the process of cube space mining.

4. Discussion

From the definition of a unified multidimensional data mining framework tailored to building analytics, this work intends on bridging the gap between the complexity endowed with big data's high-volume, high-variety and progress towards more interpretable and reproducible research for building analysts. The objective is to link applications to specific diagnostic approaches from dimensionally-reduced cube-space regions. In this context, results of the proposed mining framework implementation are here discussed while considering other possible applications as well as limitations encountered.

4.1. Insight driven mining

On the road towards more interpretable building analytics, definition of the cube dimensional space linked to application-driven insights per cuboid sub-regions has demonstrated great value. From the exploration of the 2D cube lattice, we have covered the established preeminent analytical methods, namely bottom-up {*attribute*, *time*} and top-down {*site*, *time*} approaches, all the while extending them with a temporal drill-in {*site*, *attribute*} analysis. While an identical descriptive pattern filtering mining technique was applied over the lattice, we meet each cuboid with a different analytical angle and diagnostic-objective. It then becomes interestingly relevant to contemplate the more complex analytics that would arise approaching the last, most dimensionally-dense, base-cuboid {*site*, *attribute*, *time*} region of the cube.

Given the previously defined analytical methods, one could imagine tackling this cuboid from three subsequent angles, i.e., including either multiple attributes, sites or temporal-units of

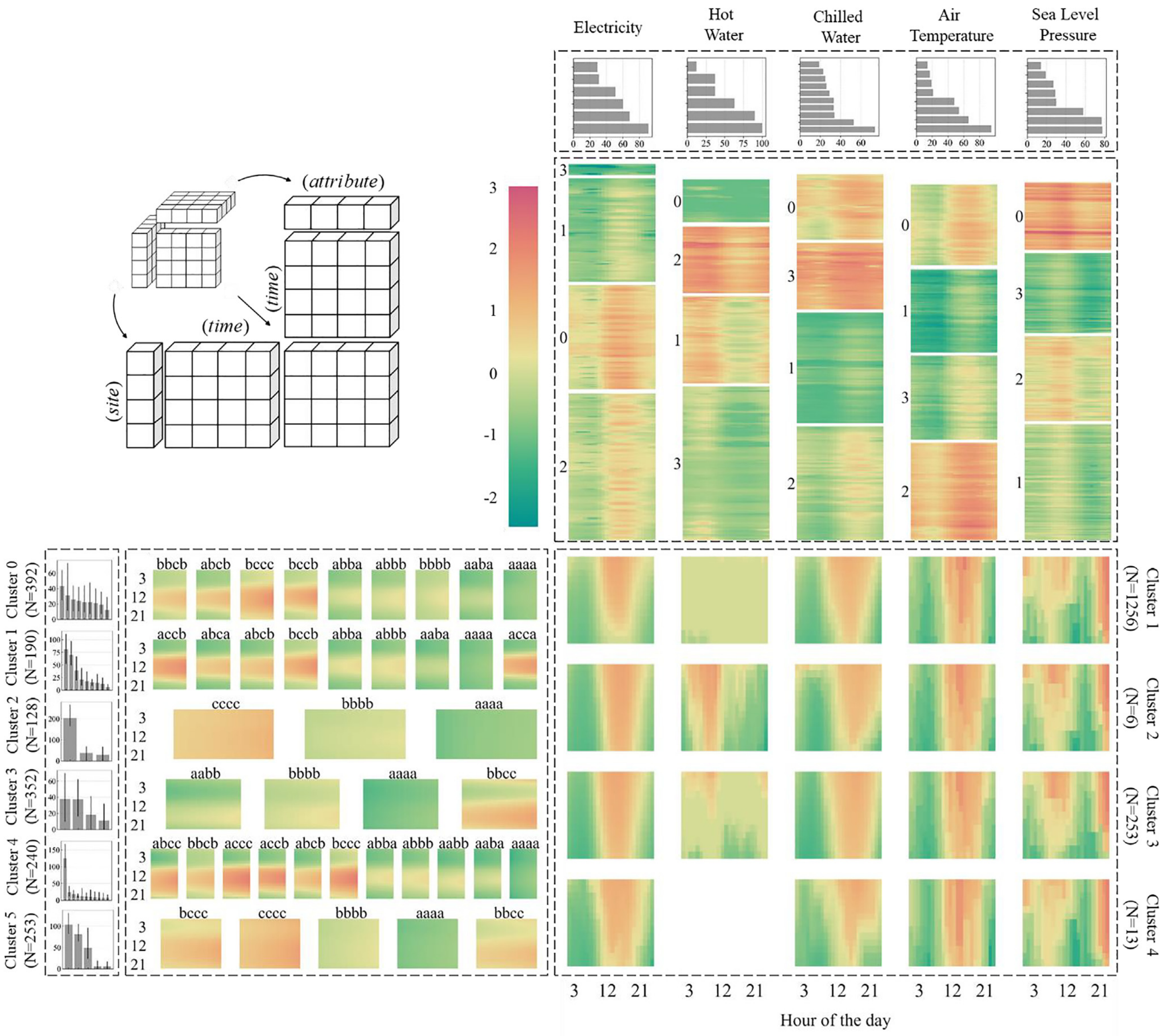

Fig. 15. Data cube base-cuboid pattern visualization from 2-dimensional cuboid lattice insights, i.e., {site, time} bottom left, {attribute, time} top right and {site, attribute} bottom right.

interest within the existing analytical frames of cuboids A, B and C respectively. As a conceptual illustration, approaching the base-cube from cuboid A, would involve a classical top-down analytical approach of building benchmarking extended through multi-attribute considerations. While descending from cuboid B, through the antagonist bottom-up approach, would imply in-site diagnostic methods extended to other buildings, e.g., testing a methods' scalability. The temporal drill-in analytical method of cuboid C, lastly, would examine additional days within its frame, adjusting time-specific insights to a larger temporal frame of interest.

Additionally, while our implementation depicts a descriptive mining technic, predictive applications share equal benefits from cube space conception. Indeed, how to effectively evaluate and select large number of feature, for example, fit naturally within OLAM supported by explicit cube dimensional-space mapping. Assessing contributions of feature combinations to the predictive learning performance over the cube-space, allows systemic optimal feature selection in the confirmatory analysis phase. Machine leaning workflows could incorporate such techniques as an a priori mining analysis to improve model performance. Employing pre-trained models within cuboids are another example of how cube space-driven mining can be practiced in predictive applications [64]. Investigation of this application, while outside the scope of this study, reveals a promising future direction for this framework.

It subsequently becomes clear that the mining process is fully application- as well as insight-driven. Applications such as energy performance benchmarking and model calibration compel to top-down approaches, while automated fault detection and diagnosis, energy saving management, or rule-base knowledge discovery entail classical bottom-up approaches. Likewise, temporal feature engineering can necessitate temporal drill-in methods for, per time-slice, cross-attribute, -building insights. These connect reduced cube-space regions of interest to undertaken applications.

4.2. Visualizing knowledge

The importance of knowledge visualization for effective and impactful result inspection is well established. However, when it comes to high-dimensions it becomes particularly complex, yet crucial to appropriately represent and link insights together. Interactive OLAP visualization tools have already been developed and widely used for data cube exploration, analysis and pattern extractions in the financial field [78], but, to the best of the authors knowledge, close to none in the building sector. The proposed 3 dimensional data cube-pattern visualization paves the way to the development of OLAM interactive visualization tools, where one could imagine iteratively scrolling through the fixed dimensional items of a cuboid. The building analyst could subsequently employ navigational tools such as *drilldown* or *rollup*, through the dimensional hierarchical relationships, e.g. the day slice width considered in cuboid C could be rolled-up to weekly slices or drilled-down to quarter days for SAX sequence analysis.

4.3. Limitations

A notable limitation encountered from cube space mining was the iterative need to reformat the data as well as adapting visualization tools to every studied dimensional frame, which are very time consuming tasks within the data mining process. In then comes into consideration that developing interactive visual tools tailored to OLAM analytics could provide interesting solutions, yet not without challenging limitations. Computational burden resulting from the mining process may render the interactivity of the visual exploration too slow to fully profit from the tool itself. Nonetheless, a priori computation of the visual cube from a set of fixed parameters could be envisioned as a means to initially coarsely characterize the cube and tackle exploratory responsiveness issues.

5. Conclusion

With this work, we have delineated a multi-dimensional, generic data mining framework tailored to big building data, effectively framing which analytical techniques to follow in a stepwise procedure. We appeal to benchmarking methods and apply the proposed DM framework to an automated pattern filtering application using a large building open data set for reproducible, comparable and empirically validated results. Furthermore, we delineate the existing underlying link between building data dimensional space and building management applications. This pushes further down the existing barriers separating building professionals from effective building data dimensional-space targeting given defined applications and insights of interest.

Future research challenges could entail in depth cube space exploration for comprehensive building management application study such as multi-automated fault diagnosis and detection. Another interesting research focus emanating from this work could undertake the determination of how dimensional analytical window frames, i.e. data granularities, window frame and horizon, influence building data analytics and their outcome.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is funded by the Dutch Research Council (NWO), in the context of the call for Energy System Integration & Big Data (ESI-BIDA), project “Small data and big data: Neighborhood Energy & Data Management Integration System” (S&B NEDMIS) for which the authors would like to express their gratitude.

References

[1] Transition to sustainable buildings: Strategies and opportunities to 2050. vol. 9789264202. Organisation for Economic Cooperation and Development (OECD). 2013.

[2] Q. T. Review. QUADRENNIAL TECHNOLOGY REVIEW AN ASSESSMENT OF ENERGY TECHNOLOGIES AND RESEARCH Chapter 5 : Increasing Efficiency of Building,” no. September, 2015.

[3] MIT. MIT Technology Review. 10 Breakthrough Technologies. 2020. [Online]. Available: http://www2.technologyreview.com/tr10/?year=2001. [Accessed: 11-May-2020].

[4] Clarivate. Web of Science [v.5.35] – Web of Science Core Collection Basic Search. Web of Science, 2020. [Online]. Available: https://apps.webofknowledge.com/WOS_GeneralSearch_ input.do?product=WOS&search_mode=GeneralSearch&SID=F5lKgrWCQYzjDto4GIO&preferencesSaved=. [Accessed: 11-May-2020].

[5] M.L.R. Oded, Data Mining and Knowledge Discovery Handbook, Springer, US, 2010.

[6] M.M. Abdelrahman, S. Zhan, C. Miller, A. Chong, Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature, Energy Build. 242 (Jul. 2021) 110885.

[7] C. Fan, F.u. Xiao, C. Yan, A framework for knowledge discovery in massive building automation data and its application in building diagnostics, Autom. Constr. 50 (2015) 81–90.

[8] F. Dalene, Technology and information management for low-carbon building, J. Renew. Sustain. Energy 4 (4) (Jul. 2012) 041402.

[9] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo, P. Chatellier, Modeling and forecasting building energy consumption: A review of data-driven techniques, Sustain. Cities Soc. 48 (2018) 101533–102019.

[10] I. Ghalehkhondabi, E. Ardjmand, G.R. Weckman, W.A. Young, An overview of energy demand forecasting methods published in 2005–2015, Energy Syst. 8 (2) (2017) 411–447.

[11] T. Ahmad, H. Chen, Y. Guo, J. Wang, A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review, Energy Build. 165 (2018) 301–320.

[12] K. Amasyali, N.M. El-Gohary. A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews. 81. Elsevier Ltd. 1192–1205. 2018.

[13] C. Fan, Y. Ding, Y. Liao, Analysis of hourly cooling load prediction accuracy with data-mining approaches on different training time scales, Sustain. Cities Soc. 51 (Nov. 2019) 101717.

[14] A. Moradzadeh, A. Mansour-Saatloo, B. Mohammadi-Ivatloo, A. Anvari-Moghaddam, Performance evaluation of two machine learning techniques in heating and cooling loads forecasting of residential buildings, Appl. Sci. 10 (11) (2020), https://doi.org/10.3390/app10113829.

[15] P. K. Sharma, T. De, S. Saha. IoT based indoor environment data modelling and prediction in 2018 10th International Conference on Communication Systems and Networks, COMSNETS 2018. 2018. 537–539.

[16] C. Xu, H. Chen, J. Wang, Y. Guo, Y. Yuan, Improving prediction performance for indoor temperature in public buildings based on a novel deep learning method, Build. Environ. 148 (2019) 128–135.

[17] C. Miller, F. Meggers, Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings, Energy Build. 156 (2017) 360–373.

[18] D. Chakraborty, H. Elzarka, Advanced machine learning techniques for building performance simulation: a comparative analysis, J. Build. Perform. Simul. 12 (2) (2019) 193–207.

[19] T. Ahmad, H. Chen, J. Shair, Water source heat pump energy demand prognosticate using disparate data-mining based approaches, Energy 152 (2018) 788–803.

[20] P.C. Sen, M. Hajra, M. Ghosh, Supervised Classification Algorithms in Machine Learning: A Survey and Review, Advances in Intelligent Systems and Computing 937 (2020) 99–111.

[21] B. Yildiz, J. I. Bilbao, A. B. Sproul. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renewable and Sustainable Energy Reviews. 73. Elsevier Ltd. 1104–1122. 2017.

[22] J. Han, M. Kamber, J. Pei. Data Mining. Concepts and Techniques. 3rd Edition (The Morgan Kaufmann Series in Data Management Systems). 2011.

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

[24] W. Kim, S. Katipamula, A review of fault detection and diagnostics methods for building systems, Sci. Technol. Built Environ. 24 (1) (2018) 3–21.

[25] Z. Shi, W. O'Brien. Development and implementation of automated fault detection and diagnostics for building systems: A review. Automation in Construction. 104. Elsevier B.V. 215–229. 2019.

[26] Y. Zhao, T. Li, X. Zhang, C. Zhang, Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future, Renew. Sustain. Energy Rev. 109 (2019) 85–101.

[27] J. Gray et al., Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals, Data Min. Knowl. Discov. 1 (1) (1997) 29–53.

[28] R. Ramakrishnan, B.-C. Chen, Exploratory mining in cube space, Data Min. Knowl. Discov. 15 (1) (2007) 29–54.

[29] C. Miller, F. Meggers, The Building Data Genome Project: An open, public data set from non-residential building electrical meters, Energy Procedia 122 (2017) 439–444.

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

[31] Z. Yu, B.C.M. Fung, F. Haghighat, Extracting knowledge from building-related data – A data mining framework, Build. Simul. 6 (2) (2013) 207–222.

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

[33] C. Fan, F. Xiao, Z. Li, J. Wang, Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review, Energy Build. 159 (2018) 296–308.

[34] BS ISO 8601 1:2019 Date and time — Representations for information interchange basic rules. 2019. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:8601:-1:ed-1:v1:en. [Accessed: 27-May-2020].

[35] A. Wagner, W. O'Brien, B. Dong (Eds.), Exploring Occupant Behavior in Buildings: Methods and challenges, Springer International Publishing, 2017.

[36] Y. Zhang, X. Bai, F. P. Mills, J. C. V. Pezzey. Rethinking the role of occupant behavior in building energy performance: A review. Energy and Buildings. 172. Elsevier Ltd. 279–294. 2018.

[37] P. Jayathissa, M. Quintana, M. Abdelrahman, C. Miller, Humans-as-a-sensor for buildings—intensive longitudinal indoor comfort models, Buildings 10 (10) (Oct. 2020) 1–22.

[38] A. Mahdavi, M. Taheri, An ontology for building monitoring, J. Build. Perform. Simul. 10 (5-6) (2017) 499–508.

[39] Project Haystack, "Project Haystack," 2016. [Online]. Available: https://project-haystack.dev/. [Accessed: 27-May-2020].

[40] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S.E. Lee, C. Sekhar, K.W. Tham, k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement, Energy Build. 146 (2017) 27–37.

[41] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, Y. Zhao, Load Profiling and Its Application to Demand Response: A Review, Tsinghua Sci. Technol. 20 (2) (2016) 117–129.

[42] H. Wang, M.J. Bah, M. Hammad, Progress in Outlier Detection Techniques: A Survey, IEEE Access 7 (2019) 107964–108000.

[43] V. J. Hodge, J. Austin. A survey of outlier detection methodologies. Artificial Intelligence Review. 22. 2. Springer. 85–126. 2004.

[44] M. Gupta, J. Gao, C.C. Aggarwal, J. Han, Outlier Detection for Temporal Data: A Survey, IEEE Trans. Knowl. Data Eng. 26 (9) (2014) 2250–2267.

[45] C. Fan, F. Xiao, H. Madsen, D. Wang, Temporal knowledge discovery in big BAS data for building energy management, Energy Build. 109 (2015) 75–89.

[46] M.S. Piscitelli, S. Brandi, A. Capozzoli, F. Xiao, A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings, Build. Simul. 14 (1) (2021) 131–147.

[47] Y. Qin, S. Zhang, X. Zhu, J. Zhang, C. Zhang. POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. Expert Syst. Appl. 36. 2 PART 2. 2794–2804. 2009.

[48] A. Farhangfar, L. Kurgan, J. Dy, Impact of imputation of missing values on classification error for discrete data, Pattern Recognit. 41 (12) (Dec. 2008) 3692–3705.

[49] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen, Methods for imputation of missing values in air quality data sets, Atmos. Environ. 38 (18) (Jun. 2004) 2895–2907.

[50] G. Hawthorne, G. Hawthorne, P. Elliott, Imputing Cross-Sectional Missing Data: Comparison of Common Techniques, Aust. New Zeal. J. Psychiatry 39 (7) (2005) 583–590.

[51] A. Plaia, A.L. Bondì, Single imputation method of missing values in environmental pollution data sets, Atmos. Environ. 40 (38) (Dec. 2006) 7316–7330.

[52] M. Di Zio, U. Guarnera, O. Luzi, Imputation through finite Gaussian mixture models, Comput. Stat. Data Anal. 51 (11) (Jul. 2007) 5305–5316.

[53] C. Chatfield, R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, J. R. Stat. Soc. Ser. A (Statistics Soc. 151 (2) (1988) 375, https://doi.org/10.2307/2982783.

[54] M. N. Norazian Ramli, A. S. Yahaya, N. A. Ramli, N. F. F. M. Yusof, and M. M. A. Abdullah, "Roles of imputation methods for filling the missing values: A review," Adv. Environ. Biol., vol. 7, no. SPEC. ISSUE 12, pp. 3861–3869, 2013.

[55] S. Morgenthaler, Exploratory data analysis, Wiley Interdiscip. Rev. Comput. Stat. 1 (1) (2009) 33–44.

[56] M. Abzalov, "Exploratory data analysis," in Modern Approaches in Solid Earth Sciences, vol. 12, 2016, pp. 207–219.

[57] B. Yildiz, J.I. Bilbao, J. Dore, A.B. Sproul, Recent advances in the analysis of residential electricity consumption and applications of smart meter data, Appl. Energy 208 (October) (2017) 402–427.

[58] M. Kottek, Jürgen Grieser, C. Beck, B. Rudolf, F. Rubel, World map of the Köppen-Geiger climate classification updated, Meteorol. Zeitschrift 15 (3) (2006) 259–263.

[59] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, J. Li, A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis, Energy Built Environ. 1 (2) (2020) 149–164.

[60] P. Bobko, R. Karren, The perception of pearson product moment correlations from bivariate scatterplots, Pers. Psychol. 32 (2) (1979) 313–325.

[61] A.C. Davison, S. Sardy, The partial scatterplot matrix, J. Comput. Graph. Stat. 9 (4) (2000) 750–758.

[62] C. Zhang, L. Cao, A. Romagnoli, On the feature engineering of building energy data mining, Sustain. Cities Soc. 39 (2018) 508–518.

[63] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data 3 (1) (2016), https://doi.org/10.1186/s40537-016-0043-6.

[64] C. Fan et al., Statistical investigations of transfer learning-based methodology for short-term building energy predictions, Appl. Energy 262 (Mar. 2020) 114499.

[65] S. Singh and A. Yassine, "Big data mining of energy time series for behavioral analytics and energy consumption forecasting," Energies, vol. 11, no. 2, 2018.

[66] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," in 2014 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2014, 2015.

[67] Multiresolution Signal Decomposition. Elsevier, 2001.

[68] Y. Wei et al., "A review of data-driven approaches for prediction and classification of building energy consumption," Renew. Sustain. Energy Rev., vol. 82, no. August 2017, pp. 1027–1047, 2018.

[69] S.B. Kotsiantis, Supervised machine learning: A review of classification techniques, Informatica (Ljubljana) 31 (3) (2007) 249–268.

[70] D. Francisci and M. Collard, "Multi-criteria evaluation of interesting dependencies according to a data mining approach," in 2003 Congress on Evolutionary Computation, CEC 2003 – Proceedings, 2003, vol. 3, pp. 1568–1574.

[71] M. Aruldoss, A Survey on Multi Criteria Decision Making Methods and Its Applications, Am. J. Inf. Syst. 1 (1) (2013) 31–43.

[72] I.P. Panapakidis, G.C. Christoforidis, Optimal selection of clustering algorithm via Multi- Criteria Decision Analysis (MCDA) for load profiling applications, Appl. Sci. 8 (2) (2018) 1–43.

[73] JoãoP. Gouveia, Júlia Seixas, Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys, Energy Build. 116 (2016) 666–676.

[74] J. Leprince and W. Zeiler, "A robust building energy pattern mining method and its application to demand forecasting," in SEST 2020 – 3rd International Conference on Smart Energy Systems and Technologies, 2020, pp. 1–6.

[75] C. Miller, A. Kathirgamanathan, B. Picchetti, P. Arjunan, J.Y. Park, Z. Nagy, P. Raftery, B.W. Hobson, Z. Shi, F. Meggers, The Building Data Genome Project 2, energy meter data from the ASHRAE Great Energy Predictor III competition, Sci. Data 7 (1) (2020), https://doi.org/10.1038/s41597-020-00712-x.

[76] R.K. Pearson, Outliers in process modeling and identification, IEEE Trans. Control Syst. Technol. 10 (1) (2002) 55–63.

[77] D. Q. Goldin and P. C. Kanellakis, "On similarity queries for time-series data: Constraint specification and implementation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1995, vol. 976, pp. 137–153.

[78] K. Techapichetvanich, A. Datta, Interactive visualization for OLAP, Lect. Notes Comput. Sci. 3482 (III) (2005) 206–214.