Energy & Buildings 242 (2021) 110885

Contents lists available at ScienceDirect

Energy & Buildings

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

Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature

Mahmoud M. Abdelrahman, Sicheng Zhan, Clayton Miller, Adrian Chong *

Department of Building, School of Design and Environment, National University of Singapore, 4 Architecture Drive, Singapore 117566, Singapore

ARTICLE INFO

Article history:
Received 16 December 2020
Revised 19 February 2021
Accepted 2 March 2021
Available online 13 March 2021

Keywords:
Reference mining
Natural language processing
Data science
Built environment
Building energy efficiency
Word embeddings

ABSTRACT

The ever-changing data science landscape is fueling innovation in the built environment context by providing new and more effective means of converting large raw data sets into value for professionals in the design, construction and operations of buildings. The literature developed due to this convergence has rapidly increased in recent years, making it difficult for traditional review approaches to cover all related papers. Therefore, this paper applies a natural language processing (NLP) method to provide an exhaustive and quantitative review.Approximately 30,000 scientific publications were retrieved from the Elsevier API to extract the relationship between data sources, data science techniques, and building energy efficiency applications across the life cycle of buildings. The text-mining and NLP analysis reveals that data sciences techniques are applied more for operation phase applications such as fault detection and diagnosis (FDD), while being under-explored in design and commissioning phases. In addition, it is pointed out that more data science techniques that are to be investigated for various applications. For example, generative adversarial networks (GANs) has potential in facilitating parametric design; transfer learning is a promising path to promoting the application of optimal building operation;

1. Introduction

With advances in information technology, buildings today are collecting ever-larger amount of real-time data from various heterogeneous sources [14,83]. The vast amount of data also have led to increased data awareness and data science applications [99]. These innovations have led to an explosion of research in this field, resulting in thousands of publications in this area (e.g., Fig. 3). It is now the case that researchers are in a position in which there are significantly more research publications available than what can be processed and digested by human [40]. Numerous literature reviews are also being produced to aggregate research literature; however, this is also not a trivial process due to the volume of research in this area.

1.1. Using data science to quantify the impact of data science on buildings

In order to address this challenge, the concept of using text-mining methods to analyze scientific literature has gained traction. The academic knowledge is exponentially expanding; thousands of research articles are authored by domain experts every day [77]. Thanks to the recent advancements in Natural Language Processing (NLP), it has been viable to extract knowledge from a large corpus of such structured text. Typically, information from the literature is extracted via traditional narrative literature review/survey of a finite number of articles (hundreds) [94], besides other methods such as questionnaire surveys and expert interviews. However, there are some challenges in applying these methods on a large scale [8]. Specifically, conducting a manual literature review on a large number of papers requires huge effort. It is even more challenging if the literature review is cross-disciplinary such as extracting relations between various data.

1.2. Similar studies

Several conventional literature reviews have been completed in recent years to capture the innovation occurring due to the convergence of data science and building energy performance research during different lifecycle phases [49,109,60,21,48,132,69]. Wang and Srinivasan explored the use of single versus ensemble-based models for building energy prediction [153]. Roth et. al explored the use of various data-driven techniques in the context of benchmarking building [134]. Colm et. al. [51] investigated Machine learning methods for maximizing measurement and verification

* Corresponding author.
E-mail address: adrian.chong@nus.edu.sg (A. Chong).

https://doi.org/10.1016/j.enbuild.2021.110885
0378-7788/

(M&V) accuracy with an application on a real building. This application concluded sufficient accuracy despite some limitations such as poor data quality and insufficient metering. In the operation phase, data science methods were found promising to tackle the challenges in building system control [96]. Fault detection and diagnosis (FDD) is another important application of improving building energy performance, where data science methods are commonly used [168]. Furthermore, energy audit and commissioning of buildings using data analytics has been investigated by Rohloff et. al. to minimize the performance testing hours and maximize the value of the test results [132]. Beyond single buildings, data-driven methods are also useful for demand response and smart grid applications [50,112,47]. On the urban scale, energy efficiency applications also have grasped the interest of researchers. For example, many researchers investigated district heating and cooling systems [88,128] and Urban Building Energy Modelling (UBEM) [11,65,127]. Most of these reviews have indicated the potentials of using big data (such as sensing data from IoT and urban building energy modelling data) and machine learning.

These reviews cover the specific application of data science to various facets of the building energy paradigm however, they are constrained by the ability of human-driven analysis to make qualitative relationships between a relatively small number of papers. Each review is only able to analyze between 100-120 publications. An emerging field of analysis of scientific literature is seeking to extract insights from quantities of publications in the tens of thousands instead of only the hundreds. These studies have been completed in fields, such as the humanities [131], bio-medicine [143], and frameworks have been built for more general text mining purposes [147].In the building domain, some studies adopted bibliometric reviews of the global trends in different building-related issues such as BIM [137,91], Green Buildings [167], Life-cycle assessment [52], Building maintenance [130] among other aspects.

Different tools and approaches of text-mining have been used in literature to conduct literature reviews. [23,41,138,158] used VOSviewer [149] to create bibliometric networks and density map between articles in different fields. Other researchers used CiteNetExplorer [150] to track the citation relations across articles in scientific research [38,142] among others. Other tools such as CiteSpaceII, BibExcel, SciMAT,Sci^2 Tool have been extensively reviewed by [110]. However, all these tools come with a graphical user interface (GUI) which limits the user ability to extend it beyond its embedded algorithms. Additionally, these tools only use the articles' metadata (title, abstract, authors, keywords, references, date ..etc) not the article body full text. Therefore, many researchers used open-sourced libraries such as the **N**atural **L**anguage **T**ool**K**it NLTK [95], Glove [120], Python/scikit-learn [119], word2Vec [102,100,103] to develop a model that performs a specific task.

This paper aims to address the challenges and deficiencies of typical literature reviews and capture the full extent of the relationships between data science and building energy performance. Given these circumstances, the current study adopts text mining survey and natural language processing to extract different segments of building data usability and their relevant users. This effort is the first text-mining and NLP review of its kind in the building energy performance research domain.

The paper is organized as follows. Section 2 provides an outline of the data extraction from the publisher's API, the text mining process, and the quantification of relationships between the different concepts being compared. Section 3 illustrates the overview graphics extracted from the mining process that show the diversity of data science techniques applied to buildings. Section 4 provides a high-level analysis of the trends and gaps found in the literature with respect to data science for building performance analysis. Finally, Section 5 concludes the analysis and provides insight on reproducibility and further analysis using the data set.

2. Methodology

The current study follows three types of research designs, namely, text-mining survey, natural language processing (NLP) semantic analysis, and relation graph extraction. Each one of these three designs is distributed across a five-phase method of data collection, preprocessing, and processing. These five phases, summarized in Fig. 1, are: 1) Identifying the querying keywords of each category, 2) Extracting the relevant articles with their corresponding metadata using ELSEVIER api, 3) Pre-processing the data, 4) Applying the NLP algorithms, 5) Extracting the relationships and creating the relation graph network.

2.1. Keyword identification

Four distinct categories of keywords are identified that were used for querying the articles for this analysis. Specifically, the categories are `data`, `data science`, `energy efficiency`, and `phase`. These keywords are meant to constitute a relational network to extract the use of different data-points, techniques, algorithms, and applications during the building life cycle phase as illustrated in Fig. 2. The analysis of the relationships between these concepts forms the foundation to understand what techniques and data sources are popular in the building energy performance domain and which ones are underutilized.

2.1.1. Definitions

To set the context, the following are more detailed definitions of each of these concept categories:

Def.1 **Data**: (`data`) refers to different types of data used in buildings, including design specifications' data such as thermal comfort and indoor environmental quality; metered data such as temperature, humidity, energy consumption, and chilled water flow rates; and spatial data such as building geometry, spaces and zones.

Def.2 **Data Science**: (`data_science`) refers to models and algorithms used by different users during different building life-cycle phases. For example, the use of energy simulation, data mining and visualization, machine learning models would be included in this category.

Def.3 **Energy Efficiency**: (`energy_efficiency`) refers to the various categories of potential application of data science in the building energy analysis domain. These techniques range from conventional approaches such as automated fault detection and diagnostics (AFDD) to more contemporary innovations such as urban-scale district energy modelling.

Def.4 **Phase**: (`phase`) refers to the building life-cycle phase/stage. We defined 5 phases found in the literature: design phase, commissioning, operation and maintenance, and retrofit.

Each of these categories consists of manually defined initial keywords. We obtained these keywords by conducting a preliminary survey over the existing literature.

2.1.2. Keywords acquisition

A preliminary literature survey was conducted to obtain the keywords of each of the categories. For example, keywords that are related to the `data` category include: `meter readings`, `energy consumption`, `load profile`, `thermal mass`, `electricity pricing`, `schedule`, `thermal comfort`, `etc`.

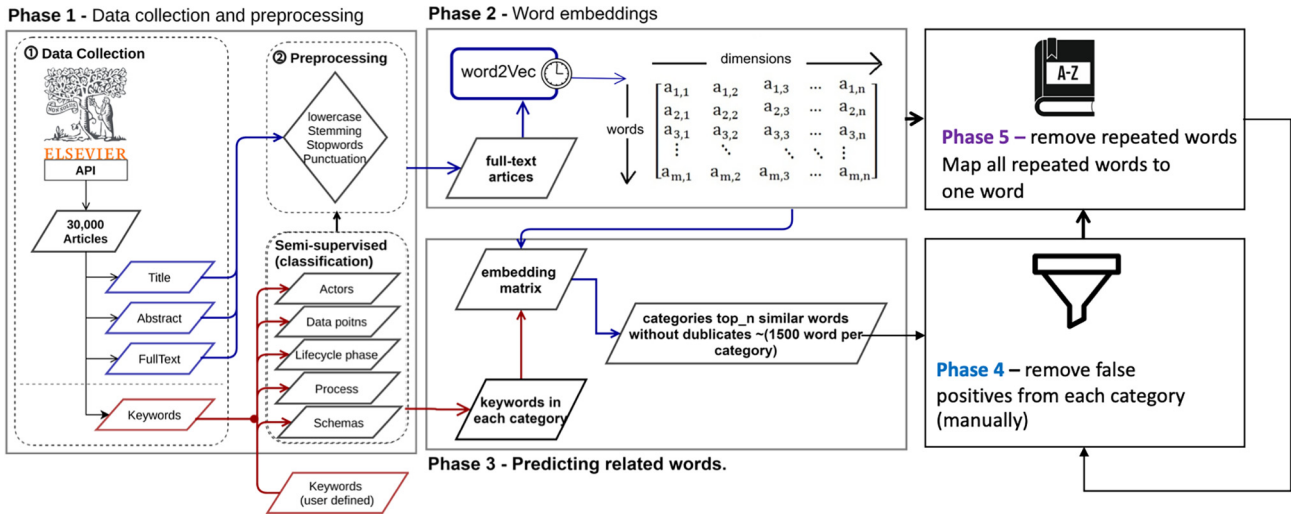

Fig. 1. The flowchart shows the methodology used in this research 1) Identifying the querying keywords of each category, 2) Extracting the relevant articles with their corresponding metadata using ELSEVIER api, 3) Pre-processing the data, 4) Applying the NLP algorithms, 5) Extracting the relationships and creating the relation graph network.

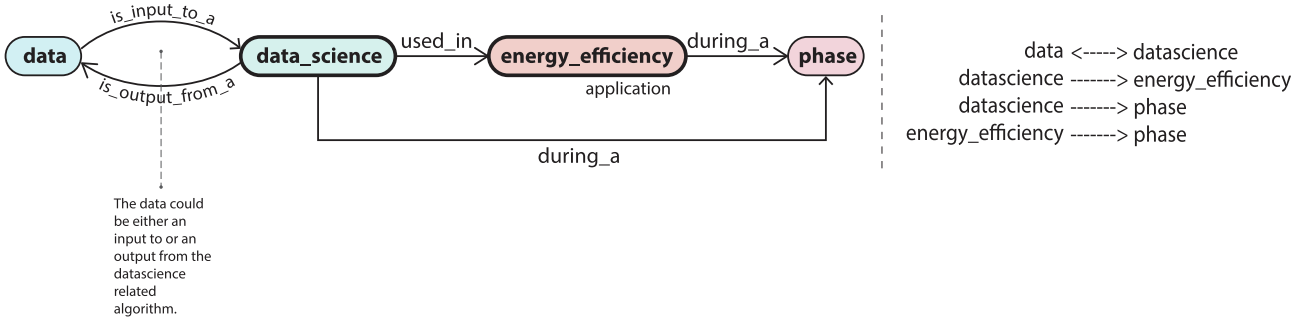

Fig. 2. Overview of the categories of concepts analysed in this text-mining analysis and their relationships with each other.

Each of the keywords has been paired with words to restrict the search query to the built environment. These restrictive words are ‘‘`building"`, `built environment`, and `buildings`. For example, using the word “Haystack” which indicates a building schema results in an irrelevant output such as “. . .Finding a needle in a haystack”.

2.2. Text mining survey

ELSEVIER is one of the largest scientific publishing and aggregation organizations. They first introduced an API for the public for text-mining research in 2014 [151]. By opening their database, researchers can extract full texts and metadata from more than 11 million research items using ELSEVIER API. In this research, the same approach was used to obtain full versions of about 30,000 papers by querying the keywords extracted from the previous step. The articles come alongside their corresponding metadata, such as date of publishing, authors and affiliation, journal (container), title, abstract, keywords, amongst others. In this analysis, we use the publications extracted from this API as a representative sample from the building energy research domain as these journals are the highest cited in energy and buildings.

2.2.1. Article filtering

The initial query process has resulted in 45,000 articles from more than 1000 journals. However, many of these articles are duplicated. Thus, after removing the duplicates, the accumulative number of articles reached around 30,000 articles. All of these articles come with a rich amount of metadata including publishing date, authors and their affiliations, keywords, number of citations, besides abstract and title. Fig. 3 illustrates the top number of papers per journal and the number of published papers per year. From this result, it is observed that the majority of the articles come from building, energy, and sensor-related journals. At this stage, the extracted articles were ready for preprocessing and preparation.

2.3. Text prepossessing

The preprocessing phase aims at preparing the extracted full text for the data mining process. The data preparation includes removing unwanted words from the articles, making the words consistent, and tokenization of words or group of words. Firstly, there are two types of unwanted words can be identified: 1) titles, subtitles, and annotations such as `introduction`, `literature review`, `figure`, `table`. These words are repeated in every article and may cause bias in the subsequent processes. 2) stop words; the term “stop-words” refers to words that are frequently repeated yet not meaningful for the context such as `the`, `a`, `in`, `of`. If these stop-words are included, they will cause bias in the NLP models. Many tools are available for removing stop words such as the Natural Language Toolkit (NLTK) [95].

Secondly, Since the NLP models are case-sensitive, they need to be consistent. For example, lowercasing letters throughout the cor-

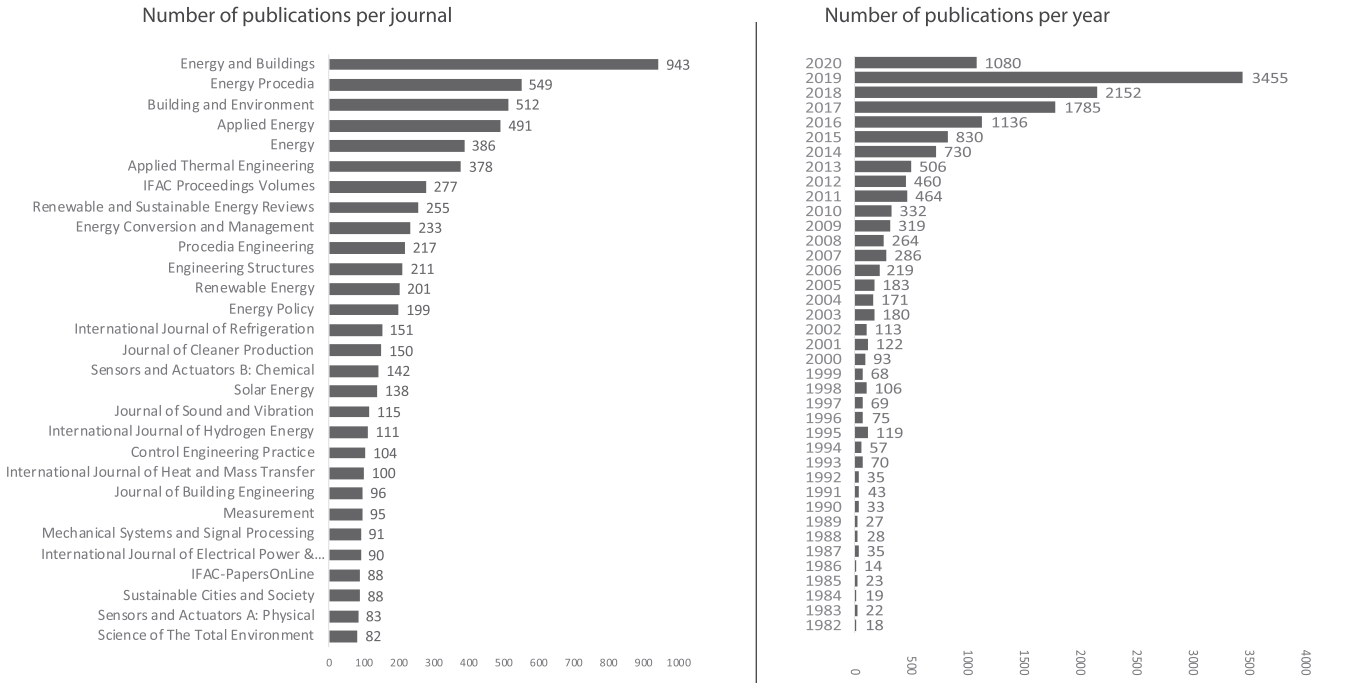

Fig. 3. Number of collected papers per journal and per year.

pus. Also, converting regular plural nouns into singular ones by removing "s". There are many other sets of tools for making the text consistent called text stemming and lemmatization. However, the current study will only use two of these methods, which have resulted in a better accuracy (Fig. 4). Firstly, the common root of different words was used. Secondly, compound words were converted into a single word with "_" separating them. The compound words, however, were extracted from each article's keyword section. We included only the keywords section as it is known to contain the main important acronyms and definitions. After making the full text consistent, it is now ready to be prepared for the NLP text mining process.

2.4. NLP text mining using Word2Vec

Word2Vec is a word embeddings algorithm that is used to extract the semantic similarities between different words in a text [100,101]. This similarity is indicated by assigning each word in the text to a multi-dimensional vector. Then the Euclidean distance between each word can be calculated using the cosine of the angle between these vectors: $sim(A,B) = cos(\theta) = \frac{A.B}{||A||||B||}$. The closer the words to each other, the more similar they are likely to be. The Word2Vec training process aims to predict a word (known as the central word) from the context within which this word falls (context words)[100,54]. This central word is initially masked, then the algorithm tries to predict it from a window of n words before and after (in our model, we used a window of 20 words). This window was decided based on hyperparameters fine tuning. After reaching a reasonable accuracy in predicting each word in the corpus, the training stops. Then, the hidden layer is extracted as an embedding vector. Deciding the dimension of the hidden layer (embedding vector) is a best-practice-driven process and is subject to hyperparameter fine-tuning. In our case, we assigned a vector of 300 dimensions to each word which was proven to give the highest accuracy for our model. The architecture of the word2vec model is illustrated in Fig. 5.

2.5. Extracting the relationship between categories

In word2Vec, two words are similar if they frequently appear in similar contexts. For example, if the word `architect` and the word `early_design_phase` are frequently appearing among similar words, then these two words will be assigned to relatively near vectors. Concurrently, two, or more, words can be added or subtracted from each other by adding/subtracting their corresponding vectors. For example, adding `artist` + `engineer` results in a vector that is closest to the word `architect`. This metric is used to extract the relationship between words from different categories in two main steps.

Firstly, there can be many words that refer to the same term. In this case, the words are mapped to that term. For example, the word `early_design_phase` is found to have many other synonyms that are similar to it such as: ''`early_design_stage`'', `conceptual_design`, `early_design_development`, `early_design`, `concept_design`, `early_design_stages`, and others. These words can differ slightly in using "_" rather than "–" or using the word `stage` rather than `phase`. Another example is the use of acronyms that refer to the same term such as `gbrs` and `green_building_rating_system` were easily captured using the word2vec similarity metric. Thus, the similarity metric is used to extract these similar words which makes it easier to implement the following step i.e. extracting the relationships.

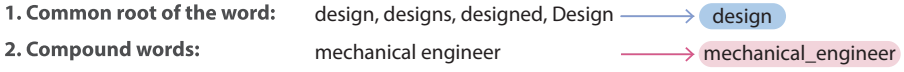

Fig. 4. Each word can have one or more similar synonyms which are mapped to the original word. The figure shows two types of text stemming and lemmatization.

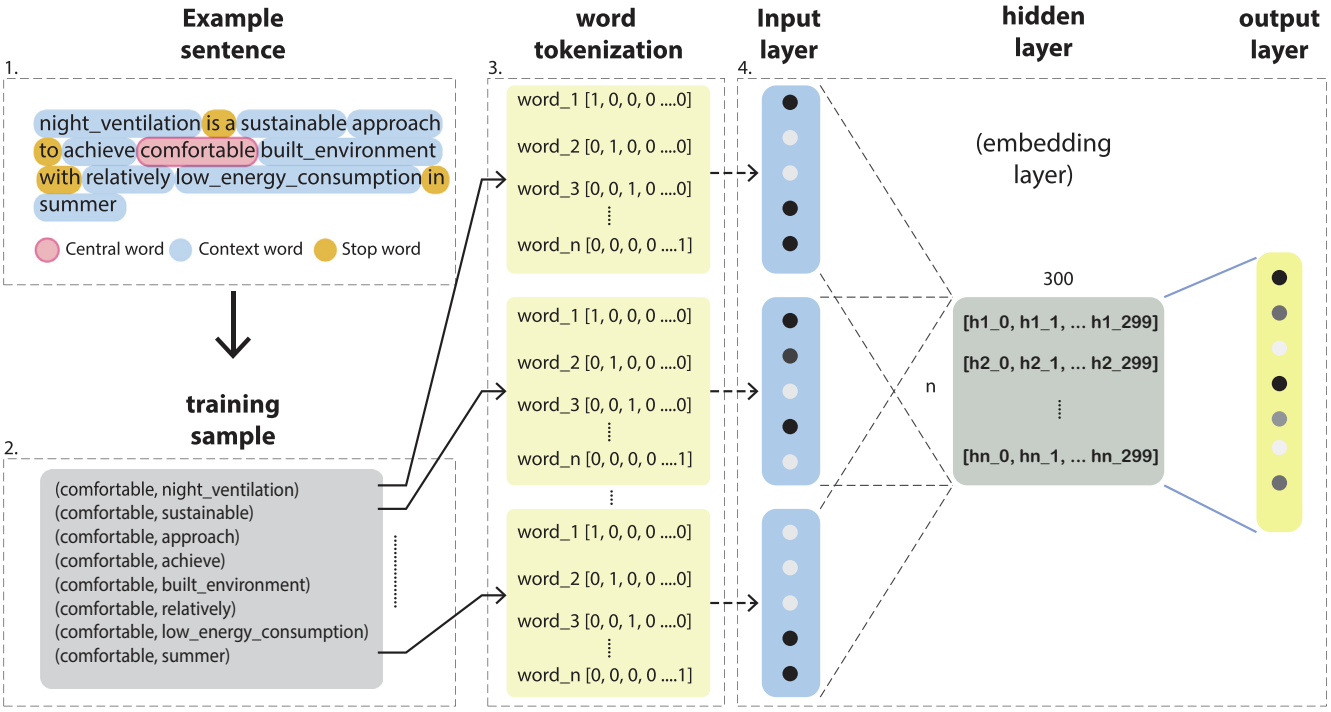

Fig. 5. Word2vec architecture. This architecture is used to extract the embedding matrix (the hidden layer) which is the vector representation of each word in the latent space.

Extracting the relationships comes after creating a list of all the words and their synonyms from each of the four categories. We used a method called **n-gram** to extract these relationships [87]. The n-gram model searches for the similarity between two words by sampling n samples from contiguous sequence of their synonyms. For example, the objective is to extract the similarity between the two words W_a and W_b such that W_a is `"energy_consumption"` which has other synonyms such as W_{a_1} (`‘‘building_energy_use"`) and W_{a_2} (`‘‘energy_consumption_data"`); and the word W_b `‘‘energy_benchmarking"`. The 1-gram model will look for the the similarity by taking one word at time, while the 2-gram model will look for the similarity by taking pairs of words at time. At the end, the total similarity between the two words is given by the average of all the similarities:

$$\bar{S}(W_a, W_b) = \frac{\sum_{n=1}^{max(len(W_a),len(W_b))} \mathrm{n-gram}(W_a, W_b)}{max(len(W_a), len(W_b))}$$

Where $\bar{S}(W_a, W_b)$ is the average similarity between two lists of words W_a and W_b and their synonyms $W_a = [w_{a_1} \dots w_{a_n}]$, $W_b = [w_{b_1} \dots w_{b_n}]$. n is the is defined by the maximum number of synonyms of the two words. If $n = 1$, then it is called unigram; if $n = 2$, it is called digram; if $n > 2$ it is referred to as n-gram. The n-gram is obtained by the cosine similarity between the two word lists W_a and W_b as follows:

$$\mathrm{n-gram}(W_a, W_b) = Sim\left(\sum_{i=1}^{n} W_{a_i}, \sum_{j=1}^{n} W_{b_j}\right)$$

An n-gram similarity is a number within the range $[-1.0, 1.0]$. If the two words are identical (e.g. w_a is the same as w_b), their similarity = 1.0, if they are perfectly semantically opposite, their similarity will be -1.0 theoretically. However, 0.0 means that there is no semantic similarity between the two words. These numbers are converted into triplets $\left\{\mathbf{W_a}, \mathbf{W_b}, \bar{\mathbf{S}}(\mathbf{W_a}, \mathbf{W_b})\right\}$ which is then converted into a directed weighted graph. The results will be explained in the following Section 3. For example, the n-gram similarity between the word `‘‘fault detection and diagnosis"` which has the synonyms : `[‘‘fault_detection_and_diagnosis", ‘‘fdd","fault_detection"]` and the word `‘‘neural_network"` which has the synonyms `[‘‘neural_networks", ‘‘deep_learning", ‘‘cnn"]` will follow 3-gram similarity which results in a value of 0.42 in this case which is relatively high.

3. Results

The methodology outlined a process of using text-mining and NLP methods to extract and process various concepts from a large corpus of research publications related to the convergence of data science and building performance. This section focuses on the detailed visualization of the aspects of drawing relationships between these categories. The key output of this work lies in the ability to quantify in relative terms the strength of relationships between the words found in the various categories being studied: the data sources, energy efficiency applications and life cycle phases of the built environment versus the data science techniques available to researchers. Fig. 6 shows the framework of this process starting with the definition of the categories and selection of words to the visualization of similarity of words and clustering of words into concepts.

3.1. Vector representation and relationships of extracted words

This first method of visualizing and drawing relationships comes in the form of a scatter plot that illustrates the various words extracted from the corpus and the directional nature and magnitude of their differences according to the vector model. Fig. 7 illustrates this situation by showing the embedding vector of words projected into a two-dimensional space. The keywords are categorized according to the four dimensions of the analysis:

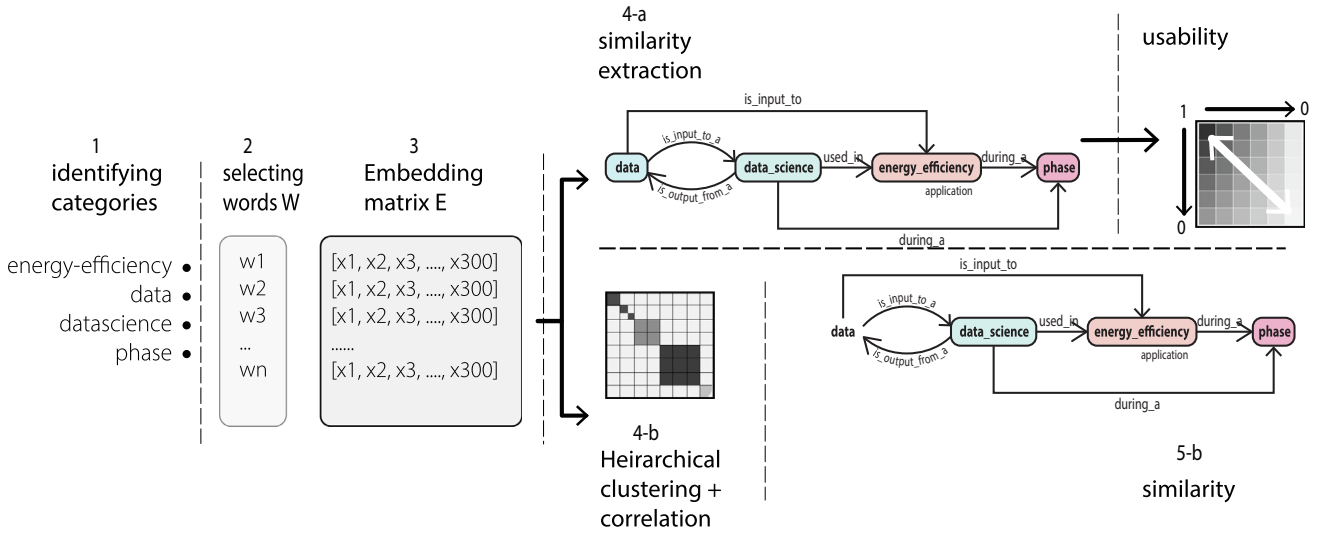

Fig. 6. Overview of the ways of showcasing the results of the text-mining and NLP process. 1) Identifying 4 categories and 2) assigning the corresponding words under each category. After that, 3) the embedding vector of each word is extracted. Then, there are two main approaches: a) is the usability relation extraction (Section 3.2) including 4-a) graph relation extraction using only the similarity metric, and 5-a) sorting the words based on its usability; and b) is the clustering of concepts (Section 3.3) including 4-b) unsupervised hierarchical clustering of words based on the embedding vector of each word (from step 3) and then 5-b) the graph relations between categories based on the clustered data.

data, data science, energy efficiency and life-cycle phase. The various words are clustered according to their relationship with each other in the vector model. The scatter plot shows how the words most closely associated with various life cycle phases of buildings can be extracted as a pattern of points from the lower left to the upper right portion of the diagram.

3.2. Usability-based similarity relation extraction

The next method to visualize relationships was the comparison of several word categories against each other to show the correlations between various concepts. These visualizations are used to illustrate the ranking of lowest to highest correlations of various data and data science concepts in both the energy efficiency applications in buildings and when those techniques are generally utilized.

3.2.1. Data sources used in building energy efficiency applications

The first comparison in this process was to show the relationship between words referring to data sources with selected energy efficiency applications. Fig. 8 shows a heat map of the various data source words extracted from the literature and their relationship strength with words extracted that related to energy efficiency applications from the life-cycle phase of the building. The horizontal axis (energy efficiency applications) is grouped according to the life cycle phases of buildings and the vertical axis (data sources) is sorted according to the average strength of relation for each data source as compared to the applications.

It can be observed that data are used mainly during the operation and maintenance and the design phases of the building lifecycle. However, data are underutilized in the commissioning phase. On the one hand, there are some energy efficiency applications that use data most frequently, such as passive design, demand-controlled ventilation, model predictive controls (MPC), fault detection and diagnosis, and retrofit analysis. On the other hand, there are other energy efficiency applications that do not use data frequently such as Measurement and verification (M&V), operation and maintenance (O&M), HVAC optimization, parametric design, and district energy systems.

Fig. 8 also shows that data sources also varies in their utilization. Some of these data are frequently used such as energy consumption data, building envelope, energy conservation measures (ECM), occupant behaviour, cost analysis, and calibrated models. Nonetheless, other data are underutilized related to HVAC design, weather and thermal comfort such as inlet/outlet temperature, condenser fan power, and mass flow rate; dew point, noise level, mean radiant temperature, and dry-bulb temperature; and clothing insulation, thermal sensation, and thermal comfort indices.

3.2.2. Data science techniques that utilize the various data sources from the built environment

The next comparison similarly uses the words related to data sources, but instead compares them to various data science techniques selected for this analysis. Fig. 9 outlines the relationship between the various data science techniques versus the data sources created in the built environment. This time both axes are sorted according to the average strength of relation for both the data science techniques (horizontal axis from right to left) and data sources (vertical axis from top to bottom).

This relationship is dominated by energy simulation, optimization, regression, and validation. However, the figure shows that there is abundant room for further data use in generative Adversarial Networks (GANs), dimensionality reduction, segmentation, and anomaly detection. There is another pattern that can be observed for applications such as factor analysis, reinforcement learning, and multi-objective optimization. These data-science applications are used frequently but with no significant relation to data sources. These relations have various observations from the data-sources perspective.

From the data source perspective, a different order from the previous heatmap can be observed. While energy consumption data is still dominating the use in data-science applications, historical data, real-time data, thermal comfort, and schedules are the highest frequently used data sources for different data-science applications. On the other side of the spectrum, HVAC design elements such as condenser fan power, inlet/outlet temperature, CAV, and fan power; as well as passive design strategies such as thermal mass are under-used in data science applications.

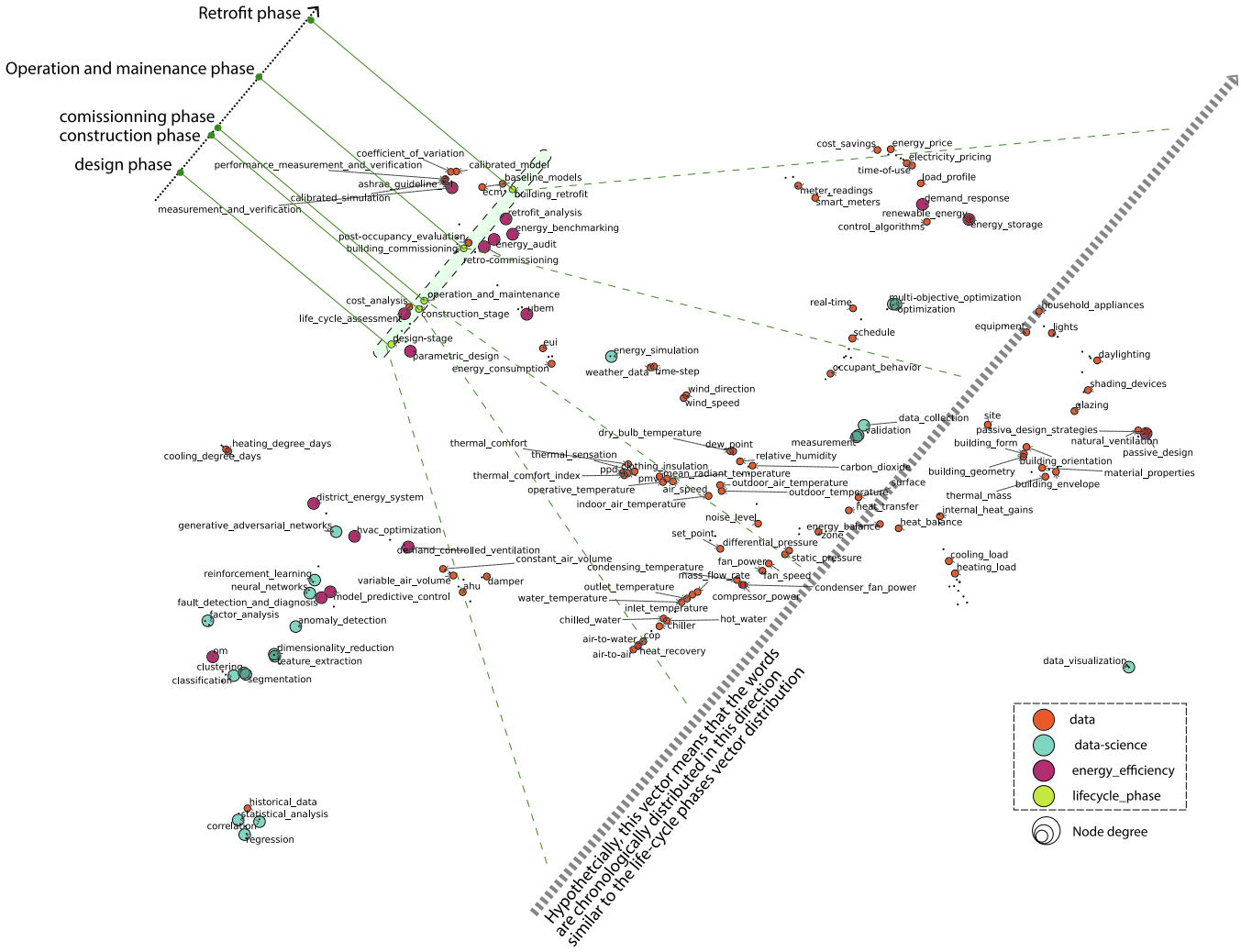

Fig. 7. The vector representation of the words from each category. These words are located based on their embedding vector. The embedding vector of each word is dimensionally reduced from 300 dimensions to 2 dimensions for the sake of visualization. The euclidean distance between words indicates the semantic similarity between these words. The degree of a specific node refers to the number of nodes connected to that specific node. The x and the y axis here represent the components of a 2D euclidean space.

3.3. Clustering of concepts

The next visualization method utilizes hierarchical clustering instead of sorting the words from strongest to weakest relation. Clustering allows for words with similarities within each category to be grouped and observed. Hierarchical Agglomerative Clustering (HAC) was used for this process using Ward's method. This algorithm is applied to the embedding vector of words in each category to group similar words together based on the euclidean distance between words in the vector space. This grouping is visualized in the form of a tree called a dendrogram (Figs. 10 and 11).

3.3.1. Hierarchical agglomerative clustering of concepts

The HAC has been applied for words in each distinct category using the Ward's method [111]. On the one hand, Fig. 10 shows the HAC of energy_efficiency category (on the left) and the HAC of the data_science category (on the right). The energy_efficiency category has been clustered into three groups. These groups are likely to be grouped based on the life-cycle phase, namely, Operation and maintenance phase, design phase, and the commissioning phase. However, the data_science category has been clustered into five different groups/subgroups. These are, Machine Learning **(ML)**, Deep Learning **(DL)**, Data pre/post-processing **(PP)**, Optimization **(OP)**, and Statistical methods **(St)**. On the other hand, Fig. 12 shows the HAC of the `data` category. In this figure, the `data` category is clustered into nine groups of keywords based on their similarities:

1. **Passive systems (PS)** which includes data that are used in passive design such as building geometry, orientation, glazing, materials, shading devices, and natural ventilation.
2. **Heat recovery ventilation data (HR)** such as air-to-air, air-to-water, and heat recovery.
3. **Building Energy Modelling data (BEM)** including heat/energy balance, zones, surfaces, and heating/cooling loads.
4. **Measurement and verification data (M&V)** including energy conservation measures (ECM), baseline model, ASHRAE guideline 14, calibrated simulation, and post-occupancy evaluation (POE).
5. **Energy consumption related data (EC)**. This includes the energy price, cost saving, time of use, energy use intensity (EUI), heat gains from appliances and equipment, load profile, smart meters, and others.
6. **HVAC** related data including **HVAC-AF** airflow data and **HVAC-T** temperature related data. HVAC airflow data include AHU, VAV, CAV, fan speed and power, etc. However, HVAC temperature related data include inlet/outlet temperature, set-point, chiller water, and hot water.

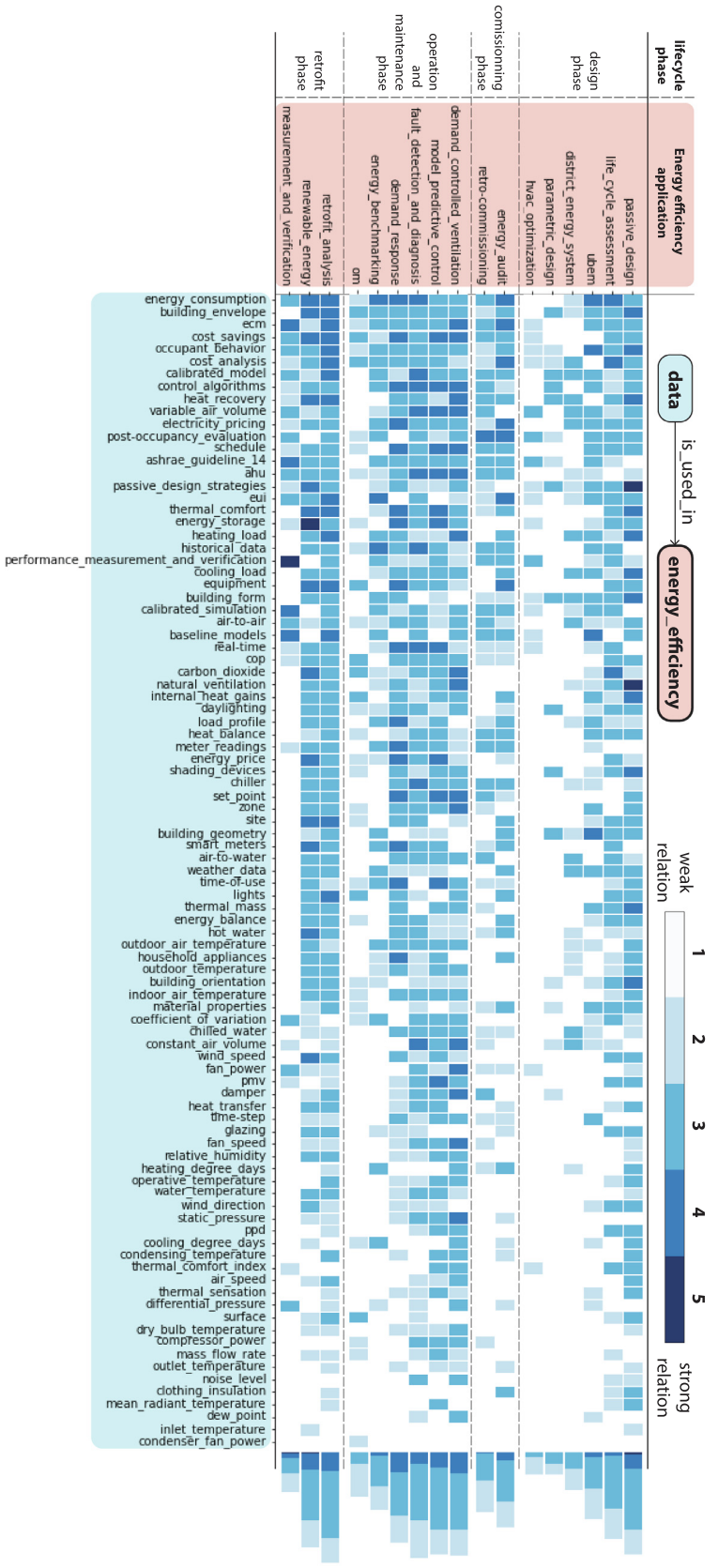

Fig. 8. The relation between data points and different energy-efficiency applications. The energy efficiency applications (shown on the Y-axis with red highlight) are chronologically grouped based on the building life-cycle phases. The data-points (shown on the X-axis blue highlight) are sorted based on their appearance frequency. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

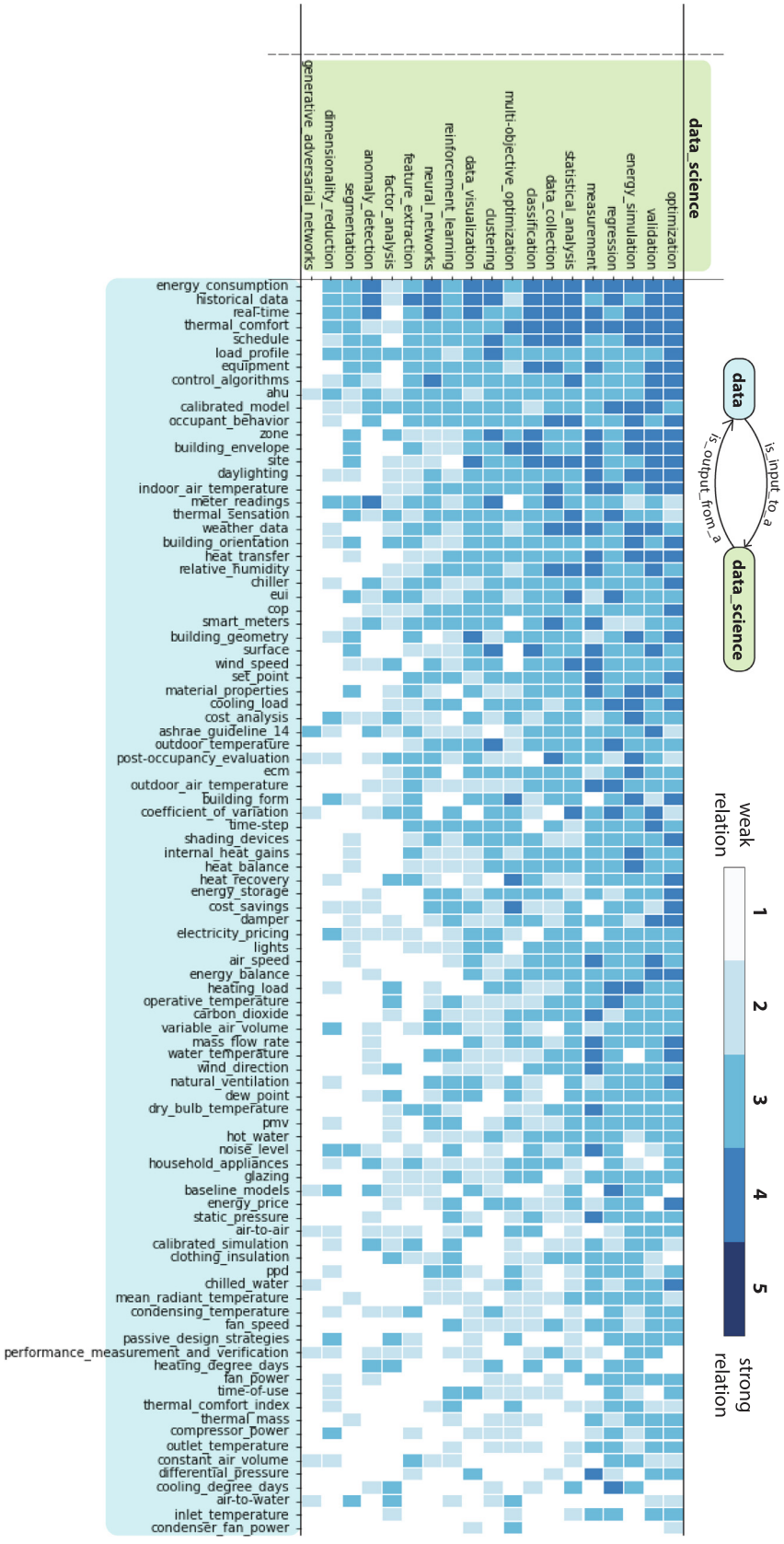

Fig. 9. The relation between data points (shown on the X-axis with blue highlight) and data-science algorithms (shown on the Y-axis with green highlight). Both of them are sorted based on the sum of the similarity per each row/column. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

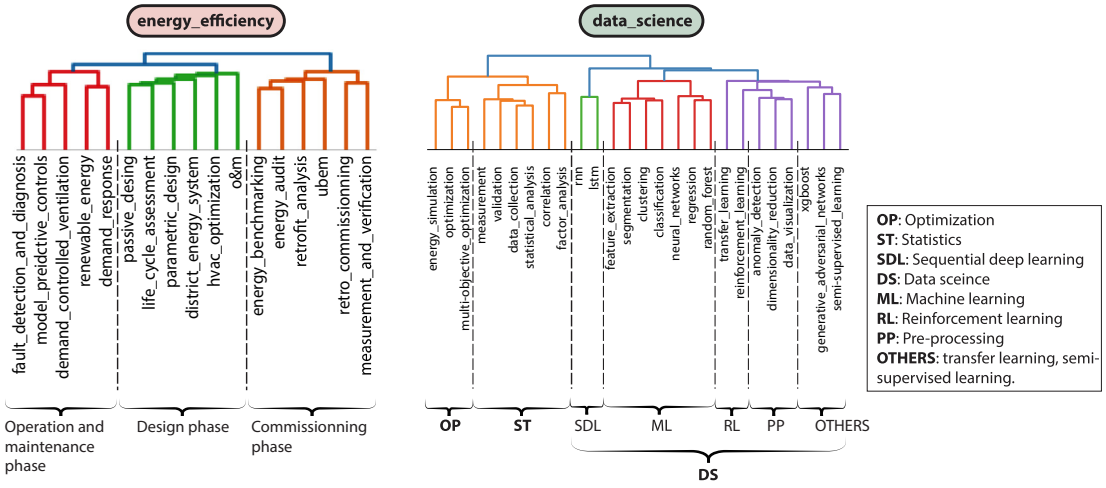

Fig. 10. The hierarchical agglomerative clustering (HAC) of the `energy_efficiency` and `data_science` categories using Ward's method.

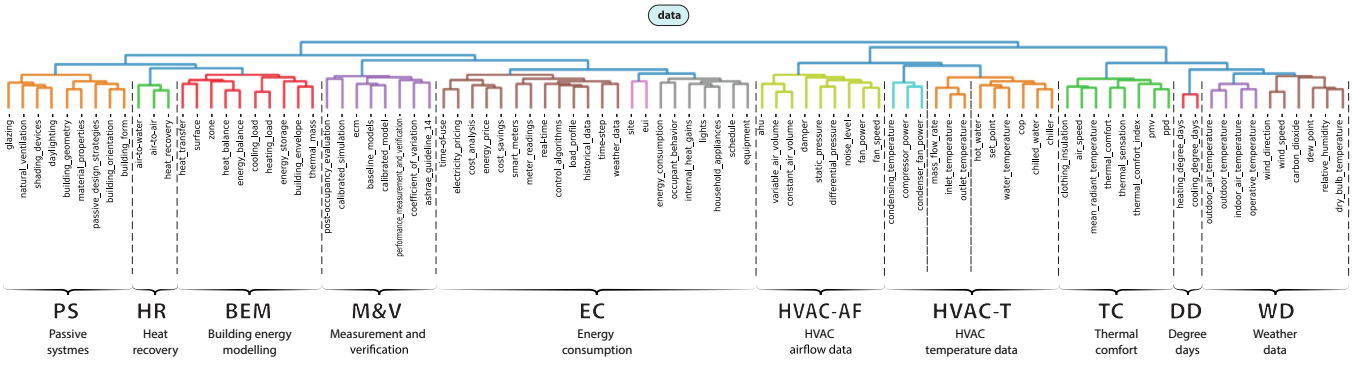

Fig. 11. The hierarchical agglomerative clustering (HAC) of the `data` keywords.

7. **Thermal comfort data (TC)** including the clothing insulation, thermal sensation, Predicted Mean Vote (PMV) and Percentage of People Dissatisfied (PPD), and Mean Radiant Temperature (MRT).
8. **Design degree days (DD)**. including heating degree days, and cooling degree days.
9. **Weather data (WD)** such as wind speed and direction, CO2, dew point, temperature and relative humidity.From Fig. 12, it can be observed that the `weather data (WD` has the highest correlation with all the other data groups. This insight could be attributed to the generic nature of weather data because it is used as input for most of the applications that use the rest of the groups except `M&V`. `M&V` on the other hand shows the lowest correlation with all the other data groups (with an average of 0.21). This conclusion might pertain to the unique nature of M&V that requires the evaluation of energy conservation measures (ECM) on building performance (in the operation phase) versus a baseline model (in the design phase). Such overlap between the operation and the design phases applications is rare. For example, `Thermal comfort` applications are used either in the design phase (i.e., Sizing, BEM, HVAC) or in the operation phase (i.e.: operation controls, Post-occupancy evaluation). Fig. 12 Also shows that `Thermal Comfort (TC)` and `Energy consumption (EC)` constitute the median of the categories with average correlations 0.67 and 0.58 respectively. These two categories falls in the area between the essential inputs/output data of energy-efficiency applications (Group1) and the fine-tuner data (Group2). Any energy efficiency application is meant to balance between these two categories, i.e. to trade-off between Thermal comfort and energy consumption. Finally, Group 2 consists of categories of data that have high potential in different energy efficiency and data science applications but are not fully matured.

3.3.2. Data use across other categories

The final visualization in this section focuses on converging the three categories of data, energy_efficiency, and data_science with the clustering techniques (Fig. 13). Data science applications such as optimization (OP), machine learning (ML), statistical methods, and sequential deep learning (SDL such as RNN and LSTM) have relatively high relation with data. Those methods require a large amount of data from various resources for more accurate results [21]. However, it is can be seen that pre and post-processing methods (PP), Reinforcement Learning and transfer learning (RL), and other emerging models such as GANs and XGBoost do not have a strong relationship with many energy efficiency applications, especially those that do not change rapidly over the building lifecycle such as passive systems and building energy modeling. These types of data are usually generated during the early design phase of the building. Fig. 15 also confirms this claim as it shows very few data-science applications utilization during the design phase of the building compared to other phases.

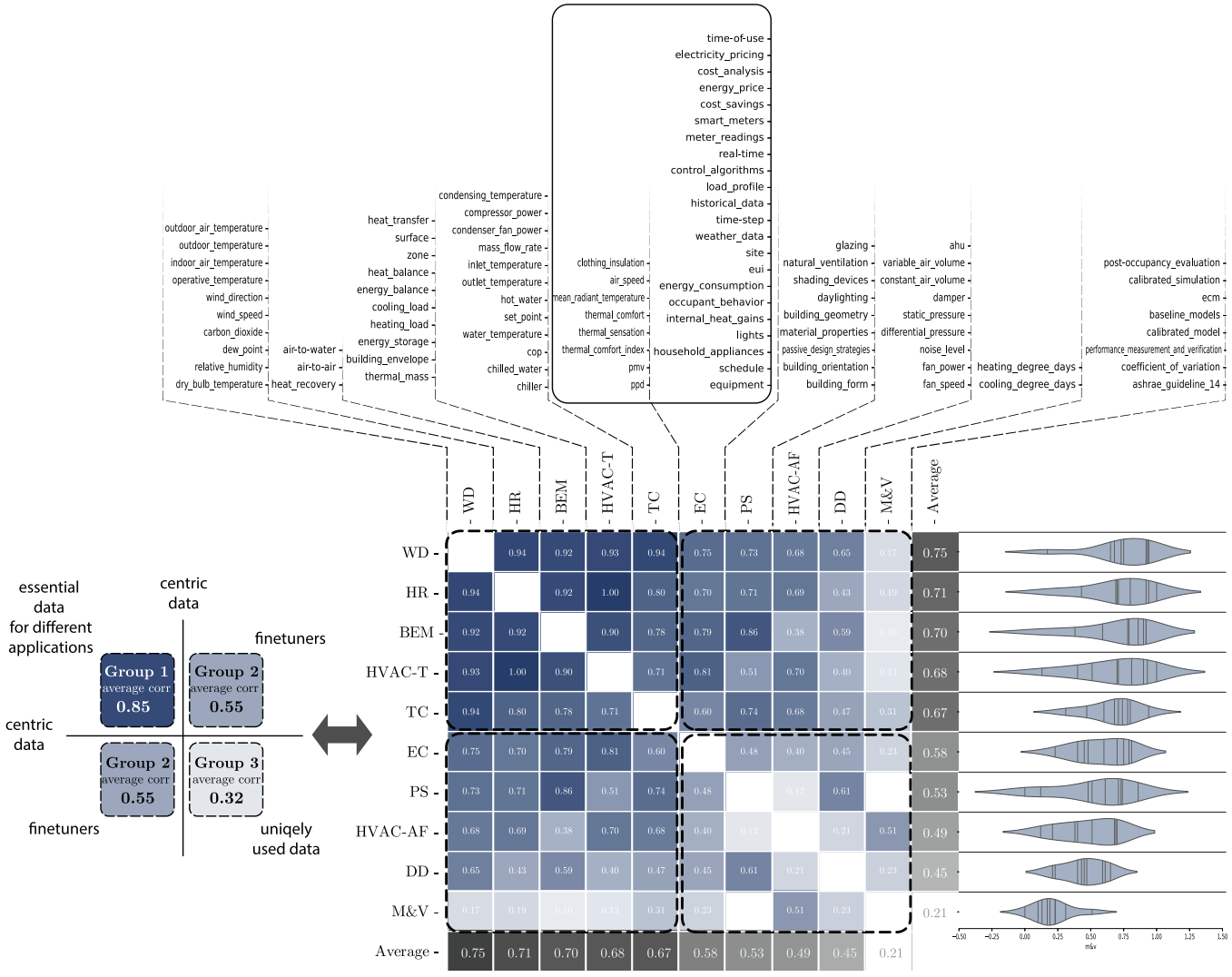

Fig. 12. The average correlation of the `data` keywords. The detailed breakdown of each individual keyword can be shown at https://doi.org/10.6084/m9.figshare.13989653.v1.

4. Discussion

The high level quantification of relationships between the data sources, data science techniques and various applications in the built environment provide the foundation for several key takeaways. This section expands upon the analysis of the results to provide high-level insights that can be used to guide future research. Each insight includes the discussion of a representative publication that illustrates the momentum or gap for that particular point.

Fig. 14 shows a comparison of the various data science techniques as compared to the energy efficiency applications for buildings as well as the life cycle phases. The following subsections outline key takeaways for the research community to consider.

4.1. What are the most common data analysis techniques?

The top five data science-related techniques found in the left side of Fig. 14 are intuitively those related to the traditional building energy domain techniques of `simulation`, `optimization`, `neural networks`, `reinforcement learning`, and `statistical analysis`.

The literature for the application of energy simulation and optimization for building energy efficiency applications is the most voluminous due to the major efforts for decades of open-source simulation projects like EnergyPlus [36] and optimization engines such as BEopt [30], GenOpt [156], and jEplus [165]. More recently, to ease the application of machine learning and statistical analysis to building simulation, there has been development on interfacing with open-source programming languages such as Python [139] and R [75]. As illustrated in Figs. 14 and 15, building simulation, also known as building performance simulation or building energy modeling, plays a vital role throughout the building's lifecycle (passive and parametric design, M&V, FDD, LCA, energy audit, and retrofit analysis). The evident developments in the discipline of building performance simulation are supported by the rapid growth of the International Building Performance Simulation Association (IBPSA) over the last two decades and research efforts under the International Energy Agency's Energy in Buildings and Communities (IEA-EBC) program. To aid applications of building performance simulation, a wide variety of tools have been developed with more than 200 software tools and programs listed on the Building Energy Software Tools directory [71]. Crawley et al. provides an overview into the capabilities of twenty major building performance simulation programs [35]. Despite its long standing history and developments, challenges remain leading to opportunities in research and development. Hong, Langevin and Sun lists

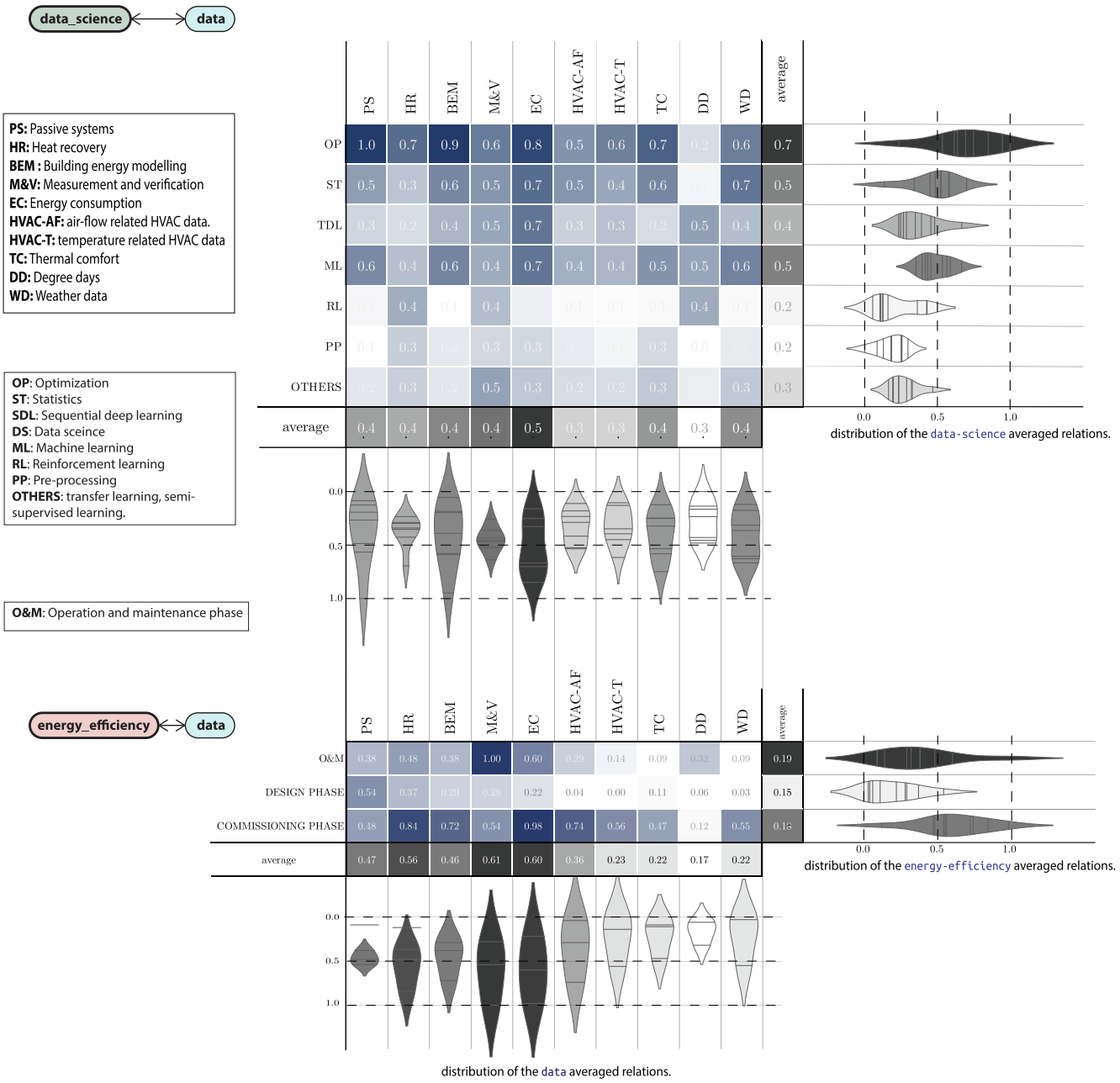

Fig. 13. The heatmap on the top shows the average relations between each pair of elements from the `data_science` and the `data` categories. The heatmap on the bottom shows the average relations between each pair of elements from the `energy_efficiency` and the `data` categories. A detailed version of the relations between every pair of keywords can be viewed at [2]..

the ten BPS challenges [67]. Table 1 lists each of these ten challenges. Additionally, we include the relevant publications and existing open-source repositories and data-sources that form the foundation in addressing these challenges.

4.2. What are the most explored building energy efficiency applications for data science?

From the perspective of applications using data-driven methods, automated `fault detection and diagnosis` followed by `retrofit analysis` , `model predictive control`, `demand response`, and `energy benchmarking` emerges as the most popular.

Fault detection and diagnosis (AFDD) is a field that has been growing rapidly since the early 1990s as a means of finding and fixing problems in building systems that result in energy waste and inefficiency. Katipamula and Brambley found the field to be maturing as early as 2006 [82]. Although matured, there have been recent developments in AFDD as a result of advancements in Artificial Intelligence techniques [168] and anomaly detection [121,122]. A challenge in the AFDD of building energy systems lies in that it is a class-imbalanced classification problem (i.e., there are few or no faulty training data). A Generative Adversarial Networks (GANs) integrated AFDD framework that generates artificial faulty samples in an adversarial way provides an innovative way to augment the training dataset, and have been shown to outperform traditional air handling unit [162] and chiller [161] AFDD methods.

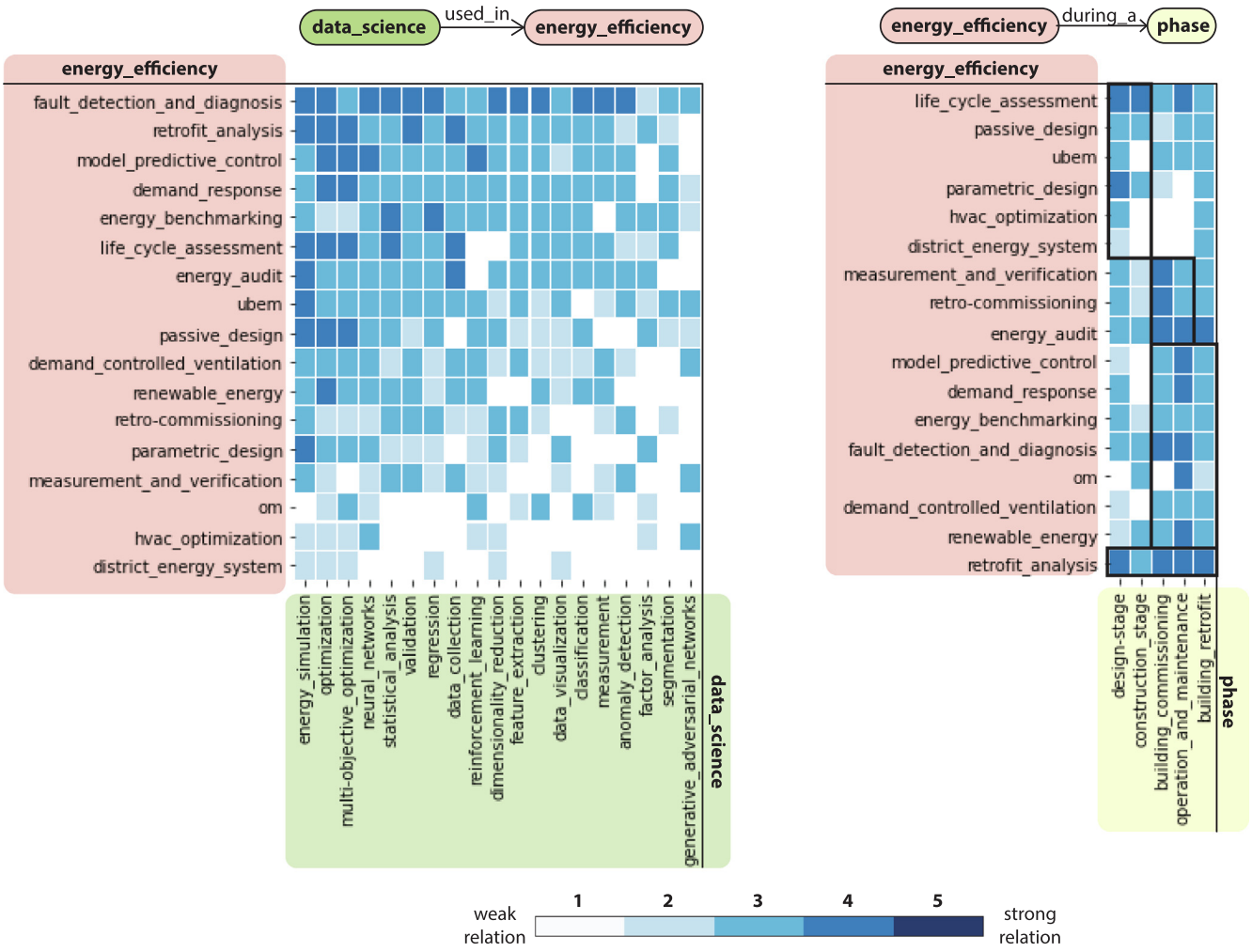

Fig. 14. The figure on the left shows the relation between data-science algorithms and energy-efficiency applications sorted based on usability. On the right, the relation between energy-efficiency applications and life-cycle phases are illustrated. The heatmaps' axes are all sorted based on strength correlation except life-cycle phase which is in a chronological order.

However, amongst the data science techniques listed in Fig. 14, GANs remain the least applied across various energy efficiency applications investigated. This is not surprising since it is a relatively new machine learning technique that might be beginning to emerge. GANs has also been applied on thermal comfort for generating balanced dataset [123,124]. Additionally, GANs recently have shown promising results in semi-real-time simulation of urban solar radiation simulation as well as urban wind simulation using Pix2Pix [32].

`Retrofit analysis` emerged as another top application across most techniques due to the influence of studies showing the large potential of upgrading the building stock [10]. As shown in Fig. 14, retrofit analysis often involves the use of physics-based simulation models, data collection and validation. The aim of retrofit analysis is to better understand the impacts of various factors on the retrofit of an existing building. However, buildings are made up of continuously changing sub-systems dynamically interacting with one another [62]. Since during a retrofit analysis training data of different scenarios is often not available, it is not surprising that retrofit analysis typically involves physics-based modeling that describes the complex dynamic interactions in buildings by a set of mathematical equations. Data collection followed by model calibration is often carried out to ensure the model's validity and thus credibility for the subsequent retrofit analysis [63]. Since building operation and characteristics may change over time, continuous model calibration and data assimilation methodologies have also been proposed to ensure the simulation model remains reasonably representative of the actual physical building system [28,155]. p`Model predictive control` (MPC) has gained traction in the last two decades through numerous case study-based implementations [5]. Data science techniques are essentially used to obtain the predictive model and to solve the receding horizon control problem. Simulation (white-box), data-driven (black-box), and hybrid (gray-box) are the three main categories of controller models [7]. Neural network models are becoming more popular due to their stronger modeling capability [6]. In addition to model identification, optimization techniques are also used to determine the optimal control actions in the coming horizon. Typical algorithms include gradient-free methods such as GA and PSO, and gradient-based methods such as NLP and MILP [43]. Over the past few years, reinforcement learning (RL) is becoming a major competitor of MPC with its advantages of lower requirements on the predictive model and better adaptability [166]. However, it also comes with other problems such as higher data requirements and lower interpretability. The comparison between these two categories of optimal control approaches will be a major research topic in the future. Another important topic for further exploration is to reduce the implementation cost and promote the application of optimal con-

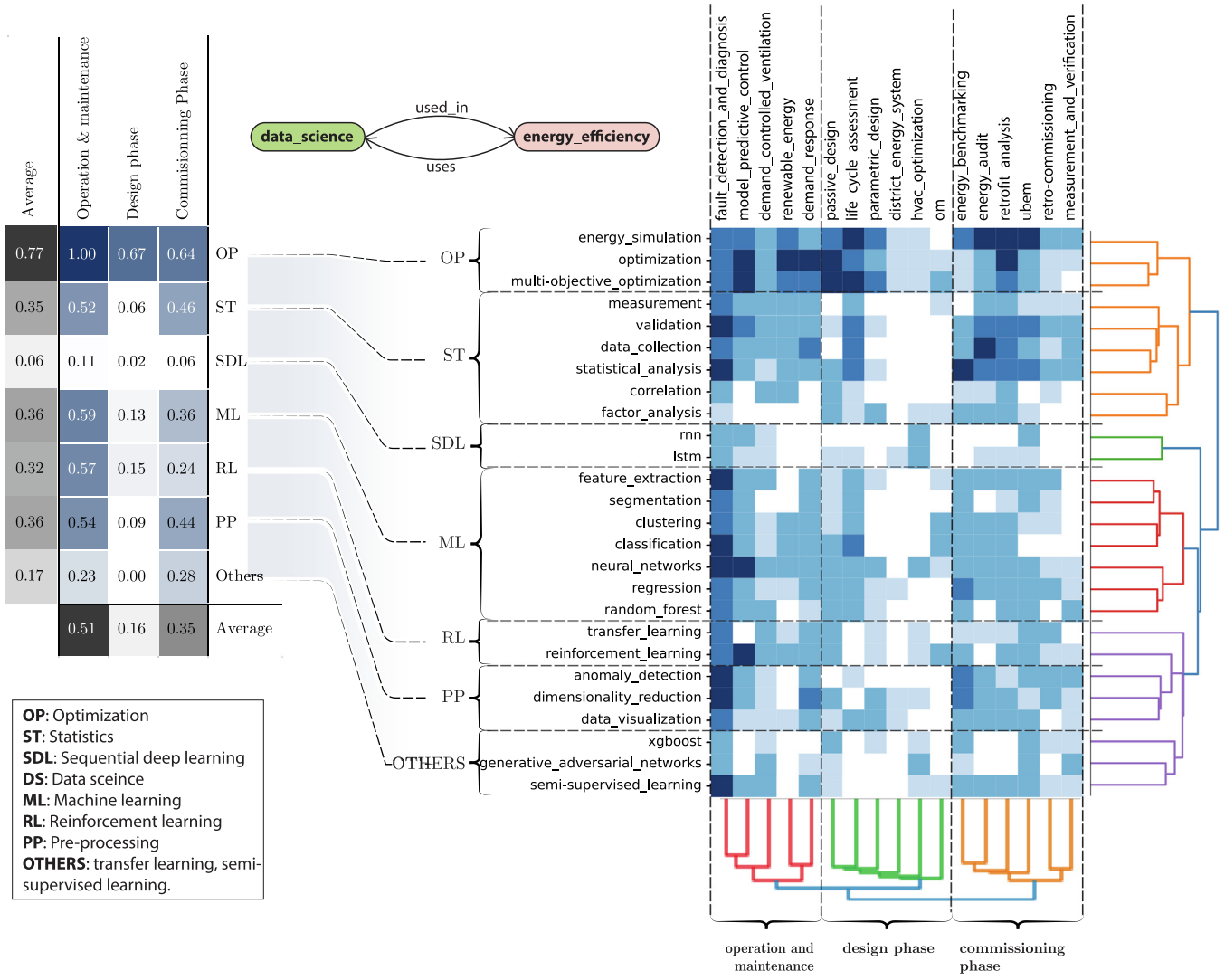

Fig. 15. The relationships between `energy_efficiency` applications and data science techniques: the HAC (on the right hand side) between energy efficiency (X-Axis) and data science (Y-Axis). The heatmap on the left hand side is a summarized version of the relation map by taking the average of each cluster.

trol. As a potential approach, transfer learning can be integrated with both MPC [25] and RL [159].

Demand response (DR), or demand-side management, is to reduce the energy cost by controlling the end-use customers' energy consumption with respect to energy prices. With the increasing penetration of renewable energy, the availability of renewable sources is another important factor to consider [129]. While reducing or shifting the electricity load, buildings still need to cater to the occupants' necessary needs such as lighting, office equipment, and environment conditioning. Thus, with no surprise, optimization is the most critical data science technique used for demand response [80]. Many other techniques are also involved in the process of designing DR programs (grid side) and helping the customers react to the programs (demand side). For example, energy simulation is a useful tool to design and evaluate the DR strategies [29]. Also, many clustering algorithms are applied to extract the typical load profiles to better understand the end users, estimate the participants' potential, and help decide the scheduling schemes [93]. Besides, at the city or grid level, the volume and variety of data generated when applying DR are both enormous. Therefore, techniques of data collection and dimension reduction are also essential in real implementation [76].

`Building energy benchmarking` is a concept which also comes up in the top five applications of data science. This field has grown based on the success of energy labeling schemes and city-wide data disclosures. Recent work in this area focuses on updating the modelling techniques [12] and even redefining the way buildings are categorized for benchmarking [118,164]. The large increase in open data sets available has created opportunities to target specific strategies to cities based on their specific needs [163] and using a combination physics-based and data-driven methods [135]. Data-driven improvements have been suggested related to generalizability [104] and interpretability [105].

4.3. What are the emerging application areas in which there are gaps?

On the other axis, it can be seen which energy efficiency techniques have the lowest relation to the data science concepts, indicating the gaps and opportunities for novelty. `District energy systems` shows up as the weakest, likely due to the only recent focus on the simulation and modelling of such techniques in the domain. Johansson et al. [78,79] looked at district energy systems and raised up some practical limitations such as the availability and quality of sensors. Also, district energy prediction is dependent

Table 1
Challenges of building performance simulation, reviews or key publications on the topic, and corresponding open-source repositories and data-sources.

Challenge	Relevant review(s) / publication(s)	Code	Open Data
1. Addressing the building performance gap	Type and definition [39]; Causes [148]; Credibility gap [16]	ObepME [81]; WinProGen (Occupant-behaviour gap) [20]	[72,108,20]
2. Modeling human-building interactions	Occupant modeling methodology in BPS [160] Challenges and opportunities [116]	Buildings.Occupants [154]	Occupant behavior [70]
3. Model calibration	Calibration methods and techniques [34,126,45]; Sensitivity analysis [146]	Bayesian calibration [27,125]; Optimization [22]	OpenStudio Calibration examples [113]
4. Modeling operation, controls and retrofits	Retrofit toolkits [89,90]; Model based commissioning [152]	Crowd-sourced ML for buildings [106]	Large, open meter data [107]
5. Modeling operational faults	Energy performance optimization [15]	Openstudio Fault Models Gem [26]	Open fault detection data [55]
6. Zero-net-energy and grid-responsive buildings	Gaps and needs [13,86]; Grid-responsive buildings [114]	BeOpt [31]	NZEB occupant behaviour [85]; Watts per person[117]
7. Urban-scale building energy modeling	Modeling methodology and workflows [127]; Challenges and future opportunities [66]	PyCity [140,144,1]	City buildings dataset [24] SynCity [135], NYC-UBEM [133]
8. Evaluating the energy-saving potential of building technologies at national or regional scales	E3 [98], Building stock energy prediction [92]	INTERDYME [9], PortableDyme [57], Scout [115]	Open data use for city-wide benchmarking [134]
9. Modeling energy efficient technology adoption	[44,73,53]	N.A.	Air-Conditioning Heating and Refrigeration Institute (AHRI) open data [61]
10. Integrated modeling and simulation	Progress, prospects, and requirements [33]	IFC[19], GBXML[42], OpenFOAM[74], EnergyPlus[36],Co-simulation e.g. obFMU[68]	Physics-based and data-driven modelling for NYC [135]

on both outdoor weather as well as control and social behaviour of consumers [58].

`HVAC optimization` and `om` (operations and maintenance) are contemporary topics, but only have small overlap with some of the more recent innovative data science techniques emerging. On the one hand, `HVAC optimization` has been reviewed by Selamat et al. [141] in three areas: HVAC operational parameters optimization, HVAC control system optimization, and building design optimization. His survey concluded that predictive optimization has more potential energy consumption reduction compared to conventional methods. Not only on the HVAC system scale, but also optimization should be done on the building design and building thermal dynamics. Other implementations of the data science for HVAC control in `om` have been conducted in recent years [59,17,46]. These issues include user security regarding data collection and storage, the lack of standardized data exchange schemes, and the lack of personnel with proper data science and domain knowledge.

`Parametric design` is also seen to be under-utilizing data science applications although it has gained much momentum in the last two decades [64]. This momentum is attributed to the advancement in Computer-Aided design CAD software as well as the emergence of user-friendly programming languages such as visual programming languages (VPL) [56]. Visual programming tools such as Revit Dynamo, and Rhino-Grasshopper has enabled end-use programmers to use data science algorithms in the design process. For example, Machine learning tools such as ANT [3] Lunchbox and OWL [84]; Optimization and multi-objective optimization such as OPOSSUM [157], octopus, Galapagos, and Optimus [37]; Energy Modelling such as Ladybug tools, [136], BuildFit [4]; Data visualization and deep learning using Gh_CPython [97]. These tools have grasped the attention of a large body of researchers and end-use programmers recently and may have a great potential for converging data science into the design process.

5. Conclusion

This paper outlined the text mining analysis of approximately 30,000 publications found in the top journals in the built environment analytics domain. This process aims to review the data science methods used in different building energy efficiency applications by mining large corpus of structured text from ELSEVIER journals. This process discovered high-level trends and potential gaps in the literature. Some data science methods have been extensively used in energy efficiency applications such as optimization, neural networks, statistical analysis, and energy simulation. However, there is still room for more opportunities of using other algorithms such as anomaly detection, factor analysis, segmentation, and GANs. Additionally, data-science methods are observed to be under-utilized during the commissioning and design phases of the building while saturated during the operation and maintenance phase. This could be attributed to the availability of ground-truth data during these lifecycle phases. Furthermore, different data sources are used frequently such as energy consumption-related data and BEM-related data. While other data sources are underutilized such as thermal-comfort related data, as well as HVAC-optimization related data. These results are extracted using a model based on the Word2Vec similarity metric. The results from this metric have shown consistency with the previous studies. Thus, researchers in this domain should utilize these results to determine which avenues are saturated, and therefore, will require much more effort to differentiate their work, and those which are emerging and have more unexplored potential.

Having said that, we acknowledge some limitations related to this method. Firstly, each paper was treated equally over the text mining procedure, neglecting the effect of those seminal studies. Correspondingly, one future work is to introduce bibliometrics as a feature to distinguish the paper's significance. Also, this method results in a non-directed relationship graph. This means that word such as "occupant" can appear as a data (e.g. number of occupants)

and can appear as a energy efficiency application (e.g. occupant behaviour modelling). There are many words that can hold different meanings such as "Operation and maintenance (O&M)" which could be used as a lifecycle phase, and energy efficiency applications. Future directions include improving the model so that it results in a directional graph (di-graph). This digraph can be drawn out using relation extraction models based on part-of-speech and stop-words. Besides, adding high-performance pre-trained models that are based on transformers such as BERT [145] and Generative Pre-trained Transformer 3 (GPT-3) [18] can be helpful to reduce ambiguity from some words that has different meanings.

5.1. Reproducibility

This analysis can be reproduced using the code and word vector data found on Github. https://github.com/ideas-lab-nus/data-science-bldg-energy-efficiency.

Declaration of Competing Interest

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

Acknowledgements

This work is supported by the Ng Teng Fong Charitable Foundation. (NTFCF) [Grant No. N-027-000-052-001].

References

[1] RWTH-EBC/pyCity: Python package for data handling and scenario generation of city districts. URL: https://github.com/RWTH-EBC/pyCity..
[2] Mahmoud AbdelRahman. Cross relation between different data-science and energy_efficiency data. 2 2021. URL: https://figshare.com/articles/figure/Cross_relation_between_different_data-science_and_energy_efficiency_data/14034563, https://doi.org/10.6084/m9.figshare.14034563.v1..
[3] Mahmoud M. Abdelrahman, Ahmed Mohamed Yousef Toutou, ANT: a machine learning approach for building performance simulation: methods and development, Acad. Res. Commun. Publ. 3 (1) (2019) 205, https://doi.org/10.21625/archive.v3i1.442.
[4] Mahmoud M. Abdelrahman, Sicheng Zhan, and Adrian Chong. A Three-Tier Architecture Visual-Programming Platform for Building-Lifecycle Data Management. SimAUD 2020, pages 439–446, 2020. URL: http://simaud.org/2020/proceedings/65.pdf..
[5] Abdul Afram, Farrokh Janabi-Sharifi, Theory and applications of HAVC control systems–a review of model predictive control (MPC), Build. Environ. 72 (2014) 343–355, https://doi.org/10.1016/j.buildenv.2013.11.016.
[6] Abdul Afram, Farrokh Janabi-Sharifi, Alan S. Fung, Kaamran Raahemifar, Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system, Energy Build. 141 (96–113) (2017), https://doi.org/10.1016/j.enbuild.2017.02.012.
[7] G.M. Zakia Afroz, Tania Urmee Shafiullah, Gary Higgins, Modeling techniques used in building HVAC control systems: A review, Renewable Sustainable Energy Rev. 83 (64–84) (2018), https://doi.org/10.1016/j.rser.2017.10.044.
[8] Herman Aguinis, Ravi S. Ramani, and Nawaf Alabduljader. Best-practice recommendations for producers, evaluators, and users of methodological literature reviews. Organizational Research Methods, page 1094428120943281, 2020. https://doi.org/10.1177/1094428120943281..
[9] Clopper Almon Jr., Interdyme: A package of programs for building interindustry dynamic macroeconomic models, version 3.10, Department of Economics, University of Maryland, 2000.
[10] Roger W. Amstalden, Michael Kost, Carsten Nathani, and Dieter M. Imboden. Economic potential of energy-efficient retrofitting in the swiss residential building sector: The effects of policy instruments and energy price expectations. Energy Pol., 35(3):1819–1829, 2007. https://doi.org/10.1016/j.enpol.2006.05.018..
[11] Yu Qian Ang, Zachary Michael Berzolla, and Christoph F. Reinhart. From concept to application: A review of use cases in urban building energy modeling. Appl. Energy, 279:115738, 2020. https://doi.org/10.1016/j.apenergy.2020.115738..
[12] Pandarasamy Arjunan, Kameshwar Poolla, Clayton Miller, EnergyStar++: Towards more accurate and explanatory building energy benchmarking, Appl. Energy 276 (October 2020), https://doi.org/10.1016/j.apenergy.2020.115413 115413.
[13] Shady Attia, Mohamed Hamdy, William O'Brien, Salvatore Carlucci, Assessing gaps and needs for integrating building performance optimization tools in net zero energy buildings design, Energy Build. 60 (2013) 110–124, https://doi.org/10.1016/j.enbuild.2013.01.016.
[14] Muhammad Rizwan Bashir and Asif Qumer Gill. Towards an IoT big data analytics framework: Smart buildings systems, in: Proceedings – 18th IEEE International Conference on High Performance Computing and Communications, 14th IEEE International Conference on Smart City and 2nd IEEE International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2016, pages 1325–1332, 2017. https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0188..
[15] Gesa A. Benndorf, Dominik Wystrcil, and Nicolas Réhault. Energy performance optimization in buildings: A review on semantic interoperability, fault detection, and predictive control. Applied Physics Reviews, 5(4), 2018. https://doi.org/10.1063/1.5053110..
[16] Bill Bordass, William Bordass Associates, and Robert Cohen. Energy Performance of Non-Domestic Buildings: Closing the Credibility Gap, in: 8th International Conference on Improving Energy Efficiency in Commercial Buildings, pages 1–10, 2004. URL: https://www.buildup.eu/en/node/1900..
[17] Silvio Brandi, Marco Savino Piscitelli, Marco Martellacci, Alfonso Capozzoli, Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings, Energy Build. 224 (110225) (October 2020), https://doi.org/10.1016/j.enbuild.2020.110225.
[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. arXiv, 2020. arXiv:2005.14165..
[19] BuildingSMART. Industry Foundation Classes (IFC) - buildingSMART Technical. 2020. URL: https://technical.buildingsmart.org/standards/ifc..
[20] Davide Calì, Mark Thomas Wesseling, Dirk Müller, WinProGen: A Markov-Chain-based stochastic window status profile generator for the simulation of realistic energy performance in buildings, Build. Environ. 136 (240–258) (2018), https://doi.org/10.1016/j.buildenv.2018.03.048.
[21] Debaditya Chakraborty, Hazem Elzarka, Advanced machine learning techniques for building performance simulation: a comparative analysis, J. Build. Perform. Simul. 12 (2) (2019) 193–207, https://doi.org/10.1080/19401493.2018.1498538.
[22] Gaurav Chaudhary, Joshua New, Jibonananda Sanyal, Piljae Im, Zheng O'Neill, Vishal Garg, Evaluation of "autotune" calibration against manual calibration of building energy models, Appl. Energy 182 (2016) 115–134, https://doi.org/10.1016/j.apenergy.2016.08.073.
[23] Qian Chen, Borja García de Soto, and Bryan T Adey. Construction automation: Research areas, industry concerns and suggestions for advancement. Automation in Construction, 94:22–38, 2018. https://doi.org/10.1016/j.autcon.2018.05.028..
[24] Yixing Chen, Tianzhen Hong, Xuan Luo, Barry Hooper, Development of city buildings dataset for urban building energy modeling, Energy Build. 183 (2019) 252–265, https://doi.org/10.1016/j.enbuild.2018.11.008.
[25] Yujiao Chen, Zheming Tong, Yang Zheng, Holly Samuelson, Leslie Norford, Transfer learning with deep neural networks for model predictive control of HVAC and natural ventilation in smart buildings, J. Clean. Prod. 254 (2020), https://doi.org/10.1016/j.jclepro.2019.119866 119866.
[26] Howard Cheung and James E. Braun. Development of Fault Models for Hybrid Fault Detection and Diagnostics Algorithm. (October 1, 2014 – May 5, 2015):59, 2015. URL: http://www.nrel.gov/docs/fy16osti/65030.pdf..
[27] Adrian Chong, Kathrin Menberg, Guidelines for the bayesian calibration of building energy models, Energy Build. 174 (2018) 527–547, https://doi.org/10.1016/j.enbuild.2018.06.028.
[28] Adrian Chong, Xu. Weili, Song Chao, Ngoc-Tri Ngo, Continuous-time bayesian calibration of energy models using bim and energy data, Energy Build. 194 (2019) 177–190, https://doi.org/10.1016/j.enbuild.2019.04.017.
[29] Despoina Christantoni, Simeon Oxizidis, Damian Flynn, Donal P. Finn, Implementation of demand response strategies in a multi-purpose commercial building using a whole-building simulation model approach, Energy Build. 131 (76–86) (2016), https://doi.org/10.1016/j.enbuild.2016.09.017.
[30] C. Christensen, R. Anderson, S. Horowitz, A. Courtney, J. Spencer, Beopt(tm) software for building energy optimization, Features Capabil. 8 (2006), https://doi.org/10.2172/891598.
[31] Craig Christensen, Adam Courtney, Scott Horowitz, Todd Givler, Greg Barker, Beopt: Software for identifying optimal building designs on the path to zero net energy, in: Proceedings of the Solar World Congress 2005: Bringing Water to the World, Including Proceedings of 34th ASES Annual Conference and Proceedings of 30th National Passive Solar Conference 1:55–60, 2005, URL: https://www.nrel.gov/docs/fy05osti/37733.pdf.
[32] Angelos Chronis, Anna Aichinger, Serjoscha Duering, Theodore Galanos, Theresa Fink, Ondrej Vesely, and Reinhard Koenig. INFRARED: An Intelligent Framework for Resilient Design ANGELOS. 25th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), pages 1–10, 2020..

[33] J.A. Clarke, J.L.M. Hensen, Integrated building performance simulation: Progress, prospects and requirements, Build. Environ. 91 (2015) 294–306, https://doi.org/10.1016/j.buildenv.2015.04.002.
[34] Daniel Coakley, Paul Raftery, Marcus Keane, A review of methods to match building energy simulation models to measured data, Renewable Sustainable Energy Rev. 37 (2014) 123–141, https://doi.org/10.1016/j.rser.2014.05.007.
[35] Drury B. Crawley, Jon W. Hand, Michaël Kummert, Brent T. Griffith, Contrasting the capabilities of building energy performance simulation programs, Build. Environ. 43(4):661–673 (2008), https://doi.org/10.1016/j.buildenv.2006.10.027.
[36] Drury B. Crawley, Linda K. Lawrie, Frederick C. Winkelmann, W.F. Buhl, Y. Joe Huang, Curtis O. Pedersen, Richard K. Strand, Richard J. Liesen, Daniel E. Fisher, Michael J. Witte, Jason Glazer, EnergyPlus: Creating a new-generation building energy simulation program, Energy Build. 33 (4) (2001) 319–331, https://doi.org/10.1016/S0378-7788(00)00114-6.
[37] Cemre Cubukcuoglu, Berk Ekici, Mehmet Fatih Tasgetiren, Sevil Sariyildiz, OPTIMUS: self-adaptive differential evolution with ensemble of mutation strategies for grasshopper algorithmic modeling, Algorithms 12(7):141 (2019), https://doi.org/10.3390/a12070141.
[38] Hanna De Vries, Victor Bekkers, Lars Tummers, Innovation in the public sector: A systematic review and future research agenda, Pub. Administr. 94 (1) (2016) 146–166, https://doi.org/10.1111/padm.12209.
[39] Pieter De Wilde, The gap between predicted and measured energy performance of buildings: A framework for investigation, Autom. Constr. 41 (2014) 40–49, https://doi.org/10.1016/j.autcon.2014.02.009.
[40] Dursun Delen and Martin D. Crossland. Seeding the survey and analysis of research literature with text mining. Expert Systems with Applications, 34 (3):1707–1720, 2008. URL: http://www.sciencedirect.com/science/article/pii/S0957417407000486, https://doi.org/10.1016/j.eswa.2007.01.035..
[41] Krisztina Demeter, Levente Szász, Andrea Kő, A text mining based overview of inventory research in the isir special issues 1994–2016, Int. J. Prod. Econ. 209 (2019) 134–146, https://doi.org/10.1016/j.ijpe.2018.06.006.
[42] Vanda Dimitriou, Steven K. Firth, Tarek M. Hassan, Farid Fouchal, BIM enabled building energy modelling: development and verification of a GBXML to IDF conversion method, Ibpsa (2016) 12–14, URL: https://dspace.lboro.ac.uk/2134/22818.
[43] Ján Drgoňa, Javier Arroyo, Iago Cupeiro Figueroa, David Blum, Krzysztof Arendt, Donghun Kim, Enric Perarnau Ollé, Juraj Oravec, Michael Wetter, Draguna L Vrabie, et al. All you need to know about model predictive control for buildings. Annual Reviews in Control, 2020. https://doi.org/10.1016/j.arcontrol.2020.09.001..
[44] US EIA. Integrating module of the national energy modeling system: Model documentation 2010, 2010. URL: https://www.eia.gov/outlooks/aeo/nems/documentation/integrating/pdf/m057(2020).pdf..
[45] Enrico Fabrizio, Valentina Monetti, Methodologies and advancements in the calibration of building energy models, Energies 8 (4) (2015) 2548–2574, https://doi.org/10.3390/en8042548.
[46] Cheng Fan, Yongjun Sun, Kui Shan, Linda F. Xiao, and Jiayuan Wang. Discovering gradual patterns in building operations for improving building energy efficiency. Appl Energy, 224:116–123, August 2018. https://doi.org/10.1016/j.apenergy.2018.04.118..
[47] Cheng Fan, Yongjun Sun, Yang Zhao, Mengjie Song, Jiayuan Wang, Deep learning-based feature engineering methods for improved building energy prediction, Appl. Energy 240 (April 2019) 35–45, https://doi.org/10.1016/j.apenergy.2019.02.052.
[48] Cheng Fan, Fu Xiao, Zhengdao Li, Jiayuan Wang, Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review, Energy Build. 159 (2018) 296–308, https://doi.org/10.1016/j.enbuild.2017.11.008.
[49] Cheng Fan, Da Yan, Fu Xiao, Ao Li, Jingjing An, Xuyuan Kang, Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches, Build. Simul. (October 2020), https://doi.org/10.1007/s12273-020-0723-1.
[50] Hannah Fontenot, Bing Dong, Modeling and control of building-integrated microgrids for optimal energy management–a review, Appl. Energy 254 (2019), https://doi.org/10.1016/j.apenergy.2019.113689 113689.
[51] Colm V. Gallagher, Kevin Leahy, Peter O'Donovan, Ken Bruton, and Dominic T. J. O'Sullivan. Development and application of a machine learning supported methodology for measurement and verification (m&v) 2.0. Energy and Buildings, 167:8–22, 2018. URL: http://www.sciencedirect.com/science/article/pii/S0378778817336630, https://doi.org/10.1016/j.enbuild.2018.02.023..
[52] Shengnan Geng, Yuan Wang, Jian Zuo, Zhihua Zhou, Huibin Du, and Guozhu Mao. Building life cycle assessment research: A review by bibliometric analysis, 2017. https://doi.org/10.1016/j.rser.2017.03.068..
[53] S.T. Gilshannon and D.R. Brown. Review of methods for forecasting the market penetration of new technologies. U.S. Department of Energy, Pacific Northwest Laboratory, PNNL-11428:1–63, 1996. URL: http://www.osti.gov/energycitations/product.biblio.jsp?osti_id=432867..
[54] Yoav Goldberg and Omer Levy. word2vec Explained: deriving Mikolov et al'.s negative-sampling word-embedding method. 2014. arXiv:1402.3722..
[55] Jessica Granderson, Guanjing Lin, Ari Harding, Piljae Im, Yan Chen, Building fault detection data to aid diagnostic algorithm creation and performance testing, Sci. Data 7 (1) (February 2020) 65, https://doi.org/10.6084/m9.figshare.11743074.
[56] T.R.G. Green, M. Petre, and Rachel K.E. Bellamy. Comprehensibility of Visual and Textual Programs: A Test of Superlativism Against the 'Match-Mismatch' Conjecture. Proceedings of the Fourth Annual Workshop on Empirical Studies of Programmers, (January):121–146, 1991. URL: https://www.researchgate.net/publication/238987815_Comprehensibility_of_visual_and_textual_programs_A_test_of_superlativism_against_the_'match-mismatch'_conjecture..
[57] A. GroBmann, F. Hohmann, K. Wiebe, Portabledyme-a simplified software package for econometric model building, Macroecon. Model. Pol. Eval. 120 (2013) 33.
[58] S. Grosswindhager, A. Voigt, and M. Kozek. Online Short-Term Forecast of System Heat Load in District Heating Networks. In Proceedings of the 31st International Symposium on Forecasting, (1):1–8, 2011. URL: http://www.forecasters.org/submissions/GROSSWINDHAGERSTEFANISF2011.pdf..
[59] Burak Gunay, Weiming Shen, Connected and distributed sensing in buildings: improving operation and maintenance, IEEE Syst. Man Cybern. Mag. 3 (4) (2017) 27–34, https://doi.org/10.1109/msmc.2017.2702386.
[60] H. Burak Gunay, Weiming Shen, Guy Newsham, Data analytics to improve building performance: A critical review, Autom. Constr. 97 (96–109) (2019), https://doi.org/10.1016/j.autcon.2018.10.020.
[61] American Heating and Refrigeration Institute (AHRI). Historical Data: Statistical information on HVACR equipment shipments, 2020. URL: http://www.ahrinet.org/resources/statistics/historical-data..
[62] Jan L. M. Hensen and Roberto Lamberts. Building performance simulation for design and operation. Routledge, 2 edition, 2019. doi:10.1201/9780429402296..
[63] Yeonsook Heo, Ruchi Choudhary, G.A. Augenbroe, Calibration of building energy models for retrofit analysis under uncertainty, Energy Build. 47 (550–560) (2012), https://doi.org/10.1016/j.enbuild.2011.12.029.
[64] Carlos Roberto Barrios Hernandez, Thinking parametric design: Introducing parametric Gaudi, Des. Stud. 27 (3) (2006) 309–324, https://doi.org/10.1016/j.destud.2005.11.006.
[65] Tianzhen Hong, Yixing Chen, Xuan Luo, Na Luo, Sang Hoon Lee, Ten questions on urban building energy modeling, Build. Environ. 168 (106508) (2020), https://doi.org/10.1016/j.buildenv.2019.106508.
[66] Tianzhen Hong, Yixing Chen, Xuan Luo, Na Luo, Sang Hoon Lee, Ten questions on urban building energy modeling, Build. Environ. 168 (106508) (2020), https://doi.org/10.1016/j.buildenv.2019.106508.
[67] Tianzhen Hong, Jared Langevin, Kaiyu Sun, Building simulation: Ten challenges, Build. Simul. 11 (5) (2018) 871–898, https://doi.org/10.1007/s12273-018-0444-x.
[68] Tianzhen Hong, Hongsan Sun, Yixing Chen, Sarah C. Taylor-Lange, Da Yan, An occupant behavior modeling tool for co-simulation, Energy Build. 117 (2016) 272–281, https://doi.org/10.1016/j.enbuild.2015.10.033.
[69] Tianzhen Hong, Zhe Wang, Xuan Luo, Wanni Zhang, State-of-the-art on research and applications of machine learning in the building life cycle, Energy Build. 212 (2020), https://doi.org/10.1016/j.enbuild.2020.109831 109831.
[70] Gesche Margarethe Huebner and Ardeshir Mahdavi. A structured open data collection on occupant behaviour in buildings. Sci. Data, 6(1):1–4, 2019. https://doi.org/10.1038/s41597-019-0276-2..
[71] IBPSA-USA. Building energy software tools (BEST) directory, formerly hosted by the US department of energy. https://www.buildingenergysoftwaretools.com, 2014. Accessed: 2020-12-03..
[72] Salah Imam, David A. Coley, Ian Walker, The building performance gap: Are modellers literate?, Build Serv. Eng. Res. Technol. 38 (3) (2017) 351–375, https://doi.org/10.1177/0143624416684641.
[73] Mark Jaccard, Margo Dennis, Estimating home energy decision parameters for a hybrid energy-economy policy model, Environ. Model. Assess. 11 (2) (2006) 91–100, https://doi.org/10.1007/s10666-005-9036-0.
[74] Hrvoje Jasak, OpenFOAM: Open source CFD in research and industry, Int. J. Nav. Arch. Ocean Eng. 1 (2) (2009) 89–94, https://doi.org/10.2478/ijnaoe-2013-0011.
[75] Hongyuan Jia and Adrian Chong. eplusr: A framework for integrating building energy simulation and data-driven analytics. Energy and Buildings, page 110757, 2021. https://doi.org/10.1016/j.enbuild.2021.110757..
[76] Anish Jindal, Neeraj Kumar, Mukesh Singh, A unified framework for big data acquisition, storage, and analytics for demand response management in smart cities, Fut. Gen. Comput. Syst. 108 (2020) 921–934, https://doi.org/10.1016/j.future.2018.02.039.
[77] Arif E. Jinha, Article 50 million: an estimate of the number of scholarly articles in existence, Learn. Publ. 23 (3) (2010) 258–263, https://doi.org/10.1087/20100308.
[78] Christian Johansson, Markus Bergkvist, Davy Geysen, Oscar De Somer, Niklas Lavesson, Dirk Vanhoudt, Operational demand forecasting in district heating systems using ensembles of online machine learning algorithms, Energy Procedia 116 (2017) 208–216, https://doi.org/10.1016/j.egypro.2017.05.068.
[79] Christian Johansson and Blekinge Tekniska Hogskola. On intelligent district heating. 2014. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:bth-00587..
[80] A. Rezaee Jordehi. Optimisation of demand response in electric power systems, a review. Renewable and sustainable energy reviews, 103:308–319, 2019. https://doi.org/10.1016/j.rser.2018.12.054..
[81] M. Jradi, K. Arendt, F.C. Sangogboye, C.G. Mattera, E. Markoska, M.B. Kjærgaard, C.T. Veje, B.N. Jørgensen, ObepME: An online building energy performance monitoring and evaluation tool to reduce energy performance

gaps, Energy Build. 166 (2018) 196–209, https://doi.org/10.1016/j.enbuild.2018.02.005.
[82] Srinivas Katipamula and Michael R. Brambley. Methods for fault detection, diagnostics, and prognostics for building systems-a review, part i. Hvac&R Research, 11(1), 3–25, 2005. https://doi.org/10.1080/10789669.2005.10391123..
[83] Azam Khan and Kasper Hornbæk. Big data from the built environment. LARGE'11 - Proceedings of the 2nd International Workshop on Research in the Large, pages 29–32, 2011. https://doi.org/10.1145/2025528.2025537..
[84] Nariddh Khean, Alessandra Fabbri, and M. Hank Haeusler. Learning Machine Learning as an Architect, How to? Computing for a better tomorrow - Proceedings of the 36th eCAADe Conference, AI for Design and Built Environment, 1:95–102, 2018. URL: http://papers.cumincad.org/data/works/att/ecaade2018_111.pdf..
[85] Mika Yagi Kim. "Watts Per Person" Paradigm To Design Net Zero Energy Buildings : Examining Technology Interventions and Integrating Occupant. https://doi.org/10.25549/usctheses-c3-281325..
[86] D. Kolokotsa, D. Rovas, E. Kosmatopoulos, K. Kalaitzakis, A roadmap towards intelligent net zero- and positive-energy buildings, Sol. Energy 85 (12) (2011) 3067–3084, https://doi.org/10.1016/j.solener.2010.09.001.
[87] Grzegorz Kondrak. N-gram similarity and distance. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3772 LNCS:115–126, 2005. https://doi.org/10.1007/11575832_13..
[88] Andrew Lake, Behnaz Rezaie, Steven Beyerlein, Review of district heating and cooling systems for a sustainable future, Renewable Sustainable Energy Rev. 67 (2017) 417–425, https://doi.org/10.1016/j.rser.2016.09.061.
[89] Sang Hoon Lee, Tianzhen Hong, and Mary Ann Piette. Review of Existing Energy Retrofit Tools. (July):38, 2014. URL: http://escholarship.org/uc/item/70p8n9x3..
[90] Sang Hoon Lee, Tianzhen Hong, Mary Ann Piette, Sarah C. Taylor-Lange, Energy retrofit analysis toolkits for commercial buildings: A review, Energy 89 (1087–1100) (2015), https://doi.org/10.1016/j.energy.2015.06.112.
[91] Xiao Li, Peng Wu, Geoffrey Qiping Shen, Xiangyu Wang, and Yue Teng. Mapping the knowledge domains of Building Information Modeling (BIM): A bibliometric approach, 2017. https://doi.org/10.1016/j.autcon.2017.09.011..
[92] Hyunwoo Lim, Zhiqiang John Zhai, Review on stochastic modeling methods for building stock energy prediction, Build. Simul. 10 (5) (2017) 607–624, https://doi.org/10.1007/s12273-017-0383-y.
[93] Shunfu Lin, Fangxing Li, Erwei Tian, Yang Fu, Dongdong Li, Clustering load profiles for demand response applications, IEEE Trans. Smart Grid 10 (2) (2017) 1599–1607, https://doi.org/10.1109/TSG.2017.2773573.
[94] Martina K. Linnenluecke, Mauricio Marrone, and Abhay K. Singh. Conducting systematic literature reviews and bibliometric analyses. Australian Journal of Management, 45(2), 175–194, 2020. https://doi.org/10.1177/0312896219877678..
[95] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. 2002. URL: http://arxiv.org/abs/cs/0205028, https://doi.org/10.3115/1118108.1118117..
[96] Emilio T. Maddalena, Yingzhao Lian, Colin N. Jones, Data-driven methods for building control-a review and promising future directions, Contr. Eng. Pract. 95 (104211) (2020), https://doi.org/10.1016/j.conengprac.2019.104211.
[97] Mahmoud Mohamed Abdelrahman. Gh_CPython: CPython plugin for grasshopper. 2017. https://doi.org/10.5281/zenodo.888148..
[98] Abbas Mardani, Dalia Streimikiene, Tomas Balezentis, Muhamad Zameri Mat Saman, Khalil Md Nor, and Seyed Meysam Khoshnava. Data envelopment analysis in energy and environmental economics: An overview of the state-of-The-Art and recent development trends. Energies, 11(8), 2018. https://doi.org/10.3390/en11082002..
[99] Paul A. Mathew, Laurel N. Dunn, Michael D. Sohn, Andrea Mercado, Claudine Custudio, Travis Walter, Big-data for building energy performance: Lessons from assembling a very large national database of building energy use, Appl. Energy 140 (2015) 85–93, https://doi.org/10.1016/j.apenergy.2014.11.042.
[100] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 2013. arXiv:1301.3781..
[101] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, Armand Joulin, Advances in pre-training distributed word representations, in: LREC 2018–11th International Conference on Language Resources and Evaluation, 2019, pp. 52–55, arXiv:1712.09405.
[102] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. Adv. Neu. Inf. Process. Syst., 26:3111–3119, 2013. arXiv:1310.4546..
[103] Tomáš Mikolov, Wen-tau Yih, Geoffrey Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.
[104] Clayton Miller, More buildings make more generalizable Models—Benchmarking prediction methods on open electrical meter data, Mach. Learn. Knowledge Extract. 1 (3) (August 2019) 974–993, https://doi.org/10.3390/make1030056.
[105] Clayton Miller, What's in the box?! Towards explainable machine learning applied to non-residential building smart meter classification, Energy Build. 199 (September 2019) 523–536, https://doi.org/10.1016/j.enbuild.2019.07.019.
[106] Clayton Miller, Pandarasamy Arjunan, Anjukan Kathirgamanathan, Chun Fu, Jonathan Roth, June Young Park, Chris Balbach, Krishnan Gowri, Zoltan Nagy, Anthony D Fontanini, and Jeff Haberl. The ASHRAE great energy predictor III competition: Overview and results. Science and Technology for the Built Environment, pages 1–21, August 2020. https://doi.org/10.1080/23744731.2020.1795514..
[107] Clayton Miller, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W Hobson, Zixiao Shi, and Forrest Meggers. The building data genome project 2, energy meter data from the ASHRAE great energy predictor III competition. Scientific Data, 7:368, October 2020. https://doi.org/10.1038/s41597-020-00712-x..
[108] Simon Moeller, Amelie Bauer, Ines Weber, Franz Schröder, and Hannes Harter. Data for: Flat specific energy performance gap - how to address internal heat shifts in multi-apartment dwellings. 1, 2020. https://doi.org/10.17632/7CVGWS3MX3.1..
[109] Miguel Molina-Solana, M. María Ros, Dolores Ruiz, Juan Gómez-Romero, María J Martín-Bautista, Data science for building energy management: A review, Renewable Sustainable Energy Rev. 70 (598–609) (2017), https://doi.org/10.1016/j.rser.2016.11.132.
[110] José A Moral-Muñoz, Antonio G López-Herrera, Enrique Herrera-Viedma, Manuel J Cobo, Science mapping analysis software tools: A review, in: Springer Handbook of Science and Technology Indicators, Springer, 2019, pp. 159–185, https://doi.org/10.1007/978-3-030-02511-3_7.
[111] Fionn Murtagh, Pierre Legendre, Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?, J Classfi. 31 (3) (2014) 274–295, https://doi.org/10.1007/s00357-014-9161-z.
[112] Cristina Nichiforov, Grigore Stamatescu, Iulia Stamatescu, Ioana Făgărăşan, Evaluation of sequence-learning models for large-commercial-building load forecasting, Information 10 (6) (June 2019) 189, https://doi.org/10.3390/info10060189.
[113] NREL. NREL/OpenStudio-analysis-spreadsheet: The OpenStudio Analysis Spreadsheet is a front-end for the OpenStudio Server, allowing for users to create large-scale cloud analyses using OpenStudio measures., 2017. URL: https://github.com/NREL/OpenStudio-analysis-spreadsheet..
[114] Monica Nuekomm, Valerie Nubbe, and Robert Fares. Grid-interactive Efficient Buildings: Overview. (April):1–36, 2019.https://doi.org/10.2172/1508212..
[115] US. Department of Energy Building Technology Office (BTO). Scout: Github repository, 2020. URL: https://github.com/trynthink/scout..
[116] William O'Brien, Andreas Wagner, Marcel Schweiker, Ardeshir Mahdavi, Julia Day, Mikkel Baun Kjærgaard, Salvatore Carlucci, Bing Dong, Farhang Tahmasebi, Da Yan, et al. Introducing iea ebc annex 79: Key challenges and opportunities in the field of occupant-centric building design and operation. Building and Environment, page 106738, 2020. https://doi.org/10.1016/j.buildenv.2020.106738..
[117] Frederick Paige, Philip Agee, and Farrokh Jazizadeh. flEECe, an energy use and occupant behavior dataset for net-zero energy affordable senior residential buildings. Scientific Data, 6(1), 2019. https://doi.org/10.1038/s41597-019-0275-3..
[118] June Young Park, Xiya Yang, Clayton Miller, Pandarasamy Arjunan, and Zoltan Nagy. Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. Appl. Energy, 236:1280–1295, February 2019. https://doi.org/10.1016/j.apenergy.2018.12.025..
[119] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. J. Mach. Learn. Res., 12:2825–2830, 2011..
[120] Jeffrey Pennington, Richard Socher, Christopher D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543, https://doi.org/10.3115/v1/D14-1162.
[121] Marco Savino Piscitelli, Silvio Brandi, and Alfonso Capozzoli. Recognition and classification of typical load profiles in buildings with non-intrusive learning approach. Appl. Energy, 255:113727, December 2019. https://doi.org/10.1016/j.apenergy.2019.113727..
[122] Marco Savino Piscitelli, Silvio Brandi, Alfonso Capozzoli, and Linda F. Xiao. A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings. Building Simulation, May 2020. https://doi.org/10.1007/s12273-020-0650-1..
[123] Matias Quintana, Clayton Miller, Towards class-balancing human comfort datasets with GANs, in: BuildSys 2019 – Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2019, pp. 391–392, https://doi.org/10.1145/3360322.3361016.
[124] Matias Quintana, Stefano Schiavon, Kwok Wai Tham, Clayton Miller, Balancing thermal comfort datasets: We gan, but should we?, in: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2020, pp 120–129, https://doi.org/10.1145/3408308.3427612.
[125] Loic Raillon, Simon Rouchier, and Sarah Juricic. pysip: an open-source tool for bayesian inference and prediction of heat transfer in buildings..
[126] T. Agami Reddy, Literature review on calibration of building energy simulation programs: uses, problems, procedures, uncertainty, and tools, ASHRAE Trans. 112 (2006) 226.

[127] Christoph F. Reinhart, Carlos Cerezo Davila, Urban building energy modeling–a review of a nascent field, Build. Environ. 97 (196–202) (2016), https://doi.org/10.1016/j.buildenv.2015.12.001.
[128] Behnaz Rezaie, Marc A. Rosen, District heating and cooling: Review of technology and potential enhancements, Appl. Energy 93 (2012) 2–10, https://doi.org/10.1016/j.apenergy.2011.04.020.
[129] Fabien Chidanand Robert, Gyanendra Singh Sisodia, and Sundararaman Gopalan. A critical review on the utilization of storage and demand response for the implementation of renewable energy microgrids. Sustainable cities and society, 40:735–745, 2018. https://doi.org/10.1016/j.scs.2018.04.008..
[130] Patricia Rocha and Rui Calejo Rodrigues. Bibliometric review of improvements in building maintenance, 2017. https://doi.org/10.1108/JQME-07-2016-0030..
[131] Danny Rodrigues Alves, Giovanni Colavizza, Frédéric Kaplan, Deep reference mining from scholarly literature in the arts and humanities, Front. Res. Metr. Anal. 3 (21) (2018), https://doi.org/10.3389/frma.2018.00021.
[132] Adam Rohloff, Data analytics from cradle to grave, ASHRAE J. 58 (2) (2016) 34.
[133] Jonathan Roth. Github: Nyc - urban building energy model for new york city. URL: https://github.com/jmr385/UBEM_NYC..
[134] Jonathan Roth, Benjamin Lim, Rishee K. Jain, and Dian Grueneich. Examining the feasibility of using open data to benchmark building energy usage in cities: A data science and policy perspective. Energy Policy, 139:111327, April 2020..
[135] Jonathan Roth, Amory Martin, Clayton Miller, Rishee K. Jain, SynCity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods, Appl. Energy 280 (2020), https://doi.org/10.1016/j.apenergy.2020.115981.
[136] Mostapha Sadeghipour Roudsari, Michelle Pak, Ladybug: A parametric environmental plugin for grasshopper to help designers create an environmentally-conscious design, in: Proceedings of BS 2013: 13th Conference of the International Building Performance Simulation Association, 2013, pp. 3128–3135.
[137] Abdullahi B. Saka, Daniel W.M. Chan, A scientometric review and metasynthesis of building information modelling (BIM) research, Africa (2019), https://doi.org/10.3390/buildings9040085.
[138] Pattarin Sanguankaew, Vichita Vathanophas Ractham, Bibliometric review of research on knowledge management and sustainability, 1994–2018, Sustainability 11 (16) (2019) 4388, https://doi.org/10.3390/su11164388.
[139] Philip Santosh. Eppy: Scripting language for e+, energyplus. https://github.com/santoshphilip/eppy, 2014. Accessed: 2020-12-03..
[140] Jan Schiefelbein, Jana Rudnick, Anna Scholl, Peter Remmen, Marcus Fuchs, Dirk Müller, Automated urban energy system modeling and thermal building simulation based on OpenStreetMap data sets, Build. Environ. 149 (Feb 2019) 630–639, https://doi.org/10.1016/j.buildenv.2018.12.025.
[141] Hazlina Selamat, Mohamad Fadzli Haniff, Zainon Mat Sharif, Seyed Mohammad Attaran, Fadhilah Mohd Sakri, and Muhammad Al'Hapis Bin Abdul Razak. Review on HVAC system optimization towards energy saving building operation. International Energy Journal, 20(3), 345–357, 2020. URL: http://rericjournal.ait.ac.th/index.php/reric/article/view/2230/pdf..
[142] Filipi N. Silva, Diego R. Amancio, Maria Bardosova, Luciano da F. Costa, and Osvaldo N. Oliveira Jr. Using network science and text analytics to produce surveys in a scientific topic. Journal of Informetrics, 10(2):487–502, 2016. https://doi.org/10.1016/j.joi.2016.03.008..
[143] Christian Simon, Kristian Davidsen, Christina Hansen, Emily Seymour, Mike Bogetofte Barnkob, and Lars Rønn Olsen. BioReader: a text mining tool for performing classification of biomedical literature. BMC Bioinformatics, 19 (Suppl 13):57, February 2019. https://doi.org/10.1186/s12859-019-2607-x..
[144] Ivelina Stoyanova, Erdem Gumrukcu, and Antonello Monti. Modular modeling concept and multi-domain simulation for smart cities. 2017 IEEE PES Innovative Smart Grid Technologies Conference Europe, ISGT-Europe 2017 - Proceedings, 2018-January:1–6, 2017. https://doi.org/10.1109/ISGTEurope.2017.8260206..
[145] Ian Tenney, Dipanjan Das, Ellie Pavlick, BERT rediscovers the classical NLP pipeline, in: ACL 2019–57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2020, pp. 4593–4601, https://doi.org/10.18653/v1/p19-1452, arXiv:1905.05950.
[146] Wei Tian, A review of sensitivity analysis methods in building energy analysis, Renewable Sustainable Energy Rev. 20 (2013) 411–419, https://doi.org/10.1016/j.rser.2012.12.014.
[147] Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. CERMINE: automatic extraction of structured metadata from scientific literature. Int. J. Doc. Anal. Recogn., 18(4):317–335, December 2015. https://doi.org/10.1007/s10032-015-0249-8..
[148] Chris van Dronkelaar, Mark Dowson, Catalina Spataru, and Dejan Mumovic. A Review of the Regulatory Energy Performance Gap and Its Underlying Causes in Non-domestic Buildings. Frontiers in Mechanical Engineering, 1, 2016. https://doi.org/10.3389/fmech.2015.00017..
[149] Nees Jan Van Eck and Ludo Waltman. Text mining and visualization using vosviewer, 2011. arXiv:1109.2058..
[150] Nees Jan Van Eck, Ludo Waltman, Citnetexplorer: A new software tool for analyzing and visualizing citation networks, J. Inform. 8 (4) (2014) 802–823, https://doi.org/10.1016/j.joi.2014.07.006.
[151] Richard Van Noorden. Elsevier opens its papers to text-mining., 2 2014. https://doi.org/10.1038/506017a..
[152] J. Verhelst, G. Van Ham, D. Saelens, L. Helsen, Model selection for continuous commissioning of HVAC-systems in office buildings: A review, Renewable Sustainable Energy Rev. 76 (2017) 673–686, https://doi.org/10.1016/j.rser.2017.01.119.
[153] Zeyu Wang and Ravi S. Srinivasan. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models, 2017. https://doi.org/10.1016/j.rser.2016.10.079..
[154] Zhe Wang, Tianzhen Hong, Ruoxi Jia, Buildings. Occupants: a Modelica package for modelling occupant behaviour in buildings, J. Build. Perform. Simul. 12 (4) (2019) 433–444, https://doi.org/10.1080/19401493.2018.1543352.
[155] Rebecca Ward, Ruchi Choudhary, Alastair Gregory, and Mark Girolami. Continuous calibration of a digital twin: comparison of particle filter and bayesian calibration approaches, 2020. arXiv:2011.09810..
[156] Michael Wetter et al. Genopt-a generic optimization program. In Seventh International IBPSA Conference, Rio de Janeiro, pages 601–608, 2001. URL: http://www.ibpsa.org/proceedings/BS2001/BS01_0601_608.pdf..
[157] Thomas Wortmann, OPOSSUM: introducing and evaluating a model-based optimization tool for grasshopper, Proc. CAADRIA 17 (2017) 283–292, URL: http://papers.cumincad.org/data/works/att/caadria2017_124.pdf.
[158] Ibrahim Y. Wuni, Geoffrey Q.P. Shen, and Robert Osei-Kyei. Scientometric review of global research trends on green buildings in construction journals from 1992 to 2018. Energy and Buildings, 190:69–85, 2019. https://doi.org/10.1016/j.enbuild.2019.02.010..
[159] Shichao Xu, Yixuan Wang, Yanzhi Wang, Zheng O'Neill, Qi Zhu, One for many: Transfer learning for building hvac control, in: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2020, pp. 230–239, https://doi.org/10.1145/3408308.3427617.
[160] Da Yan, William O'Brien, Tianzhen Hong, Xiaohang Feng, H. Burak Gunay, Farhang Tahmasebi, and Ardeshir Mahdavi. Occupant behavior modeling for building performance simulation: Current state and future challenges. Energy Build., 107:264–278, 2015. https://doi.org/10.1016/j.enbuild.2015.08.032..
[161] Ke Yan, Adrian Chong, Yuchang Mo, Generative adversarial network for fault detection diagnosis of chillers, Build. Environ. 172 (2020), https://doi.org/10.1016/j.buildenv.2020.106698 106698.
[162] Ke Yan, Jing Huang, Wen Shen, Zhiwei Ji, Unsupervised learning for fault detection and diagnosis of air handling units, Energy Build. 210 (2020), https://doi.org/10.1016/j.enbuild.2019.109689 109689.
[163] Zheng Yang, Jonathan Roth, Rishee K. Jain, DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis, Energy Build. 163 (58–69) (March 2018), https://doi.org/10.1016/j.enbuild.2017.12.040.
[164] Sicheng Zhan, Zhaoru Liu, Adrian Chong, Da Yan, Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking, Appl. Energy 269 (2020), https://doi.org/10.1016/j.apenergy.2020.114920 114920.
[165] Yi Zhang, Ivan Korolija, Performing complex parametric simulations with jeplus, in: SET2010-9th International Conference on Sustainable Energy Technologies, 2010, pp. 24–27.
[166] Zhiang Zhang, Adrian Chong, Yuqi Pan, Chenlu Zhang, Khee Poh Lam, Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning, Energy Build. (199:472–490, 2019.), https://doi.org/10.1016/j.enbuild.2019.07.029.
[167] Xianbo Zhao, Jian Zuo, Wu. Guangdong, Can Huang (Eds.), A bibliometric review of green building research 2000–2016, 2019, https://doi.org/10.1080/00038628.2018.1485548.
[168] Yang Zhao, Tingting Li, Xuejun Zhang, Chaobo Zhang, Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future, Renewable Sustainable Energy Rev. 109 (2019) 85–101, https://doi.org/10.1016/j.rser.2019.04.021.