Manuscript for *Building Simulation*

Topical issue on *Advanced Data Analytics for Building Energy Modeling and Management*

Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches

Cheng Fan[1], Da Yan[2, *], Fu Xiao[3, **], Ao Li[3], Jingjing An[4], Xuyuan Kang[2]

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

[2]Building Energy Research Center, School of Architecture, Tsinghua University, Beijing, China

[3]Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong, China

[4]School of Environment and Energy Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China

Emails of the corresponding authors:

* yanda@tsinghua.edu.cn

** linda.xiao@polyu.edu.hk

Acknowledgement

The authors gratefully acknowledge the support of this research by the Research Grant Council of Hong Kong SAR (152075/19E), the National Natural Science Foundation of China (No. 51908365), and the National Natural Science Foundation of China (No. 51778321).

Abstract

Buildings have a significant impact on global sustainability. During the past decades, a wide variety of studies have been conducted throughout the building lifecycle for improving the building performance. Data-driven approach has been widely adopted owing to less detailed building information required and high computational efficiency for online applications. Recent advances in information technologies and data science have enabled convenient access, storage, and analysis of massive on-site measurements, bringing about a new big-data-driven research paradigm. This paper presents a critical review of data-driven methods, particularly those methods based on larger datasets, for building energy modeling and their practical applications for improving building performances. This paper is organized based on the four essential phases of big-data-driven modeling, i.e., data preprocessing, model development, knowledge post-processing, and practical applications throughout the building lifecycle. Typical data analysis and application methods have been summarized and compared at each stage, based upon which in-depth discussions and future research directions have been presented. This review demonstrates that the insights obtained from big building data can be extremely helpful for enriching the existing knowledge repository regarding building energy modeling. Furthermore, considering the ever-increasing development of smart buildings and IoT-driven smart cities, the big data-driven research paradigm will become an essential supplement to existing scientific research methods in the building sector.

Keywords

Advanced data analytics; Big-data-driven; Building energy modeling; Building operational data; Building performance.

1 Introduction

Buildings represent a significant amount of total energy consumption in the world. According to the World Energy Balances (International Energy Agency (IEA), 2019), the building sector accounts for more than 30% of the final energy consumption globally and contributes to nearly 40% of global carbon-dioxide emissions. Energy consumption and carbon emissions are expected to continue increasing in upcoming years (IEA, 2019). As such, building energy performance modeling has proven to be an essential technique for evaluation and optimization of building design and operation (Harish & Kumar, 2016), thus improving the management of building energy system.

Traditional physics-based building performance simulation (BPS) has been well researched and developed over the past 40 years (Foucquier et al., 2013), and widely applied to the assessment and optimization of building energy system design (Attia et al., 2012), the development and evaluation of operational control and optimization strategies of building systems (Coakley et al., 2014; Li & Wen, 2014), and policy making on building regulations and power grid operations (Chung, 2011). A series of BPS software programs have also emerged during the past several decades, including EnergyPlus (Crawley et al., 2001), DeST (Yan et al., 2008), and ESP-r (Strachan, Kokogiannakis & Macdonald, 2008). BPS typically builds upon physical principles and thermodynamics as well as heat and mass transfer, and relies heavily on meteorological data and detailed building information, including the building

envelope configuration and properties, the air conditioning system design and operational parameters, and energy-related behaviors of occupants. Although BPS has made significant contributions to improvements in building performance in terms of building energy efficiency (Fesanghary et al., 2012), indoor environments (Tian et al., 2018), and policy effectiveness (Gao et al., 2014), it has encountered new challenges, facing increasingly large and complex buildings and building energy systems, along with a characterization of realistic occupant behaviors.

With the development of urbanization and advancements in building technologies, large high-rise buildings with complicated structures and multiple functions have emerged in recent years. The preparation of inputs for the BPS of such buildings has become an overwhelming and time-consuming task (Amasyali & El-Gohary, 2018). Meanwhile, large buildings are served by complex energy systems to provide desired indoor environment (Zhao & Magoulès, 2012). The increasing complexity of the coupled effects of the building envelope, energy systems (e.g., air conditioning and thermal storage), automated control systems, and climate conditions have brought about a significant challenge to efficient building energy modeling (Xiao & Fan, 2014). Moreover, occupants' energy related behaviors are among the most essential factors in overall building energy performance modeling (Yan et al., 2015). Previous studies on building occupant behavior has improved general understanding of its impact on building energy (Yan et al., 2017); however, at the same time exhibits the complexity of building-occupant interactions (Hong et al., 2017). In this context, little

confirmed knowledge regarding occupant behavior in large and new buildings is available. For advanced building systems with feedback control, the integration of occupant behavior and building components has been significantly enhanced (Day & Gunderson, 2015), resulting in more complicated nonlinear building dynamics and significant difficulties in building energy performance modeling.

The ever-growing complexity of building energy systems and continually enhanced interactions between occupant behavior and building components have brought major challenges to building performance modeling. Under these circumstances, data-driven approach is of particular interest as it requires little priori knowledge of building and energy system configurations and integrations, and the building energy behavior can be quickly learned from the building operation data (Amasyali & El-Gohary, 2018; Bourdeau et al., 2019). The development of sensing technologies and building automation systems has provided reliable sources for big data on building operations, and advanced data mining and machine learning algorithms offer significant technical support for big data analytics of building energy use. Big-data-driven analytics is attracting growing interest in terms of building energy performance modeling (Wang & Chen, 2019) for building design (Ahmad et al., 2018; Wei et al., 2018), operation control (Maddalena et al., 2020; Schmidt & Åhlund, 2018; Mehmood et al., 2019; Fan et al., 2018), and policy making process (Hu et al., 2020).

In view of the challenges and perspectives mentioned above, this paper offers a review of big-data-driven modeling and analysis of building performance. First, the

general structure and framework of data-driven analytics are described, followed by a review of the data sources and data pre-processing techniques. After a technical review of big-data-driven modeling and the associated model evaluation and interpretation, the current applications regarding the building design, operation control, and policy making, are reviewed. The final section provides critical discussions from different perspectives and offers an outlook of the development and enhancement of data analytics in building energy performance analytics.

2 Framework for data-driven model development

A significant amount of R&D has been conducted on data-driven modeling of the building performance. It was found that the majority of studies have followed a general framework, as shown in Figure 1, which consists of four major phases. The first is a data pre-processing, which serves as an initial step to transforming raw data into useful information for predictive modeling. Typical tasks for data pre-processing include data cleaning, reduction, transformation, and partitioning. The second step is to develop data-driven models using different machine learning algorithms and training schemes. The main target is to develop robust and reliable models with sufficient capabilities in terms of modeling static and dynamic relationships. Once the data-driven models are constructed, a knowledge post-processing is conducted to evaluate the model generalization performance and its underlying inference

mechanisms. The insights obtained will be helpful in evaluating the validity of the model and extending its practicality. Finally, data-driven models are applied to facilitate the decision making of various building energy management tasks, such as fault detection and diagnosis and optimal controls. In light of this general framework, this paper reviews the representative studies conducted on the four phases above.

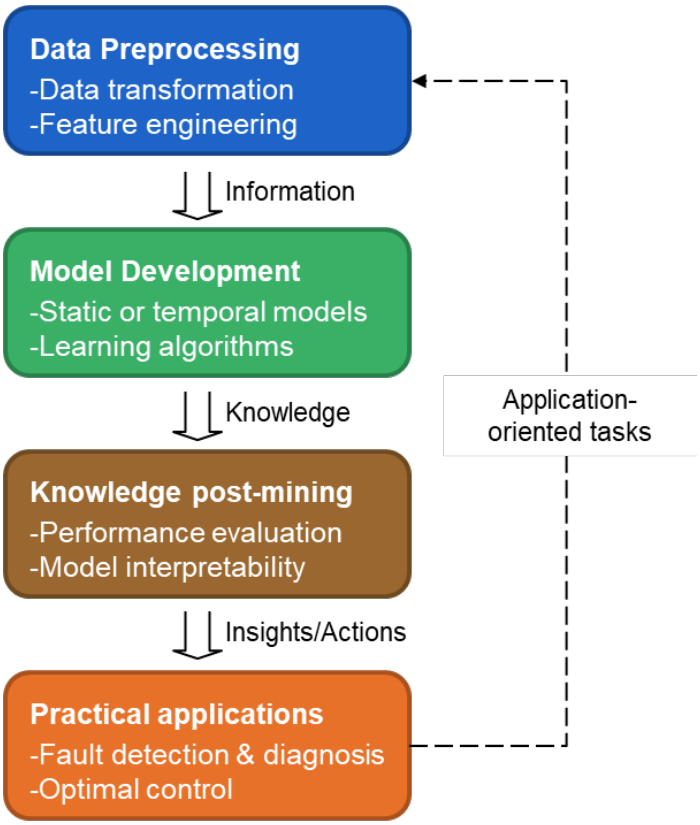

Figure 1 General framework for data-driven model development

3 Data sources and data preprocessing

3.1 Data source

Data are the fuel of all data-driven approaches and techniques. In existing studies on data-driven building energy analysis and modeling, the data sources can be classified into two main categories, i.e., measured and simulated data, as shown in Table 1.

Table 1 Data sources used in building energy modeling

	Measured data	Simulated data
Private data	• Experimental data • Monitored data e.g., BASs, weather stations, energy meters, IoT sensors (Bottaccioli et al. 2017; Brundu et al. 2016), onsite surveys (Raftery et al. 2011)	**DeST** (Li et al. 2009a, 2009b), **Energyplus** (Zhao & Magoulès, 2010; Yezioro et al. 2008; Wong et al. 2010), **TRNSYS** (Du et al. 2014), **Ecotect** (Tsanas & Xifara 2012), **eQuest** (Yezioro et al. 2008), etc
Public-available data	Great Building Energy Predictor Shootout (Karatasou et al. 2006), Building data genome project (Miller & Meggers 2017), UCI machine learning repository (Tsanas et al. 2012; Marino et al. 2016), NOAA online climate data (NOAA, 2013),etc.	Open AI (U.S. Department of energy) (https://openei.org/doe-opendata/dataset)

Measured data can be obtained from experiments and on-site measurements. On-site measurement data are directly collected from building automation systems (BASs), energy meters, weather stations, on-site surveys, and IoT sensors. Measurement data can reveal and reflect the real operational conditions of buildings and their energy systems. However, the quality of the measured data is typically low owing to the presence of measurement noises, uncertainties, sensor faults, and insufficient calibrations. Sensor-based data collection approaches require examining and verifying the quality of the data, which are the main tasks of data cleaning described below. Simulation data are collected from physics-based models and simulation tools of

either real or virtual buildings. Simulation tools commonly used in previous research have adopted a data-driven approach, and include DeST (Li et al., 2009a, 2009b), TRNSYS (Du et al., 2014), Energyplus (Zhao & Magoulès, 2010; Yezioro et al., 2008; Wong et al., 2010), Ecotect (Tsanas & Xifara, 2012), and eQuest (Yezioro et al., 2008). The simulated data should be noise free and without measurement errors, operational mistakes, or faults. However, the modeling accuracy, simulation assumptions (e.g., occupancy schedule), and ideal operational conditions (e.g., without considering a performance degradation) indicate that the simulation data do not represent the actual building operations or performance in a meaningful way. Li et al. (2015) found that current building energy simulation tools have limited reliability in terms of a performance assessment of energy conservation measures, considering using the assumed occupancy data and adopting a single energy model for cross-estimation.

An increasing number of publicly available datasets (sometimes called benchmarking datasets), which may consist of either measured or simulation data, have been created in recent years by a number of research institutions, companies, and academics. The open-source building energy datasets employed in previous research into building energy analytics include ASHRAE's Great Building Energy Predictor Shootout (Karatasou et al., 2006), the Building Data Genome Project (Miller & Meggers, 2017), and the UCI machine learning repository (Tsanas et al., 2012; Marino et al., 2016). In addition, the National Oceanic and Atmospheric Administration provides public access to high-quality historical weather data all around the world (NOAA, 2013).

These datasets allow researchers to make comparisons of different algorithms or models using the same datasets, and obtain more general conclusions and insight.

Once the raw data are collected, the data must be provided in the proper amount, structure, and format that perfectly suit each data analytic task (Garcia et al., 2015). Data pre-processing aims to fulfill this requirement, for which four tasks are mainly carried out, i.e., data cleaning, data transformation, data reduction, and data partitioning, as shown in Figure 2. Data cleaning aims to enhance the data quality by filling in missing values and removing outliers. Data transformation is conducted when a proper data attribute (e.g., numerical and categorical) or data scale is required by specific modeling algorithms. Data reduction aims to identify the most relevant/influential factors/variables in modeling, reduce the dimensions of the datasets, and improve the calculation efficiency. Data partitioning aims to divide a large dataset into several small datasets, which can be analyzed separately to improve the sensitivity and robustness of the model. It is worth mentioning that these four data pre-processing tasks are not compulsory for big-data-driven analytics. Researchers can design their own data pre-processing procedures based on individual requirements and conditions. The following sections introduce each task successively.

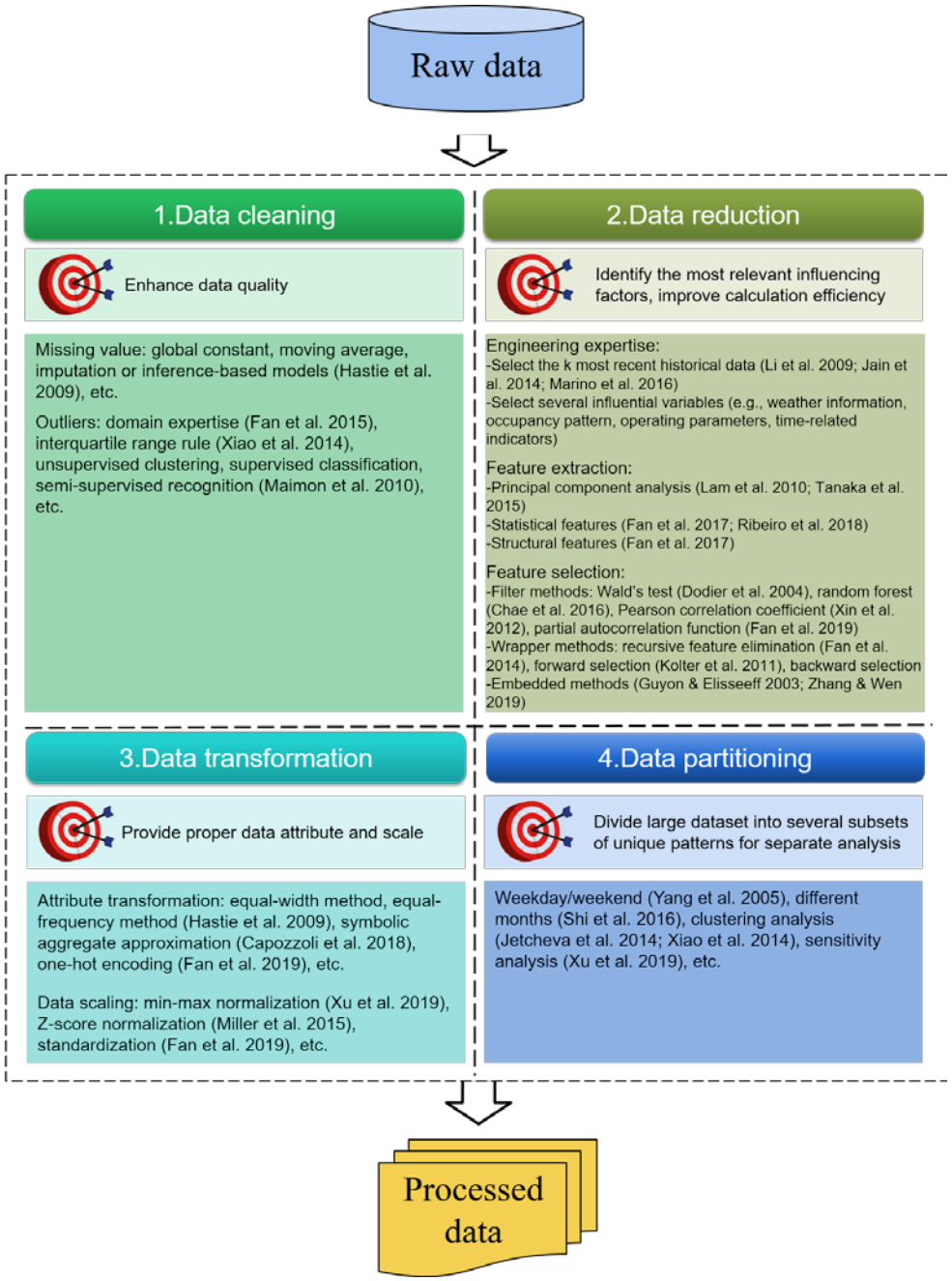

Figure 2 Typical data preprocessing tasks for building energy modeling

Once the raw data are obtained, the next step is to preprocess the data for use in the development of predictive models. The input data must be provided in the proper amount, structure, and format that perfectly suit each data analytic task (Garcia et al., 2015).

3.2 Data cleaning

Data cleaning aims to enhance the data quality. The accuracy and reliability of data-driven modelling are largely determined by the quality of the data. Two typical problems with automatically measured data are missing values and outliers, and thus two main tasks of data cleaning are the handling of missing values and outlier detection and removal.

Missing values, or missing data, occur when no data values are stored for the variable for a short time period owing to sensor faults or communication problems. Missing values can be filled in using the global constant, moving average, imputation, or inference-based model (Hastie et al., 2009). Outliers are observations which appear to be inconsistent with the remainder of a specific dataset (Barnett & Lewis, 1994). Outliers may arise for various reasons, such as human mistakes, instrument errors, and a sudden change in the system behavior. Outliers can be identified based on domain expertise (Fan et al., 2015), or using unsupervised clustering, supervised classification, or semi-supervised recognition (Maimon & Rokach, 2010). Fan et al. (2014) employed the generalized extreme studentized deviate algorithm to detect outliers in a feature space. In addition, Xiao and Fan (2014) used the interquartile range rule to detect outliers in raw BAS datasets.

3.3 Data transformation

Data transformation consists of data attribute/type transformation and data scaling.

Building operational data consist of both numerical (quantitative) and categorical (qualitative) data. Typical examples of numerical data include temperature measurements, power consumption, the flow rate, and water pressure. Typical examples of categorical data are the ON/OFF control and state signals and time-related indicators. Many data-driven models and techniques have special requirements in terms of the data format required. For example, association rule mining (ARM) algorithms, such as *a priori* and frequent-pattern growth algorithms, can only handle categorical data, and numerical data should therefore be transformed into categorical data before applying the ARM. Numerous methods for discretizing data from a numeric form into a categorical form are available. Equal-width and equal-frequency methods have been widely used owing to their simplicity and reliability (Hastie et al., 2009). An equal-width binning method divides the data into m intervals of equal size, whereas an equal-frequency method divides the data into m groups containing approximately the same number of observations. Capozzoli et al. (2018) also adopted a symbolic aggregate approximation (SAX) to transform a time series into a symbolic string. Some machine learning algorithms cannot operate on categorical data directly (e.g., an artificial neural network), and require all input and output variables to be numeric. One-hot encoding is usually adopted to transform categorical data into a numerical form (Fan et al., 2019). Some predictive data-driven techniques (e.g., a support vector machine and an artificial neural network) perform better if the input data have similar scales. However, the scales of BAS data are

extremely different owing to the different units used. For instance, the power measurements may change for 0 to 4,000 kW, the temperature measurements may change from 0 °C to 40 °C, and a typical control signal usually changes from 0 to 1. Therefore, scaling methods should be applied during the data preprocessing. Commonly used scaling methods include max–min normalization (Xu et al., 2019), Z-score normalization (Miller et al., 2015), and standardization (Fan et al. 2019).

3.4 Data reduction

Building operational data are usually stored in such a format that each column represents the values of a variable at consecutive time instants, and each row represents an observation sampled at a specific instant in time (Fan et al., 2015). Current BASs monitor and control hundreds and even thousands of devices and items of equipment used in buildings. The volume of the stored data continues to increase over time during the building lifecycle. As a result, the building operational data become highly dimensional in temporal (number of rows) and spatial spaces (number of columns). Using all influential variables as model inputs might increase the risk of over-fitting and result in unaffordable computational costs. Redundant variables in the input dataset will decrease the accuracy, stability, and effectiveness of the model.

Data reduction (also called feature engineering in big data analytics) aims to identify the most relevant/influential factors/variables, reduce the dimensions of the datasets, minimize the risk of over-fitting, improve the calculation efficiency, and meanwhile retain or improve the model performance. In general, there are three commonly used

approaches to selecting a model input.

The first is to select the variables of interests based on domain knowledge and engineering expertise. In a typical engineering method, several influential variables (e.g., weather information, occupancy pattern, or HVAC operational signal), or the k-most recent historical data, are selected as model inputs, or a candidate input pool is formed for further feature selection/extraction.

The second approach is to adopt a feature extraction method, such as a principal component analysis, in which the new low-dimensional variables are linear combinations of the original high-dimensional variables. By projecting onto the first few principal directions, a new set of data with lower dimensions is obtained through a linear combination of the original data. Fan et al. (2017) adopted four feature extraction methods, namely, engineering, statistical, structural, and deep learning feature extraction, for the measurements taken during the previous 24 h for comparison. Ribeiro et al. (2018) also extracted statistical features (the maximum, mean, and minimum values of the weather variables) as model input. The newly extracted features can be directly used as model input or added into the candidate input pool for feature selection. One disadvantage of a feature extraction method is that none of the original data can be abandoned, and it may be difficult to interpret the inputs (Guyon et al., 2003).

The third approach is to use a feature selection method in which the variables most relevant to the current problem are chosen (Xu et al., 2019). This type of method

relies on the concept of subset selection. Commonly used feature selection methods can be further classified into filter, wrapper, and embedded methods. With a filter method, features are ranked and selected according to certain univariate metrics, such asPearson's correlation coefficient. Dodier et al. (2004) used Wald's test to evaluate the relevance of the input variables, including environmental variables, time-related variables, and time-lag variables, for building energy prediction. Chae et al. (2016) adopted a random forest algorithm to assess the importance of the variables by measuring the candidate parameters in terms of their impact on the prediction response, and ranked the variables based on both the permutation importance and Gini importance. Fan et al. (2019) employed partial autocorrelation functions to select the maximal time lag considered. The disadvantage of a filter method lies in the possible redundancy of the subset selected. A wrapper method is used to evaluate the usefulness of a subset by considering a certain learning algorithm. Fan et al. (2014) employed a recursive feature elimination (RFE) algorithm to extract 12 out of a total of 96 features to represent the daily energy consumption. Kolter et al. (2011) adopted a forward selection method in which the features were selected based on how much they will reduce the root mean square error (RMSE) of a linear regression predictor for the forecasting of building energy consumption. Because exhaustive searches of the subsets must be conducted, the wrapper method may incur a dramatic increase in the computational costs. Alternatively, an embedded method, which also applies a variable selection based on a certain learning algorithm, may be more efficient

because it is carried out by directly optimizing a two-part objective function with a goodness-of-fit and a penalty for a large number of input variables (Guyon et al., 2003).

3.5 Data partitioning

Most building service systems are highly dynamic and inter-correlated (Fan et al., 2015). The values of the variables and the relationships between them may vary significantly under different occupancy patterns and operating and weather conditions. Therefore, analyzing massive amounts of building operational data simultaneously may result in significant information loss. Data partitioning, or data sub-setting, is used to separate a large BAS dataset into several subsets of unique patterns, which is important for enhancing the efficiency and reliability of the knowledge discovery by separately analyzing the data in each subset.

Some researchers partition the datasets into weekdays and weekends (Yang et al., 2005), or into different months (Shi et al., 2016), based on their understanding of the building operational patterns. However, this approach may be unreliable if the building is a multi-functional complex that does not exhibit such periodic operational patterns. A clustering analysis is frequently used in data partitioning. Jetcheva et al. (2014) adopted k-means clustering to cluster the daily building load profile and temperature data and train a neural network for each cluster to find the best performing neural network. Xiao and Fan (2014) used entropy weighted k-means clustering to identify typical building energy consumption profiles and group similar

profiles for further analysis. In addition, Xu et al. (2019) conducted a sensitivity analysis and a linear regression successively to partition the load dataset.

4 Data analytics for model development

4.1 Single- and ensemble-model based approaches

The increased interest in machine learning has provided numerous algorithms for data-driven model development. Machine learning is a rather broad category which consists of statistical algorithms ranging from conventional linear algorithms (e.g., multiple linear regression and autoregressive models) to complicated nonlinear algorithms (e.g., decision trees and support vector machines) [James et al., 2017; Hastie et al., 2010]. Nonlinear machine learning algorithms are more capable of capturing complicated and dynamic relationships in building systems and therefore, have been widely adopted in recent studies to achieve better generalization performance. The following section mainly reviews studies based on nonlinear machine learning algorithms.

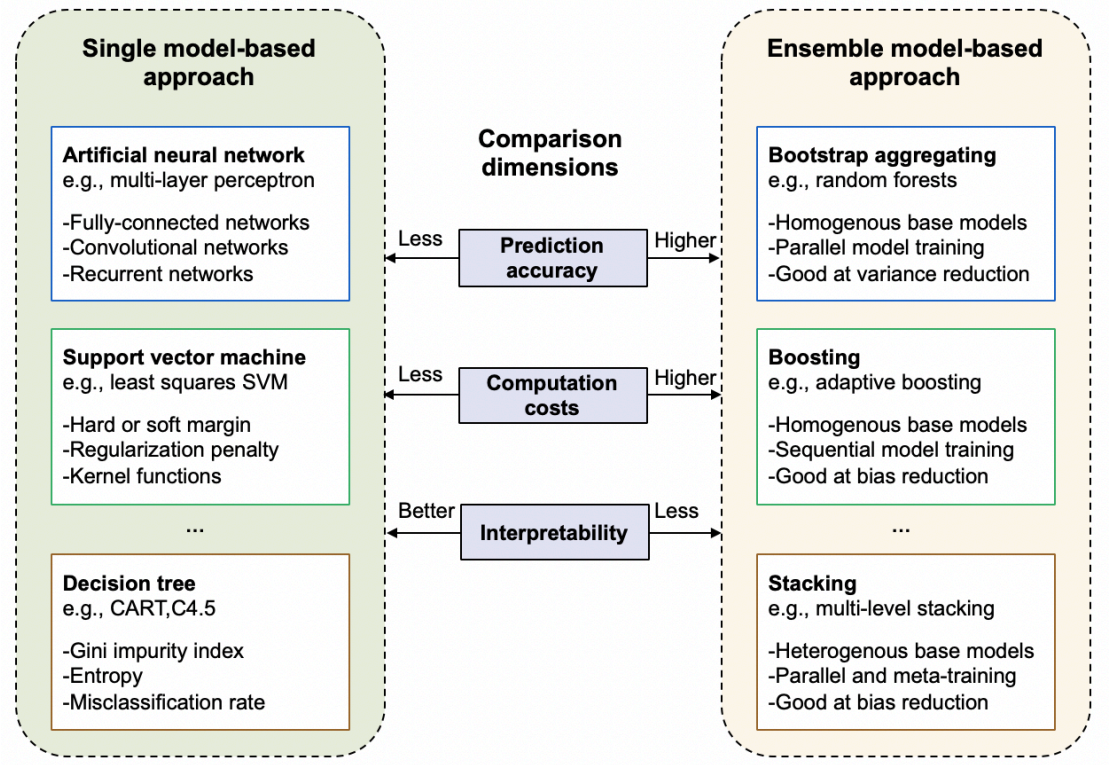

Figure 3 Single- and ensemble-model based approaches

As shown in Figure 3, the research trend in the building sector is in accordance with the developments in computer science and artificial intelligence, i.e., the modeling approach has gradually changed from single-model based approaches to ensemble-model based approaches. At the early stage, the majority of studies in the building sector adopted a single-model based approach (Wei et al., 2018; Amasyali & El-Gohary, 2018). In other words, such an algorithm will only derive a single model based on the training data. Artificial neural networks have gained great popularity owing to their wide applicability in analyzing different types of data (Goodfellow, Bengio & Courville, 2016; Fan et al., 2020). In addition to conventional fully connected networks, convolutional networks can be used to analyze image data, whereas recurrent networks are capable of analyzing sequential and time series data.

A support vector machine is another popular machine learning technique that can be used for both regression and classification problems (Cortes & Vapnik, 1995). Such models can be developed based on a hard or soft margin mechanism. Different kernel functions can be utilized to enhance the predictive power for nonlinear relationships, e.g., polynomial and radial basis kernel functions. A decision tree model adopts a tree-like graph to present the inference mechanism and can be used for either classification or regression problems (Breiman, 2001). Different metrics can be applied for splitting the tree model, e.g., the Gini impurity index, entropy, or misclassification rate (Hastie et al., 2009). Despite the encouraging results obtained, the generalization performance of a single-model based approach can be poor when applied to new datasets because each algorithm has its own intrinsic limitations and data assumptions. For instance, the decision boundary of a single decision tree model can only be rectangular and thus may be unsuitable for problems with an intrinsically smooth decision boundary (Hastie et al., 2009).

Ensemble modeling has also been proposed to enhance the robustness and reliability in data-driven models (Dietterich, 2000; Fan et al., 2014). The main idea here is to develop a set of base models, based upon which the final prediction is made. There are two general methods for developing a base model. The first method is to artificially manipulate the training data for base model development. A prominent technique is called bootstrap aggregating, which utilizes the bootstrap sampling method to create training samples for parallel base model development (Hastie et al.,

2009). The main aim is to reduce the variance in the predictions. The popular random forest algorithm can be regarded as a variation of the bootstrap aggregating technique (Breiman, 2001). Here, each base model is a decision tree model developed based on the bootstrapped training data. Extra randomness is introduced by considering a random subset of variables when applying node splitting. Another popular technique belonging to this category is called boosting, which aims to reduce the biases occurring in the predictions (Hastie et al., 2009). In such a case, base models are developed in a sequential manner, each with the aim to reduce the prediction errors resulting from the previous model. Some representative algorithms include adaptive boosting trees (Freund & Schapire, 1999) and extreme gradient boosting trees (Chen & Guestrin, 2016).

The second method is to adopt different supervised learning algorithms for the base model development. As a result, a set of heterogenous base models can be obtained based on the same training dataset. The final prediction can be obtained by either simply averaging the predictions from the base models or using a so-called stacking strategy to develop a meta-model for the final predictions.

4.2 Static and temporal relationship modeling strategies

There are two general types of modeling tasks. The first focuses on revealing the static relationships between the model inputs and outputs, while neglecting the temporal dependencies in the building operations. One such example is building energy consumption predictions based on the building-level variables, e.g., the

physics of the building envelopes, the primary building usage type, and the indoor occupancy (Zhao et al., 2020). Similar studies have been carried out at the system or component level. For instance, the power consumption of a chiller can be described by its operating parameters at the same time, e.g., the temperatures of the chilled water supplied and the returned condensing water (Chou, Hsu & Lin, 2014).

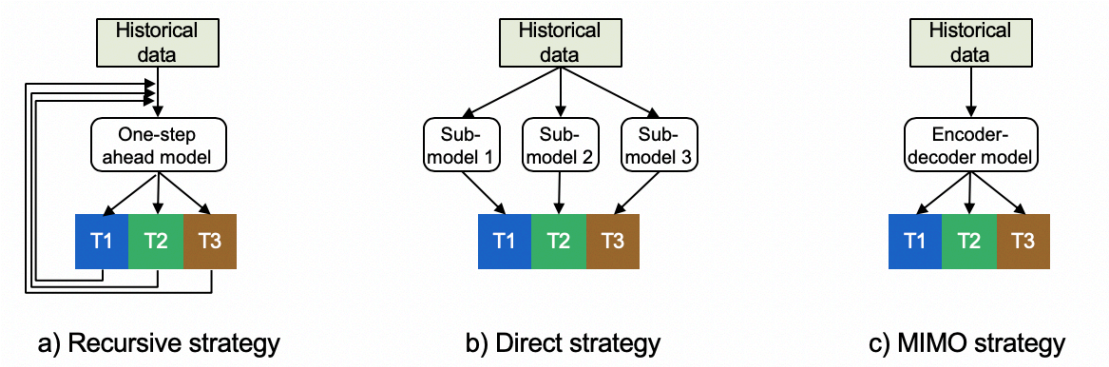

Figure 4 Strategies for multi-step-ahead predictions

By contrast, the second type focuses on revealing the temporal relationships in the building operational data. A typical task is to apply one-step or multi-step ahead predictions on the building operations (Rahman, Srikumar & Smith, 2018; Fan et al., 2019). Taking the building energy consumption as an example, the modeling task is to predict the building energy consumptions in the next m timesteps (e.g., m = 1 if one-step ahead is used) given the historical measurements of the previous n timesteps. The most essential key in dynamic modeling is to accurately capture the temporal data dependencies. As illustrated in Figure 4, there are three strategies for multi-step ahead predictions in general. The first is called a recursive strategy. The main idea here is to develop a one-step ahead prediction model and use it recursively for generating multi-step ahead predictions (Fan, Xiao & Zhao, 2017; Deb et al., 2016). Such a

strategy is easy to implement. However, the strategy suffers from an error accumulation problem, i.e., the prediction made at time T will be used as an input for following predictions, and thus prediction errors will gradually accumulate along the prediction time horizon. The second is called a direct strategy, where separate models are developed for each time step along the prediction horizon (Taieb et al., 2012). Such a strategy is highly compatible with algorithms allowing multiple outputs. For instance, an artificial neural network with m neurons at the output layer can be designed for m-step ahead predictions. Compared with a recursive strategy, a direct strategy is less affected by the error accumulation problem. The main limitation is that the predictions are made in a parallel manner and are thus essentially independent from each other and may seem to be incoherent or disconnected. The third strategy is called multi-input and multi-output (MIMO), and has been proposed to better describe the stochastic dependencies in multi-step ahead predictions (Bontempi, 2008). In theory, it can avoid the error accumulation problem in a recursive strategy while overcoming the conditional independency assumption used in a direct strategy (Taieb et al., 2012). Recurrent neural networks, which are specially designed for analyzing sequential data, have been widely used in a MIMO strategy (Chollet & Allaire, 2018). One popular MIMO learning scheme is called encoder–decoder learning. The main idea here is to develop two recurrent neural networks for input encoding and output decoding, respectively. The input data are transformed into a hidden state, based upon which the decoder is used to generate multi-step ahead predictions. Fan et al.

conducted a study to investigate the performance of these three strategies in 24-h ahead building energy predictions (Fan et al., 2019). Recurrent neural networks have also been used for the development of a prediction model. The results show that a direct strategy can achieve the best performance, whereas a recursive strategy leads to the worst performance.

5 Knowledge post-mining: Model evaluation and interpretation

One of the most essential advantages of a data-driven model is its flexibility in terms of model development. Indeed, given sufficient data measurements, functional data-driven models can be developed in a fairly straightforward manner with little domain expertise on the building physics. Nevertheless, it also imposes greater challenges in a model performance evaluation and model interpretation, which are the two main tasks in knowledge post-mining.

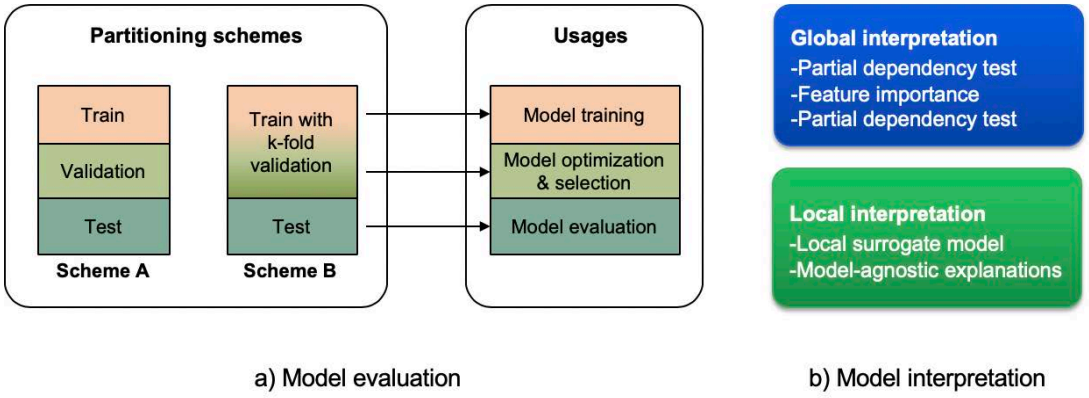

Figure 5 Data-driven model evaluation and interpretation

5.1 Model evaluation

The performance of a data-driven model is mainly evaluated based on the accuracy

metrics. To ensure the unbiasedness of the accuracy metrics in reflecting the actual generalization performance, it is essential to apply data partitioning before the model development.

A common approach is to divide the data into three separate datasets, namely, training, validation, and testing datasets. In such a case, a number of predictive algorithms are applied on the training data to derive a set of candidate models. The validation data are then used for a model comparison, where the best predictive algorithm along with its optimal parameter settings are determined. The testing data are used for an accuracy metric calculation, and the results serve as estimates for the generalization performance. Such a data partitioning approach may not be optimal when the data at hand are limited and the data acquisition costs are relatively high. One possible solution is to integrate a k-fold cross validation technique into the model training process. In such a case, the entire dataset is divided into two sets, i.e., training and testing sets. The training data are further divided into k equal-sized data folds, where k-1 folds are used for model training and the remaining serve as the validation data. The process is repeated k times such that each data sample is used in both model training and validation. Once the best predictive algorithm and its parameters are determined, the generalization can be assessed using the remaining testing data. It should be mentioned that any accuracy metrics reported based on the training and validation datasets should be regarded as invalid because they are optimistic estimates for practical applications (Hastie et al., 2009).

The accuracy metrics can be broadly divided into two groups, one for regression problems and the other for classification. The accuracy metrics for regression tasks can be further divided into two types, i.e., scale-dependent and scale-independent metrics. Scale-dependent metrics have the same units as the target variables. The most commonly used scale-dependent metrics include the root mean squared error (RMSE) and the mean absolute error (MAE). Such metrics are helpful for reflecting the error scales. However, they are inapplicable for comparing the model performance using different datasets. By contrast, scale-independent metrics, which present prediction errors using relative proportions, are more suitable for evaluating data-driven models derived from different datasets. Example metrics include the mean absolute percentage error (MAPE) and the coefficient of variance of the root mean squared error (CV-RMSE). Another popular metric is the coefficient of determination (R2), which has been widely adopted to reflect the fitting performance of linear regression models. It can be calculated based on Eq. (6), where TSS is the total sum of squares of the target variable, ESS is the explained sum of squares, and RSS is the residual sum of squares. It should be noted that such a metric is generally invalid for nonlinear models because the TSS may not be equivalent to the sum of the ESS and RSS.

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \tag{1}$$

$$MAE = \frac{\sum | y_i - \hat{y}_i |}{n} \tag{2}$$

$$MAPE = \frac{\sum | y_i - \hat{y}_i | / y_i}{n} \tag{3}$$

$$CV(RMSE) = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}}{\sum \frac{y_i}{n}} \tag{4}$$

$$F - measure = \frac{2 \times recall \times precision}{recall + precision} \tag{5}$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \overline{y})^2} \tag{6}$$

There are a variety of accuracy metrics for evaluating the classification performance. The basic metric is the accuracy, which is defined as the ratio between the correct predictions and all predictions. Such a metric can only provide a high-level description of the classification model. In-depth metrics are often needed to provide a comprehensive model evaluation. Taking the two-class classification problem as an example, a confusion matrix is typically used, as shown in Table 2. Four in-depth metrics can be formulated to describe the classification performance for each class. The positive predicted value (PPV) is also known as the precision, which describes the ratio between the true positive examples and the total number of predicted positive samples. The sensitivity or recall defines the model capability in predicting true

samples out of all actual true samples. These metrics are particularly useful when dealing with imbalanced datasets. For instance, a high recall and low precision indicate that most of the actual positive examples can be successfully identified, yet at the cost of a high false-positive rate. By contrast, a low recall and high precision indicate that the model cannot adequately identify actual positive examples and thus should have a high number of false negatives. The F-measure is formulated by considering both the recall and precision. As shown in Eq. (5), this measure is the harmonic mean of the recall and precision and should always be closer to the smaller value. Compared with conventional accuracy metrics showing one aspect of the model performance, the F-measure can provide a more comprehensive evaluation on the classification performance.

Table 2 An example confusion matrix for binary classification

Predicted vs. Actual	Actual values		Metrics
	True	False	
Predicted As True	True positive (TP)	False positive (FP)	Positive predicted value $PPV = \frac{TP}{TP + FP}$
Predicted as False	False negative (FN)	True negative (TN)	Negative predicted value $NPV = \frac{TN}{TN + FN}$
Metrics	Sensitivity or Recall $= \frac{TP}{TP + FN}$	Specificity $= \frac{TN}{TN + FP}$	Accuracy $= \frac{TP + TN}{TP + TN + FP + FN}$

5.2 Model interpretation

Compared with conventional statistical methods, advanced machine learning

algorithms are more capable of capturing complicated nonlinear relationships, although at the cost of a poor model interpretability. Model interpretability plays an essential role in knowledge post-mining and can significantly influence the model applicability in practice. First, accuracy metrics alone cannot fully justify the model validity. For instance, a naïve classification model, which simply makes predictions based on the majority class, can achieve an extremely high classification accuracy if the data are highly imbalanced. Nevertheless, such a model cannot be applied to practical applications. Second, building professionals cannot fully trust data-driven models unless the underlying inference mechanisms match their domain expertise. Therefore, it is essential to develop tools or methods to interpret the patterns or relationships learned through data-driven models.

One possible solution is to adopt algorithms with high transparency for model development, e.g., multiple linear regression and decision trees (Lipton, 2016; Doshi-Velez & Kim, 2017). Such models are easy to interpret owing to their straightforward model architectures. However, the prediction accuracy may not be satisfactory, particularly when the relationships among variables are nonlinear and complicated.

To tackle the intrinsic trade-off between the model complexity and model interpretability, an emerging research field called interpretable machine learning has become increasingly popular (Molnar, 2018). The main idea here is to develop methods for describing complicated data-driven models at different levels, i.e., global

and local levels. At the global level, the model is described in terms of its overall structure and parameters. Representative methods include partial dependency tests, individual conditional expectations, and the feature importance (Molnar, 2018). Global explanations are helpful for understanding the general model behaviors and the impact of individual variables to model the outputs. For instance, the random forest algorithm has significantly enhanced the accuracy of a single decision tree model. The importance of each input variable can be calculated based on the decrease in accuracy resulting from a random permutation. Such a variable importance can be used as a global explanation on the impact of model inputs to outputs.

At the local level, explanations are provided to describe why a certain prediction is made for an individual observation. One representative technique is a local interpretable model-agnostic explanation (LIME) (Ribero, Singh & Guestrin, 2016). This method is called “model-agnostic” because it can be integrated with any supervised learning algorithm. The key idea is to build a local surrogate model using an algorithm with high transparency to describe the simplified local relationships among the data variables. Fan et al. developed a LIME-based method to provide explanations regarding building energy prediction models (Fan et al., 2019). The local surrogate model was developed based on the permuted data in interpretable representations. Based on the local interpretation results, the authors also developed a novel metric to evaluate the quality of each individual prediction. The method helps ensure the prediction accuracy while providing additional evidence for building

professionals to justify the model validity.

From the authors' perspective, model interpretability is a key challenge in integrating a state-of-the-art machine learning technique in the building field. A model-agnostic approach is one of most promising research directions in interpretable machine learning, because it helps break the trade-off between the model complexity and model interpretability. Compared with a global explanation, local explanations are more sophisticated yet helpful because they can provide in-depth insight into the underlying data structures and inference mechanisms.

6 Application of data-driven models for improving building performances

Data-driven models are applied in various aspects throughout the building lifecycle, including building design, operation, control, and policy making. At every different stage, data-driven models serve different purposes, focusing on different types of buildings, presenting different spatial and temporal scales, utilizing data from different sources, conducting different types of machine learning methods, and delivering different outputs. Table 3 summarizes the main features of the development and application of data-driven building energy models.

Table 3 Applications of data-driven methods in building performance modeling

Application	Building type	Spatial scale	Temporal scale	Data source	Data volume	Methods	Outputs	References
Parametric analysis in design	Mostly residential	Household	Mostly hourly	Smart meter data	One year's data	Clustering methods	Typical energy use profiles	Wen et al., 2019; Yang et al., 2018; Fu et al., 2018; Popoola & Chipango, 2020; Escobar et al., 2020; Quintana et al., 2020; Satre-Meloy et al., 2020; Zhou et al., 2017; Sala et al., 2019; An et al., 2018; Rhodes et al., 2014; Zhou et al., 2017;
Optimal simulation models in design	Mostly non-residential	Building	Mostly hourly	BMS data	One year's data	ANN, SVM, etc.	Simulated energy use or EUIs	Neto & Fiorelli, 2008; Karatasou et al., 2006; Zhan et al., 2020; Yu, et al., 2010; Sha et al., 2019; Kalogirou, 2000; Tian et al., 2020;
Fault detection and diagnosis	Mostly commercial	Mostly on equipment	Second	Experiment data from laboratories	Test data with 7-8 faults conditions	PCA, SVM, etc.	Faulty samples and faulty types	Andriamamonjy et al., 2018; Bonvini et al., 2014; Cotrufo & Zmeureanu, 2016; Beghi et al., 2016; Li et al., 2016; Li & Wen, 2014; Tran et al., 2015; Han et al., 2011; He et al., 2016; Li et al., 2016; Xia et al., 2020; Yan et al., 2018; Du et al., 2014; Yoshida & Kumar, 2001;
Thermal-comfort-based environment control	Mostly office	Room	-	ASHRAE Database	21,000 sets of data	ANN, SVM, etc. or probability-based methods	Predicted thermal comfort and thermostat setpoint	Ghahramani et al., 2015; Kim et al., 2018; Ghahramani et al., 2018; Zhou et al., 2020; Wang & Hong, 2020; Dai et al., 2017; Chaudhuri et al., 2018;
Building energy system	Mostly	System	Hourly	Building	Several	Reinforcement	Control	Tang et al., 2020; Dalamagkidis et al., 2007;

control	non-residential			BMS data	month's data	learning	strategy of cooling plants and TES	Zakula et al., 2014; Lee & Braun, 2008; Yuan et al., 2020; Vázquez-Canteli et al., 2019; Chen et al., 2020; Yu & Dexter, 2010; Liu & Henze, 2006; Luo et al., 2017
Retrofit analysis	Mixed type	Building	Yearly	Building EUI survey database	Several hundred to thousand samples	Clustering methods, classification methods	Energy saving potential of retrofit measures	Sanhudo et al., 2018; Re Cecconi et al., 2019; Marasco & Kontokosta, 2016; Geyer et al., 2017;
Benchmarking	Mostly non-residential	Building	Yearly	Building EUI survey database	Several hundred to thousand samples	Clustering methods, PCA, etc.	Ratings of building energy efficiency	Pérez-Lombard et al., 2009; Chung et al., 2006; Yang et al., 2018; Yalcintas, 2006; Chung, 2012; Wang, 2015; Papadopoulos & Kontokosta, 2019;
Pricing mechanism	Mostly residential	Household	Hourly	Smart meter data	One year's data	Clustering methods	Optimal time-of-use tariff	Fu et al., 2018; Yilmaz et al., 2019

6.1 Design Optimization

Building performance simulation has always been an essential step in the building design phase. The current design process usually involves building performance simulation based on physics principles. With advances in big data in the building sector, data-driven analytics can support the detailed inputs for building simulation or optimize the entire simulation process. Enhancement of big data analytics applied to the building design process is achieved from two perspectives: supporting parametric analysis, and implementing data-driven approaches in the building design process.

6.1.1 Parametric analysis at the design phase

Building performance simulation requires detailed inputs, including building geometry, building envelope properties, occupancy schedules, etc. Among these aspects, occupancy schedule and appliance use schedule are two essential inputs requiring data-driven analysis of the typical profiles. Pattern identification is a major approach to understanding the occupant-behavior-related energy profiles in buildings (Wen et al., 2019; Yang et al., 2018; Quintana et al., 2020; Popoola & Chipango, 2020), and is widely applied in both residential and non-residential buildings. This provides insight for distinguishing different types of users or understanding the distributions of different schedules at different time periods.

Numerous researchers have studied the typical patterns of energy use. Most have focused on residential building profiles (Escobar et al., 2020; Sala et al., 2019). The promotion of smart meters in households has made it possible to acquire hourly or

sub-hourly data from massive number of households, which are used to support pattern analysis (Zhou et al., 2017). To acquire such data, some researchers also use electricity recorder installed in test households. All types of clustering-based algorithms are applied for the typical pattern analysis (Zhou et al., 2017), the most popular of which is K-means clustering (An et al., 2018; Rhodes et al., 2014). Many other types of big data algorithms have also been introduced to assist with the analysis, such as principle component analysis (PCA) and random forest classifier (RF) (Fu et al., 2018; Satre-Meloy et al., 2020). The output of such studies usually involves typical energy use profiles, which is a set of average energy use curves to improve the general understanding of the energy consumption of certain types of buildings, and are then used as the input schedule of the building energy models in simulation of the design phase.

6.1.2 Data-driven models for building design

In the building design process, most designers adopt physics-based building simulation models to predict and evaluate the energy efficiency of the design. Traditional physics-based model requires detailed inputs and is extremely time-consuming. Thus, some researchers have introduced data-driven models into the design phase as a substitute to physics-based models. To simplify the model, researchers conduct correlation analysis or sensitivity analysis to discover the most relevant features that affect the building energy consumption for a specific case. Then, data-driven models based on machine learning algorithms are trained and validated to

regress the energy consumption with selected features (Sha et al., 2019; Neto & Fiorelli, 2008; Karatasou et al., 2006). The training of the model utilizes building management system (BMS) data of existing buildings (Zhan et al., 2020). The proposed model will then be used to support the evaluation of energy efficiency in the design phase (Yu, Haghighat et al., 2010; Kalogirou, 2000), focusing on a single building or at an urban scale (Tian et al., 2020). It should be noted that, because the information and knowledge in the building design phase is quite limited, transfer learning is an optimal method used to store knowledge from existing buildings and applying it to new buildings under a similar context. Using transfer learning to build data-driven models for energy consumption prediction at the building design phase is a possible future research perspective.

6.2 Benchmarking analysis

Benchmarking refers to the evaluation and rating of the energy use efficiency of buildings by comparison with buildings of the same type (Pérez-Lombard et al., 2009). Benchmarking analytics are usually applied to commercial or public buildings (Chung et al., 2006; Yang et al., 2018), and requires understanding of current status of the building energy distribution. Big data analysis provides insight into this perspective. By regressing the building energy use indicators (EUIs) with independent variables such as the building properties and meteorological variables, the trained model can be used to predict the expected value or range of EUIs of the target building, based on which the energy efficiency of the building is rated (Papadopoulos & Kontokosta,

2019; Yalcintas, 2006). Chung (2012) also developed a fuzzy linear regression model to benchmark the energy efficiency of commercial buildings, and Wang (2015) developed a benchmarking model for residential buildings using PCA, multiple linear regression, and k-means clustering techniques. Big data used as building information and energy use indicators provide sources for data analytics, whereas regression methods support the development of models for benchmarking buildings at a regional scale.

6.3 Control optimization

Control optimization has always been a key issue in building energy system management. The control strategy, particularly a real-time control strategy, relies highly on big data analytics from the building energy systems. Control strategies are then applied to achieve both environmental comfort and energy efficiency.

6.3.1 Thermal-comfort-based environment control

One of the essential applications of building system control is real-time equipment management used to achieve a higher quality of indoor environment control and human comfort. Owing to the diversity and complexity of occupants' thermal comfort preference, physics-based thermal balance model may not achieve significant accuracy in thermal comfort prediction. Data-driven models, however, predict the thermal comfort of the occupants from real monitoring data and environmental parameters (Dai et al., 2017; Ghahramani et al., 2015; Kim et al., 2018), and are capable of learning and correcting the models under a different application context

(Ghahramani et al., 2018).

Most models use machine learning algorithms to regress the thermal sensation vote (TSV) with multiple features including environmental factors (e.g., temperature and humidity), metabolic rate, clothing condition, and air-conditioning modes (Chaudhuri et al., 2018). One of the most popular datasets used for such an analysis is the ASHRAE RP-884 dataset, which is a global thermal comfort database. Researchers have proposed different machine learning algorithms to improve the prediction accuracy of TSV in comparison with the traditional PMV-PPD model. Zhou et al. (2020) proposed a support vector machine (SVM)-based model to predict the thermal comfort of the occupants. Wang et al. (2020) introduced a Bayesian inference approach to predict the indoor thermal comfort and determine comfortable temperature for the occupants. Well-tuned data-driven thermal comfort models can be adopted in various types of buildings to determine the set-point temperature of thermostats and help improve the operation of HVAC systems (Delcroix et al., 2020).

6.3.2 Energy-efficient-oriented system control optimization

Control optimization based on building energy prediction models is adaptable to the predictive control of building energy systems. Building energy prediction requires a prior step before applying a strategic optimization, which has been a hot topic in recent studies. Many energy prediction models are data-driven (Tang et al., 2020), taking advantage of machine learning algorithms and integrating data with physics knowledge in buildings to improve the accuracy of the prediction, thus offering

reliable information for the next step of the control optimization (Zakula et al., 2014). Specifically, for control optimization, most studies have used deep reinforcement learning to build a step-forward prediction and optimization mechanism under practical control scenarios (Dalamagkidis et al., 2007; Lee & Braun, 2008; Yuan et al., 2020; Chen et al., 2020).

Researchers have developed numerous types of real-time controllers for building energy systems. Yu et al. (2010) proposed a fuzzy rule-based controller with reinforcement learning optimization to balance the energy costs and thermal discomfort in buildings. The results showed significant improvement in terms of low-energy building system performance. Vázquez-Canteli et al. (2019) also proposed a deep reinforcement learning approach for online learning and tuning of the controller for the heat pump system. The automatic control model also achieved moderate energy saving under different scenarios.

Control optimization is specifically applicable for thermal storage system, such as ice-based cooling storage and battery-based electricity storage. The introduction of thermal storage or battery storage in buildings aims to utilize low-cost electricity during the night (valley period) and reduce the peak-period energy consumption, thus creating economic benefits from energy costs. This time-of-use tariff for electricity is applied as a stimulus for electricity users and benefits the power grid from shifting the peak loads to valley period. Within this procedure, precise prediction of the next-day energy consumption is the most essential part, followed by a control optimization

using reinforcement learning (Q-learning).

For control optimization of the thermal storage system, Liu et al. (2006) proposed a simulated reinforcement learning controller to learn the pre-cooling mechanism of the cooling supply system and the control of the charging and discharging of the thermal energy storage system according to the utility rate. Luo et al. (2017) also developed an optimized sequential quadratic programming algorithm and several control strategies to minimize the energy cost of ice-based thermal energy storage system. One-day-ahead building cooling load prediction is an essential data analysis for control optimization of thermal energy storage system.

6.4 Fault detection and diagnosis

Fault detection and diagnosis (FDD) is a popular engineering application used to identify the type and location of the fault in a system. In building sector, this is essential for real-time building energy system management and is highly related to big data derived from the sensors of the system, including but not limited to the chiller, pumps, fans, AHUs, and indoor environmental parameters. Major applications of big data analysis of FDD involves two levels: the component level and building level.

6.4.1 FDD at the component level

The application of FDD at the component level refers to the identification of malfunctioning on energy system equipment. The objective of such research usually focuses on one single component itself, such as chiller or AHUs (Andriamamonjy et al., 2018), without consideration of the connections between each component or

within the system. Thus, FDD at the component level is usually applied to evaluate the values of the parameters independently, taking no account of the conflicts in the relations among the parameters of the different components. Popular FDD methods at component level include machine-learning algorithms (Bonvini et al., 2014). PCA (Cotrufo & Zmeureanu, 2016; Beghi et al., 2016; Li et al., 2016; Li & Wen, 2014) and an SVM (Tran et al., 2015; Han et al., 2011) are two of the most widely applied algorithms, considering their unique features of labeling and identifying samples.

Most studies on chillers have utilized the ASHRAE project RP-1043 dataset (He et al., 2016; Li et al., 2016; Xia et al., 2020). This dataset is a universal evaluation tool for chiller FDD, and the data are acquired from experiments conducted in laboratory (Yan et al., 2018). The major aspects of faults for chillers include fouling, refrigerant overcharging, refrigerant leakage, excessive oil, and reduced flow.

Some studies have also utilized simulated data to test the FDD methods. TRNSYS is a commonly used simulation tool for component simulation. Numeral experiments are first conducted to acquire both normal and malfunctioning data. The simulated data are then used to train and test the FDD models (Du et al., 2014). Compared with the analysis of real experimental data in laboratories, numerical studies are more likely to be a theoretical analysis.

6.4.2 FDD at the building level

FDD at building level usually considers the energy consumption or indoor environment as a whole. Studies have mainly focused on identifying whether the

energy consumption or indoor environment is normal compared with the previous status (Du et al., 2014).

One perspective on whole building FDD is the use of clustering-based methods to understand the typical energy use profiles (Yoshida & Kumar, 2001). The researchers tested whether the energy use curve of a certain day is identified as an outlier among the known typical profiles. A fault is detected by detecting outliers from the dataset.

It should be noted that current studies have seldom focused on the FDD of integrated system or at the building level, taking into consideration the correlated faults of the parameters from different equipment. Moreover, few researchers have studied the application of FDD for the entire building management, which could therefore be a promising future research topic.

6.5 Retrofit analysis

Retrofitting analysis of existing buildings at a district or city scale has always been an important perspective for urban planning. Evaluation of different retrofitting measures requires cross comparison of different building properties. Researchers have trained data-driven models using building EUIs and building properties from massive samples to learn their correlations (Sanhudo et al., 2018). The energy saving potential can then be evaluated using the trained model.

In the scope of a regional-scale retrofit analysis, Re Cecconi (2019) introduced different data-driven methods to support regional energy retrofit policy when applied to school buildings. The author used clustering-based algorithms to identify

homogenous classes of buildings, then trained artificial neural networks to evaluate the energy saving of possible retrofit cases. Marasco et al. (2016) introduced a machine learning classifier to evaluate the eligibility of various energy conservation measures for buildings in New York City. Geyer et al. (2017) also analyzed the application of clustering methods for building retrofitting measures.

To conclude, the use of data-driven method is extremely cost-effective in evaluating energy saving potential of retrofitting measures for large numbers of buildings at regional scale. As the advantage of big data analytics, the model is trained using real energy consumption data of existing buildings, avoiding the gap between the simulated results and the real consumption.

6.6 Pricing mechanism

In the power grid system of certain cities, time-of-use tariff is an effective measure for the demand-side management of electricity consumption. The determination of the utility rates relies on the understanding of the current energy use profiles from the demand side. Thus, typical energy use profile analysis by clustering methods performs as the first step to determine the electricity pricing tariff. Fu et al. (2018) discussed a clustering-based load pattern analysis and machine-learning based short-term load prediction model for the evaluation of an increasing-block pricing tariff placed on electricity. Yilmaz et al. (2019) also developed a clustering-based model for residential electricity load profile characterization, based upon which the authors discussed its policy implementation for a time-of-use utility tariff determination.

Data-driven models perform efficiently in recognizing energy use characteristics and can significantly support an energy-related policy making process.

7 Discussion

7.1 Why is big data analytics being used as a new scientific method?

Big data analytics has been widely adopted as a new research method in many fields, such as healthcare and the medical sector, business and finance, the Internet and social media, and smart cities. With a dramatic increase in data volumes from smart buildings and IoT-enabled environmental control devices, as well as the advancement of data mining and machine learning techniques, will big data analytics become a valuable and credible scientific method parallel to theory, simulation, and experimentation in the building sector? In this section, the essential differences and similarities compared with conventional scientific methods are discussed, including theoretical, experimental, and simulation methods.

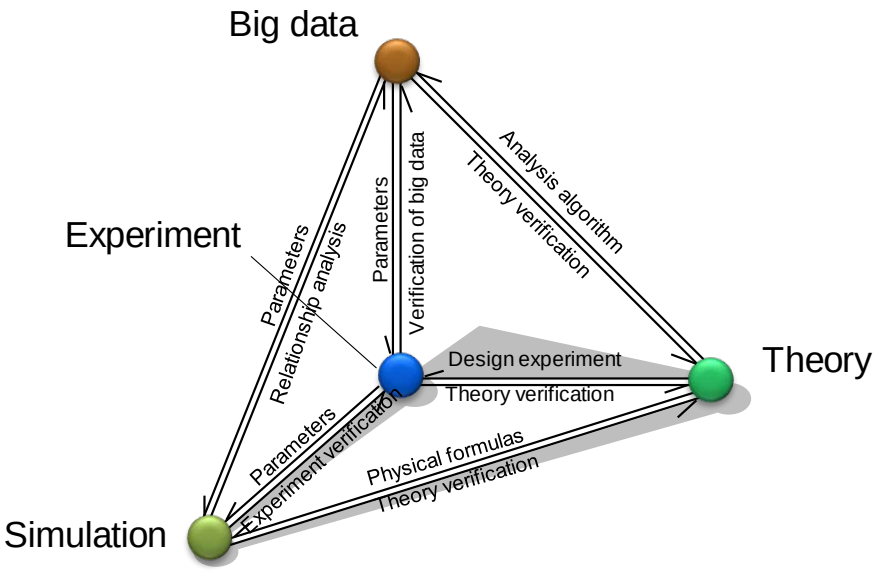

Figure 6 Relationship among big data analytics and three conventional research

methods

The knowledge obtained from big data analytics is a valuable supplementation to the exiting physics-based theory repository in the building sector, which has been proven in exiting research on big data from buildings. The data-driven knowledge obtained from big data analytics include the energy consumption patterns and building occupancy patterns at different time scales, e.g., hourly, daily, weekly, monthly, and annual (Wen et al., 2019; Satre-Meloy et al., 2020; Rhodes et al., 2014; Zhou et al., 2017). With the help of big data analytics, we can better understand the dynamic correlations among building energy use across different systems, such as air conditioning, lighting, and lift systems (Fan et al., 2015; Xiao et al., 2014; Ren et al., 2015), as well as the actual temporal and spatial distributions of occupant behavioral parameters and the equipment system performance, such as the distribution of air-conditioning setting temperatures across China (Hu et al., 2017; An et al., 2018), which are difficult to discover and quantify through a theoretical analysis.

Big data analytics is used to analyze huge amounts of building data from various sources, which significantly exceed the data size available from experiments including both laboratory and on-site/field experiments. Buildings are extremely complex objects, the operational performances of which are influenced by numerous uncertainties, such as the configuration of the building envelope, occupant behavior, and equipment performance. Owing to the time and labor required, as well as various technical difficulties, it is extremely difficult to carry out experiments covering the

full variations of all influential parameters. Therefore, the experiment results cannot fully reflect the building/system/equipment performance in reality. However, big data on building operations, for example, operational data from the past several years, retrieved from a building automation system (or building management system) and IoT devices naturally cover almost all possible operating conditions, contributing to a better understanding of a building performance at different temporal and spatial scales.

Simulations can cover all possible operational conditions but involves too many assumptions and simplifications, which have difficulty reflecting the actual behavior and performance, incurring uncertainties and inevitable errors. For example, typical hourly occupancy schedules (ASHRAE, 2019) have been widely adopted in simulation programs to represent the number of occupants in a space at different times. Because the occupant presence and behavior have a significant influence on the building energy performance, using more realistic occupancy patterns learned from big data on the building operations (Jiefan et al., 2018) enables a BPS software to produce more reliable estimations of the building energy performance. Big data analytics can provide more realistic parameters with probabilistic distributions as simulation inputs and settings (Feng et al., 2016; Wilke et al., 2013; Foteinaki et al., 2019).

As illustrated in Figure 6, theory, experimentation, and simulations together with big data analytics support each other, greatly enriching the existing knowledge repository

of the building sector.

7.2 What are the differences between big data-driven and data-driven modeling?

Data-driven modeling has been widely adopted toward the development of building and building system models for several decades (Bourdeau et al., 2019; Ruch et al., 1993). However, the data used have usually been from a short time period, i.e., from serval days to several months, as indicated in Section 6. The data series used to develop such models has typically been less than 1 month long when the sampling interval is between 1 and 10 min. If a data series of longer than 1 year is used, the sampling interval is typically hourly or daily. Data sampled hourly, or at even longer intervals, have difficulty revealing the thermal dynamics of real buildings. However, longer datasets with shorter intervals adversely increase the computational load, which may exceed the capability of most existing data analysis algorithms or cause an overfitting problem. In addition, when the operating conditions are outside the range of the training data, the models become unreliable (Kramer, et al., 2012; Afram, et al., 2017). When increasing the size of the dataset, e.g., 1 year of operational data, the models suffer from a degraded accuracy owing to the large variations in the data, as well as a low computational efficiency from the inability of conventional data analytics to deal with large datasets. Moreover, the generalization of the models and modeling methods remains questionable.

With the significantly increasing volume of building data and the rapid advances in big data analytics, a general transition can be observed in the evolution of the data-driven modeling of a building performance, i.e., the entire process becomes increasingly data-driven and involves less domain knowledge. The selection of model inputs is becoming more data-driven. Earlier, domain knowledge was heavily involved. For example, domain knowledge tells us that the outdoor air temperature significantly influences the cooling load, and thus it was chosen as a single input. However, in recent studies, the input selection process has relied more on feature extraction from big data on the building operations, bringing forth new and valuable information. Features may be individual variables or a combination of multiple variables (Fan et al., 2014). For example, unsupervised deep learning models, such as autoencoders, have been used to develop features as inputs for the predictive models used in the building sector (Zou et al., 2018). Such an approach is particularly useful for constructing high-level features from long and noisy time series data (Bonfigli et al., 2018) and significantly reduces the computational load. By contrast, when big data are used for model development, the model performance evaluation becomes a challenging issue, and thus more research on data-driven modeling has focused on performance metrics, as reviewed in Section 5.1. As more powerful data analytics algorithms, such as deep learning and ensemble learning, are adopted to develop building models based on big data, the models become “blacker,” which means they are more difficult to understand. Therefore, a model interpretation has become a hot

research topic in recent years and the major studies in this area are summarized in Section 5.2.

7.3 What are the scientific contributions of big data-driven methods when using different data sources?

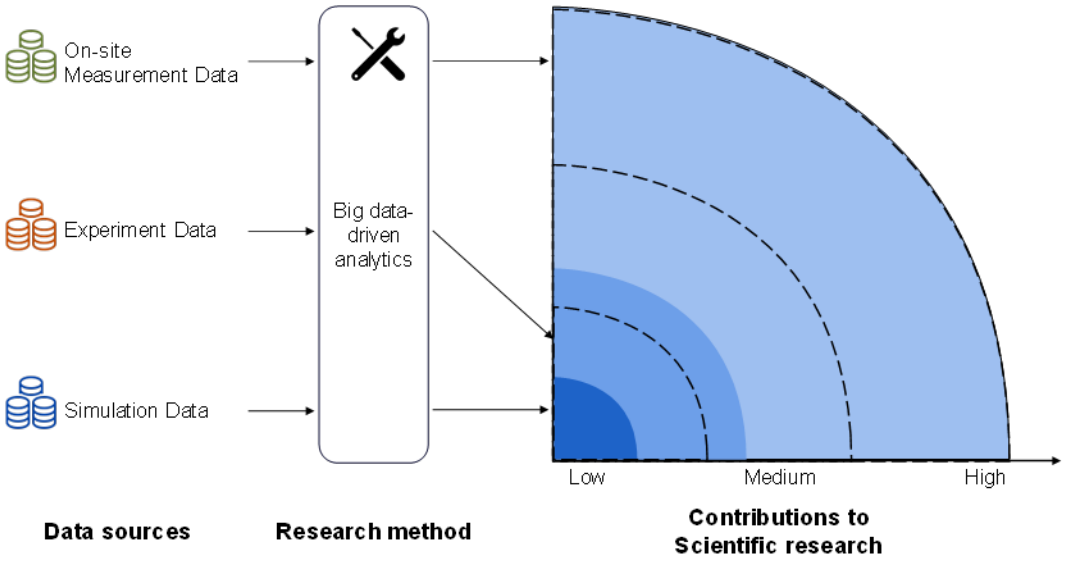

Figure 7 Comparison among different data sources for big data-driven analytics on scientific research contributions

As mentioned in section 3, there are three types of data, i.e., simulation data, experiment data, and on-site measurement data, which are widely used in data-driven building energy analyses.

Some researchers have utilized simulated data from BPS tools, such as TRNSYS and EnergyPlus, to carry out a statistical analysis (Du et al., 2014; Edwards et al., 2017). These tools usually adopt physics-based models, which are based on existing knowledge including physical equations and empirical parameters, and thus the simulation results can be clearly understood and explained. Therefore, what we get

from big data-driven analytics of simulation data is the regression results of physics-based models, which are more useful when building performance prediction requires fast speed yet less accuracy. Overall, we can hardly acquire new knowledge when using simulation data for big data-driven analytics.

In addition, most studies on FDD have utilized experiment data in HVAC systems for model training with machine learning algorithms, and the trained models have been used for a fault diagnosis of equipment during the operational stage and energy use prediction (D. Li et al. 2016; G. Li et al. 2016; Tran et al. 2015; Yan et al. 2018). In general, the amount of experiment data has been limited, and such data have been unable to reflect a degradation in the equipment performance or the behavior differences during a practical operation. Therefore, big data-driven models derived from experiment data cannot guarantee the prediction accuracy for complex practical processes even when sufficient experimental data have been available.

Compared with the other two data sources, on-site measurement data can reveal and reflect the real operational conditions of buildings (or systems and equipment), which are useful in big data-driven analytics in terms of scientific discovery.

7.4 What are the influences of data veracity on big data-driven analytics?

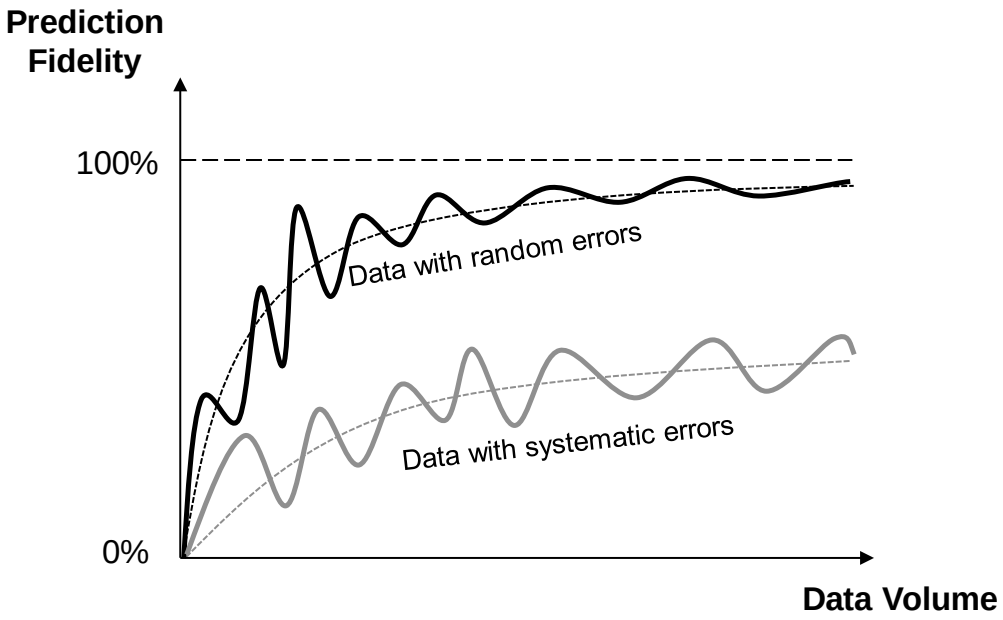

Figure 8 Influence of data veracity and volume on the predicted results

With the rapid development of sensing and communication technologies, it has become easier and less expensive to obtain and store massive amounts of data. Big data have also brought about a new problem, i.e., a larger dataset has a greater possibility to contain errors. Therefore, the data quality is becoming an important issue. Mayer-Schönberger and Cukier (2013) concluded that even massive error-prone datasets are more reliable than accurate but small samples. If there is only one sample applied, we must confirm its accuracy because any errors will result in a poor outcome; however, when the data volume is extremely large, even if some incorrect data are present, the aggregation of massive data will provide a result closer to reality. However, this conclusion is based on the assumption that the data do not have systematic errors. In general, the data related to building energy performance are uncertain, such as the heat transfer coefficient of the envelope and the outdoor

temperature, and these uncertainties are usually random. For data having only random errors, a larger data size can contribute to more accurate results. However, there also exist numerous data with systematic errors owing to the measurement devices used and the participants collecting the data, thereby producing incorrect results regardless of how large the data volume is. IBM first brought up veracity as an important data characteristic in the big data field, which has attracted wide attention from the research community (Sivarajah et al. 2017). Data veracity is critical to the prediction fidelity. Therefore, we should improve the data veracity and eliminate false and erroneous data before conducting a data analysis to make the results more accurate and reliable.

8 Conclusion

Building performance simulations have proven to be an extremely important tool for improving the energy efficiency, indoor environment quality, and thermal comfort of buildings, as well as the reliability and efficiency of a building-grid eco-system. Data-driven modeling plays a significant role in a BPS when facing large complex buildings with limited information. Most modern non-residential buildings are equipped with an advanced BAS capability for real-time monitoring and control, allowing huge amounts of building operational data to be stored. In IoT-driven smart cities, numerous IoT sensors monitor and collect data from distributed environmental

control devices, such as residential air conditioners. As a result, the data volume available for building performance modeling is becoming much larger and continues to increase. Big data analytics, with challenges in terms of the volume, variety, and velocity of the data applied, is a valuable way to develop more powerful and computationally efficient data-driven models for a BPS. From the comprehensive review and critical discussions presented in this paper, a new paradigm for data-driven modeling, i.e., big-data-driven modeling, is emerging as a valuable supplementation to existing scientific research methods in the building sector, and is greatly enriching the exiting knowledge repository for improving the performance of modern buildings.

Reference

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

Amasyali, K., & El-Gohary, N. (2018). A review of data-driven building energy consumption prediction studies. Renewable And Sustainable Energy Reviews, 81, 1192-1205.

An, J., Yan, D., & Hong, T. (2018). Clustering and statistical analyses of air-conditioning intensity and use patterns in residential buildings. Energy And Buildings, 174, 214-227.

Andriamamonjy, A., Saelens, D., & Klein, R. (2018). An auto-deployed model-based

fault detection and diagnosis approach for Air Handling Units using BIM and Modelica. Automation In Construction, 96, 508-526.

ASHRAE. (2019). ANSI/ASHRAE/IES Standard 90.1-2019 -- Energy Standard for Buildings Except Low-Rise Residential Buildings.

Attia, S., Gratia, E., De Herde, A., & Hensen, J. (2012). Simulation-based decision support tool for early stages of zero-energy building design. Energy And Buildings, 49, 2-15.

Barnett, V. & Lewis, T. (1994) Outliers in statistical data. 3th ed. New York: John Wiley & Sons Inc.

Ben Taieb, S., Bontempi, G., Atiya, A., & Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Systems With Applications, 39(8), 7067-7083.

Bonfigli, R., Felicetti, A., Principi, E., Fagiani, M., Squartini, S., & Piazza, F. (2018). Denoising autoencoders for Non-Intrusive Load Monitoring: Improvements and comparative evaluation. Energy And Buildings, 158, 1461-1474.

Bontempi, G. (2008). Long term time series prediction with multi-input multi-output local learning. The 2nd European Symposium on Time Series, 145-54.

Bonvini, M., Sohn, M., Granderson, J., Wetter, M., & Piette, M. (2014). Robust on-line fault detection diagnosis for HVAC components based on nonlinear state estimation techniques. Applied Energy, 124, 156-166.

Bottaccioli, L., Aliberti, A., Ugliotti, F., Patti, E., Osello, A., Macii, E., & Acquaviva, A. (2017, July). Building energy modelling and monitoring by integration of iot devices and building information models. In 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 914-922). IEEE.

Bourdeau, M., Zhai, X., Nefzaoui, E., Guo, X., & Chatellier, P. (2019). Modeling and forecasting building energy consumption: A review of data-driven techniques. Sustainable Cities And Society, 48, 101533.

Breiman, L. (2001). Random forests. Machine Learning 45: 5-32.

Brundu, F. G., Patti, E., Osello, A., Del Giudice, M., Rapetti, N., Krylovskiy, A., ... & Acquaviva, A. (2016). IoT software infrastructure for energy management and simulation in smart cities. IEEE Transactions on Industrial Informatics, 13(2), 832-840.

Capozzoli, A., Piscitelli, M., Brandi, S., Grassi, D., & Chicco, G. (2018). Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. Energy, 157, 336-352.

Chae, Y., Horesh, R., Hwang, Y., & Lee, Y. (2016). Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. Energy And Buildings, 111, 184-194.

Chaudhuri, T., Zhai, D., Soh, Y., Li, H., & Xie, L. (2018). Random forest based thermal comfort prediction from gender-specific physiological parameters using wearable sensing technology. Energy And Buildings, 166, 391-406.

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. arXivL 1603.02754.

Chen, Z., Xu, P., Feng, F., Qiao, Y., & Luo, W. (2020). Data mining algorithm and framework for identifying HVAC control strategies in large commercial buildings. Building Simulation. doi: 10.1007/s12273-019-0599-0

Chollet, F. & Allaire, J. (2018). Deep learning with R. 1st ed. Manning Publications, Shelter Island, New York, USA.

Chou, J., Hsu, Y., & Lin, L. (2014). Smart meter monitoring and data mining techniques for predicting refrigeration system performance. Expert Systems With Applications, 41(5), 2144-2156.

Chung, W. (2011). Review of building energy-use performance benchmarking methodologies. Applied Energy, 88(5), 1470-1479.

Chung, W. (2012). Using the fuzzy linear regression method to benchmark the energy efficiency of commercial buildings. Applied Energy, 95, 45-49.

Chung, W., Hui, Y., & Lam, Y. (2006). Benchmarking the energy efficiency of commercial buildings. Applied Energy, 83(1), 1-14.

Coakley, D., Raftery, P. and Keane, M., 2014. A review of methods to match building energy simulation models to measured data. Renewable and Sustainable Energy Reviews, 37, pp.123-141.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20: 273-97.

Crawley, D., Lawrie, L., Winkelmann, F., Buhl, W., Huang, Y., & Pedersen, C. et al. (2001). EnergyPlus: creating a new-generation building energy simulation program. Energy And Buildings, 33(4), 319-331.

Dai, C., Zhang, H., Arens, E., & Lian, Z. (2017). Machine learning approaches to predict thermal demands using skin temperatures: Steady-state conditions. Building And Environment, 114, 1-10.

Dalamagkidis, K., Kolokotsa, D., Kalaitzakis, K., & Stavrakakis, G. (2007). Reinforcement learning for energy conservation and comfort in buildings. Building And Environment, 42(7), 2686-2698.

Day, J., & Gunderson, D. (2015). Understanding high performance buildings: The link between occupant knowledge of passive design systems, corresponding behaviors, occupant comfort and environmental satisfaction. Building And Environment, 84, 114-124.

Deb, C., Eang, L., Yang, J., & Santamouris, M. (2016). Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks. Energy And Buildings, 121, 284-297.

Delcroix, B., Ny, J., Bernier, M., Azam, M., Qu, B., & Venne, J. (2020). Autoregressive neural networks with exogenous variables for indoor temperature prediction in buildings. Building Simulation. doi: 10.1007/s12273-019-0597-2

Dietterich, T. (2000) Ensemble methods in machine learning. Proceedings of the first international workshop on multiple classifier systems, Springer; p. 1–15.

Dodier, R., & Henze, G. (2004). Statistical Analysis of Neural Networks as Applied to Building Energy Prediction. Journal Of Solar Energy Engineering, 126(1), 592-600.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608v2.

Du, Z., Fan, B., Jin, X., & Chi, J. (2014). Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. Building And Environment, 73, 1-11.

Edwards, R., New, J., Parker, L., Cui, B., & Dong, J. (2017). Constructing large scale surrogate models from big data and artificial intelligence. Applied Energy, 202, 685-699.

Escobar, P., Martínez, E., Saenz-Díez, J., Jiménez, E., & Blanco, J. (2020). Modeling and analysis of the electricity consumption profile of the residential sector in Spain. Energy And Buildings, 207, 109629.

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

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

Fan, C., Wang, J., Gang, W., & Li, S. (2019). Assessment of deep recurrent neural

network-based strategies for short-term building energy predictions. Applied Energy, 236, 700-710. doi: 10.1016/j.apenergy.2018.12.004

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

Fan, C., Xiao, F., & Yan, C. (2015). A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Automation In Construction, 50, 81-90.

Fan, C., Xiao, F., & Zhao, Y. (2017). A short-term building cooling load prediction method using deep learning algorithms. Applied Energy, 195, 222-233.

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. doi: 10.1016/j.enbuild.2017.11.008

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

Feng, X., Yan, D., Wang, C., & Sun, H. (2016). A preliminary research on the derivation of typical occupant behavior based on large-scale questionnaire surveys. Energy And Buildings, 117, 332-340.

Fesanghary, M., Asadi, S., & Geem, Z. (2012). Design of low-emission and energy-efficient residential buildings using a multi-objective optimization

algorithm. Building And Environment, 49, 245-250.

Foteinaki, K., Li, R., Rode, C., & Andersen, R. (2019). Modelling household electricity load profiles based on Danish time-use survey data. Energy And Buildings, 202, 109355.

Foucquier, A., Robert, S., Suard, F., Stéphan, L., & Jay, A. (2013). State of the art in building modelling and energy performances prediction: A review. Renewable And Sustainable Energy Reviews, 23, 272-288.

Freund, Y. & Schapire, R. (1999). A short introduction to boosting. Journal of Japanese Society for artificial intelligence, 14(5): 771-80.

Fu, X., Zeng, X., Feng, P., & Cai, X. (2018). Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China. Energy, 165, 76-89.

Gao, Y., Xu, J., Yang, S., Tang, X., Zhou, Q., & Ge, J. et al. (2014). Cool roofs in China: Policy review, building simulations, and proof-of-concept experiments. Energy Policy, 74, 190-214.

García, S., Luengo, J. & Herrera, F. (2015) Data preprocessing in data mining. Cham, Switzerland: Springer International Publishing.

Geyer, P., Schlüter, A., & Cisar, S. (2017). Application of clustering for the development of retrofit strategies for large building stocks. Advanced Engineering Informatics, 31, 32-47.

Ghahramani, A., Castro, G., Karvigh, S., & Becerik-Gerber, B. (2018). Towards

unsupervised learning of thermal comfort using infrared thermography. Applied Energy, 211, 41-49.

Ghahramani, A., Tang, C., & Becerik-Gerber, B. (2015). An online learning approach for quantifying personalized thermal comfort via adaptive stochastic modeling. Building And Environment, 92, 86-96.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). Deep learning. 1st ed. MIT press, Cambridge, London, England.

Guyon, I. & Elisseeff, A. (2003) An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–82.

Han, H., Gu, B., Hong, Y., & Kang, J. (2011). Automated FDD of multiple-simultaneous faults (MSF) and the application to building chillers. Energy And Buildings, 43(9), 2524-2532.

Harish, V., & Kumar, A. (2016). A review on modeling and simulation of building energy systems. Renewable And Sustainable Energy Reviews, 56, 1272-1292.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer Series in Statistics

He, S., Wang, Z., Wang, Z., Gu, X., & Yan, Z. (2016). Fault detection and diagnosis of chiller using Bayesian network classifier with probabilistic boundary. Applied Thermal Engineering, 107, 37-47.

Hong, T., Yan, D., D'Oca, S., & Chen, C. (2017). Ten questions concerning occupant

behavior in buildings: The big picture. Building And Environment, 114, 518-530.

Hu, S., Yan, D., Azar, E., & Guo, F. (2020). A systematic review of occupant behavior in building energy policy. Building And Environment, 175, 106807.

Hu, S., Yan, D., Guo, S., Cui, Y., & Dong, B. (2017). A survey on energy consumption and energy usage behavior of households and residential building in urban China. Energy And Buildings, 148, 366-378.

International Energy Agency (IEA). (2019). World energy balances. Retrieved from https://www.iea.org/subscribe-to-data-services/world-energy-balances-and-statistics

International Energy Agency (IEA). (2019). World Energy Outlook 2019. Retrieved from https://www.iea.org/reports/world-energy-outlook-2019

Jain, R., Smith, K., Culligan, P., & Taylor, J. (2014). Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Applied Energy, 123, 168-178.

Jetcheva, J., Majidpour, M., & Chen, W. (2014). Neural network model ensembles for building-level electricity load forecasts. Energy And Buildings, 84, 214-223.

Jiefan, G., Peng, X., Zhihong, P., Yongbao, C., Ying, J., & Zhe, C. (2018). Extracting typical occupancy data of different buildings from mobile positioning data. Energy And Buildings, 180, 135-145.

Jovanovic, R., Sretenovic, A., & Zivkovic, B. (2016). Multistage ensemble of

feedforward neural networks for prediction of heating energy consumption. Thermal Science, 20(4), 1321-1331.

Kalogirou, S. (2000). Artificial neural networks for the prediction of the energy consumption of a passive solar building. Energy, 25(5), 479-491.

Karatasou, S., Santamouris, M., & Geros, V. (2006). Modeling and predicting building's energy use with artificial neural networks: Methods and results. Energy And Buildings, 38(8), 949-958.

Kim, J., Zhou, Y., Schiavon, S., Raftery, P., & Brager, G. (2018). Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning. Building And Environment, 129, 96-106.

Kolter, J. & Ferreira, J. (2011) A large-scale study on predicting and contextualizing building energy usage. Twenty-fifth AAAI conference on artificial intelligence.

Lam, J., Wan, K., Lam, T., & Wong, S. (2010). An analysis of future building energy use in subtropical Hong Kong. Energy, 35(3), 1482-1490.

Lee, K., & Braun, J. (2008). Model-based demand-limiting control of building thermal mass. Building And Environment, 43(10), 1633-1646.

Li, D., Zhou, Y., Hu, G., & Spanos, C. (2016). Fault detection and diagnosis for building cooling system with a tree-structured learning method. Energy And Buildings, 127, 540-551.

Li, G., Hu, Y., Chen, H., Shen, L., Li, H., & Hu, M. et al. (2016). An improved fault detection method for incipient centrifugal chiller faults using the PCA-R-SVDD

algorithm. Energy And Buildings, 116, 104-113.

Li, N., Yang, Z., Becerik-Gerber, B., Tang, C., & Chen, N. (2015). Why is the reliability of building simulation limited as a tool for evaluating energy conservation measures?. Applied Energy, 159, 196-205.

Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Applying support vector machine to predict hourly cooling load in the building. Applied Energy, 86(10), 2249-2256.

Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. Energy Conversion And Management, 50(1), 90-96.

Li, S., & Wen, J. (2014). A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform. Energy And Buildings, 68, 63-71.

Li, X., & Wen, J. (2014). Review of building energy modeling for control and operation. Renewable And Sustainable Energy Reviews, 37, 517-537.

Lipton, Z. (2016). The mythos of model interpretability. ICML Workshop on Human Interpretability in Machine Learning, New York, USA. arXiv: 1606.03490.

Liu, S., & Henze, G. (2006). Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory. Energy And Buildings, 38(2), 142-147.

Luo, N., Hong, T., Li, H., Jia, R., & Weng, W. (2017). Data analytics and optimization

of an ice-based energy storage system for commercial buildings. Applied Energy, 204, 459-475.

Maimon, R. (2010) Data Mining and Knowledge Discovery Handbook, 2nded., Springer, New York.

Marasco, D., & Kontokosta, C. (2016). Applications of machine learning methods to identifying and predicting building retrofit opportunities. Energy And Buildings, 128, 431-441.

Marino, D., Amarasinghe, K. & Manic, M. (2016) Building energy load forecasting using deep neural networks. IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society. IEEE, 2016: 7046-7051.

Mayer-Schönberger, V. and K. Cukier. (2013). Big Data: A Revolution That We Transform How We Live, and Think. London.

Mehmood, M., Chun, D., Zeeshan, Han, H., Jeon, G., & Chen, K. (2019). A review of the applications of artificial intelligence and big data to buildings for energy-efficiency and a comfortable indoor living environment. Energy And Buildings, 202, 109383.

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.

Miller, C., Nagy, Z., & Schlueter, A. (2015). Automated daily pattern filtering of measured building performance data. Automation In Construction, 49, 1-17.

Molnar, C. (2018). Interpretable machine learning: A guide for making black box models explainable. 2018. Retrieved from: https://christopha.github.io/interpretable-ml-book/

Neto, A., & Fiorelli, F. (2008). Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. Energy And Buildings, 40(12), 2169-2176.

Nguyen, A., Reiter, S., & Rigo, P. (2014). A review on simulation-based optimization methods applied to building performance analysis. Applied Energy, 113, 1043-1058.

NOAA, N. (2013). Climate data online.

Papadopoulos, S., & Kontokosta, C. (2019). Grading buildings on energy performance using city benchmarking data. Applied Energy, 233-234, 244-253.

Pérez-Lombard, L., Ortiz, J., González, R., & Maestre, I. (2009). A review of benchmarking, rating and labelling concepts within the framework of building energy certification schemes. Energy And Buildings, 41(3), 272-278.

Popoola, O., & Chipango, M. (2020). Improved peak load management control technique for nonlinear and dynamic residential energy consumption pattern. Building Simulation. doi: 10.1007/s12273-020-0601-x

Quintana, M., Arjunan, P., & Miller, C. (2020). Islands of misfit buildings: Detecting uncharacteristic electricity use behavior using load shape clustering. Building Simulation. doi: 10.1007/s12273-020-0626-1

Raftery, P., Keane, M., & O'Donnell, J. (2011). Calibrating whole building energy models: An evidence-based methodology. Energy and Buildings, 43(9), 2356-2364.

Rahman, A., Srikumar, V., & Smith, A. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Applied Energy, 212, 372-385.

Re Cecconi, F., Moretti, N., & Tagliabue, L. (2019). Application of artificial neutral network and geographic information system to evaluate retrofit potential in public school buildings. Renewable And Sustainable Energy Reviews, 110, 266-277.

Ren, X., Yan, D., & Hong, T. (2015). Data mining of space heating system performance in affordable housing. Building And Environment, 89, 1-13.

Rhodes, J., Cole, W., Upshaw, C., Edgar, T., & Webber, M. (2014). Clustering analysis of residential electricity demand profiles. Applied Energy, 135, 461-471.

Ribeiro, M., Grolinger, K., ElYamany, H., Higashino, W., & Capretz, M. (2018). Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. Energy And Buildings, 165, 352-363.

Ribeiro, M., Singh, S. & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. ICML Workshop on Human Interpretability in Machine Learning, New York, USA. arXiv: 1606.05386.

Ribero, M., Singh, S. & Guestrin, C. (2016) "Why should I trust you": Explaining the predictions of any classifier. arXiv:1602.04938v3.

Ruch, D., Chen, L., Haberl, J., & Claridge, D. (1993). A Change-Point Principal Component Analysis (CP/PCA) Method for Predicting Energy Usage in Commercial Buildings: The PCA Model. Journal Of Solar Energy Engineering, 115(2), 77-84.

Sala, J., Li, R., & Christensen, M. (2019). Clustering and classification of energy meter data: A comparison analysis of data from individual homes and the aggregated data from multiple homes. Building Simulation. doi: 10.1007/s12273-019-0587-4

Sanhudo, L., Ramos, N., Poças Martins, J., Almeida, R., Barreira, E., Simões, M., & Cardoso, V. (2018). Building information modeling for energy retrofitting – A review. Renewable And Sustainable Energy Reviews, 89, 249-260.

Satre-Meloy, A., Diakonova, M., & Grünewald, P. (2020). Cluster analysis and prediction of residential peak demand profiles using occupant activity data. Applied Energy, 260, 114246.

Sha, H., Xu, P., Hu, C., Li, Z., Chen, Y., & Chen, Z. (2019). A simplified HVAC energy prediction method based on degree-day. Sustainable Cities And Society, 51, 101698.

Shi, G., Liu, D., & Wei, Q. (2016). Energy consumption prediction of office buildings based on echo state networks. Neurocomputing, 216, 478-488.

Sivarajah, U., Kamal, M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal Of Business Research, 70,

263-286.

Strachan, P., Kokogiannakis, G., & Macdonald, I. (2008). History and development of validation with the ESP-r simulation program. Building And Environment, 43(4), 601-609.

Tanaka, Y., Iwamoto, K., & Uehara, K. (2005). Discovery of time-series motif from multi-dimensional data based on MDL principle. Machine Learning, 58(2-3), 269-300.

Tang, R., Wang, S., & Sun, S. (2020). Impacts of technology-guided occupant behavior on air-conditioning system control and building energy use. Building Simulation. doi: 10.1007/s12273-020-0605-6

Tian, W., Han, X., Zuo, W., & Sohn, M. (2018). Building energy simulation coupled with CFD for indoor environment: A critical review and recent applications. Energy And Buildings, 165, 184-199.

Tian, W., Zhu, C., Sun, Y., Li, Z., & Yin, B. (2020). Energy characteristics of urban buildings: Assessment by machine learning. Building Simulation. doi: 10.1007/s12273-020-0608-3

Tran, D., Chen, Y., Chau, M., & Ning, B. (2015). A robust online fault detection and diagnosis strategy of centrifugal chiller systems for building energy efficiency. Energy And Buildings, 108, 441-453.

Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools.

Energy And Buildings, 49, 560-567.

Vázquez-Canteli, J., Ulyanin, S., Kämpf, J., & Nagy, Z. (2019). Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities. Sustainable Cities And Society, 45, 243-257.

Wang, E. (2015). Benchmarking whole-building energy performance with multi-criteria technique for order preference by similarity to ideal solution using a selective objective-weighting approach. Applied Energy, 146, 92-103.

Wang, Z., & Chen, Y. (2019). Data-driven modeling of building thermal dynamics: Methodology and state of the art. Energy And Buildings, 203, 109405.

Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., & Wu, J. et al. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. Renewable And Sustainable Energy Reviews, 82, 1027-1047.

Wen, L., Zhou, K., & Yang, S. (2019). A shape-based clustering method for pattern recognition of residential electricity consumption. Journal Of Cleaner Production, 212, 475-488.

Wilke, U., Haldi, F., Scartezzini, J., & Robinson, D. (2013). A bottom-up stochastic model to predict building occupants' time-dependent activities. Building And Environment, 60, 254-264.

Wong, S., Wan, K., & Lam, T. (2010). Artificial neural networks for energy analysis of office buildings with daylighting. Applied Energy, 87(2), 551-557.

Xia, Y., Ding, Q., Li, Z., & Jiang, A. (2020). Fault detection for centrifugal chillers

using a Kernel Entropy Component Analysis (KECA) method. Building Simulation. doi: 10.1007/s12273-019-0598-1

Xiao, F., & Fan, C. (2014). Data mining in building automation system for improving building operational performance. Energy And Buildings, 75, 109-118.

Xin, Y., Lu, S., Zhu, N., & Wu, W. (2012). Energy consumption quota of four and five star luxury hotel buildings in Hainan province, China. Energy and Buildings, 45, 250-256.

Xu, L., Wang, S., & Tang, R. (2019). Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load. Applied Energy, 237, 180-195.

Yalcintas, M. (2006). An energy benchmarking model based on artificial neural network method with a case example for tropical climates. International Journal Of Energy Research, 30(14), 1158-1174.

Yan, D., Hong, T., Dong, B., Mahdavi, A., D'Oca, S., Gaetani, I., & Feng, X. (2017). IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. Energy And Buildings, 156, 258-270.

Yan, D., O'Brien, W., Hong, T., Feng, X., Burak Gunay, H., Tahmasebi, F., & Mahdavi, A. (2015). Occupant behavior modeling for building performance simulation: Current state and future challenges. Energy And Buildings, 107, 264-278.

Yan, D., Xia, J., Tang, W., Song, F., Zhang, X., & Jiang, Y. (2008). DeST — An

integrated building simulation toolkit Part I: Fundamentals. Building Simulation, 1(2), 95-110.

Yan, K., Ma, L., Dai, Y., Shen, W., Ji, Z., & Xie, D. (2018). Cost-sensitive and sequential feature selection for chiller fault detection and diagnosis. International Journal Of Refrigeration, 86, 401-409.

Yang, J., Rivard, H., & Zmeureanu, R. (2005). On-line building energy prediction using adaptive artificial neural networks. Energy And Buildings, 37(12), 1250-1259.

Yang, T., Ren, M., & Zhou, K. (2018). Identifying household electricity consumption patterns: A case study of Kunshan, China. Renewable And Sustainable Energy Reviews, 91, 861-868.

Yang, Z., Roth, J., & Jain, R. (2018). DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis. Energy And Buildings, 163, 58-69.

Yezioro, A., Dong, B., & Leite, F. (2008). An applied artificial intelligence approach towards assessing building performance simulation tools. Energy And Buildings, 40(4), 612-620.

Yilmaz, S., Chambers, J., & Patel, M. (2019). Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management. Energy, 180, 665-677.

Yoshida, H., & Kumar, S. (2001). Development of ARX model based off-line FDD

technique for energy efficient buildings. Renewable Energy, 22(1-3), 53-59.

Yu, Z., & Dexter, A. (2010). Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning. Control Engineering Practice, 18(5), 532-539.

Yu, Z., Haghighat, F., Fung, B., & Yoshino, H. (2010). A decision tree method for building energy demand modeling. Energy And Buildings, 42(10), 1637-1646.

Yuan, X., Pan, Y., Yang, J., Wang, W., & Huang, Z. (2020). Study on the application of reinforcement learning in the operation optimization of HVAC system. Building Simulation. doi: 10.1007/s12273-020-0602-9

Zakula, T., Armstrong, P., & Norford, L. (2014). Modeling environment for model predictive control of buildings. Energy And Buildings, 85, 549-559.

Zhan, S., Chong, A., & Lasternas, B. (2020). Automated recognition and mapping of building management system (BMS) data points for building energy modeling (BEM). Building Simulation. doi: 10.1007/s12273-020-0612-7

Zhang, L., & Wen, J. (2019). A systematic feature selection procedure for short-term data-driven building energy forecasting model development. Energy And Buildings, 183, 428-442.

Zhao, H., & Magoulès, F. (2010). Parallel Support Vector Machines Applied to the Prediction of Multiple Buildings Energy Consumption. Journal Of Algorithms & Computational Technology, 4(2), 231-249.

Zhao, H., & Magoulès, F. (2012). A review on the prediction of building energy

consumption. Renewable And Sustainable Energy Reviews, 16(6), 3586-3592.

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

Zhou, K., Yang, C., & Shen, J. (2017). Discovering residential electricity consumption patterns through smart-meter data mining: A case study from China. Utilities Policy, 44, 73-84.

Zhou, K., Yang, S., & Shao, Z. (2017). Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study. Journal Of Cleaner Production, 141, 900-908.

Zou, H., Zhou, Y., Yang, J., & Spanos, C. (2018). Towards occupant activity driven smart buildings via WiFi-enabled IoT devices and deep learning. Energy And Buildings, 177, 12-22.