BIM and BAS Data Integration and Big Data Analytics for Building Energy Management

Fu Xiao [1] and Cheng Fan [2]

[1] Department of Building Services Engineering, the Hong Kong Polytechnic University, Hong Kong SAR, China

[2] Sino-Australia Joint Research Center in BIM and Smart Construction, College of Civil and Transportation Engineering, Shenzhen University, Shenzhen, China

Abstract

Data are continuously generated during the lifetime of the building, and mainly stored in Building Information Models (BIMs) and Building Automation Systems (BASs). BIMs store the static and spatial design and construction data, while BASs store the real-time dynamic/temporal operational data. The data from these two e-resources are highly complementary. Effective data integration can provide a more complete spatiotemporal description of a building and bridge information gaps among different stages of a building life cycle. The data integration also facilitates the transition of the AEC industry in the pervasive big data and AI revolution.

This chapter first reviews BIM and BAS data exchanges and integration schemas and their applications in building energy management. Then, the major challenges in analyzing big building data are identified. After that, a big data analysis framework incorporating machine learning for utilizing big building data is proposed and demonstrated. Finally, this chapter is concluded with remarks on prospects and challenges.

Keywords

Building Automation System, Building Information Models, Building energy management, Data integration, Big data analytics

1. Introduction

Today's buildings are not only energy intensive, but also information/data intensive. Buildings consume energy in their life cycles. According to IEA (IEA, 2019), building construction and operations account for 36% of the global final energy use and 39% of energy-related carbon dioxide (CO_2) emissions in 2018. Building energy performance has a great impact on the sustainable development of the world. Meanwhile, data are continuously generated at different stages during the lifetime of the building. With the advances in information technologies, a huge volume of building data can be collected and stored nowadays, which are valuable for better understanding and enhancing building energy performance. However, the heterogeneous data are usually stored in different format and databases and there is little interaction and interoperability among them. Building information models (BIMs) and building automation systems (BASs) are two main electronic resources of building data. BASs mainly collect and store dynamic temporal operational data at the operation stage, while BIMs store static and spatial data at the design and construction stage, such as design parameters of a building and its energy systems and the installation locations of chillers, fans, air ducts and sensors. These two types of data resources perfectly complement each other in this regard. Integration data from the two e-resources will generate a more complete spatiotemporal representation of physical, functional and operational characteristics of a building. This kind of representation is alive as it continuously updates with the real-time operational data and changes made on building systems during maintenance and retrofits in a building life cycle. The integration also contributes to bridge the information gaps between the design, construction and operation stages of the building life cycle. The gaps were considered a key obstacle to improve building life cycle energy performance and cost-effectiveness (Oti, A. H. et al., 2016). Furthermore, a number of recent published research papers showed that the integration of BIMs and BASs data can dramatically increase the data volume and variety, which enables the AEC industry to greatly benefit from the powerful big data analytics and machine learning technologies (Gopalakrishnan, K. et al., 2017). However, data exchange and integration between these two e-resources is still inconvenient even impossible.

BIM is considered as a knowledge resource represented by a digital model of physical and functional characteristics of a facility and shared by different stakeholders at different phases of the life cycle of a facility. However, research suggests that the current applications of BIM mainly focus on design and construction stages, and its applications at building operation and maintenance stage is being left behind (Oti, A. H. et al., 2016). A more widely used commercialized platform for managing building operations is Building Automation System (BAS), also known previously as Building Management System (BMS) or Building Energy Management System (BEMS), which is a network of sensors,

controllers and actuators with wired and wireless connections for automatic monitoring and control of various building services systems and devices (Wang, S., 2009). Almost all large commercial and public buildings are managed by BASs nowadays. BIMs have more user-friendly interface and powerful 3-D visualization capability. BASs can directly communicate with numerous local measurement instruments and controllers distributed all over the building; however, they do not know or show the exact locations of the sensors and devices which is not informative and convenient for facility management staff. BIMs are often work together with design assessment and optimization tools, like daylight simulation (Kota, S. et al., 2014), OTTV (Natephra, W. et al., 2018) and LEED (Jalaei, F. et al., 2020), while BASs can facilitate the implementation of real-time operation control, diagnosis and optimization tools/strategies (Fan, C. et al., 2019a; Shan, K. et al., 2016). BIM and BAS have their own pros and cons of complementary in different aspects. Technology development on BIM and BAS has been proceeding in parallel for many years without considering the integration and interoperability between them.

This chapter first reviews BIM and BAS data exchanges and integration schemas and their applications in building energy management, after a brief overview of BIM and BAS data types and formats. Then, the major challenges in analyzing big building data from the two e-resources are identified. After that, a big data analysis framework incorporating advanced data mining and machine learning for utilizing big building data is proposed. Applications of the framework are demonstrated in several real cases for building energy consumption prediction and building performance diagnosis and optimization. Finally, this chapter is concluded with a summary of prospects and challenges of BIM and BAS data integration and utilization in future smart and energy efficient buildings.

2. Data exchange and integration schemas and applications in the AEC industry

Since the AEC industry was aware of the painful consequence of information gaps among different stages in a building life cycle and the interoperability problem among the various design, simulation and operation software tools, there have been tremendous endeavor to improve the interoperability and fill the information gaps from various aspects, including to achieve convenient BIM and BAS data exchange and integration. The data from different sources inherently possess different types and formats. To effectively utilize the data from BIMs and BASs, it is critically important that the data have unified representation which can be interpreted by both machines and humans. This section first provides an overview of BIM and BAS data, and then summarizes the representative schemas and technologies for data exchange and integration and their applications in the AEC industry.

2.1 Overview of BIM and BAS data

BIMs contain primarily semantic, geometric and parametric data which are usually space-related static data and mainly in the text and numerical data format, for example, the name, type, height, width, orientation and materials of building walls and windows (as entities), the name and location of air ducts as well as the design thermal temperature of spaces and rooms. BIMs can also provide the relational representation between different entities, for example, each VAV box entity has an association relationship with its supply duct and the room it serves. The Industry Foundation Class (IFC) (buildingSMART, 2020) and the Green Building eXtensible Markup Language (gbXML, 2020) are two open data standards/data schemas which define the data representation and data structure for information exchange in the AEC industry. Dong, et al. presented a detailed comparison between these two data schemas (Dong, B. et al., 2007). In recent years, although BIM made huge effort to enable its applications in the operation stage, relevant data in the building operation stage in general and in BEMS specifically are still insufficiently represented, for example, a sensor (datapoint) of a BMS can only be rudimentarily modeled in IFC (Petrushevski, F. et al., 2018).

Building operational data in BAS are typically multivariate time series data, which can be easily retrieved from database like MySQL and stored in spreadsheets in CSV or EXCEL files. Each row represents the values of multiple variables collected at a certain sampling instant and each column represents the values of a specific variable collected at each sampling instants. The data collection intervals range from seconds to hours, providing detailed electronic descriptions of building dynamic operations at various temporal scales. The data included in conventional building automation systems can be divided into four major categories.

(1) Energy consumptions: The energy consumptions of individual building equipment different types of building equipment can be measured through the direct use of energy meters, such as the electricity power consumptions of chillers, pumps and fans. Besides energy consumptions of individual equipment, the overall building energy consumptions are also available in BAS. It should be mentioned that data standardization or normalization should be conducted before data modeling to ensure data analysis reliability and validity. The energy consumption data from BAS are highly valuable in building energy performance assessment.

(2) Operating parameters: The operating conditions of different types of building equipment can be described using real-time physical measurements and control signals. The actual operating conditions of different equipment can be described using physical measurements, such as temperatures, water flowrates, pressures, and frequencies the on/off status, motor frequency set points and air duct damper positions control signals as well as indoor air temperature set-points. The variables may be numerical or categorical, such as ON and OFF or High and Low. Some operating parameters exhibit large

variations, such as the supply air flowrate to an air-conditioned space, but others may be more stable, such as the fixed control set-points and the control signal of a two-speed fans which has two values.

(3) Environmental conditions: Building operations are subject to both indoor and outdoor environmental conditions, such as the indoor and outdoor dry-bulb temperature and humidity, which are usually recorded in BASs. With the advance in information technologies, Furthermore, a wide variety of real-time and predicted measurements on outdoor environment parameters are publicly have become available nowadays from the website of observatory, such as wind speeds, solar radiations levels and CO_2 concentrations, which are also valuable for smart building management. Such data are mostly numerical data with different scales.

(4) Miscellaneous data: Due to the widespread use of IoT and personal mobile devices, more data which cannot be collected conveniently before are now readily available in smart buildings, like the real-time indoor occupancy and occupant behaviors which are valuable to smart human-in-the-loop control of indoor environment. Today's BASs integrates more and more functions or automation sub-systems, such as security control and computerized maintenance management system. As a result, an increasing volume of non-conventional data, such as text data of maintenance records and videos for security control. Such data bring both opportunities and challenges to using advanced data analytics for smart building management.

BAS data are massive and informative in terms of building operations. However, BAS data don't contain geometric and spatial information about the building, for example, the wall materials and a sensor location. In addition, it always involves significant manual work to understand the massive data retrieved from the database of a proprietary BAS due to the lack of a unified semantic description of building operation data. To make the BAS data have a unified semantic representation, the ASHRAE BACnet Committee, who is responsible for the development of the widely adopted BAS open communication protocol (i.e., Building Automation and Control Networks (BACnet)), Project Haystack and the Brick initiative announced they are actively collaborating to integrate Haystack tagging and Brick data modeling concepts into the new proposed ASHRAE Standard 223P for semantic tagging of building data (ASHRAE, 2018). The Project Haystack semantic data model is used to represent the various equipment and relationships in automation, control, energy, HVAC, lighting, and other environmental systems. A tag defines a fact or attribute about an entity, i.e., a physical object in a building. Brick is an open-source effort to develop a uniform metadata schema for buildings. It incorporates the concept of tags in Haystack and combines it with an underlying ontology to describe the physical, logical and virtual assets in buildings and the relationships between them using a linked data structure, i.e., graphs represented with Semantic Web technology.

Building data are massive and heterogeneous due to the huge number and diversity of data acquisition devices as well as the different semantic representations/languages adopted by different data information systems, which results in the difficulties in data exchange and integration as well as challenges in analyzing the data.

2.2 Preparing input data for Building Energy Simulation

Building energy simulation (BES) is a widely adopted effective technology in code compliance, green building certificate, assessment of design alternatives as well as fault diagnosis and control optimization for improved building energy performance in a building life cycle. BES requires a lot of information as inputs, including the building geometry and materials as well as system design parameters available in BIMs and operational data available in BASs. The preparation of input data for establishing a BES model could be a cumbersome and time-consuming process as it highly relies on manual or semi-manual mapping and translation from architecture and building services design specifications, models and drawings to the simulation model. Errors are easily introduced in such an information/data intensive process. If the data already available in BIMs and BASs can be used conveniently and effectively by BES software tools, it could save a significant amount of time and effort in establishing BES models while reducing information missing and errors. However, importing data from BIMs and BAS for BES is still a big challenge.

Chong, et al. proposed a framework for the continuous Bayesian calibration of whole building BES models utilizing data from BIM and BEMS (Chong, A. et al., 2019). The study populated the BIM exported gbXML with information that included construction layers and material properties, internal loads (lighting, plug, and occupancy), as well as HVAC systems information, for generating an EnergyPlus input data file (IDF). Test cases indicated good geometric agreement between the native BIM and the gbXML-based BES model. In this study, the 3 years of monthly electricity energy consumption data were used to calibrate and evaluate the model, which were prepared in advanced rather than retrieved from BEMS in a real-time manner. Asl, et al., presented an integrated performance optimization framework, BPOpt, for building designers to explore design alternatives (Asl, M. R et al., 2015). BPOpt aims to achieve interoperability among the various software applications, including BIM (Revit), energy simulation (Green Building Studio), daylighting simulation (Autodesk Rendering Service), and optimization (Optimo). The project information, the geometry data, and the thermal properties of construction materials stored in the BIM model were used to generated energy model data in the gbXML open schema from BIM using Autodesk®Revit®'s API. Kim, et al., presents the development of a Modelica library for BIM-based building energy simulation (ModelicaBIM library) (Kim, J. B. et al., 2015). Instead of using IFC or gbXML, the authors proposed a BIM API method to

access the BIM data directly from the BIM authoring tools in order to take advantages of the parametric modeling capability of BIM and a more seamless integration with less data conversions. In view that thermal property information of the BIM project cannot be directly exchanged between BIM-compatible applications through IFC, Natephra, et al., used Python scripting to access the thermal properties of building envelop from the BIM material assets (Natephra, W. et al., 2018). *ThermalAssetClass* in API for Revit was used. It can directly use the required data from the BIM database and realize a seamless integration between BIM and the application tools. The physical properties, such as window to wall ratio, were obtained from the BIM database using visual programming language Dynamo.

Existing R&D on using BIM as the data resource for BES mainly focused on automatic preparation of the building energy model and model inputs for various energy simulation tools such as DOE-2 (Kim, H. et al., 2016), EnergyPlus (Chong, A. et al., 2019), TRNSYS (Cormier, A. et al., 2011), and Green Building Studio (Asl, et al., 2015), and Modelica (Kim, J. B. et al., 2015). Many studies showed that information sharing between BIM authoring tools and BEM tools currently relies on open exchange schemas like IFC and gbXML both of which provide means of storing geometry with attributed data; however, this information is often not accurately exported by the BIM tools or interpretable by the BES tools (Pezeshki, Z. et al., 2019). In addition, relevant data in the building operation phase in general and in BEMS specifically are still insufficiently represented in IFC and gbXML (Petrushevski, F. et al., 2018). Although open data exchanges schemas are available, the mapping and translation of the BIM information/data to a BES software are still inconvenient and require substantial time. The automatic seamless integration of BIM and BES data has not yet appeared.

2.3 Importing and visualizing BAS data in BIM

BASs are proprietary systems and the technical details are seldom disclosed, which hinders data exchanges between BASs and other software tools. it is rarely seen to design BAS (i.e. construct 3D BAS models in BIM tools) or exchange BAS information (i.e. exchange BAS information with IFC) in different project stages using BIM tools (Tang, S. et al., 2020; Gao, X. & Pishdad-Bozorgi, P., 2019). The interface of BASs is not as user-friendly and informative as BIMs which can show the 3D details of the building structure and the locations of building services equipment, but BIMs are usually static. Some researchers attempted to make BIMs alive and dynamic (Petrushevski, F. et al., 2018), which means the BIMs keep updating in the lifetime of a building by importing and visualizing real-time sensor measurements and updating with any changes made during maintenance and retrofit. Sensors are connected in a complicated network system in buildings, i.e. the BAS network, sensor measurements are transmitted via network communication. In order to visualize sensor measurement

data, BIM can reply on its add-on network communication functions or communication with BAS database.

Lee, et al. the BIM of a test-bed constructed by Revit was transferred to the web browser (Lee, D., et al., 2016). The building shape model in BIM was converted to a file format to show in the 3D web browser using the 'web published model' program, which is an add-on program in Revit. BACnet Building Automation and Control Networks (BACnet) is the well-established BAS open communication protocol. In this study, the BAS BACnet data expressed in XML was transferred through BAS gateway and saved in the server which can be visualized in the web browser to support facility manager's decision making on operating the building.

Quinn, et al. presented a database architecture integrating IoT sensor data to a facility management enabled BIM by referring to the linked data structure for semantic web and visualizing the data in the BIM (Quinn, C. et al., 2020). The most common integration technique for static and dynamic data is referred to as linked data (Heath, T., & Bizer, C, 2011). A case study of the proposed integration on a university building was presented to demonstrate how each of the sensor data streams can be mapped to fields within the FM-BIM using Dynamo, a Visual Programming Language (VPL).

Oti, et al. proposed a framework for utilizing feedback loops from building energy consumption data from BMSs of other similar buildings to inform and improve design and facility management of a new building in a BIM environment (Qti, A. H., et al., 2016). This paper proposed two approaches to integrating BMS data into BIM, i.e., using energy analysis tools which reply on database management systems, and BIM external applications like Energy Consumption Plug-in. Both approaches prepare BMS data in Excel files to be used as the input data to BIM to improve design. The two approaches are sophisticated and not technologically matured, the authors had to manually programed spreadsheet calculations to transform the BMS data to the required level and format for BIM in the case test.

Tang, S. et al. took a fundamental step to facilitate information exchange for BIM assisted BAS design and operation using BACnet and open BIM standard Industry Foundation Class (IFC) (Tang, S., et al., 2020) . Their work leveraged Information Delivery Manual (IDM) and Model View Definition (MVD) methodologies to define an IFC subset schema (a BACnet MVD) so that BAS information conforming to the BACnet protocol can be represented in IFC data model for information exchange throughout various project stages with BIM tools. Revit and a web browser were used to demonstrate the implementation of the BACnet MVD for BAS information exchange.

In spite of the efforts made on the semantic description of data to be integrated into BIMs, an increasing amount of projects and research groups are working on developing semantic description of BA in

recent years (Butzin, B. et al., 2017), which is important to capture data in a structured and machine readable manner and support data-driven decision making at the operation stage. Although these ontologies are become popular in BA, they cannot model relationships between spatial elements, such as the list of rooms in a floor (Bhattacharya, A. et al., 2015) and their application in direct link of BAS data to BIM was rarely reported. A promising metadata schema based on Semantic Web technology is expected to be developed owing to the joint efforts by BACnet, Haystack and Brick (ASHRAE, 2018) in recent years.

Most of current research on importing BAS data to BIMs aims to visualize the data and some aim to use the data to support facility management to make data-driven decisions. 8 use cases that were believed to guide the development of a dynamic BIM for BEMS concept in which facility data are combined with BEMS data were presented (Petrushevski, F. et al., 2018). The first 6 use cases were to visualize data-point (sensor) values, alarms, preprocessed data and logging data in a spatial context, visualize radiant cooling/heating elements and report energy consumption. The last two use cases were verification of a ventilation system using the provided layout and configuring lighting system. Only simple data analysis functions were implemented in BIMs which is not beneficial to utilizing the big building data.

2.4 Utilization of BIM and BAS data for building energy management

Although a large number of data semantics, schemas, ontologies as well as databased management and web-based technologies were proposed for BAS and BIM data exchange and integration as reviewed above, very few were implemented in practical building energy management at the operation stage. Convenient seamless integration of BIM and BAS data is still missing and utilization of BIM and BAS data for building energy management is rarely reported.

Dong, et al. proposed a BIM enabled information infrastructure for real-time building energy fault detection and diagnosis (FDD), to tackle the two main technology challenges preventing smooth transitions from an offline to an online real-time building FDD, i.e., lack of data integration from the design stage to the operation stage and lack of a scalable information infrastructure for the purpose of integrating various technologies (e.g., BMS, FDD, building energy modeling/simulation and visualization) in real-time (Dong, B. et al., 2014). The authors used both gbXML and IFC files for static (design) information acquisition from BIM, and the BACnet reading and storing utilities for dynamic (operation) information acquisition. The data/information are saved in SQL database which serves as the hub of all information exchanges among BIM, BAS and building simulation and FDD modules. Although the infrastructure developed in this study was successfully implemented in a real-

time building for energy FDD, it involved tremendous customizing and manual work. It is a valuable attempt in integrating BAS, BIM and BES data for building energy management at the operation stage.

Lu, et al. presented a study using Digital Twin (DT) based anomaly detection for a centrifugal pump in HVAC systems (Lu, Q. et al., 2020). The DT is a digital model, which is a dynamic representation of an asset (i.e., the building) and mimics its real-world behavior, which consists of five layers: data acquisition layer, transmission layer, digital modelling layer, data/model integration layer and service layer. The DT was developed to meet the requirement of cross-referencing of multiple data sources including BIM, BAS and facility management system. The data integration method was developed based on the primary IFC file and added additional IFC entities to fulfill O&M information representation. The matching tables for other database (e.g., BAS) integration were created which involved comprehensive customization work. More use cases to define the reasons to deploy digital twins within the built environment by answering some obvious business case questions were established (David Q. et al., 2020). Although the DT technology is promising in data integration for practical applications, it is still in its infancy for building energy management.

In summary, data exchanges among various software tools used by different building project teams through the whole building life cycle in the AEC industry gradually become possible with the widespread implementation of machine-readable semantics and data exchange schemas like IFC, gbXML, Haystack and Brick. However, data integration of BIM, BAS and BES is still inconvenient. BIM is a standalone information system with limited data storage and data analysis capabilities which are insufficient to embrace the upcoming big data era in the AEC industry. To make effective utilization of the big building data, web-based and cloud-based data integration technology is promising as the powerful data storage and analysis platforms and tools like *Hadoop* and *MapReduce* as well as *R* and *Python* which are all deployed or deployable in cloud. In addition, the cloud-based data management platform can be shared by different professionals involving in different stages of a building life cycle. There is no technical barrier for BAS connecting with Internet and cloud as BAS is essentially a network-based planform. The joint efforts professionals and researchers on developing convenient and reliable BIM and BAS data integration technology speed up the transition of the traditional AEC industry in the big data era. Big building data brings about both opportunities and challenges.

3 An analytic framework for big building data analysis

3.1 Challenges in big building data analysis

Big building data are heterogeneous which results in challenges in big data analytics (Wang, L., 2017). With the increasing efforts made on unified metadata schemas for buildings, the heterogeneity problem can be partially solved by adopting the same or interoperable open data exchange and integration schemas at different building stages and in various tools. However, integration of the static data from BIMs and dynamic data from BASs generate unstructured data with a diversity of data types and formats which impost great challenges to big building data analytics.

Building operations typically present high complexity considering the dynamic interactions of a large amounts of electrical and mechanical equipment components, the uncertainties in occupant behaviors and dynamic operating conditions. The correlations among data are very strong and dynamic. Building operational data are mainly collected by sensors and transmitted by wired or wireless network. The data quality is restricted by the accuracy of sensors and suffers from deterioration due to measurement errors, uncertainties and transmission problems. In practice, the raw building operational data contains a large number of missing values and outliers as well as drifting and stagnant measurements. Besides, building data are high-dimensional, consisting of hundreds of variables with different data types and scales. From a statistical point of view, high-dimensional data analysis is always challenging considering that most of learning algorithms are relied on distance measures, which become increasingly ineffective as data dimension arises. In addition, building variables have different data types which may be intriguing for practical applications. For instance, the motor frequency of a 2-speed fans may appear to be numerical and continuous, but it should be treated as ordinal for more stringent data analysis.

3.2 Big data analysis framework

Big building data analysis is not just about a method or an algorithm for analyzing the data. It is a complex process starting from data preparation to the applications of the knowledge discovered from the big data sets and integrating various powerful data mining and machine learning algorithms specifically selected for different tasks. A generic framework is needed to guide the complex process and support the AEC professionals to develop big data analysis tools for building energy management. An analytic framework for big building data analysis is proposed in Fig. 1 considering the characteristics of the big integrated BIM and BAS data and practical requirements.

Extensive data preprocessing is often needed before feeding the data to advanced data mining and machine learning algorithms for knowledge discovery. A survey showed that data preparation accounts for about 80% of the work of data scientists (Press, G., 2016). Data preprocessing is highly demanding in using big building data due to its heterogeneity and poor data quality. Considerately designed data

preprocessing procedures to unify heterogeneous data and improve the quality of big building data are essentially needed. Data preprocessing is conducted with the aim of enhancing data quality while ensuring the data format compatible with advanced learning algorithms.

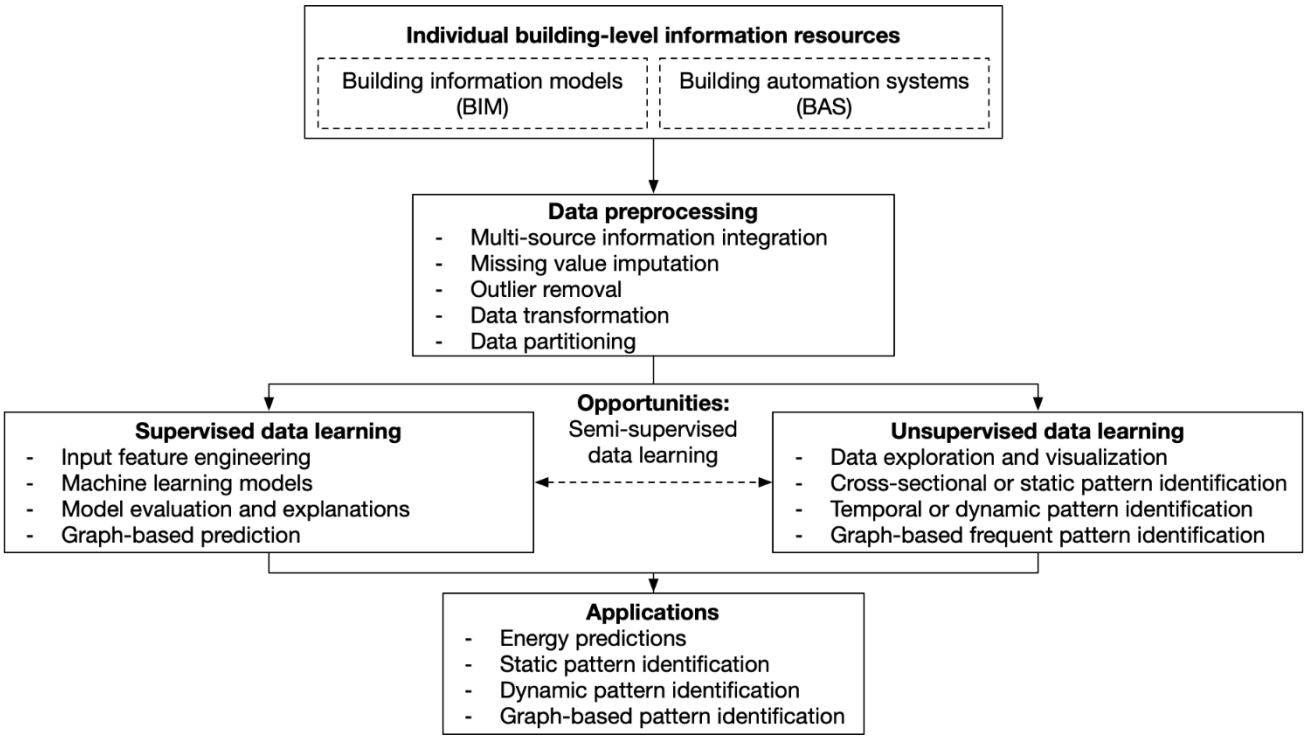


Fig.1 The analytic framework for big building data analysis

There are five typical tasks in preprocessing building data. The first is multi-source information integration, which unifies data/information from various sources for data analysis. It is of great significance considering BIM and BAS data are very different in terms of semantic representation and data formats. The second is missing value imputation, which can be achieved using simple statistical methods (e.g., replacing missing values based on means or medians of the same variable) or model-based methods (e.g., replacing missing values based on values predicted by models developed from other data variables). The third is outlier removal, which can be done either using domain expertise or statistical methods, such as the simple three-sigma method. The fourth is data transformation, which aims to transform the scales or types of different variables. Considering that building variables have a wide range of data scales, standardization or normalization is a must before developing regression or classification models. The values of building data vary largely, from smaller than 1 like control signals to several thousand like power measurement. Data standardization or normalization can prevent data analysis results from being dominated by the variables with larger values. Data type transformation aims to ensure the data format is compatible with learning algorithms. For instance, the conventional association rule mining algorithms, such as the A-priori, only work with categorical variables and therefore, numerical variables should be discretized before feeding to the A-priori algorithm. The fifth

is data partitioning, which aims to enhance the sensitivity and reliability of data analysis by dividing the big data into several subsets considering their intrinsic data structures or similarity.

Once the big data is preprocesses, knowledge discovery can be conducted by adopting data analysis methods/algorithms, including data mining, graph mining, machine learning, and etc. The widely adopted data analysis methods in the building field can be roughly categorized into supervised and unsupervised learning algorithms. More specifically, supervised data learning aims to develop predictive models to describe the relationships between input and output variables. There are several major tasks in supervised data learning. The first is feature engineering, which aims to select or construct input features for predictive modeling. One popular approach is to select variables from existing data based on domain expertise or statistical measures (e.g., correlation). A more advanced approach is to apply dimensionality reduction methods (e.g., principal component analysis and autoencoders) to construct new features from existing data. For instance, principal component analysis can be used to transform existing data into a set of principal components, which can be used as input features for predictive modeling. The second major task is to develop prediction models based on machine learning algorithms. There is a plethora of supervised learning algorithms for prediction model development, ranging from conventional linear ones (e.g., multiple linear regression) to nonlinear ones (e.g., decision trees, support vector machine and artificial neural networks). To further enhance the generalization performance, advanced ensemble methods, such as bagging, boosting and stacking, have been developed. In practice, users should make their choices based on the size of available data samples, the accuracy requirement and the need for model interpretability. Theoretically, there is a trade-off between accuracy and interpretability. The third task aims to break such trade-off by developing customized methods for evaluating and interpreting complicated black-box models. For instance, the local interpretable model-agnostic explanations (LIME) method is a popular approach for revealing the inference mechanism for local predictions. Furthermore, supervised learning algorithms can be applied to graph data for predictions. Example tasks include within-graph link prediction and between-graph classification.

The data used for supervised learning should have sufficient labels, e.g., normal and faulty, which may not be the case in practice as labelling is typically consuming and costly. For instance, it is often not possible to collect sufficient data labelled as fault for fault detection and diagnosis in buildings and building energy systems as faults are usually rare. In such a case, semi-supervised learning can be applied to leverage the value of unlabeled data for predictive modeling. Unlike supervised or semi-supervised learning algorithms which have explicit analysis targets, unsupervised learning algorithms are mainly applied for data exploration (Fan, C. et al., 2018). It can be used for visualizing intrinsic

data structures, mining cross-sectional, temporal, and graph mining-based frequent patterns in data. Such methods are more flexible to use and may obtained unexpected yet useful knowledge from massive data. It should be mentioned that the raw data analysis results can be redundant and useless and therefore, extensive post-mining efforts are needed to transform the raw knowledge into actionable measures for building energy management.

The final step aims to utilize the knowledge discovered to facilitate the decision making of building professionals. The representative applications include building energy predictions, operation pattern identification, rare event identification and temporal association/dynamics identification, which are all highly valuable for developing building operation management strategies for improving building energy efficiency, indoor thermal comfort, power grid supply-demand balance and reliability, and etc.

4 Application demonstration in building energy management

4.1 Short-term building energy prediction

Short-term building energy predictions aims to forecast energy usage profiles over the next few hours or days, which are essential to the development of model-based or model predictive control and optimization strategies for buildings and smart grids (Shan, K. et al., 2016; Xue, X. et al., 2014). The following sections present the development of a short-term building energy prediction model by referring to the proposed big data analytics framework and adopting advance machine learning algorithms. The key and challenging aspects of short-term building energy predictions are highlighted.

4.1.1 Feature engineering for preparing inputs

The fundamental task in predictive modeling is to determine the input variables for model development. Building energy consumptions are subject to various influential variables, such as outdoor environment, occupant schedules and system working conditions. Considering that each building has its own operating characteristics and data availability, it is not possible to come up with a fixed subset of input variables as universal solutions. In addition, building operations present intrinsic temporal dependencies and therefore, it is useful to include historical measurements for enhancing prediction performance. Nevertheless, it can be very challenging to determine the accurate time lag for historical measurements across different variables. Traditional methods mainly rely on domain expertise to develop customized feature solutions for individual buildings, making it almost impossible to fully automate the process of short-term building energy predictions.

To tackle the above-mentioned problems, statistical methods and unsupervised deep learning are adopted to automatically construct useful features for a certain prediction task, e.g., building cooling

load or the indoor air quality . The methodology contains two main steps. The first is to apply spectral density estimation to the time series of total building energy data to determine maximal time lags of historical data. The second is to develop unsupervised deep learning-based methods for automatic feature engineering. As an example, both autoencoder and generative adversarial network (GAN)-based methods have been used for constructing useful features for short-term building energy predictions.

Autoencoder-based feature engineering

An autoencoder aims to reconstruct the input data through multi-layer neural network operations. It typically has two parts, i.e., the encoder and decoder. The encoder transforms the input into latent representations, based on which the decoder is used to reconstruct the original inputs. If converged, the model should present sufficient capability in data reconstruction and thereby, the fewer hidden activations in the middle of the autoencoder can be used as features for information compression and knowledge representations. As shown in Fig. 2, the feed-forward autoencoders with bottle-neck architectures are used for feature engineering, where X, H and Z represent the original inputs, hidden activations and extracted features respectively. The potential of both fully connected and one-dimensional convolutional autoencoders in representing temporal building energy data has been exploited (Fan, C. et al., 2019b).

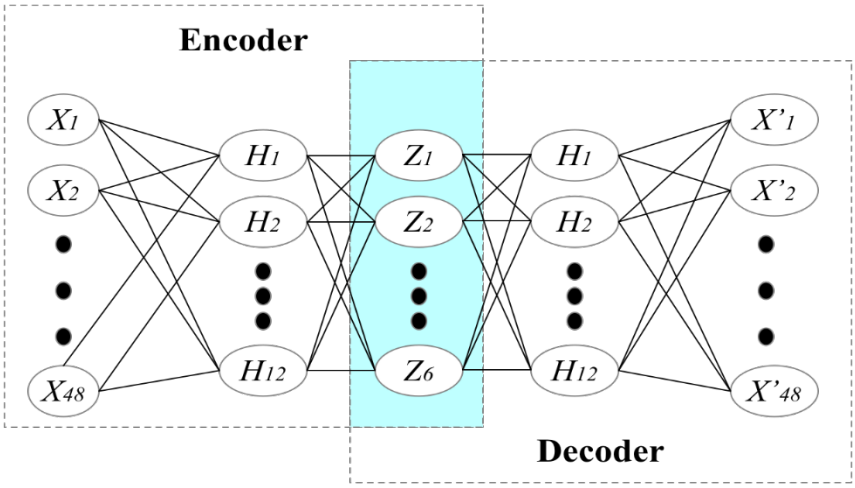


Fig. 2 The schematic of autoencoder-based feature engineering (Fan, C. et al., 2019b)

Generative adversarial network (GAN)-based feature engineering

GAN is one of the most promising techniques in the field of deep generative modeling. The concept of GAN was firstly introduced by Goodfellow et al. in 2014 (Goodfellow, I. et al., 2014). As shown in Fig. 3., a GAN model consists of two neural network models, one as the generator and the other as the discriminator. The generator takes random noises from certain distributions as inputs and output

synthetic data samples. The discriminator takes both synthetic and real data as inputs with the aim of making correct classifications. These two networks are trained in an adversarial way, i.e., the generator tries to create highly realistic synthetic data to fool the discriminator, while the discriminator tries to make correct classifications between real and fake data. Once the GAN learning converges, the discriminator should be capable of describing the intrinsic data characteristics in real data and therefore, the activations before the output layer can be used as useful features for predictive modeling.

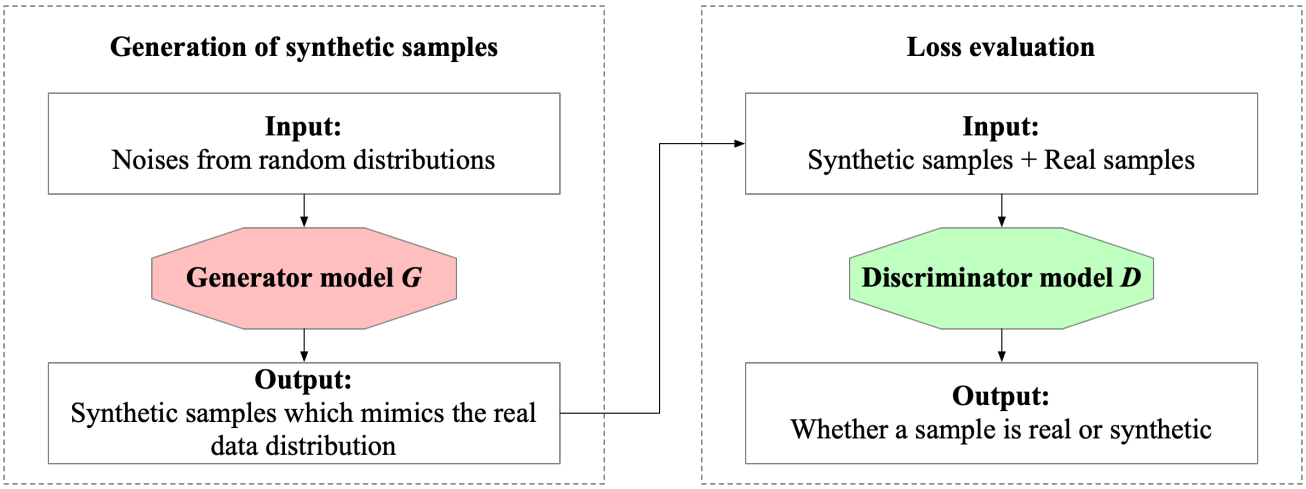


Fig. 3 The GAN model architecture (Fan, C. et al., 2019b)

4.1.2 Strategies for short-term prediction model development

Short-term building energy predictions are in essence multi-step ahead predictions. For instance, when the prediction horizon is set as 24 hours, there should be 24 predictions given hourly time resolution. The performances of different strategies for multi-step ahead building energy prediction models have be compared (Fan, C. et al., 2019c). As shown in Fig. 4, there are three major strategies for multi-step ahead predictions. The first is called the recursive strategy which applies one-step ahead prediction models multiple times for predictions. In such a case, the prediction generated at time step T will be used as inputs for prediction at time step $T+1$. Such strategy is easy to implement. The prediction model is relatively simple as it only outputs prediction at the next time step. The main drawback is that the resulting multi-step ahead predictions may suffer from the error accumulation problems, as the prediction error will propagate along the prediction horizon. The second strategy, i.e., direct strategy, has been proposed to tackle such drawbacks. The main idea is to develop individual models for prediction at each time step. Assuming that the prediction horizon is k corresponding to predictions at time step $T+1$, $T+2$, …, $T+K$, k sub-models will be developed for predictions at each time step. The main drawback of the direct strategy is that the computational resources required for model development can be very large and the predictions may not preserve temporal dependencies along the

prediction horizon. It should be mentioned that the neural network models with multiple neurons at the output layer can be used for the efficient implementation of the direct strategy. To better preserve the stochastic temporal dependency among predictions along the prediction horizon, the multi-input and multi-output (MIMO) strategy has been proposed using recurrent models and sequence-to-sequence learning. The performance of these three strategies in 24-hour ahead building cooling loads based on recurrent neural network models have been investigated (Fan, C. et al., 2019c). The results indicate that the direct strategy can achieve the most accurate predictions without significantly increasing the computation load. The MIMO strategy has the second-best performance in terms of prediction accuracy while the recursive strategy results in the worst performance.

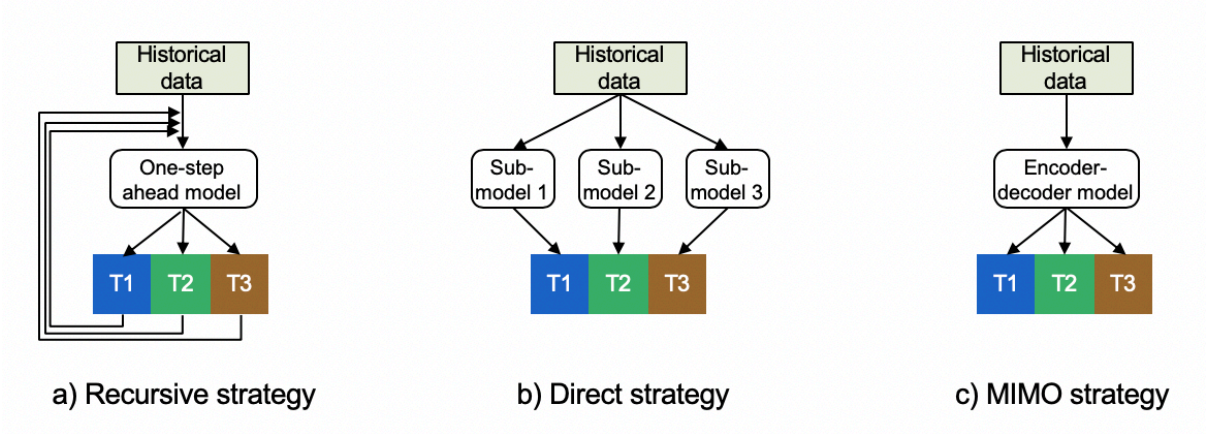


Fig. 4 Three major strategies for multi-step ahead predictions

4.1.3 Model evaluation and explanation

One of the major tradeoffs to be considered in predictive modeling is between accuracy and interpretability. On the one hand, to accurately describe complex and nonlinear data relationship, advanced machine learning algorithms have to be used for model development. On the other hand, complicated machine learning models are often regarded as "black-box models" as underlying inference mechanisms are typically too complicated to be understood and interpreted by ordinary professionals. In practice, it is hard for building professionals to put their full trust on complicated machine learning models without roughly understanding the inference mechanisms.

To tackle this challenge, a novel method to explain and evaluate the performance of data-driven models was developed (Fan, C. et al., 2019d). The main idea is to apply linear models with high transparency to provide explanations on every single step of predictions generated by a sophisticated machine learning algorithm. The methodology contains five key steps and an example is shown in Fig. 5.

Data permutation

A permutation input data (i.e., denoted as X_{perm}) is generated by performing randomly sampling based on kernel density estimation obtained from the actual input training data (i.e., denoted as X_{train}). The

X_{perm} is then feed to the complicated machine learning models to generate predictions denoted as Y_{perm}.

Similarity calculation

For each observation in the input testing data (i.e., denoted as X_{test}), a vector of similarity scores can be obtained by calculating the Gower's diusingssimilarity coefficient to each sample in X_{perm}.

Interpretable representation transformation

Each variable in X_{perm} and X_{test} are transformed into interpretable representations. For instance, numerical variables can be discretized into categorical variables for the ease of interpretation. The resulting interpretable data are denoted as X_{interp} and $X_{test\text{-}interp}$ for X_{perm} and X_{test} respectively.

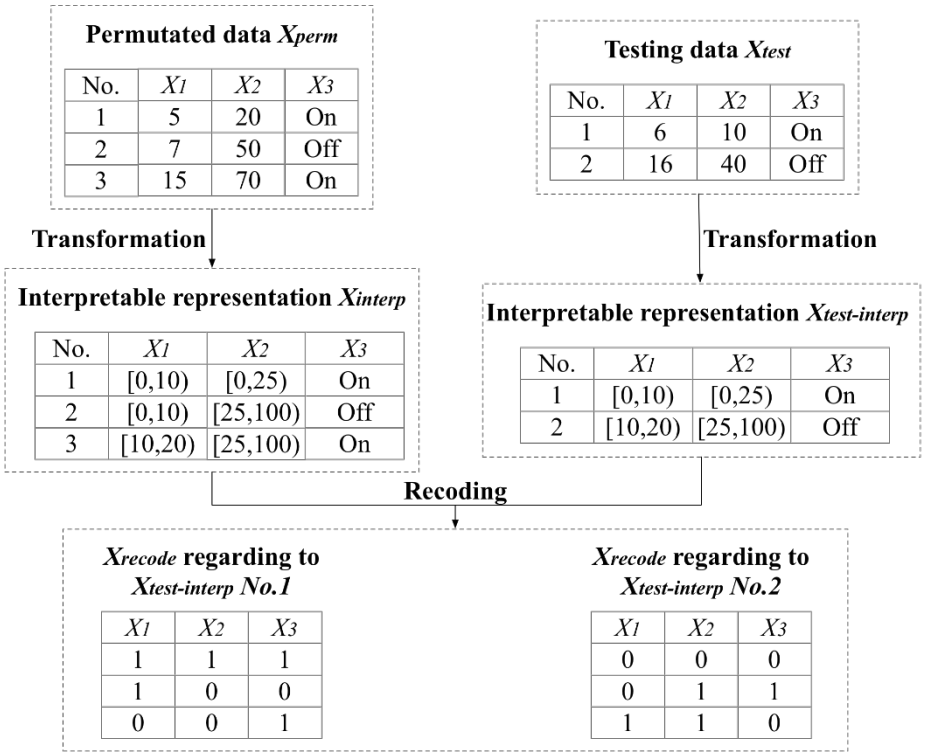


Fig. 5 An example of data transformation and recoding (Fan, C. et al., 2019d)

Data recoding

A recoded input data (i.e., denoted as X_{recode}) is created by comparing each sample in $X_{test\text{-}interp}$ and X_{interp}. The elements of X_{recode} consist of zeros and ones only, indicating that the values in X_{interp} are identical to those in each testing sample or not. Similarly, Y_{perm} is transformed to Y_{recode} for each sample in the testing data. If the prediction matches, the recoded value is 1 and 0 otherwise.

Local surrogate model development

The elastic net model which combines Lasso and Ridge penalty terms are developed using X_{recode} as the input and Y_{recode} as the output. The input variables with the top-k largest coefficient values are selected as evidence for decision making.

In addition, a novel metric, i.e., trust, is proposed to quantify the validity of individual predictions

based on the top-k coefficients of the local surrogate model. The metric is developed with the following two considerations. Firstly, the larger the number of positive coefficients among the top-k model coefficients, the more reliable the prediction is. Secondly, the larger the absolute values of the positive coefficients, the more reliable the prediction is and vice versa. The trust metric is formulated as shown in Eq. 1.

$$Trust = \left(1 - e^{-\frac{N_s+1}{N_c+1}}\right) \times \frac{\sum_{i=1}^{N_s} \theta_{N_s,i}}{\sum_{i=1}^{N_s} \theta_{N_s,i} + \sum_{i=1}^{N_c} |\theta_{N_c,i}|} \qquad (Eq.1)$$

More specifically, N_s and N_c represent the numbers of supporting and conflicting evidences, $\theta_{N_s,i}$ and $\theta_{N_c,i}$ refer to the values of the i^{th} positive and negative coefficients respectively. The first part represents the influence of the numbers of supporting and conflicting evidences on prediction validity. The plus-one Laplace smoothing is used to prevent the denominator from being zero. The second part represents the relative strengths of supporting and conflicting evidences. As a result, the whole trust metric for each individual prediction ranges from 0 to 1. The higher the trust value, the more trustworthy or reliable the prediction is. The methods have been utilized to provide explanations for binary classification models for chiller coefficient of performance (i.e., COP) prediction, which is the most widely used index for evaluating the chiller energy performance and in developing chiller operation optimization strategies. As shown in Figs. 6 and 7, two examples are presented to illustrate the most trustworthy and untrustworthy predictions, respectively. The heights of green and red bars are used to represent strengths of supporting and conflicting evidences respectively. It is observed that data sample No. 2980 is predicted to be "COP=High" with a very high trust value, i.e., 0.965. All the top-5 evidences are supporting such prediction. For instance, the most supporting evidence is that the returned chilled water temperature (i.e., denoted as *CHW_RT_Main*) falls in the interval between 13.2°C and 24°C. Other supporting evidences are about the supplied chilled water temperature, the building cooling load, the chilled water flowrate and the No. 5 condensing water pump on/off status. By contrast, the prediction on sample No. 4593 is *Low* with a very low trust score, i.e., 0.079. It is observed that three of the top-5 data evidences are voting against such prediction.

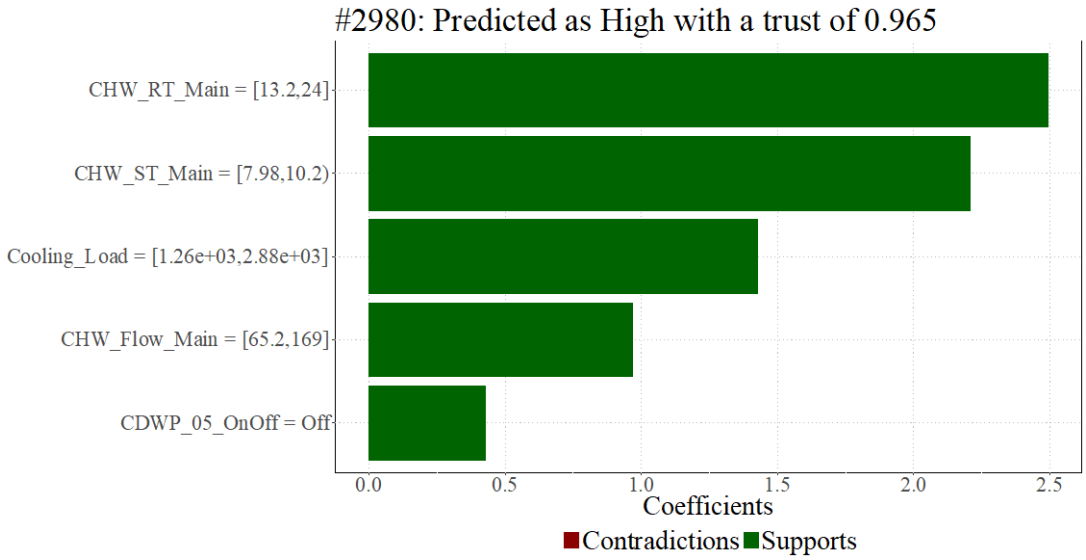


Fig. 6 The most trustworthy prediction on Chiller COP classifications (Fan, C. et al., 2019d)

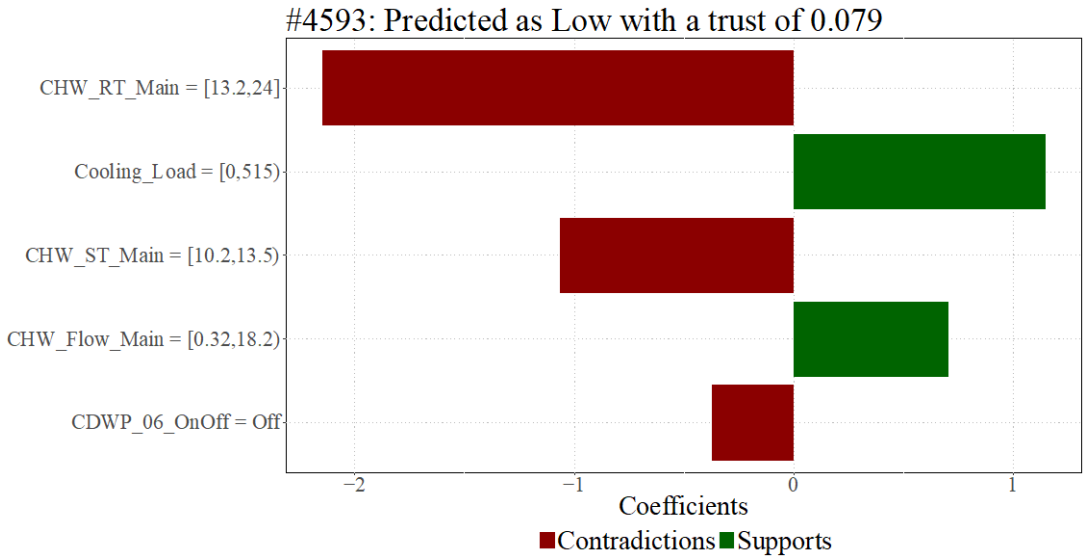


Fig. 7 The most untrustworthy prediction on Chiller COP classifications (Fan, C. et al., 2019d)

4.2 Graph mining-based methods for knowledge discovery in integrated data

Conventional data analytics are usually suitable for analyzing data with simple data structures, e.g., two-dimensional data tables widely used for BAS data. However, the BIM and BAS integrated data are more diverse in data type, format and structure, e.g., the static spatial and hierarchical data/information and the dynamic time-series data. Data variety is a persistent challenge in big data analytics. Recent progress made in BIM and BAS data integration provides a promising solution to analyze the integrated data. Linked data structure, i.e., graphs represented with Semantic Web technology, were adopted by both BIM (Quinn, C. et al., 2020; Hu, S. et al., 2018) and BAS (Brickschema) to represent relationships among data. The graphs can represent a broad spectrum of

information taking a variety of formats ranging from traditional vectors to time-series, spatial information, hierarchical structures and etc. They can integrate and represent complicated relationships among building variables and therefore, providing a promising data format (i.e., graph) for discovering novel spatiotemporal insights by adopting graph mining techniques. The framework proposed is applicable to knowledge discovery from graphs by using graph mining techniques. The data analysis process is similar while the specific data analysis methods need to be specifically selected and customized. There are two essential steps in carrying out graph-mining based data analysis, i.e., graph generation in the data preparation and graph mining for the knowledge discovery/learning.

4.2.1 Overview of graph mining

Graphs are considered as one of the most generic and interpretable formats for representing various kinds of information (Cook, D. J., & Holder, L. B., 2000). A graph consists of a set of vertices/nodes and edges/links. A vertex/node typically represents an entity, while edges are used for describing relationships among vertices. Graphs can be either directed or undirected, depending on whether the edges have directions or not. In addition, the edges can be used to represent both categorical and continuous relationships, providing great flexibilities for information representations.

Graph mining techniques can extract useful insights from graph data. In general, there are two types of graph mining techniques, one for predictive modeling and one for pattern identifications. The graph mining-based predictive modeling can be applied at two different levels. At the lower level, predictions can be made to predict whether there will be new edges among currently unlinked vertices, i.e., link prediction. Such methods have been widely used in social network analysis, where each node represents a person and the link between nodes represent their interaction levels. The link prediction in this case is typically used to predict whether two persons will become friends or connected given their current networks. At the higher level, predictions can be made to classify graph-level characteristics. Taking the biochemical industry as examples, a chemical compound can be represented as a graph and graph-level prediction can be conducted to classify whether the chemical compound is toxic or not. The other groups of graph mining techniques aim to identify frequent common sub-structures in graphs and therefore, are generally denoted as the frequent sub-structure mining (FSM) techniques. Similarly, FSM can be conducted on two levels, i.e., either at the lower level to find common substructures in a single graph or at the higher level to find common sub-structures among a number of graphs. The latter typically has a wider range of applications and the FSM algorithms can be classified into different groups based on their search strategies, either inexact or exact strategies. Inexact algorithms, such as the *SUBDUE* (Cook, D. J., & Holder, L. B., 1993) and *CREW* (Kuramochi, M., & Karypis, G., 2004), have higher mining efficiency as s one does not need to perform exact

comparisons among graph sub-structures. Nevertheless, it may not be able to identify all common sub-structures, leading to non-deterministic results in data analysis. By contrast, exact algorithms, such as the *MoFa* (Borgelt, C., & Berthold, M. R., 2002), *gSpan* (Yan, X., & Han, J., 2002), *CloseGraph* (Yan, X., & Han, J., 2003) and *GASTON* (Nijssen, S., & Kok, J. N., 2004), are more commonly used for practical applications. Previous studies have showed that the *gSpan* and *CloseGraph* had better performance than the others in terms of computational loads and mining efficiency (Wörlein, M. et al., 2005; Yan, X., & Han, J., 2003).

4.2.2 Graph generation methods for preparing inputs

A building variable-based method was developed to generate graphs (Fan, C. et al., 2019a). In such a case, each node represents a subsystem, a component or a space, and each link represents the hierarchical information among graph nodes. Graphs are generated with two considerations. The first is to minimize the number of vertices and edges used so that the computational burden of graph-mining based analysis is minimized. The second is to ensure the connectivity between any two vertices in a labeled graph, making it compatible with FSM algorithms for pattern identification.

A radiating layout is adopted for graph generation. More specifically, each building variable is denoted as a vertex and the hierarchical relationships among building variables are preserved using the radiating layout, i.e., the central vertex represents building-level variables (e.g., the total building cooling load or power consumptions), the first layer of vertices represents system-level variables, and the second layer denotes component-level variables (e.g., individual chillers or water pumps). A third layer can be created to represent the physical measurement of individual components, such as the supplied and returned chilled water temperature of an individual chiller.

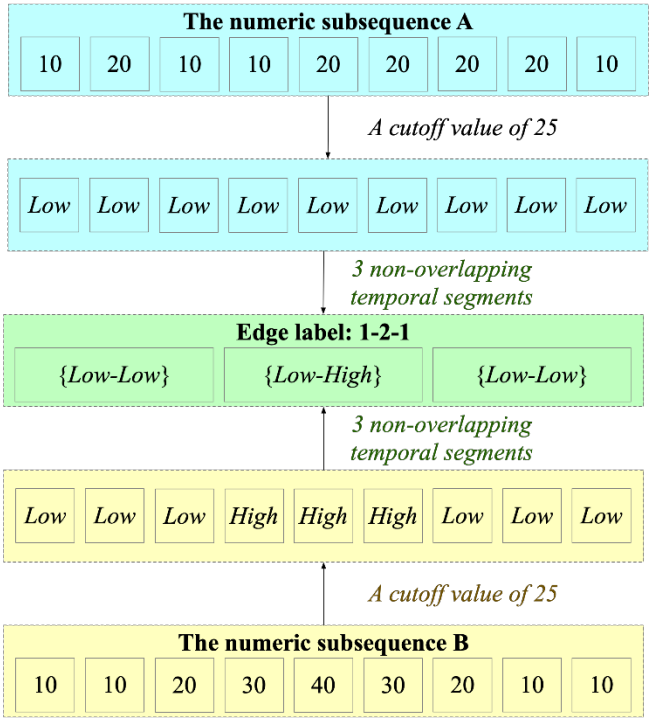


Fig.8 An example on edge labeling (Fan, C. et al., 2019a)

A novel edge labelling scheme has been proposed to describe the high-level interactions among building variables. It is compatible with numerical variables and can be summarized as a three-step approach. Firstly, each numerical building variable is transformed into categorical variable using certain data discretization techniques. Secondly, each subsequence is divided into *k* non-overlapping segments. Thirdly, the most frequent interaction mode between two variables in each segment is extracted and used for creating edge labels. A labeling example is shown in Fig.8 and Table-1, where two temporal subsequences are firstly discretized into two levels (i.e., *Low* and *High*) and three temporal segments, based on which the most frequent interaction modes are extracted and transformed into an edge label of {*1-2-1*}.

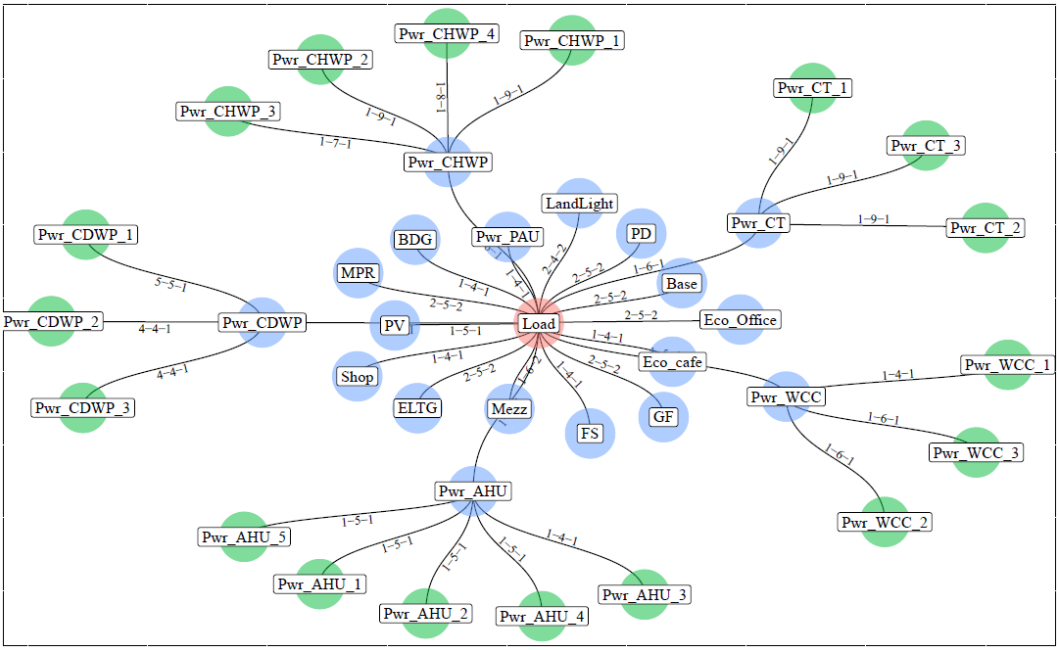
Table-1 an example notation scheme for interaction modes (Fan, C. et al., 2019a)

Inner variable A	Outer variable B	Interaction mode	Notation
Low	Low	{Low, Low}	1
Low	High	{Low, High}	2
High	Low	{High, Low}	3
High	High	{High, High}	4

4.2.3 Graph mining-based knowledge discovery and applications

The methods have been applied to analyze building operational data retrieved from an exhibition building in Hong Kong. Graphs have been generated on a daily basis considering the power consumptions of different spatial rooms and services systems, including three water-cooled chillers (*WCC-1 to 3*), four chilled water pumps (*CHWP-1 to 4*), three condenser water pumps (*CDWP-1 to 3*), three cooling towers (*CT-1 to 3*), five air-handling units (*AHU-1 to 5*), one primary air-handling unit (*PAU*), the power consumptions of outdoor landscape lighting (*LandLight*), the normal power and lighting consumptions of the eco-areas (*Eco-office* and *Eco-café*), basement area (*Base*), G/F common area (*GF*), multi-purpose room (*MPR*) and mezzanine area (*Mezz*). Each numerical variable has been discretized into three levels, i.e., *Idle*, *Low* and *High* and therefore, there are 9 interaction modes between two variables. An example daily graph on July 16, 2013 is shown in Fig.9. The central vertex represents the total building cooling load. The first layer represents system-level variables and is shown as blue circles. The second layer represents component-level variables and is shown in green circles. The daily subsequence of each variable is divided into three non-overlapping segments, based on which pairwise interactions are extracted as edge labels. The graph is capable of representing both temporal, spatial and hierarchical information among building variables.

Afterwards, FSM algorithms have been applied to find frequent sub-structures in daily operations. Daily graphs were divided into four sets accordingly to *Month* and *Day Type*, which represent effects of seasonalities and occupancy schedules. Once the frequent subgraphs are obtained, a filtering process can be applied to find unmatched parts in daily graphs, based on which anomaly detection is achieved. As an example, the daily graph on October 30, 2013 (i.e., shown in Fig.10) has been identified as abnormal given the frequent subgraph shown in Fig.11. The matched and unmatched parts are depicted



in green and red circles respectively. Abnormal interactions are observed between *Load* and *LandLight*, *Pwr_AHU* and *Pwr_AHU_2*. The normal interaction modes between the total building cooling load and landscape lighting should be "2-4-2", representing {*Idle*, *Low*}, {*Low*, *Low*} and {*Idle*, *Low*} respectively. However, the interaction mode between 0 a.m. to 8 a.m. changed to {*Idle*, *Idle*} on October 30, 2013, indicating that there is no landscape lighting at all. It was later found that the landscape lighting should be automatically switched on from 7 p.m. to 7 a.m. Such anomaly might be caused by manual faults or maintenance activities.

Fig.9 An example daily graph generated on July 16, 2013 (Tuesday) (Fan, C. et al., 2019a)

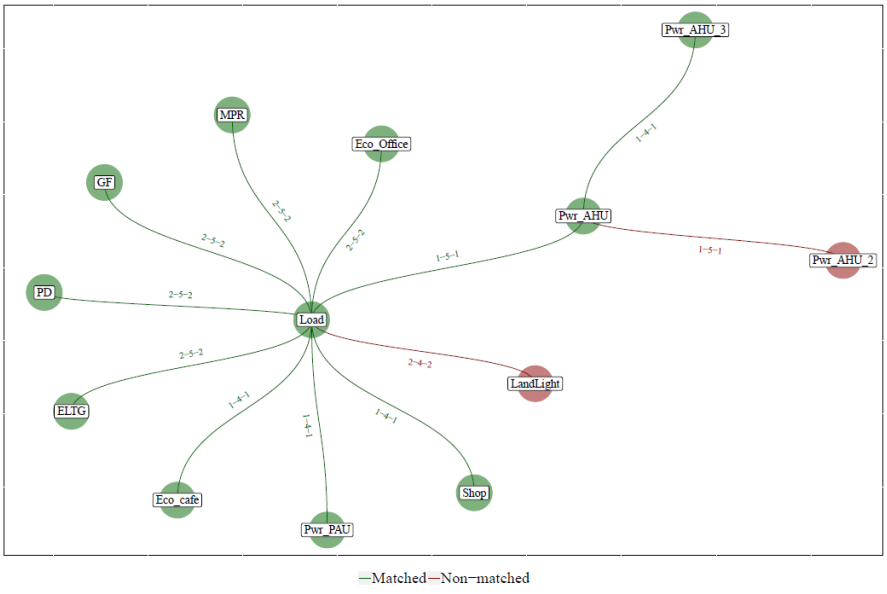


Fig.10 The daily graph on October 30, 2013 (Wednesday) (Fan, C. et al., 2019a)

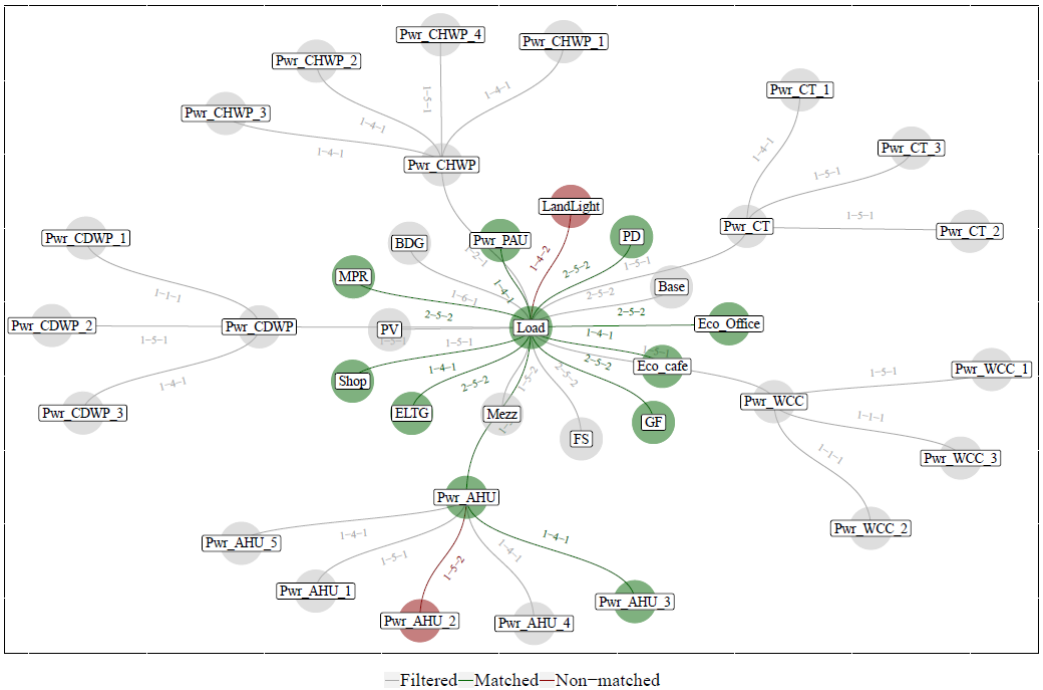


Fig.11 The reference frequent subgraph used for the anomaly detection (Fan, C. et al., 2019a)

5 Conclusive remarks

The imperative needs to fill the information gaps between various stages of a building life cycle and to improve the interoperability of various design, simulation and management software tools in the AEC industry stimulates the R&D on BIM and BAS data integration. In view of the technical maturities of BIM and BAS as well as the deep market penetrations in their own application domains, it is impossible that one of them can replace the other in the foreseeable future. However, data integration will become technically feasible in the near future given the tremendous efforts made on it. A diversity of data schemas has been proposed to define the semantic representations of big building data; however, universally accepted open metadata schemas have yet to emerge. Although IFC and gbXML are widely adopted to define BIM data and achieve data exchanges between BIM and BES, their compatibility in the popular software tools for building operation and maintenance (like BAS and Computer Aided Facility Management System) is very low. Using BIM data for building operation management always involved significant manual data operations like data point mapping. In recent year, with the development of smart buildings, BAS also encountered the big data challenge. BACnet is the open communication protocol for building automation and control networks, also an international standard (ISO 16484-5), which has been widely adopted by BAS manufacturers. BACnet, Haystack and Brick are making joint endeavor in developing a uniform machine-readable semantic metadata schema for buildings, which is a very important and essential step to data integration and automated big building analytics. To achieve convenient and reliable data exchange and integration, BIM and BAS make their own efforts. Fortunately, the efforts are not undertaken completely in parallel but take each other into consideration. Semantic Web technology for data integration is one potential solution for BIM and BAS data integration and big building data analytics. Linked data structure can effectively link the static BIM data and dynamic BAS data.

Big data analytics for analyzing integrated BIM and BAS data can be carried out in two ways. As BIM data are static, a fixed data file or table can be generated by BIM which mainly contains the information about geometry of building envelops, thermal properties of building materials, locations and design parameters of major equipment, and etc., which do not change after design and installation. Big data analytics can be carried out on the massive building operational data from BAS only. In case the data in the BIM data file is needed, date query will be performed to retrieve the data. In this way, the difficulty in integrating static data and dynamic data can be avoided, but it is not beneficial to discover spatiotemporal knowledge. The second way is to apply advanced machine learning and data mining algorithms to the analysis of integrated data with proper format directly, like graphs. Emerging graph mining techniques, such as frequent subgraph mining and graph neural networks, can be adopted to

discover novel spatiotemporal insights from the graphs, a representation of the relationships among linked data. In either way, uniform machine-readable semantics and data schemas for BIM and BAS data are needed for convenient data integration, and a generic big data analysis framework is needed to guide the process of building big data analysis for AEC professions, the majority of whom are not familiar with big data analytics. To this regard, the data analytics used should also present high interpretable features, which helps AEC professionals to better understand the knowledge discovered. This chapter proposes a big data analysis framework which embraces advance data mining and machine learning algorithms for utilizing heterogeneous big building data. Two applications implementing the framework in building energy management are demonstrated. The first is short-term building energy predictions, which serve as the foundation for typical control tasks in building operation management. The methodology proposed can achieve automatic, accurate and interpretable predictions based on the use of unsupervised, supervised and interpretable machine learning techniques. The second is to discover high-level interactions among building systems using a novel graph-based knowledge discovery methodology. Graphs have been used to integrate and represent temporal and hierarchical information in building operations, based on which frequent subgraph mining techniques are used for knowledge discovery. The knowledge discovered have been used to detect faults and anomalies in building operations. It can be foreseen that with the advances in open building data standards/schemas and big data analytics, the AEC industry is embracing innovations in building energy management in the era of big data.

References

ASHRAE (2018), ASHRAE's BACnet Committee, Project Haystack and Brick Schema Collaborating to Provide Unified Data Semantic Modeling Solution, accessed 12 August at https://www.ashrae.org/about/news/2018/ashrae-s-bacnet-committee-project-haystack-and-brick-schema-collaborating-to-provide-unified-data-semantic-modeling-solution

Asl, M. R., Zarrinmehr, S., Bergin, M., & Yan, W. (2015). BPOpt: A framework for BIM-based performance optimization. Energy and Buildings, 108, 401-412.

BACnet, accessed 12 August 2020 at http://www.bacnet.org/

Bhattacharya, A., Ploennigs, J., & Culler, D. (2015, November). Short paper: Analyzing metadata schemas for buildings: The good, the bad, and the ugly. In Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (pp. 33-34).

Borgelt, C., & Berthold, M. R. (2002, December). Mining molecular fragments: Finding relevant substructures of molecules. In 2002 IEEE International Conference on Data Mining, 2002. Proceedings. (pp. 51-58). IEEE.

Brick, accessed 12 August at https://brickschema.org/

buildingSMART, IFC., accessed 12 August 2020 at http://www.buildingsmart-tech.org

Butzin, B., Golatowski, F., & Timmermann, D. (2017). A survey on information modeling and ontologies in building automation. In IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society (pp. 8615-8621). IEEE.

Chong, A., Xu, W., Chao, S., & Ngo, N. T. (2019). Continuous-time Bayesian calibration of energy models using BIM and energy data. Energy and Buildings, 194, 177-190.

Cook, D. J., & Holder, L. B. (1993). Substructure discovery using minimum description length and background knowledge. Journal of Artificial Intelligence Research, 1, 231-255.

Cook, D. J., & Holder, L. B. (2000). Graph-based data mining. IEEE Intelligent Systems and Their Applications, 15(2), 32-41.

Cormier, A., Robert, S., Roger, P., Stephan, L., & Wurtz, E. (2011, November). Towards a BIM-based service oriented platform: application to building energy performance simulation. In Proceedings of the 12th conference of international building performance simulation association, Sydney, Australia (Vol. 1416).

David Q., John L., Neil C., (2020). Digital Twins: Answering The Hard Questions, ASHRAE Journal, Aug, 2020.

Dong, B., Lam, K. P., Huang, Y. C., & Dobbs, G. M. (2007, December). A comparative study of the IFC and gbXML informational infrastructures for data exchange in computational design support environments. In Building Simulation 2007, BS 2007.

Dong, B., O'Neill, Z., & Li, Z. (2014). A BIM-enabled information infrastructure for building energy Fault Detection and Diagnostics. Automation in Construction, 44, 197-211.

Fan, C., Sun, Y., Xiao, F., Ma, J., Lee, D., Wang, J., & Tseng, Y. C. (2020). Statistical investigations of transfer learning-based methodology for short-term building energy predictions. Applied Energy, 262, 114499.

Fan, C., Xiao, F., Song, M., & Wang, J. (2019a). A graph mining-based methodology for discovering and visualizing high-level knowledge for building energy management. Applied Energy, 251, 113395.

Fan, C., Sun, Y., Zhao, Y., Song, M., & Wang, J. (2019b). Deep learning-based feature engineering methods for improved building energy prediction. Applied Energy, 240, 35-45.

Fan, C., Wang, J., Gang, W., & Li, S. (2019c). Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. Applied Energy, 236, 700-710.

Fan, C., Xiao, F., Yan, C., Liu, C., Li, Z., & Wang, J. (2019d). A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. Applied Energy, 235, 1551-1560.

Fan, C., Xiao, F., Li, Z., & Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy and Buildings, 159, 296-308.

Gao, X. & Pishdad-Bozorgi, P. (2019). BIM-enabled facilities operation and maintenance: a review, Advanced Engineering Informatics. 39, 227-247.

gbXML, Green Building XML (gbXML) Schema, accessed 12 August 2020 at https://www.gbxml.org/index.html

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

Gopalakrishnan, K., Agrawal, A., & Choudhary, A. (2017). Big Data in building information modeling

research: survey and exploratory text mining. MOJ of Civil Engineering, 3(6), 00087.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, 1(1), 1-136.

Hu, S., Corry, E., Horrigan, M., Hoare, C., Dos Reis, M., & O'Donnell, J. (2018). Building performance evaluation using OpenMath and Linked Data. Energy and Buildings, 174, 484-494.

IEA (2019), Global Status Report for Buildings and Construction 2019, IEA, Paris

Jalaei, F., Jalaei, F., & Mohammadi, S. (2020). An integrated BIM-LEED application to automate sustainable design assessment framework at the conceptual stage of building projects. Sustainable Cities and Society, 53, 101979.

Kim, H., Shen, Z., Kim, I., Kim, K., Stumpf, A., & Yu, J. (2016). BIM IFC information mapping to building energy analysis (BEA) model with manually extended material information. Automation in Construction, 68, 183-193.

Kim, J. B., Jeong, W., Clayton, M. J., Haberl, J. S., & Yan, W. (2015). Developing a physical BIM library for building thermal energy simulation. Automation in construction, 50, 16-28.

Kota, S., Haberl, J. S., Clayton, M. J., & Yan, W. (2014). Building Information Modeling (BIM)-based daylighting simulation and analysis. Energy and Buildings, 81, 391-403.

Kuramochi, M., & Karypis, G. (2004, November). Grew-a scalable frequent subgraph discovery algorithm. In Fourth IEEE International Conference on Data Mining (ICDM'04) (pp. 439-442). IEEE.

Lee, D., Cha, G., & Park, S. (2016). A study on data visualization of embedded sensors for building energy monitoring using BIM. International journal of precision engineering and manufacturing, 17(6), 807-814.

Lu, Q., Xie, X., Parlikad, A. K., & Schooling, J. M. (2020). Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance. Automation in Construction, 118, 103277.

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

Natephra, W., Yabuki, N., & Fukuda, T. (2018). Optimizing the evaluation of building envelope design for thermal performance using a BIM-based overall thermal transfer value calculation. Building and Environment, 136, 128-145.

Nijssen, S., & Kok, J. N. (2004, August). A quickstart in frequent structure mining can make a difference. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 647-652).

Oti, A. H., Kurul, E., Cheung, F., & Tah, J. H. M. (2016). A framework for the utilization of Building Management System data in building information models for building design and operation. Automation in Construction, 72, 195-210.

Petrushevski, F., Montazer, M., Seifried, S., Schiefer, C., Zucker, G., Preindl, T., ... & Kastner, W. (2018, June). Use Cases for Improved Analysis of Energy and Comfort Related Parameters Based on BIM and BEMS Data. In Workshop of the European Group for Intelligent Computing in Engineering (pp. 391-413). Springer, Cham.

Pezeshki, Z., Soleimani, A., & Darabi, A. (2019). Application of BEM and using BIM database for BEM: A review. Journal of Building Engineering, 23, 1-17.

Press, Gil. (2016) "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says." Forbes, Forbes Magazine, 23 Mar. 2016,

www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/.

Project Haystack, accessed 12 August 2020 at https://project-haystack.org/

Quinn, C., Shabestari, A. Z., Misic, T., Gilani, S., Litoiu, M., & McArthur, J. J. (2020). Building automation system-BIM integration using a linked data structure. Automation in Construction, 118, 103257.

Shan, K., Wang, S., Gao, D. C., & Xiao, F. (2016). Development and validation of an effective and robust chiller sequence control strategy using data-driven models. Automation in Construction, 65, 78-85.

Tang, S., Shelden, D. R., Eastman, C. M., Pishdad-Bozorgi, P. & Gao, X. (2020). BIM assisted Building Automation System information exchange using BACnet and IFC, Automation in Construction, 110, 103049.

Wang, L. (2017). Heterogeneous data and big data analytics. Automatic Control and Information Sciences, 3(1), 8-15.

Wang, S. (2009). Intelligent buildings and building automation. Routledge.

Wörlein, M., Meinl, T., Fischer, I., & Philippsen, M. (2005, October). A quantitative comparison of the subgraph miners MoFa, gSpan, FFSM, and Gaston. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 392-403). Springer, Berlin, Heidelberg.

Xue, X., Wang, S., Sun, Y., & Xiao, F. (2014). An interactive building power demand management strategy for facilitating smart grid optimization. Applied Energy, 116, 297-310.

Yan, X., & Han, J. (2002, December). gspan: Graph-based substructure pattern mining. In 2002 IEEE International Conference on Data Mining, 2002. Proceedings. (pp. 721-724). IEEE.

Yan, X., & Han, J. (2003, August). CloseGraph: mining closed frequent graph patterns. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 286-295).