A review of data-driven building performance analysis and design on big on-site building performance data

Zhichao Tian, Xinkai Zhang, Shen Wei, Sihong Du, Xing Shi

PII: S2352-7102(21)00564-7

DOI: https://doi.org/10.1016/j.jobe.2021.102706

Reference: JOBE 102706

To appear in: *Journal of Building Engineering*

Received Date: 5 January 2021

Revised Date: 6 May 2021

Accepted Date: 7 May 2021

A review of data-driven building performance analysis and design on big on-site building performance data

Zhichao Tian[1,2], Xinkai Zhang[1,2], Shen Wei[3], Sihong Du[1,2], Xing Shi[1,2,*]

1. College of Architecture and Urban Planning, Tongji University, Shanghai, P.R. China;
2. Key Laboratory of Ecology and Energy-saving Study of Dense Habitat (Tongji University), Ministry of Education, P.R. China
3. The Bartlett School of Construction and Project Management, University London College, WC1E 7HB, London, UK

Xing Shi, 20101@tongji.edu.cn

Abstract

Building performance design (BPD) is a crucial pathway to achieve high-performance buildings. Previous simulation-based BPD is being questioned due to the performance gaps between simulated and measured values. In recent years, accumulated on-site building performance data (OBPD) make it possible to analyze and design buildings with data-driven methods. This article makes a review of previous studies that conducted data-driven building performance analysis and design on a large amount of OBPD. The covered studies are summarized by the applied techniques, i.e., statistics, regression, classification, and clustering. The data used by these studies are compared and discussed emphasizing the data size and public availability. A comprehensive discussion is given about the achievements of existing studies, and challenges for boosting data-driven BPD from three aspects, i.e., developing data-driven models, the availability of building performance data, and stimulation of industrial practices. The review results indicate that data-driven methods were commonly applied to estimate energy consumptions, and explore energy trends, determinant features, and reference buildings. Identifying determinant features is one of the most successful applications. This study highlights the future research gaps for boosting data-driven building performance design.

Keywords

Building performance design, data-driven, building energy, building performance data.

Abbreviations

Acronym	Full Name
ANN	Artificial Neural Network
ANOVA	Analysis of variance
BPD	Building Performance Database
BPNN	Back propagation neural network
CART	Classification and Regression Tree
CBECS	Commercial Building Energy Consumption Survey
CHAID	Chi -Square Automatic Interaction Detection
CVMSE	Cross-validated Mean Square Error
CV-RMSE	Coefficient of Variation, Root Mean Square Deviation
DD-BPD	Data-driven Building Performance Design
EUI	Energy Usage Intensity
GLR	General Linear Regression
GRNN	General Regression Neural network
HEED	Homes Energy Efficiency Database
HVAC	heating, ventilation, and air conditioning
ID3	Iterative Dichotomiser 3
LEED	Leadership in Energy and Environmental Design
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLP	Multi-layer Perceptron
MLR	Multiple Linear Regression
NEED	National Energy Efficiency Data-Framework
OBPD	On-site Building Performance Data
OLS	Ordinary Least Squares
R^2	Coefficient of Determination
RBFNN	Radial Basis Function Neural Network
RECS	Residential Energy Consumption Survey
REPT	Reduced Error Pruning Tree
RF	Random Forest
RMSE	Root Mean Square Error
SHAP	SHapley Additive exPlantions
SVM	Support Vector Machine
XGBoost	Extreme gradient boosting

1. Introduction

Building energy has been marked as a key section of carbon emissions [1]. Buildings contribute to nearly 40% of society's total energy consumption in developed countries [2]. Decision made in the design stage is crucial for reducing building energy consumption. Building performance design (BPD) aims at finding out the best design solution considering several criteria including energy, cost, and thermal comfort, etc. To design low-energy buildings, researchers have dedicated themselves to developing effective design methods. In the past two decades, simulation-based BPD has become a significant method for designing high-performance buildings [3, 4]. Detailed energy simulation tools, such as EnergyPlus, are used to evaluate the energy consumption of competing design solutions. Before making design decisions, engineers need to build and calibrate detailed energy models which may take several days' work [5]. Besides, the reliability of this method is always being questioned due to the performance gap, which refers to the difference between simulated performance during the design stage and actual performance during the operation stage [6-8].

With the rapid development of green building certification and energy monitoring projects, large-scale data in terms of the actual performance of buildings have been accumulated. Due to the availability of several on-site building performance datasets, data-driven building performance design (DD-BPD) is becoming a hot research topic. Fig. 1 depicts the procedure of implementing the DD-BPD. There are two key parts to fulfill DD-BPD, i.e., the database and specially developed data-driven approaches. The database should contain a large amount of OBPD other than simulated or any other generated data. Data-driven methods usually refer to statistical [6, 9] or machine learning methods [10]. As for a specific design scenario, engineers need to adopt suitable algorithms to build models.

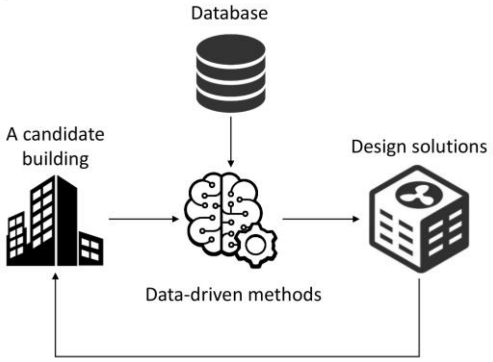


Fig. 1 The general procedure of the DD-BPD

It is not a new phenomenon that applying data-driven approaches to building energy prediction. However, due to the lack of big-OBPD, most previous studies were carried out on simulated data or the sensor data of a building. Those studies have been well-reviewed by Amasyali et al. [11], Wei et al.[12], Bourdeau et al.[13], and Miller et al. [14]. Until now, however, there is still no comprehensive review about data-driven building performance analysis and design on OBPD. So, it remains unclear what the status quo, trend, and future research directions are. This review article fills this gap by answering the following questions:

1. What kinds of data-driven approaches have been adopted in the DD-BPD and the accuracy of existing data-driven prediction models?
2. What is the current status of existing building performance databases?
3. What are the future research directions of the DD-BPD?

2. Methodology

First, related papers were screened out by keywords searching on widely known academic databases, including ScienceDirect and Google Scholar, using combined keywords from machine learning and building performance domains, for example, ‘SVM, building energy’, ‘data mining, building performance’, and ‘Machine learning building, retrofit’. Then, much more literature was found by examining the cited and citing papers of the preliminary works. After that, a thorough collection of core publications to the proposed topic has been filtered out using the following three criteria:

1) The selected publications should address one or more matters relevant to building performance, including energy, thermal environment, visual environment, acoustic environment, and indoor air quality;
2) The dataset used by the existing study should be collected from many buildings. Studies that were carried out on time-series, simulated, or other generated data are excluded;
3) One of the data-driven methods, related to statistics, regression, clustering, and classification, was adopted in the study.

Finally, a total of 91 core papers have been picked out. The following part of this paper will address the three questions proposed above in turn in Section 3 to 5, respectively. At the end of the paper, a conclusion is given to summarize the main findings.

3. Data-driven applications for analysis and design

3.1. With Statistical methods

3.1.1. Simple statistical analysis

A large amount of OBPD is a valuable resource to validate the effectiveness of energy policies, energy-efficient measures, and design methods. Many studies have compared the energy consumption between different building groups to evaluate the effectiveness of the design methods. In recent years, LEED (Leadership in Energy and Environmental Design) has become a widespread green building certification system in North America. Newsham et al. [9] statistically analyzed the energy consumption of 100 LEED-certified buildings. They found out that LEED buildings averagely consumed 18-39% less energy, but 28-35% of LEED buildings used more energy than their traditional counterparts. Further analysis showed that LEED buildings did not live up to the expectation of performance set in the design stage. In 2013, Scofield [15] statistically analyzed the energy consumption of 953 office buildings in New York City. He concluded that LEED gold buildings outperformed non-LEED ones, but LEED silver and certified underperformed non-LEED buildings. Scofield and Doane [6] compared the energy consumption between LEED-certified school buildings and conventional school buildings in Chicago. The results indicate that LEED-certified buildings consumed 17% more source energy than other buildings. Household electricity use for heating and cooling was taken by Wang et al. [16] as the metric to evaluate the effectiveness of China Building Energy Efficiency Standards on residential buildings in Chongqing, China. It turned out that households that adopted the standards saved about 41% more energy than those who didn't.

The actual effectiveness of energy retrofit measures can be unveiled by comparing the energy of pre- and post-retrofitted buildings. Liang et al. [17] analyzed the pre-and post-renovation energy bills of 201 residential buildings and 636 commercial buildings. It turned out that those energy-efficient measures saved 30-50% less than the engineering models expected. With the energy efficiency data of nearly 30000 households in Michigan, Fowlie et al. [18] found that energy savings calculated by energy efficiency programs were more than three times the actual savings. Filippidou et al. [19] investigated the outcomes of various energy-efficient retrofit measures with the Dutch non-profit house data. The results showed that dwellings implemented with three or more measures achieved significant energy performance improvement. But, only less than 3% of dwellings were retrofitted with three or more measures.

Another meaningful job is to identify potential energy retrofit opportunities for buildings within a specific region by comparing the energy differences for buildings with and without an energy-efficient measure. With data of approximately 90,000 dwellings in the Netherlands, Brom et al. [20] calculated the energy differences between dwellings retrofitted with a measure and buildings without any renovation. In this way, they explored whether building types and occupant behavior had influenced the energy savings for different renovation measures. By statistically analyzing the thermal performance data of about 10400 houses, Streicher et al. [21] pointed out that approximately three quarters had not reached the latest building thermal performance requirements. Besides, they identified the renovation requirements of opaque envelopes, windows, and oil-fired boilers. Shahrokni et al. [22] compared the energy-efficient potentials of buildings in different age ranges and drew a conclusion that if existing buildings were retrofitted to satisfy current codes, the heating energy would be reduced by one-third. Moreover, buildings constructed between 1946 and 1975 were verified to have the largest energy reduction potentials. Besides energy consumption, Calero et al. [23] statistically examined the CO_2 emissions and energy costs of many multi-family residential buildings that adopted various energy-saving measures. The results indicated that solar/biomass energy stood out for the most significant benefits. With OBPD obtained by walk-through surveys of energy end-uses, Lee et al. [24] built a set of design criteria as a substitute for the traditional design requirements, which were identified as an effective method to mitigate the impacts of oversized cooling.

3.1.2. Advanced statistical analysis

Identifying dominant factors of building performance is an essential job for better predictions, understanding of building performance, and policy-making. When determinant features are targeted, engineers can adjust these parameters to achieve high-performance design and operation. If operation patterns were identified, once be informed, occupants would be interested in and willing to change their behavior to pursue high-performance operations [25]. When identifying determinant factors, several studies, like [26], merely compared the performance difference between different groups, but the majority of existing studies adopted statistical and machine learning methods.

Correlation coefficient

The correlation coefficient is defined to measure the correlation between two variables. As a commonly used one, the Pearson correlation coefficient accesses the linear relationships between two variables. Eq. (1) shows the mathematical formula of the popular Pearson correlation coefficient. Some researchers applied this method to test the correlation between each feature and annual energy consumption [10, 27-32]. When testing the relationship between two variables, Spearman's rank correlation coefficient, adopted by [33], is equal to the Pearson correlation coefficient. As an exemplified study, with the data of 20,802 residential buildings in the city of Basel, Aksoezen et al. [27] identified that gas consumption is closely related to building volume, gross residential floor area, exposed surface area, number of people, and exposed elevation area using Pearson correlation coefficients.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \qquad \text{Eq. (1)}$$

where:

- cov(X,Y) is the covariance of two variable X, and Y;
- σ_X is the standard derivation of X;
- σ_Y is the standard derivation of Y.

Chi-square test

The Chi-square test is a statistical test to evaluate the differences in the distribution of categorical data in different sets. Eq. (2) gives the chi-square statistic χ^2 calculation equation. P-value is used to quantify whether a null hypothesis is accepted or rejected in a chi-square test. Out of the reviewed literature, Kuo et al. [32] took advantage of the Pearson correlation coefficient and Chi-square test methods to identify determinant features of the energy consumption of convenience stores.

$$\chi^2 = \sum_j \frac{(o_j - e_j)^2}{e_j} \qquad \text{Eq. (2)}$$

where:

- o_j is the observed frequencies in the jth cell;
- e_j is the expected frequencies in the jth cell.

Analysis of variance

Analysis of variance (ANOVA) is a group of statistical methods to test the significance of discrepancies among more than two groups, including t-test, F-test, and one-way ANOVA, etc. ANOVA models have been frequently adopted to analyze dominant factors in the reviewed literature [16, 20, 34-40]. Bartusch et al. [35] have adopted t-tests and one-way ANOVA methods to examine the impact of household features on the electricity consumption of residential buildings. Heating system types, building area, family members, year of construction, and service water heater types were found to impose a significant influence on electricity consumption. On the other hand, in terms of the effect on different subgroups, insulation of external walls, ventilation heat exchanger, indoor temperature control system and supplementary heating installations did not show apparent variance. Despite so, it was pointed out that the operation behavior of these household appliances is worthy of careful consideration in the future.

3.2. With machine learning methods

DD-BPD refers to any procedure that allows rising high-performance design solutions for a proposed building with statistical or machine learning methods. The applied methods are traditionally classified into regression, classification, and clustering.

3.2.1. Regression

Regression encompasses a variety of processes of developing models to predict the target-dependent variables with one or more variables. In the building performance field, many machine learning algorithms have been applied to evaluate building energy consumption. Table 1 lists the main literature that utilized regression techniques to investigate building performance. Fig. 2 and 3 demonstrate the distributions of regression studies by algorithms and target variables. Fig. 2 demonstrates the distribution of algorithms adopted by those studies, where each type stands for a category of algorithms. For instance, decision tree includes the J48 decision tree, C4.5 decision tree, ID3, random forest, CART, CHAID, and MARS. With Fig. 3, it can be perceived that linear regression models are the most commonly used algorithms, followed by ANN, decision tree, and SVM models. Fig. 3 shows that a majority of studies focused on the prediction of the EUI, heating energy, and electricity. Only 15% of studies [10, 41-47] probed into the renovation opportunities, HVAC selection, or energy-saving measures.

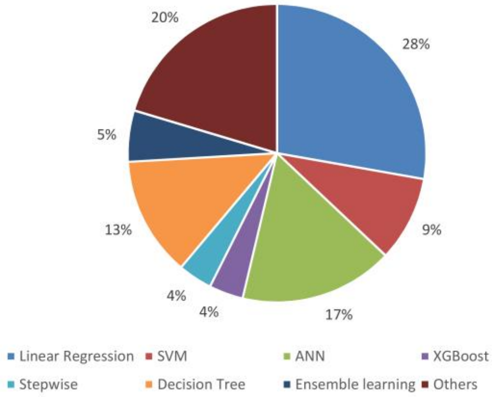


Fig. 2 Distribution of reviewed regression papers by algorithms

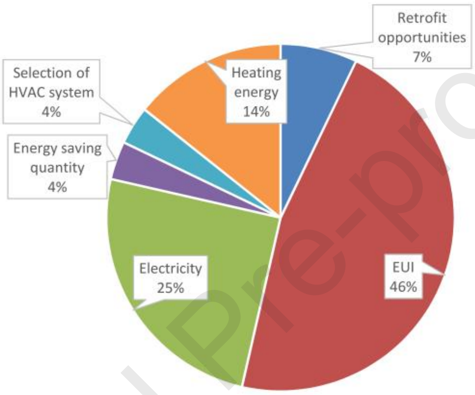


Fig. 3 Distribution of reviewed regression papers by target variables

High prediction accuracy is an essential requirement for BPD. Typical criteria of regression results include R^2 (coefficient of determination), Adj. R^2, MAPE, CVMSE, CV-RMSE, and Std. Error etc. referring [11] for their detailed, definitive equations. As for prediction accuracy, we found advanced machine learning, like ANN and Random Forest, achieved higher accuracy than Multiple Linear Regression (MLR) [33, 48-50]. In Table 1, if R^2 or Adj. R^2 is taken as the assessment criterion, the values of almost half of these studies could exceed 0.5 including [10, 32, 38, 51-54]. However, several prediction objectives are not traditional EUI, but energy consumption [10, 51] which is linear to building area, or logarithmic transformed EUI [52, 53] which makes the predictions more like classification than regression, or EUI changes after retrofitting [50]. Most puzzling of all, Pan and Zhang [54] predicted EUI with subentry energies as input variables. The study conducted by Martinez and Choi [38] lacks credibility as it remains open to doubt how it was possible to predict EUI merely with façade information achieving such high accuracy.

Besides modeling building performance, regression algorithms have also been applied to predict decision-making processes. For example, Gamtessa [55] tried to model residential energy retrofit decisions from an economics perspective with pre- and post-retrofit audit data. The results showed the proposed model plays a part in proposing the most significant energy-saving measures for buildings with specific characteristics.

Table 1

A summary of the representative regression studies

Literature	Methods	Number of buildings	Building type	Feature No.	Designable feature No.	Objective	Accuracy
[10]	MLR	74	School	7	0	Electricity, EUI	Adj. R^2: 0.948 for university
[30]	MLR, BPNN	30	Office	7	5	Electricity	MAPE : 3.1% (BPNN)
[32]	Gaussian processes, MLR, SMOreg, M5P, M5Rules, Decision trees.	723	Convenience stores	33	N.A.	EUI	Correlation coefficient: 0.85 (M5 Rules)
[38]	Stepwise regression	92	Miscellaneous	29	6	EUI	R^2: 0.83
[39]	Robust regression	134K	Residential	5	2	Heating demand	Std. error: 0.027
[41]	MLP, Nonlinear principal component analysis	4767	Office	5	4	Energy retrofit index	R^2: 0.42(MLP)
[43]	MLR	926	Commercial	10	6	Retrofit savings	R^2: 0.40
[48]	Regularization, Hierarchical group lasso regularization	4748	Residential and office	257	52	EUI	MSE: 0.46 (Multifamily); 0.40 (Office).
[49]	XGBooost, Decision tree, SHAP,	7487	Residential	15	0	Weather normalized source EUI	R^2: 0.31 (XGBoost)
[50]	MLR, ANN	56	Office	4	1	Retrofit EUI changes	R^2: 0.744 (ANN).
[52]	Transformed linear regression	12K	Residential	13	10	Electricity	Adj. R^2: 0.86
[56]	Clusterwise regression, MLR,	3902	Residential	250	70+	EUI	CVMSE: 0.30
[57]	Lasso regression	845	Households	20	2	Electricity	Adj. R^2:0.34
[58]	MLR	713	Miscellaneous	16	13	Energy intensity	Adj. R^2: 0.526
[59]	ANN	1872	School	11	1	Electricity and heating energy	MAPE: 34.0% (Electricity)； MAPE: 25.1% (Heating energy).
[60]	RF, MLR and Lasso, SVM	3640	Residential	171	9	EUI	MSE: 0.773(RF); RMSE: 0.879(RF).
[61]	SVM, BPNN, RBFNN, GRNN,	59	Residential	15	14	Electricity	RMSE: 2.40% (SVM); MRE: 1.90 (SVM).
[62]	GLR	3446	Residential	14	4	Daily electricity	MAPE: 3.9%; MAE: 0.81; MSE 1.87; RMSE 1.37.
[63]	MLR	49K	Residential	30	N.A.	Gas use	Adj. R^2: 0.238.
[64]	MLR	72	Bank	6	2	EUI	R^2: 0.253; Adj. R^2: 0.185.
[65]	ANN, REPT, RF, SVM	90K	Residential	10	3	Heating energy	MAPE: 16.4%; MAE: 22.2.

N.A.: not available. The accuracy column describes the best accuracy of each study, Unit of MAE and RMSE: KWh/m^2.

3.2.2. Classification

Classification aims at solving the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, based on a training set of data containing observations (or instances) whose category memberships are known. Table 2 summarizes the main studies that applied classification methods to building performance analysis, all of which achieved quite high classification accuracies. Two of these papers focused on classifying energy consumption levels [32, 66], while the remaining two aimed at identifying renovation strategies [67, 68].

As for classification, a foremost work is to define the target categories based on building performance. In statistical and machine learning, classification categories are often two-fold. For example, data scientists categorize bank loan applications as safe or risky. Kuo et al. [32] binned data into 2, 3, and 5 subsets based on buildings' EUI. Their results showed that the more categories the data has, the worse the classification accuracy will be. Gupta et al. [69] divided buildings into two groups: efficient and inefficient. Yu et al. [66] also classified residential buildings into two categories, i.e., high and low ones with their total energy consumption per area. EUI is the most commonly used feature to indicate buildings' energy performance [27, 32, 45, 66, 70]. However, it fails to reflect the influence of building operation hours, types, and weather conditions, etc. [71]. NABERS, an Australian green building certification standard, regards operation hours as a critical variable except building area when rating the star level of a building [72]. In other minority studies, energy was not taken as a criterion for the classification of buildings, but cost and schedule performance [73].

Table 2
A summary of the representative classification studies

Literature	Intention	Method	Number of buildings	Building type	Feature No.	Prediction accuracy
[32]	To develop a new energy model for convenience stores other than traditional energy models.	Gaussian processes, Decision trees	723	Convenience stores	33	87%
[66]	To estimate residential building EUI levels.	Decision tree	67	Residential	10	92%
[67]	To utilize available data to target ECMs across a city's entire building stock.	User-facing falling rule list	1100	Miscellaneous	23	ROC AUC: 0.72–0.86
[68]	To get understanding the decision-making processes of energy efficiency investments.	Logistic regression	763	Office	7	ROC AUC: 0.85.

3.2.3. Clustering

Clustering refers to a general process of grouping a set of objects into different groups wherein objects are similar to each other. A straightforward application of clustering is to cluster buildings into several groups based on building attributes [74]. Petcharat et al. [34] conducted a clustering analysis of a set of actual lighting power density data extracted from an energy audit database. They found that the clustering analysis achieved much higher accurate results than that from general prediction methods. As a consequence, this method can be used to access different lighting systems.

In the building performance field, clustering is commonly applied to identify reference buildings and performance patterns [64, 75-82]. Wong et al. [64] grouped 72 bank buildings into four types based on their shapes, sizes, geographical layouts, and construction ages. Their review of existing works that applied clustering techniques showed that buildings were always grouped by geometry, climate zone, age, thermal performance, usage, and HVAC. Gaitani et al. [76] applied the K-mean algorithm to cluster 5 main building classes of 1100 Greece schools. Then, principal components analysis was used to figure out typical school buildings and to access the energy-saving potentials of each building group. Gao and Malkawi [77] proposed to cluster reference buildings for benchmarking based on the multi-dimensional domain of building features other than just on building types.

Clustering also commonly services as a pre-analysis of other data-driven analyses [64, 83]. As one kind of clustering, similarity analysis aims at sifting buildings similar to the design building. The data-driven design strategies are supposed to inspire buildings similar to the proposed buildings [45]. Fig. 4 shows the distribution of clustering algorithms in these reviewed papers, which shows that even as the most straightforward method, k-means was the most frequently used clustering method.

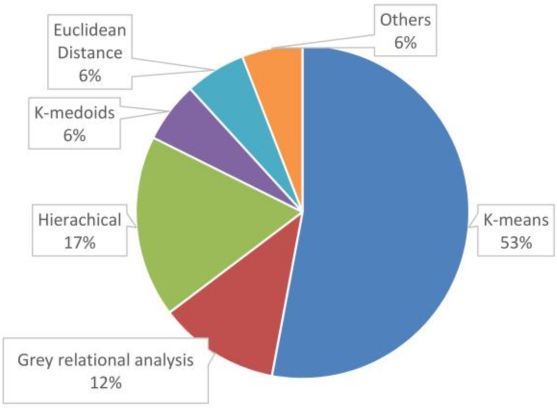


Fig. 4 Distribution of clustering algorithms in the reviewed papers

3.3. A summary of data-driven methods

Table 3 summarizes the key applications, strengths, and weaknesses of different data-driven methods. In DD-BPD practical terms, existing studies have concentrated too much on the prediction of energy other than the energy-efficient design which can be reflected by Fig. 3. More critical discussion would be given in Section 5.

Table 3

A summary of the applications, strengths, weaknesses of existing data-driven methods

DD-BPD methods	Key applications	Strengths	Weaknesses
Simple statistical	Feature distributions. Comparison of different building groups.	Easy to understand, outstanding visualization.	Too intuitive. Cannot quantify latent reasons.
Advanced statistical	Identification of determinant features. Pre-processing of DD-BPD.	Easy to do.	Hard to define the thresholds of significances.
Regression	Prediction of building energy consumption.	Infer the relationships between building features and energy uses.	Low prediction accuracy. Less resilience.
Classification	Classification of buildings by their energy performance.	Prediction of categorical features. Relatively high prediction accuracy.	
Clustering	Exploring reference buildings.	Easy to understand and build models.	Hard to give the weight of different features.

4. Building performance databases

Building performance data are the foundation of data-driven analysis. In other words, without a versatile building performance database, it would be difficult to develop a successful model. Table 4 summarizes the major building performance databases used in the core literature. As shown in Table 4, these databases were initially set up for energy benchmarking, performance certification, and energy statistics. Energy benchmarking relates to a comparison of energy consumption between one building and its similar buildings [84]. In practice, energy benchmarking programs only involve several building features. For example, the New York City benchmarking dataset has six building features, i.e. building area, principle activity, year built, number of buildings, occupancy percentage, metered areas [85]. What’s more, most of the databases in Table 4 have less than 50 features. As for the data size, many databases contain millions of buildings, while some only have several thousand pieces of data.

Databases developed in the U.S. and the U.K. are mostly public accessible due to disclosure laws, such as NYC Benchmarking Law [85] or the conscience of the avant-garde groups such as the energy-saving trust in the U.K. [86]. There are many other government-supported databases, for example, the Residential Energy Consumption Survey (RECS) dataset built by the U.S. Energy Information Administration [87].

Table 4

A summary of main databases adopted in the core literature

Database or program	Adopted by	Country	Initial intention	Number of buildings	Feature No.	Open access?	Main characteristic s
CBECS[88]	[33, 53, 89]	US	To unveil the current energy consumption status of non-residential buildings in the United States.	6,720	>502	Y	• Contains a large number of features. • Lacks a detailed description of building components, for example, the U-value of opaque enclosures.
BPD[90]	[43, 91]	US	To build an open accessible benchmarking platform.	>870,000	Unclear	P	• Comes from more than 30 sources. • Less 2% data include HVAC system information.
SHAERE	[19, 20, 92]	Netherlands	To investigate energy efficiency evaluation of building stock in the Netherlands.	>1.7M	Unclear	N	• Contains the energy consumption data several years in a row. • Includes physical characteristics, heating and ventilation installations, theoretical energy consumption, the average energy intensity and more. • Monitors the progress of energy-saving measures in the social house sector.
CECB	[21]	Switzerland	To provide a clear state of certified buildings in terms of energy efficiency.	20,919	Unclear	N	• Certified experts collect building characteristics, heating system information and some behaviour-related aspects.
Danish EPC database	[39]	Denmark	To issue an energy performance certification.	134,093	6	N	• Contains information about the physical properties, e.g. U- values and areas of all external walls, heated floor areas, types of heat supply, ownership, and geographical location, and other related characteristics.
NYC's benchmarking	[49, 56, 60, 67, 78]	US	To benchmark buildings in New York City.	7,500	6	Y	• Mainly includes building energy consumption but lacks building characteristic features.
HEED[86]	[46]	UK	To record the energy efficiency installations in the domestic building stock.	>168,998	42	Y	• Contains information about energy performance and the installation of energy efficiency retrofits.
NEED[93]	[94]	UK	To examine the changes in domestic gas and total energy consumptions for the dwellings.	>3.0M	Unclear	Y	• Based on a rich, well-structured and reliable data of good quality, enabling a robust estimation of pro- and post-retrofit with different energy efficiency measures.
2006 ABS census data	[52]	Australia	To quantify the relationship between the energy consumption and the households and dwellings characteristics.	11,967	19	N	• There is a total of 249 fields to describe building characteristics.

P: partial accessible. Y: open accessible. N: not available for public.

In addition, researchers have customized their datasets to fulfill specific studies. In Table 1 and 2, datasets with over 1000 buildings were derived from various government-supported energy projects. Except for these from government-supported projects, 67% of the remaining datasets have less than 100 buildings. However, those studies also covered many interesting topics. For instance, in order to predict the energy savings of new air-conditioning systems, it is necessary to record energy consumption pre- and post-retrofitting [50].

5. Discussion

5.1. Achievements

5.1.1. Overview

Statistical methods have been deployed to analyze building performance patterns, explore the effectiveness of performance policies and design methods, such as benchmarking and green building certification. To delve into determinant factors, several advanced statistical algorithms, like correlation coefficient, Chi-square test, and ANOVA were often adopted but did not take the interrelation between features into consideration. The results of correlation coefficient analyses show that highly correlated parameters did not coincide with other analysis results [10, 27], and imply that less correlated parameters may be crucial for building a regression or classification model. Existing studies have analyzed the relationship between energy consumption and various building characteristics, including building envelopes [70], HVAC systems [44], HVAC operation and maintenance [58], lighting [30], human behavior [95], capita income [96].

Several studies have tried to build models for BPD of building envelops [70], HVAC system selection [44], cooling capacities [30], cooling plant efficiency [50], air-conditioning operation and maintenance [58]. As for the prediction accuracy, Turner and Frankel [7] investigated the measured and predicted energy performance of 121 LEED new construction buildings. The R^2, MSE, and RMSE of those predictions are 0.505, 19.1kBtu/ft^2, and 23.8kBtu/ft^2, respectively. Using the CBECS database, Dong et al. [33] compared the performance of several statistical and machine learning algorithms in predicting building energy performance. The best prediction results for total EUI were achieved by SVM with MSE of 18.2kBtu/ft^2, 18.1kBtu/ft^2, and 25.7kBtu/ft^2 for training, validation, and testing dataset, respectively. It is quite possible that energy prediction accuracy achieved by data-driven models will bypass the energy simulation method.

5.1.2. Identifying determinant features

Section 2.1.2 summarizes several studies that applied statistical algorithms to raise determinant features. Besides, many studies have adopted regression or classification learning algorithms to identify determinant factors. These algorithms include linear regression models [32, 43, 47, 97], quantile regression [96, 98, 99], hierarchical group-lasso regularization models [56] , decision tree [38, 58, 66, 100, 101], random forest [60, 102], step regression [38, 103], ANN [59], ordinary least squares (OLS) regression and lasso regression [57], and k-means [104].

Identifying determinant features of building energy consumption is the most commonly and successfully applied strategy. Table 5 summarizes several representative studies on identifying determinant features. To identify determinant features of residential electricity consumption, Esmaeilimoakher et al. [105] surveyed many factors mainly related to building size, occupants, and their behavior in Western Australia. They adopted one-variable linear regression models to depict the linear correlation between each factor with electricity consumption. The results showed that floor area, household size, disposable household income, and Head of Household have a significant influence on electricity consumption, while the number of children and window-opening behavior show little effect. Zhang et al. [106] conducted three different feature engineering approaches, i.e., exploratory data analysis, random forest, and principal component analysis for feature visualization, feature selection, and feature extraction, respectively. The three feature engineering methods were found to share some common features vital to machine learning, however, which can be hardly explained by experts.

In a conventional experiment, to test the influence of one parameter, it is a standard practice to keep all other settings the same in different test groups. In this sphere, similarity analysis is a preliminary work before analyzing the influence of one parameter upon the building performance. To qualify the impact of occupant behavior on energy performance, Ashouri et al. [42] proposed a two-level clustering framework wherein buildings were clustered into different groups using the k-means algorithm. Papadopoulos [49] used SHAP (SHapley Additive exPlanation) to evaluate the effects of single features on EUI.

Table 5

A summary of the representative studies on identifying determinant factors

Literature	Building Type	Number of buildings	Feature No.	Method	Dependent variable	Determinant factors (sorted by influence)
[35]	Residential house	595	8	t-tests and ANOVA	Electricity for heating	Electric boiler, heat pump, non-electric, electric, and combined boiler.
[58]	Miscellaneous	713	16	Decision tree	Energy intensity	Chiller condition, AC cleaning frequency, luxury housing, and commercial use.
[70]	Household	3688	20	Decision tree	Heating energy	Window heat-transmission coefficient, window and roof heat-transmission coefficient,
[98]	Household	Unclear	18	Quantile regression	Electricity consumption	Higher income, larger household size, and more elderly members consumed more electricity
[102]	School	41	10	Random Forest	Heating energy	Thermal transmittance of the windows, external walls, ground floor, roof and capacity of the heating system.

5.2. Challenges

Even though data-driven models attract much attention, many challenges are still existing. As Fig.5 shows, the subsequent part of this section would discuss these obstacles from three aspects, i.e., practical data-driven models, availability of data, and DD-BPD industrial practices.

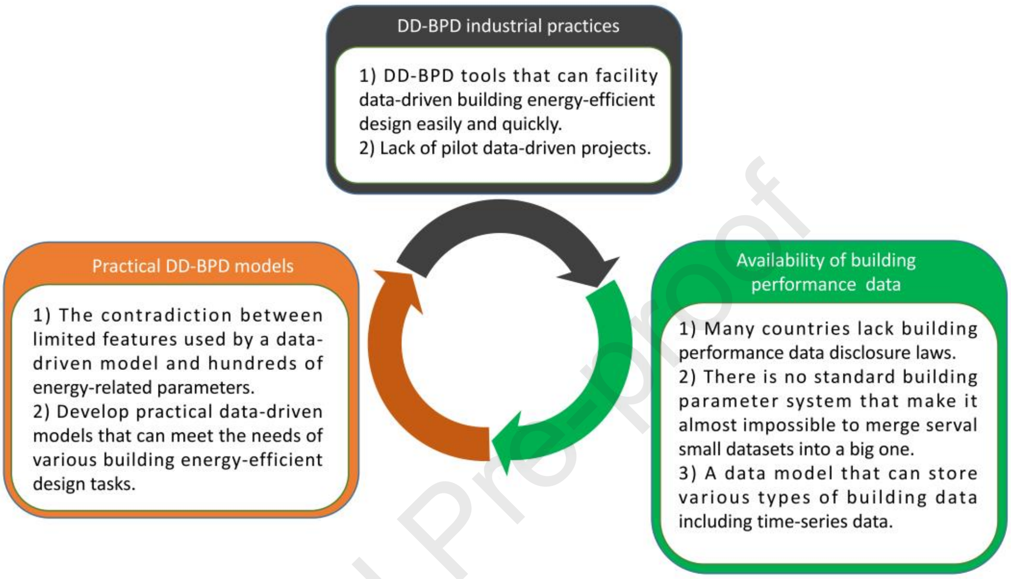

DD-BPD industrial practices
1) DD-BPD tools that can facility data-driven building energy-efficient design easily and quickly.
2) Lack of pilot data-driven projects.
Practical DD-BPD models
1) The contradiction between limited features used by a data-driven model and hundreds of energy-related parameters.
2) Develop practical data-driven models that can meet the needs of various building energy-efficient design tasks.
Availability of building performance data
1) Many countries lack building performance data disclosure laws.
2) There is no standard building parameter system that make it almost impossible to merge serval small datasets into a big one.
3) A data model that can store various types of building data including time-series data.

Fig. 5 Main challenges of current DD-BPD technique

5.2.1. How to develop practical data-driven models?

Comparing with the building energy simulation techniques that have been applied to building energy predictions and energy-efficient design projects, seldom data-driven models were used in actual projects [12, 107]. For one reason, data-driven models can hardly cope with data outside the training data [108]. In other words, the wisdom excavated from some buildings cannot be applied to other projects [109]. For another reason, previous data-driven models emphasized energy prediction other than energy-efficient design [110].

Embedded many calculation programs, energy simulation software can calculate the impacts of many features. In reverse, these features can be designed by energy simulation tools. However, a supervised learning model for energy prediction usually contains less than 10 features. Besides, these models mainly may be used to design important designable features that account for a small proportion of hundreds of energy-related features [110]. In Table 1, only 23.8% of features were designable. Many regression models, such as in [10, 49], did not contain any designable features. It is an urgent demand to develop practical data-driven models that can meet the needs of various building energy-efficient design tasks. For example, data-driven designs of building envelopes, shadings, insulation, and ventilation are still insufficient. Specific models should be developed for particular design works.

Data-driven studies about energy predictions demonstrated the reliability of this technique not only in rapid modeling but also in accuracy. However, along with the revealed problems of existing data-driven energy prediction in section 3.2.1, many more solid studies should be carried out to dispel the doubts in the future.

5.2.2. How to promote the availability of building performance data?

First, the disclosure laws of building performance data are still missing in many countries. In the U.S., the disclosure requirements have boosted the formulation of several public accessible building performance datasets including the CBECS survey data and New York City benchmarking data [111]. In contrast, in China, large-scale public buildings are required to install building energy monitoring systems, and a large amount of building energy data is being collected [112]. However, due to the lack of disclosure laws, researchers and engineers cannot acquire these data.

Second, there is no standard building parameter system that hinders the merging of small datasets. At present, due to this reason, although several building performance datasets are available, they can hardly be merged into a large database [91].

Third, the integration of multi-source data is still weak especially between descriptive data and time-series data. A well-designed data model is a key factor to successfully promote engineering practices [113]. Considering the wide-available of the internet of things (IoTs), the data model also needs to consider the integration of time-series data in the future. For a long time, data-driven models for time-series prediction were always carried out on data from a sensor or a single building [114-116]. The weak integration of large time-series building data of many buildings hinders the exploring hidden information with big data.

5.2.3. How to stimulate DD-BPD industrial practices?

From the perspectives of engineers and designers, to fulfill industrial projects, DD-BPD tools are necessary foundations. Built by the U.S. Department of Energy, the Building Performance Database is not only a large building performance dataset but also an online platform that allows users to compare the energy consumption of different building groups [90]. The U.S. Department of Energy also sponsored the development of a data-driven monthly building energy prediction tool, named BETTER [117]. However, these tools can only conduct limited statistical analyses or predictions. In addition, pilot projects and data-driven building design competitions would stimulate DD-BPD industrial practices. DD-BPD projects can promote the accumulation of building data if the inputs can be saved with the permission of users.

6. Conclusion

This paper focused on studies of applying data-driven approaches for building performance analysis and design on big on-site building performance data and analyzed existing building performance databases covering current statuses and challenges. A comprehensive discussion highlights the achievements and challenges which indicate future research directions.

Several remarkable findings need to be depicted as follows. In the application fields, most studies concentrated on exploring energy trends, energy prediction, determinant features, and referring buildings. Few studies aimed at developing data-driven models for BPD. As for data-driven approaches, comparing with regression, classification approaches have received much less attention. Identification of determinant features is a common but successful application in existing studies. To thoroughly boost data-driven methods for BPD, several challenges should be tackled relating to practical data-driven models, availability of building performance data, and stimulation of DD-BPD industrial practices.

Acknowledgements

This research is supported by Shanghai Super Post-doctoral Program.

Reference

[1]. Ürge-Vorsatz, D. and A. Novikova, *Potentials and costs of carbon dioxide mitigation in the world's buildings.* Energy Policy, 2008. **36**(2): 642-661.

[2]. Perez-Lombard, L., J. Ortiz, and C. Pout, *A review on buildings energy consumption information.* Energy and Buildings, 2008. **40**(3): 394-398.

[3]. Shi, X., *Design optimization of insulation usage and space conditioning load using energy simulation and genetic algorithm.* Energy, 2011. **36**(3): 1659-1667.

[4]. Hensen, J.L. and R. Lamberts, *Building performance simulation for design and operation*. 2012: Routledge.

[5]. Mustafaraj, G., D. Marini, A. Costa, and M. Keane, *Model calibration for building energy efficiency simulation.* Applied Energy, 2014. **130**: 72-85.

[6]. Scofield, J.H. and J. Doane, *Energy performance of LEED-certified buildings from 2015 Chicago benchmarking data.* Energy and Buildings, 2018. **174**: 402-413.

[7]. Turner, C. and M. Frankel, *Energy performance of LEED for new construction buildings.* New Buildings Institute, 2008. **4**: 1-42.

[8]. van den Brom, P., A. Meijer, and H. Visscher, *Performance gaps in energy consumption: household groups and building characteristics.* Building Research & Information, 2018. **46**(1): 54-70.

[9]. Newsham, G.R., S. Mancini, and B.J. Birt, *Do LEED-certified buildings save energy? Yes, but ...*

Energy and Buildings, 2009. **41**(8): 897-905.

[10]. Wang, J.C., *A study on the energy performance of school buildings in Taiwan.* Energy and Buildings, 2016. **133**: 810-822.

[11]. Amasyali, K. and N.M. El-Gohary, *A review of data-driven building energy consumption prediction studies.* Renewable & Sustainable Energy Reviews, 2018. **81**: 1192-1205.

[12]. Wei, Y.X., X.X. Zhang, Y. Shi, L. Xia, S. Pan, J.S. Wu, M.J. Han, and X.Y. Zhao, *A review of data-driven approaches for prediction and classification of building energy consumption.* Renewable & Sustainable Energy Reviews, 2018. **82**: 1027-1047.

[13]. Bourdeau, M., X. qiang Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, *Modeling and forecasting building energy consumption: A review of data-driven techniques.* Sustainable Cities Society, 2019. **48**: 101533.

[14]. Miller, C., Z. Nagy, and A. Schlueter, *A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings.* Renewable and Sustainable Energy Reviews, 2018. **81**: 1365-1377.

[15]. Scofield, J.H., *Efficacy of LEED-certification in reducing energy consumption and greenhouse gas emission for large New York City office buildings.* Energy and Buildings, 2013. **67**: 517-524.

[16]. Wang, X., W. Feng, W. Cai, H. Ren, C. Ding, and N. Zhou, *Do residential building energy efficiency standards reduce energy consumption in China?–A data-driven method to validate the actual performance of building energy efficiency standards.* Energy Policy, 2019. **131**: 82-98.

[17]. Liang, J., Y.M. Qiu, T. James, B.L. Ruddell, M. Dalrymple, S. Earl, and A. Castelazo, *Do energy retrofits work? Evidence from commercial and residential buildings in Phoenix.* Journal of Environmental Economics and Management, 2018. **92**: 726-743.

[18]. Fowlie, M., M. Greenstone, and C. Wolfram, *Do energy efficiency investments deliver? Evidence from the weatherization assistance program.* The Quarterly Journal of Economics, 2018. **133**(3): 1597-1644.

[19]. Filippidou, F., N. Nieboer, and H. Visscher, *Are we moving fast enough? The energy renovation rate of the Dutch nonprofit housing using the national energy labelling database.* Energy Policy, 2017. **109**: 488-498.

[20]. van den Brom, P., A. Meijer, and H. Visscher, *Actual energy saving effects of thermal renovations in dwellings-longitudinal data analysis including building and occupant characteristics.* Energy and Buildings, 2019. **182**: 251-263.

[21]. Streicher, K.N., P. Padey, D. Parra, M.C. Burer, and M.K. Patel, *Assessment of the current thermal performance level of the Swiss residential building stock: Statistical analysis of energy performance certificates.* Energy and Buildings, 2018. **178**: 360-378.

[22]. Shahrokni, H., F. Levihn, and N. Brandt, *Big meter data analysis of the energy efficiency potential in Stockholm's building stock.* Energy and Buildings, 2014. **78**: 153-164.

[23]. Calero, M., E. Alameda-Hernandez, M. Fernandez-Serrano, A. Ronda, and M.A. Martin-Lara, *Energy consumption reduction proposals for thermal systems in residential buildings.* Energy and Buildings, 2018. **175**: 121-130.

[24]. Lee, W.L., F.W.H. Yik, P. Jones, and J. Burnett, *Energy saving by realistic design data for commercial buildings in Hong Kong.* Applied Energy, 2001. **70**(1): 59-75.

[25]. Mansouri, I., M. Newborough, and D. Probert, *Energy consumption in UK households: Impact of domestic electrical appliances.* Applied Energy, 1996. **54**(3): 211-285.

[26]. Melois, A.B., B. Moujalled, G. Guyot, and V. Leprince, *Improving building envelope knowledge*

from analysis of 219,000 certified on-site air leakage measurements in France. Building and Environment, 2019. **159**: 106145.

[27]. Aksoezen, M., M. Daniel, U. Hassler, and N. Kohler, *Building age as an indicator for energy consumption.* Energy and Buildings, 2015. **87**: 74-86.

[28]. Baker, K.J. and R.M. Rylatt, *Improving the prediction of UK domestic energy-demand using annual consumption-data.* Applied Energy, 2008. **85**(6): 475-482.

[29]. Theodoridou, I., A.M. Papadopoulos, and M. Hegger, *Statistical analysis of the Greek residential building stock.* Energy and Buildings, 2011. **43**(9): 2422-2428.

[30]. Jing, R., M. Wang, R.X. Zhang, N. Li, and Y.R. Zhao, *A study on energy performance of 30 commercial office buildings in Hong Kong.* Energy and Buildings, 2017. **144**: 117-128.

[31]. Yuan, L., Y.J. Ruan, G. Yang, F. Feng, and Z.W. Li, *Analysis of Factors Influencing the Energy Consumption of Government Office Buildings in Qingdao.* Clean Energy for Clean City: Cue 2016 - Applied Energy Symposium and Forum: Low-Carbon Cities and Urban Energy Systems, 2016. **104**: 263-268.

[32]. Kuo, C.F.J., C.H. Lin, and M.H. Lee, *Analyze the the energy consumption characteristics and affecting factors of Taiwan's convenience stores-using the big data mining approach.* Energy and Buildings, 2018. **168**: 120-136.

[33]. Deng, H.F., D. Fannon, and M.J. Eckelman, *Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata.* Energy and Buildings, 2018. **163**: 34-43.

[34]. Petcharat, S., S. Chungpaibulpatana, and P. Rakkwamsuk, *Assessment of potential energy saving using cluster analysis: A case study of lighting systems in buildings.* Energy and Buildings, 2012. **52**: 145-152.

[35]. Bartusch, C., M. Odlare, F. Wallin, and L. Wester, *Exploring variance in residential electricity consumption: Household features and building properties.* Applied Energy, 2012. **92**: 637-643.

[36]. Filippin, C., F. Ricard, and S.F. Larsen, *Evaluation of heating energy consumption patterns in the residential building sector using stepwise selection and multivariate analysis.* Energy and Buildings, 2013. **66**: 571-581.

[37]. Hong, S.H., T. Oreszczyn, I. Ridley, and W.F.S. Grp, *The impact of energy efficient refurbishment on the space heating fuel consumption in English dwellings.* Energy and Buildings, 2006. **38**(10): 1171-1181.

[38]. Martinez, A. and J.H. Choi, *Exploring the potential use of building facade information to estimate energy performance.* Sustainable Cities and Society, 2017. **35**: 511-521.

[39]. Brogger, M., P. Bacher, H. Madsen, and K.B. Wittchen, *Estimating the influence of rebound effects on the energy-saving potential in building stocks.* Energy and Buildings, 2018. **181**: 62-74.

[40]. Fournier, E.D., F. Federico, E. Porse, and S. Pincetl, *Effects of building size growth on residential energy efficiency and conservation in California.* Applied Energy, 2019. **240**: 446-452.

[41]. Khayatian, F., L. Sarto, and G. Dall'O, *Building energy retrofit index for policy making and decision support at regional and national scales.* Applied Energy, 2017. **206**: 1062-1075.

[42]. Ashouri, M., F. Haghighat, B.C.M. Fung, and H. Yoshino, *Development of a ranking procedure for energy performance evaluation of buildings based on occupant behavior.* Energy and Buildings, 2019. **183**: 659-671.

[43]. Walter, T. and M.D. Sohn, *A regression-based approach to estimating retrofit savings using the Building Performance Database.* Applied Energy, 2016. **179**: 996-1005.

[44]. Yamaguchi, Y., Y. Miyachi, and Y. Shimoda, *Stock modelling of HVAC systems in Japanese commercial building sector using logistic regression.* Energy and Buildings, 2017. **152**: 458-471.

[45]. Tian, Z., B. Si, X. Shi, and Z.J.J.o.B.E. Fang, *An application of Bayesian Network approach for selecting energy efficient HVAC systems.* 2019. **25**: 100796.

[46]. Hamilton, I.G., A.J. Summerfield, D. Shipworth, J.P. Steadman, T. Oreszczyn, and R.J. Lowe, *Energy efficiency uptake and energy savings in English houses: A cohort study.* Energy and Buildings, 2016. **118**: 259-276.

[47]. Ruan, H.Q., X. Gao, and C.X. Mao, *Empirical Study on Annual Energy-Saving Performance of Energy Performance Contracting in China.* Sustainability, 2018. **10**(1666): 1-25.

[48]. Hsu, D., *Identifying key variables and interactions in statistical models of building energy consumption using regularization.* Energy, 2015. **83**: 144-155.

[49]. Papadopoulos, S. and C.E. Kontokosta, *Grading buildings on energy performance using city benchmarking data.* Applied Energy, 2019. **233**: 244-253.

[50]. Deb, C., S.E. Lee, and M. Santamouris, *Using artificial neural networks to assess HVAC related energy saving in retrofitted office buildings.* Solar Energy, 2018. **163**: 32-44.

[51]. Capozzoli, A., D. Grassi, and F. Causone, *Estimation models of heating energy consumption in schools for local authorities planning.* Energy and Buildings, 2015. **105**: 302-313.

[52]. Boulaire, F., A. Higgins, G. Foliente, and C. McNamara, *Statistical modelling of district-level residential electricity use in NSW, Australia.* Sustainability Science, 2014. **9**(1): 77-88.

[53]. Robinson, C., B. Dilkina, J. Hubbs, W.W. Zhang, S. Guhathakurta, M.A. Brown, and R.M. Pendyala, *Machine learning approaches for estimating commercial building energy consumption.* Applied Energy, 2017. **208**: 889-904.

[54]. Pan, Y. and L. Zhang, *Data-driven estimation of building energy consumption with multi-source heterogeneous data.* Applied Energy, 2020. **268**: 114965.

[55]. Gamtessa, S.F., *An explanation of residential energy-efficiency retrofit behavior in Canada.* Energy and Buildings, 2013. **57**: 155-164.

[56]. Hsu, D., *Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data.* Applied Energy, 2015. **160**: 153-163.

[57]. Huebner, G., D. Shipworth, I. Hamilton, Z. Chalabi, and T. Oreszczyn, *Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes.* Applied Energy, 2016. **177**: 692-702.

[58]. Lin, M., A. Afshari, and E. Azar, *A data-driven analysis of building energy use with emphasis on operation and maintenance: A case study from the UAE.* Journal of Cleaner Production, 2018. **192**: 169-178.

[59]. Hawkins, D., S. Hong, R. Raslan, D. Mumovic, and S. Hanna, *Determinants of energy use in UK higher education buildings using statistical and artificial neural network methods.* International Journal of Sustainable Built Environment, 2012. **1**(1): 50-63.

[60]. Ma, J. and J.C.P. Cheng, *Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests.* Applied Energy, 2016. **183**: 193-201.

[61]. Li, Q., P. Ren, and Q. Meng. *Prediction model of annual energy consumption of residential buildings*. in *2010 international conference on advances in energy engineering*. 2010. IEEE.

[62]. Fan, H., I.F. MacGill, and A.B. Sproul, *Statistical analysis of driving factors of residential energy demand in the greater Sydney region, Australia.* Energy and Buildings, 2015. **105**: 9-25.

[63]. Majcen, D.A., L. Itard, and H. Visscher, *Statistical model of the heating prediction gap in Dutch*

dwellings: Relative importance of building, household and behavioural characteristics. Energy and Buildings, 2015. **105**: 43-59.

[64]. Wong, I.L., E. Kruger, A.C.M. Loper, and F.K. Mori, *Classification and energy analysis of bank building stock: A case study in Curitiba, Brazil.* Journal of Building Engineering, 2019. **23**: 259-269.

[65]. Attanasio, A., M.S. Piscitelli, S. Chiusano, A. Capozzoli, and T. Cerquitelli, *Towards an Automated, Fast and Interpretable Estimation Model of Heating Energy Demand: A Data-Driven Approach Exploiting Building Energy Certificates.* Energies, 2019. **12**(7): 1273.

[66]. Yu, Z., F. Haghighat, B.C.M. Fung, and H. Yoshino, *A decision tree method for building energy demand modeling.* Energy and Buildings, 2010. **42**(10): 1637-1646.

[67]. Marasco, D.E. and C.E. Kontokosta, *Applications of machine learning methods to identifying and predicting building retrofit opportunities.* Energy and Buildings, 2016. **128**: 431-441.

[68]. Kontokosta, C.E., *Modeling the energy retrofit decision in commercial office buildings.* Energy and Buildings, 2016. **131**: 1-20.

[69]. Gupta, A., M. Kohli, and N. Malhotra, *Classification based on Data Envelopment Analysis and supervised learning: A case study on energy performance of residential buildings*, in *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems*. 2016, IEEE: Delhi, India.

[70]. Zhou, H., B.R. Lin, J.Q. Qi, L.H. Zheng, and Z.C. Zhang, *Analysis of correlation between actual heating energy consumption and building physics, heating system, and room position using data mining approach.* Energy and Buildings, 2018. **166**: 73-82.

[71]. Sharp, T.R. *Benchmarking energy use in schools*. in *Proceedings of the ACEEE 1998 Summer Study on Energy Efficiency in Buildings*. 1998. Citeseer.

[72]. NABERS, *NABERS Energy and Water for Offices -The Rules*. 2018.

[73]. Son, H. and C. Kim, *Early prediction of the performance of green building projects using pre-project planning variables: data mining approaches.* Journal of Cleaner Production, 2015. **109**: 144-151.

[74]. Chang, C., N. Zhu, K. Yang, and F. Yang, *Data and analytics for heating energy consumption of residential buildings: The case of a severe cold climate region of China.* Energy and Buildings, 2018. **172**: 104-115.

[75]. Santamouris, M., G. Mihalakakou, P. Patargias, N. Gaitani, K. Sfakianaki, M. Papaglastra, C. Paviou, P. Doukas, E. Primikiri, V. Geros, M.N. Assimakopoulos, R. Mitoula, and S. Zerefos, *Using intelligent clustering techniques to classify the energy performance of school buildings.* Energy and Buildings, 2007. **39**(1): 45-51.

[76]. Gaitani, N., C. Lehmann, M. Santamouris, G. Mihalakakou, and P. Patargias, *Using principal component and cluster analysis in the heating evaluation of the school building sector.* Applied Energy, 2010. **87**(6): 2079-2086.

[77]. Gao, X.F. and A. Malkawi, *A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm.* Energy and Buildings, 2014. **84**: 607-616.

[78]. Papadopoulos, S., B. Bonczak, and C.E. Kontokosta, *Pattern recognition in building energy performance over time using energy benchmarking data.* Applied Energy, 2018. **221**: 576-586.

[79]. Almeida, R.M.S.F., N.M.M. Ramos, M.L. Simoes, and V.P. de Freitas, *Energy and Water Consumption Variability in School Buildings: Review and Application of Clustering Techniques.* Journal of Performance of Constructed Facilities, 2015. **29**(6).

[80]. Salvalai, G., L.E. Malighetti, L. Luchini, and S. Girola, *Analysis of different energy conservation strategies on existing school buildings in a Pre-Alpine Region.* Energy and Buildings, 2017. **145**: 92-106.

[81]. Lara, R.A., G. Pernigotto, F. Cappelletti, P. Romagnoni, and A. Gasparella, *Energy audit of schools by means of cluster analysis.* Energy and Buildings, 2015. **95**: 160-171.

[82]. Famuyibo, A.A., A. Duffy, and P. Strachan, *Developing archetypes for domestic dwellings—An Irish case study.* Energy and Buildings, 2012. **50**: 150-157.

[83]. Ashouri, M., F. Haghighat, B.C.M. Fung, A. Lazrak, and H. Yoshino, *Development of building energy saving advisory: A data mining approach.* Energy and Buildings, 2018. **172**: 139-151.

[84]. Perez-Lombard, L., J. Ortiz, R. Gonzalez, and I.R. Maestre, *A review of benchmarking, rating and labelling concepts within the framework of building energy certification schemes.* Energy and Buildings, 2009. **41**(3): 272-278.

[85]. Sustainability, N.M.s.O.o. *About LL84*. 2019 [cited 2019 23-Apr.]; Available from: https://www1.nyc.gov/html/gbee/html/plan/ll84_about.shtml.

[86]. trust, E.s. *Introduction to HEED*. 2019 [cited 2019 21-Apr]; Available from: https://www.energysavingtrust.org.uk/scotland/businesses-organisations/data-services/heed.

[87]. Administration, E.I. *Residential Energy Consumption Survey*. 2019 July 31 2019 [cited 2019 October 7]; Available from: https://www.eia.gov/consumption/residential/index.php.

[88]. EIA. *COMMERCIAL BUILDINGS ENERGY CONSUMPTION SURVEY (CBECS)*. 2019 [cited 2019 21-Apr.]; Available from: https://www.eia.gov/consumption/commercial/data/2012/index.php?view=microdata.

[89]. Azar, E. and C.C. Menassa, *A comprehensive framework to quantify energy savings potential from improved operations of commercial building stocks.* Energy Policy, 2014. **67**: 459-472.

[90]. Bergmann, H. *Building Performance Database*. 2019 [cited 2019 21-Apr.]; Available from: https://www.energy.gov/eere/buildings/building-performance-database-bpd.

[91]. Mathew, P.A., L.N. Dunn, M.D. Sohn, A. Mercado, C. Custudio, and T. Walter, *Big-data for building energy performance: Lessons from assembling a very large national database of building energy use.* Applied Energy, 2015. **140**: 85-93.

[92]. Filippidou, F., N. Nieboer, and H. Visscher, *Energy efficiency measures implemented in the Dutch non-profit housing sector.* Energy and Buildings, 2016. **132**: 107-116.

[93]. Department for business, E.I.S.o.U. *National Energy Efficiency Data-Framework (NEED)*. 2019 [cited 2019 21-Apr]; Available from: https://data.gov.uk/dataset/473afefd-9028-48d1-a959-c865c1387a9d/national-energy-efficiency-data-framework-need.

[94]. Adan, H. and F. Fuerst, *Do energy efficiency measures really reduce household energy consumption? A difference-in-difference analysis.* Energy Efficiency, 2016. **9**(5): 1207-1219.

[95]. Yu, Z., B.C.M. Fung, F. Haghighat, H. Yoshino, and E. Morofsky, *A systematic procedure to study the influence of occupant behavior on building energy consumption.* Energy and Buildings, 2011. **43**(6): 1409-1417.

[96]. Niu, S.W., Y.Q. Jia, L.Q. Ye, R.Q. Dai, and N. Li, *Does electricity consumption improve residential living status in less developed regions? An empirical analysis using the quantile regression approach.* Energy, 2016. **95**: 550-560.

[97]. Laicāne, I., A. Blumberga, M. Rošā, and D. Blumberga, *Determinants of household electricity consumption savings: A Latvian case study.* Agronomy Research, 2014. **12**(2): 527-542.

[98]. Huang, W.H., *The determinants of household electricity consumption in Taiwan: Evidence from quantile regression.* Energy, 2015. **87**: 120-133.

[99]. Roth, J. and R. Rajagopal, *Benchmarking building energy efficiency using quantile regression.* Energy, 2018. **152**: 866-876.

[100]. Yoon, Y.R. and H.J. Moon, *Energy consumption model with energy use factors of tenants in commercial buildings using Gaussian process regression.* Energy and Buildings, 2018. **168**: 215-224.

[101]. Zhou, X., D. Yan, T.Z. Hong, and X.X. Ren, *Data analysis and stochastic modeling of lighting energy use in large office buildings in China.* Energy and Buildings, 2015. **86**: 275-287.

[102]. Pistore, L., G. Pernigotto, F. Cappelletti, A. Gasparella, and P. Romagnoni, *A stepwise approach integrating feature selection, regression techniques and cluster analysis to identify primary retrofit interventions on large stocks of buildings.* Sustainable Cities and Society, 2019. **47**(101438).

[103]. He, Y., Y. Zheng, and Q. Xu, *Forecasting energy consumption in Anhui province of China through two Box-Cox transformation quantile regression probability density methods.* Measurement, 2019. **136**: 579-593.

[104]. Deb, C. and S.E. Lee, *Determining key variables influencing energy consumption in office buildings through cluster analysis of pre-and post-retrofit building data.* Energy and Buildings, 2018. **159**: 228-245.

[105]. Esmaeilimoakher, P., T. Urmee, T. Pryor, and G. Baverstock, *Identifying the determinants of residential electricity consumption for social housing in Perth, Western Australia.* Energy and Buildings, 2016. **133**: 403-413.

[106]. Zhang, C., L.W. Cao, and A. Romagnoli, *On the feature engineering of building energy data mining.* Sustainable Cities and Society, 2018. **39**: 508-518.

[107]. Wang, Z. and R.S. Srinivasan, *A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models.* Renewable and Sustainable Energy Reviews, 2017. **75**: 796-808.

[108]. Li, X. and J. Wen, *Review of building energy modeling for control and operation.* Renewable & Sustainable Energy Reviews, 2014. **37**: 517-537.

[109]. Mirnaghi, M.S., F. Haghighat, and Buildings, *Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review.* Energy and Buildings, 2020. **229**: 110492.

[110]. Tian, Z., S. Wei, and X. Shi, *Developing data-driven models for energy-efficient heating design in office buildings.* Journal of Building Engineering, 2020. **32**: 101778.

[111]. Palmer, K. and M. Walls, *Using information to close the energy efficiency gap: a review of benchmarking and disclosure ordinances.* Energy Efficiency, 2017. **10**(3): 673-691.

[112]. Wei, N., W. Yong, S. Yan, and D. Zhongcheng, *Government management and implementation of national real-time energy monitoring system for China large-scale public building.* Energy Policy, 2009. **37**(6): 2087-2091.

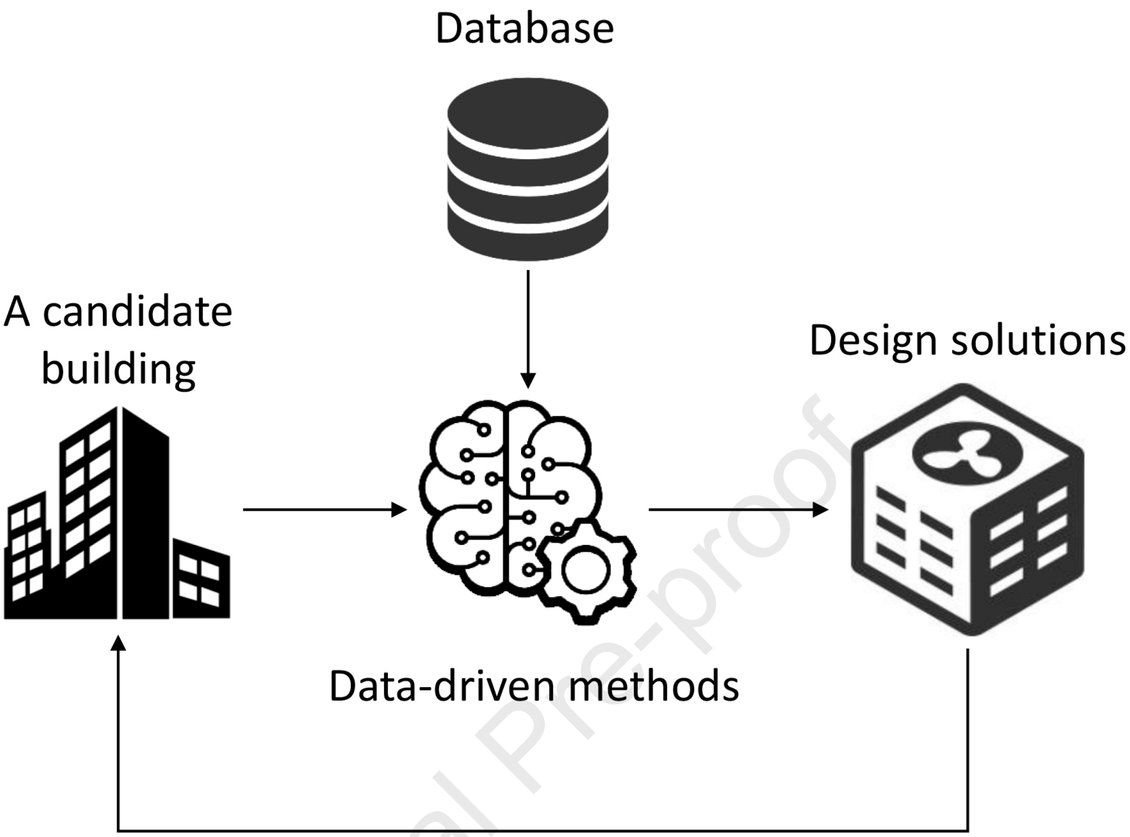
[113]. Fischer, P.M., M. Deshmukh, V. Maiwald, D. Quantius, A.M. Gomez, and A. Gerndt, *Conceptual data model: A foundation for successful concurrent engineering.* Concurrent Engineering, 2018. **26**(1): 55-76.

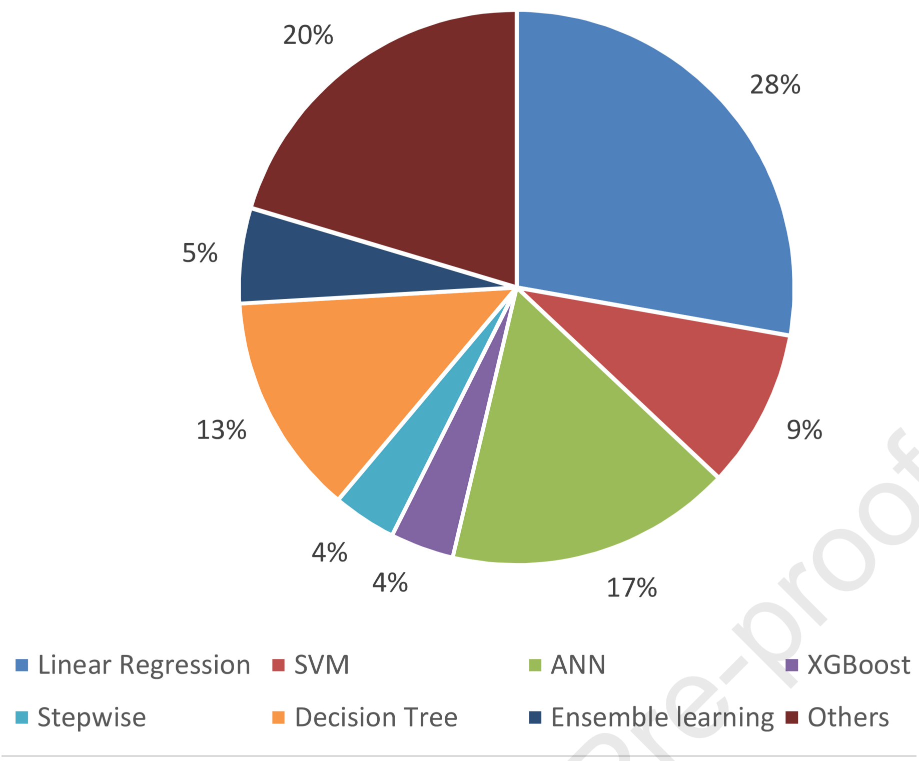
[114]. Bünning, F., P. Heer, R.S. Smith, and J. Lygeros. *Sensitivity analysis of data-driven building energy demand forecasts*. in *Journal of Physics: Conference Series*. 2019. IOP Publishing.

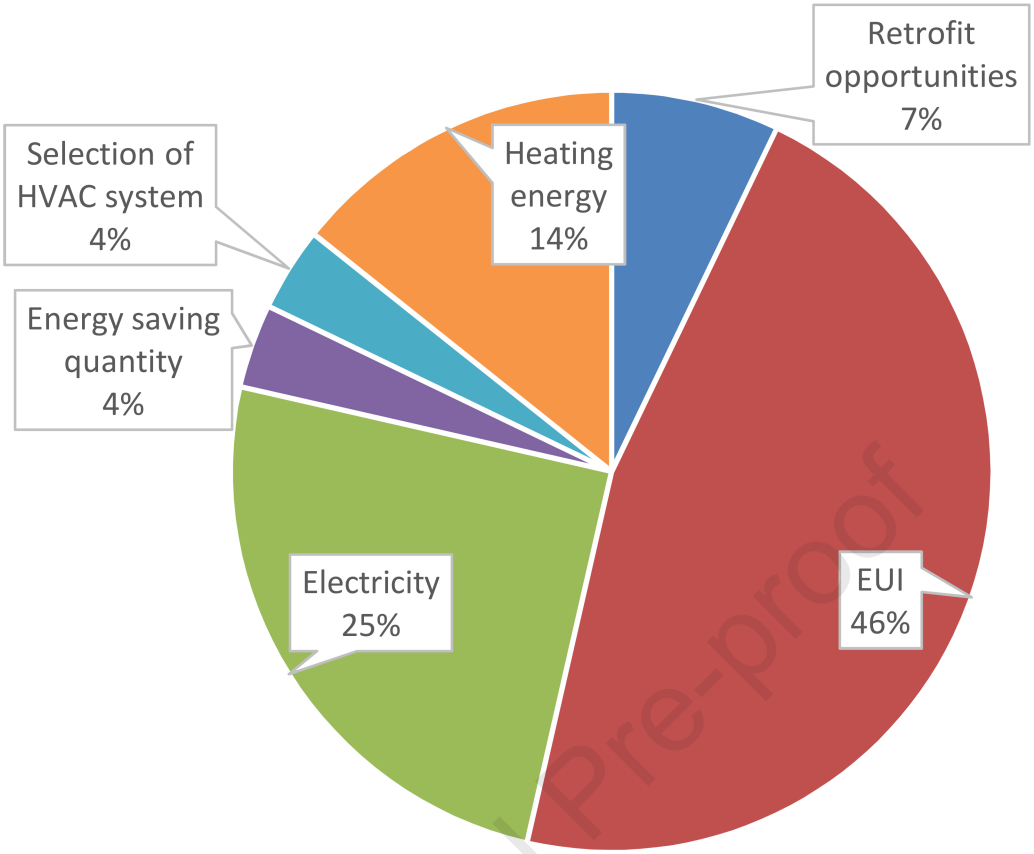
[115]. Edwards, R.E., J. New, and L.E. Parker, *Predicting future hourly residential electrical consumption: A machine learning case study.* Energy and Buildings, 2012. **49**: 591-603.

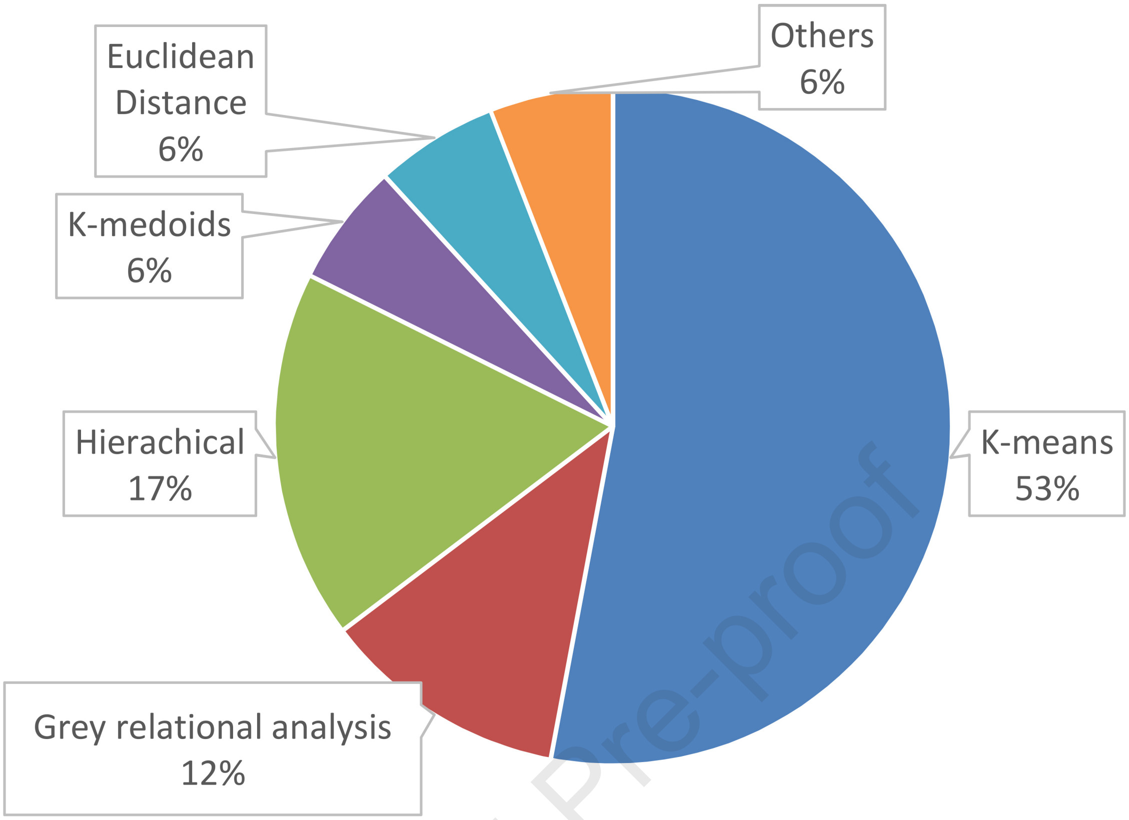
[116]. Dong, B., Z. Li, S.M. Rahman, and R. Vega, *A hybrid model approach for forecasting future residential electricity consumption.* Energy and Buildings, 2016. **117**: 341-351.

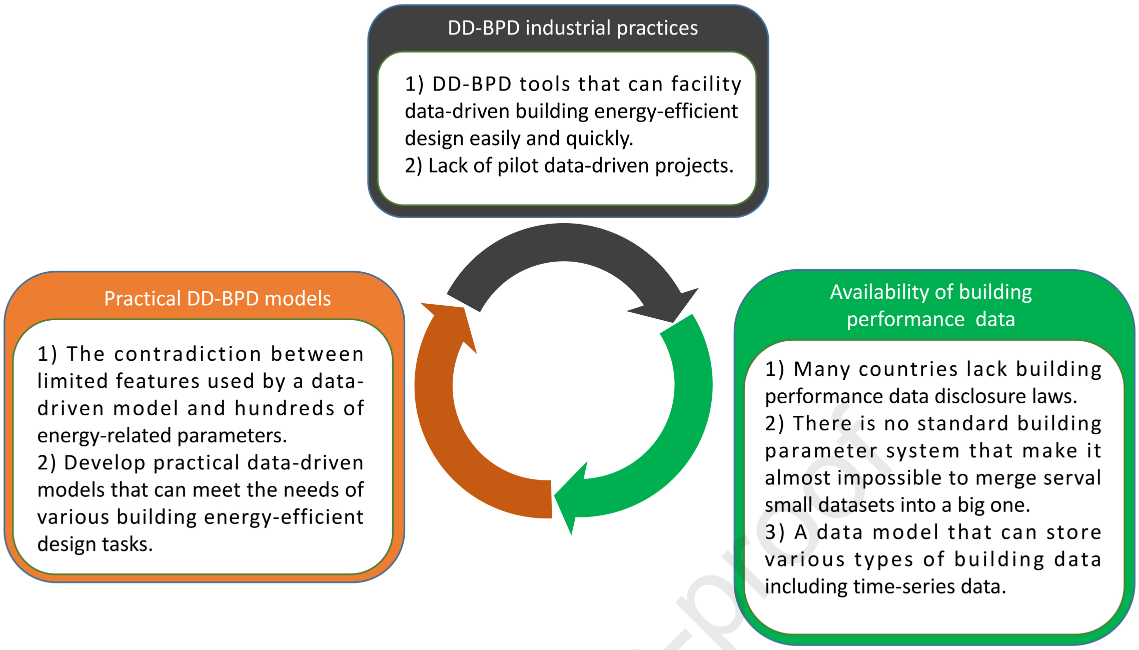
[117]. EERE. *BETTER - Building Efficiency Targeting Tool for Energy Retrofits*. 2020 [cited 2020 Dec. 18]; Available from: https://better.lbl.gov/.


Database
A candidate building
Design solutions
Data-driven methods


20%
28%
5%
13%
9%
4%
4%
17%
Linear Regression
SVM
ANN
XGBoost
Stepwise
Decision Tree
Ensemble learning
Others


Retrofit opportunities 7%
Heating energy 14%
Selection of HVAC system 4%
Energy saving quantity 4%
Electricity 25%
EUI 46%


Others
6%
Euclidean
Distance
6%
K-medoids
6%
Hierachical
17%
K-means
53%
Grey relational analysis
12%



DD-BPD industrial practices

1) DD-BPD tools that can facility data-driven building energy-efficient design easily and quickly.
2) Lack of pilot data-driven projects.

Availability of building performance data

1) Many countries lack building performance data disclosure laws.
2) There is no standard building parameter system that make it almost impossible to merge serval small datasets into a big one.
3) A data model that can store various types of building data including time-series data.

Practical DD-BPD models

1) The contradiction between limited features used by a data-driven model and hundreds of energy-related parameters.
2) Develop practical data-driven models that can meet the needs of various building energy-efficient design tasks.

Highlights:

- This study focused on data-driven studies that fulfilled exclusively on on-site building data.
- Data-driven studies were classified by statistics, regression, classification, clustering.
- Major on-site building performance databases were summarized.
- Challenges were discussed on data-driven models, availability of data, industrial practices.
- Exploring determinant features and prediction of energy consumption are usual applications.

None

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: