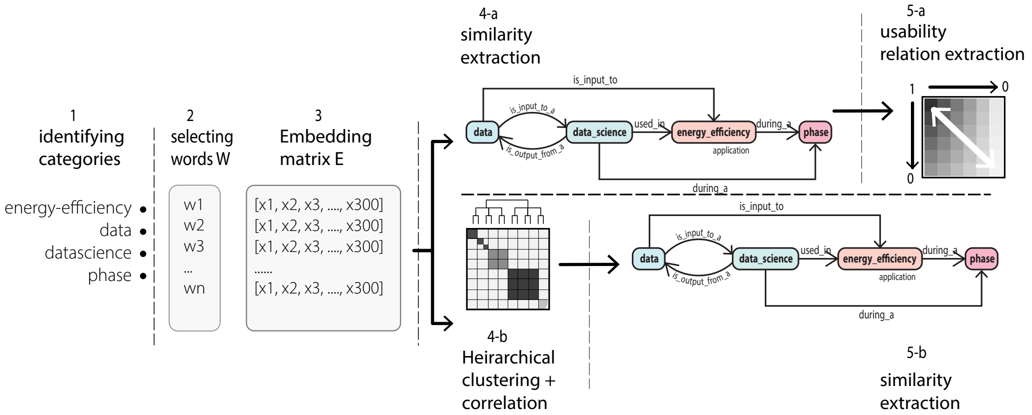
Graphical Abstract

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

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



Highlights

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

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

- 30,000 full-text building-related articles have been extracted for text mining.
- Data-science, energy efficiency, and lifecycle phases relations have been drawn.
- Word embeddings model has been used to extract the relationship between keywords.
- Gaps, opportunities, and potential future directions have been discussed.

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

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

[a] *Department of Building, School of Design and Environment, National University of Singapore, 4 Architecture Drive, Singapore 117566, Singapore*

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 and this paper outlines a process of quantifying the impact of various data science topics using natural language processing. Approximately 30,000 scientific publications were extracted from the Elsevier API and a process of natural language processing (NLP) extracted the relationship between data sources, data science techniques, and building energy efficiency applications across the life cycle of buildings. The data-driven analysis identifies significant application areas especially during the operation and maintenance phase such as fault detection and diagnosis (FDD) that are saturated. Meanwhile, gaps and under-explored techniques are also revealed especially during the commissioning and the design phases such as generative adversarial networks (GANs). This paper covers the first NLP-driven scientific literature analysis for the built environment.

Keywords:
Reference mining, natural language processing, data science, built environment, building energy efficiency, word embeddings

1. Introduction

With advances in information technology, buildings today are collecting ever-larger amount of real-time data from various heterogeneous sources [13, 80]. The huge amount of data also have led to increased data awareness and data science applications [95]. These innovations have led to an explosion of research in this field, resulting in thousands of publications in this area (e.g. Figure 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 [38]. 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.

[*] Corresponding authors
Email address: adrian.chong@nus.edu.sg (Adrian Chong)

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

In order to address this challenge, the concept of using data-driven 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 [74]. These articles, mostly, pass through different rounds of peer-review processes. The peer-review processes are conducted by one or more people known to be knowledgeable enough of the field. The aim of the peer-review process is to ensure the high quality of publications regarding content (e.g. context, significance, objectives, novelty, data, and results), organization, and language (proof-reading). This process ensures that the final product is well-written and well-structured (in terms of natural language), and contains the correct information in the form of text. Thanks to the recent advancements in Natural Language Processing (NLP), it has been viable to extract knowledge from large corpus of such structured text. Normally, information from the literature is extracted via traditional narrative literature review/survey of a finite number of articles (hundreds) [90]. Besides other methods such as questionnaire surveys and expert interviews. However, there are some challenges in applying these methods on large scale [7]. 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 different data. The same challenge applies to expert interview and questionnaires.

1.2. Background

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 [47, 105, 57, 20, 46, 127, 66]. Wang and Srinivasan explored the use of single versus ensemble-based models for building energy prediction [147]. Roth et. al explored the use of various data-driven techniques in the context of benchmarking building [129]. Colm et. al. [49] investigated Machine learning methods for maximizing measurement and verification (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 [92]. Fault detection and diagnosis (FDD) is another important application of improving building energy performance, where data science methods are commonly used [160]. 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 [127]. Beyond single buildings, data-driven methods are also useful for demand response and smart grid applications [48, 108, 45]. On the urban scale, energy efficiency applications also have grasped interest of researchers. For example, many researchers investigated district heating and cooling systems [85, 124] and Urban Building Energy Modelling (UBEM) [10, 62, 123]. 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 [126], biomedicine [137], and frameworks have been built for more general text mining purposes [141].

Different tools and approaches of text-mining have been used in literature to conduct literature reviews. [22, 39, 132, 152] used VOSviewer [143] to create a bibliometric networks and density map between articles in different fields. Other researchers used CiteNetExplorer [144] to track the citation relations across articles in scientific research [36, 136] among others. Other tools such as CiteSpaceII, BibExcel, SciMAT,Sci^2 Tool have been extensively reviewed by [106]. 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 metadata of the articles (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 [91], Glove [116], Python/scikit-learn [115], word2Vec [98, 96, 99] to develop a model that performs a specific task.

The objective of this paper is to address the challenges and the 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 data-driven 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 Figure 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 Figure 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:

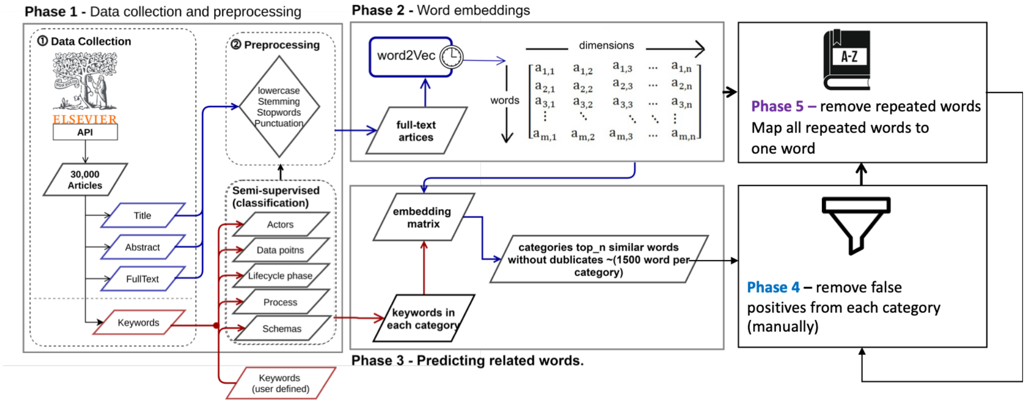

Phase 1 - Data collection and preprocessing
① Data Collection
ELSEVIER
API
30,000 Articles
Title
Abstract
FullText
Keywords
② Preprocessing
lowercase
Stemming
Stopwords
Punctuation
Semi-supervised (classification)
Actors
Data poitns
Lifecycle phase
Process
Schemas
Keywords (user defined)
Phase 2 - Word embeddings
word2Vec
full-text articles
dimensions
words
$a_{1,1}$ $a_{1,2}$ $a_{1,3}$ ... $a_{1,n}$
$a_{m,1}$ $a_{m,2}$ $a_{m,3}$ ... $a_{m,n}$
Phase 5 – remove repeated words Map all repeated words to one word
embedding matrix
categories top_n similar words without dublicates ~(1500 word per category)
keywords in each category
Phase 4 – remove false positives from each category (manually)
Phase 3 - Predicting related words.

Figure 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

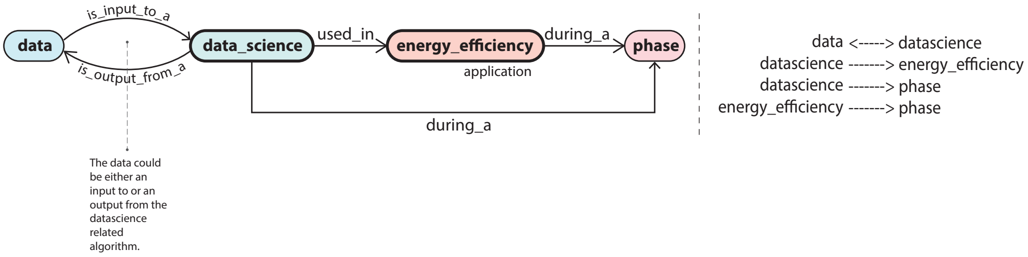

is_input_to_a
is_output_from_a
data
data_science
used_in
energy_efficiency
application
during_a
phase
during_a
The data could be either an input to or an output from the datascience related algorithm.
data <-----> datascience
datascience -------> energy_efficiency
datascience -------> phase
energy_efficiency -------> phase

Figure 2: Overview of the categories of concepts analysed in this data-driven analysis and their relationships with each other

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 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.`

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 organization. They first introduced an API for the public for text-mining research in 2014 [145]. 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. *Data collection*

Elsevier text mining API [1] provides a rich amount of methods for querying and retrieval. For example, the query can include Boolean operators such as `AND, OR,` and `NOT`. Also, it allows retrieving articles by their Document Object Identifier (DOI), Publication Item Identifier (PII), Electronic Identifier (EID) and others. Additionally, it supports searching for articles by keywords that appear in the title. Furthermore, it is possible to restrict the query to a specific journal name and/or the year of publication. This method was used to collect all the possible articles that contain the query keywords.

2.2.2. *Article filtering*

The initial query process has resulted in 45,000 articles from more than 1,000 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. Figure 3 illustrates the top number of papers per journal, and the number of published paper 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

[1] https://dev.elsevier.com/documentation/FullTextRetrievalAPI.wadl

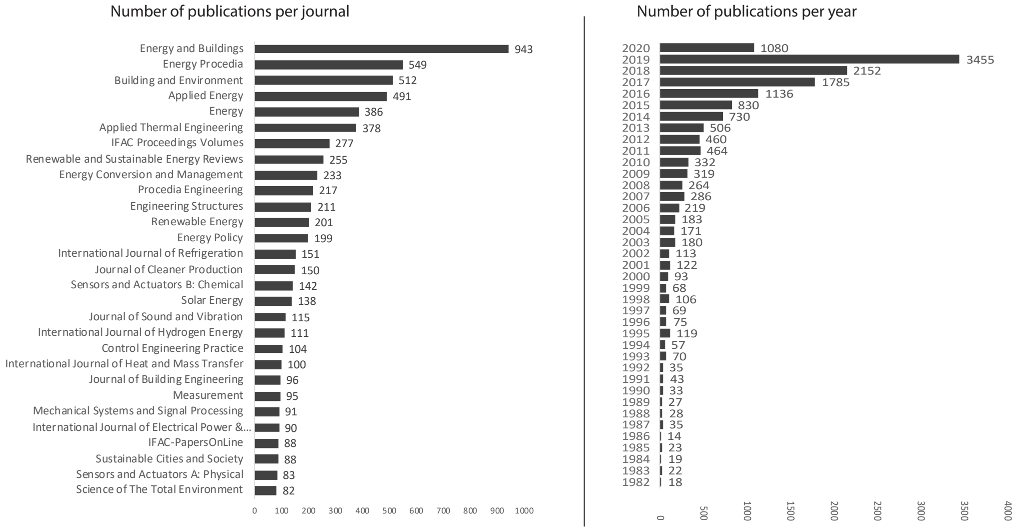


Figure 3: Number of collected papers per journal and per year.

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) [91].

Secondly, Since the NLP models are case-sensitive, they need to be consistent. For example, lowercasing letters throughout the corpus. Also, converting regular plural nouns into singular ones by removing ”s”. There are many other set 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 (Figure 4). Firstly, the common root of different words were 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.

To prepare the articles as an input for the NLP text mining algorithms, they should be combined into a one corpus of text. This process includes removing the line-breaks, indentations, and other punctuation marks in addition to removing tables, equations, and other non-paragraph-like text. By doing this, the text is now ready for the text mining model.

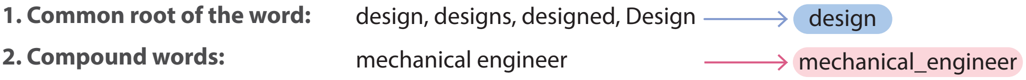


Figure 4: Each word can have one or more similar synonyms which are mapped to the original word. The figure shows two types of text stemming and lemmatization.

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 [96, 97]. 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 process of assigning a vector to each word starts by tokenizing each word from the full text. Tokenizing means that each unique word is assigned to a unique on-hot-encoder. Our full text consists of about 39,000 unique words, these words appears in a frequency range between 5 to 680,000 times.

The Word2Vec training process aims to predict a word (known as the central word) from the context within which this word falls (context words)[96, 51]. 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). 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 hyper-parameter fine-tuning. In our case, we assigned a vector of 300 dimensions to each word. The architecture of the word2vec model is illustrated in figure 5.

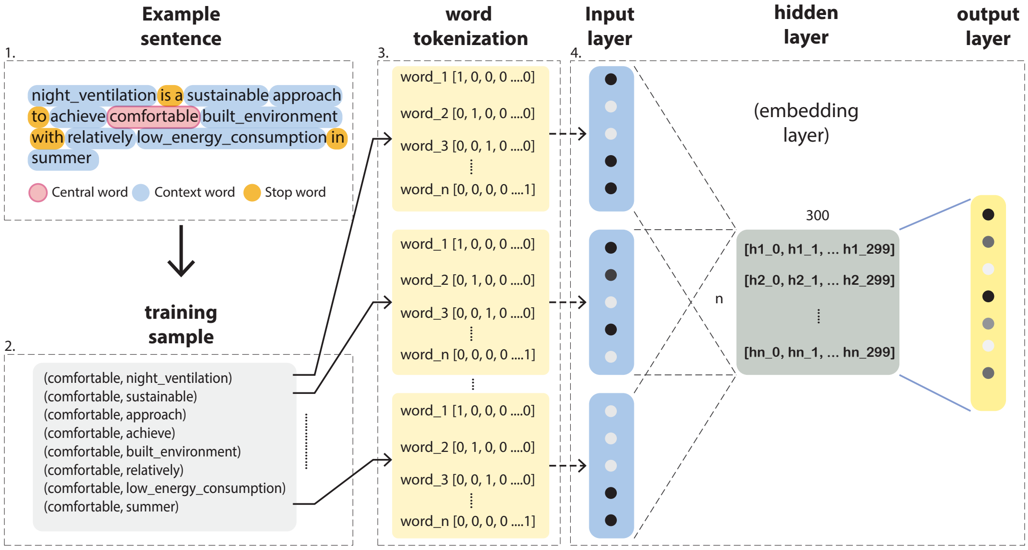

Example sentence
1.
night_ventilation is a sustainable approach to achieve comfortable built_environment with relatively low_energy_consumption in summer
Central word
Context word
Stop word
training sample
2.
(comfortable, night_ventilation)
(comfortable, sustainable)
(comfortable, approach)
(comfortable, achieve)
(comfortable, built_environment)
(comfortable, relatively)
(comfortable, low_energy_consumption)
(comfortable, summer)
word tokenization
3.
word_1 [1, 0, 0, 0 ....0]
word_2 [0, 1, 0, 0 ....0]
word_3 [0, 0, 1, 0 ....0]
word_n [0, 0, 0, 0 ....1]
Input layer
4.
hidden layer
(embedding layer)
300
n
[h1_0, h1_1, ... h1_299]
[h2_0, h2_1, ... h2_299]
[hn_0, hn_1, ... hn_299]
output layer

Figure 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.

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.

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 [84]. 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:

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

Where $\bar{S}(W_a, W_b)$ is the average similarity between two lists of words W_a and W_b and their synonyms $W_a = [w_{a_1}...w_{a_n}]$, $W_b = [w_{b_1}...w_{b_n}]$. 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:

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

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 $\{W_a, W_b, \bar{S}(W_a, W_b)\}$ which is then converted into a directed weighted graph. The results will be explained in the following section 3.

3. Results

The methodology outlined a process of using data-driven 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

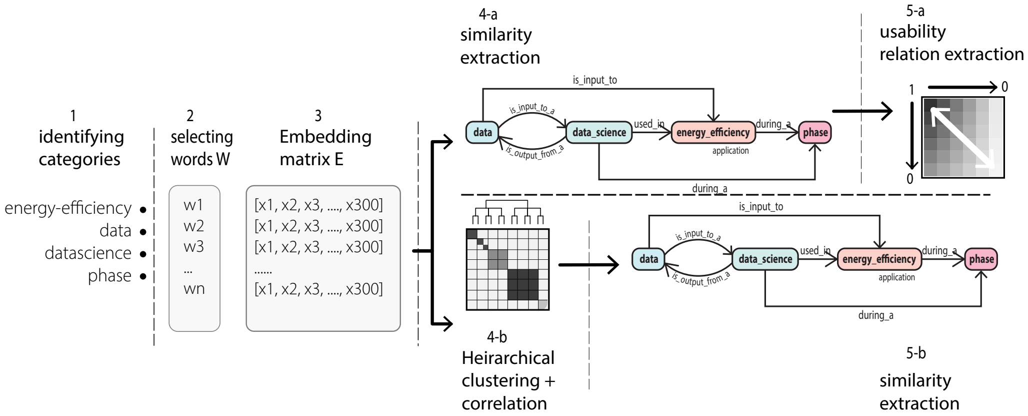


Figure 6: Overview of the ways of showcasing the results of the data-driven process. 1) Identifying 4 categories and 2) assigning the corresponding words under each category. After that, 3) the embedding vector of each word is extracted. Then, there are two main approaches: a) is the usability relation 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.

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. Figure 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. Figure 7 illustrates this situation by showing the embedding vector of words projected into a two-dimensional space. The key words are categorized according to the four dimensions of the analysis: 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.

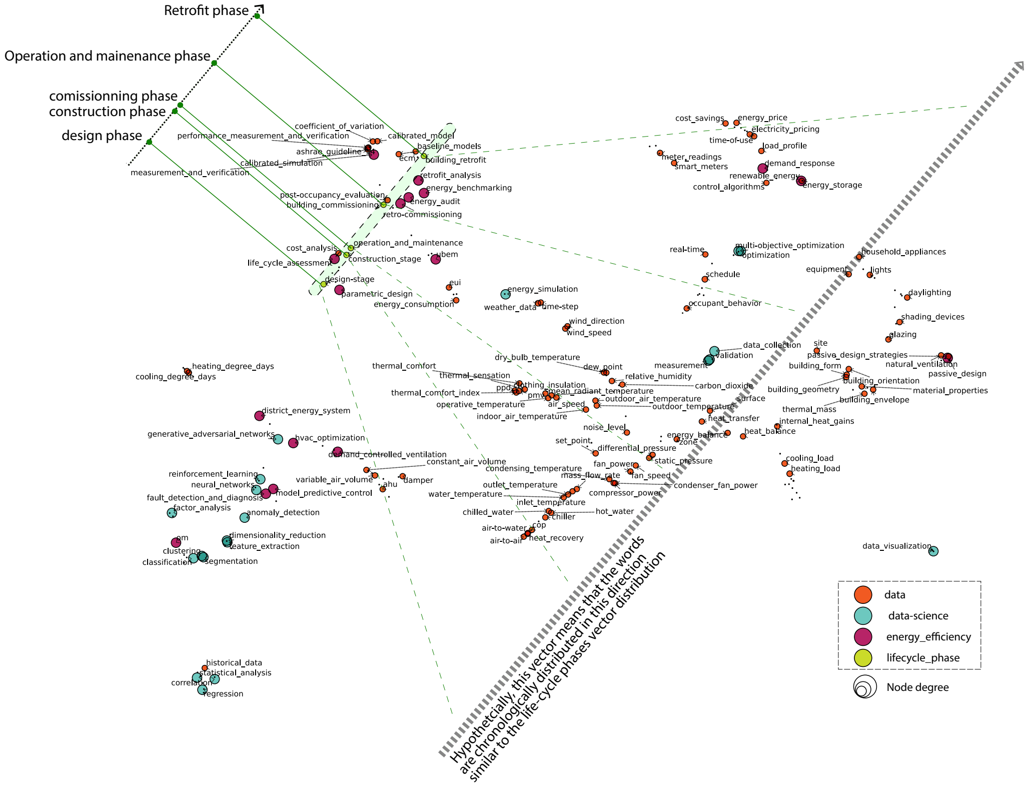


Figure 7: The vector representation of the words from each category. These words are located based on their embedding vector. The embedding vector of each word is dimensionally reduced from 300 dimensions to 2 dimensions for the sake of visualization. The euclidean distance between words indicates the semantic similarity between these words. The degree of a specific node refers to the number of nodes connected to that specific node

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. Figure 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 average strength of relation for each data source as compared to the applications.

It can be observed that data are largely used 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 uses 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.

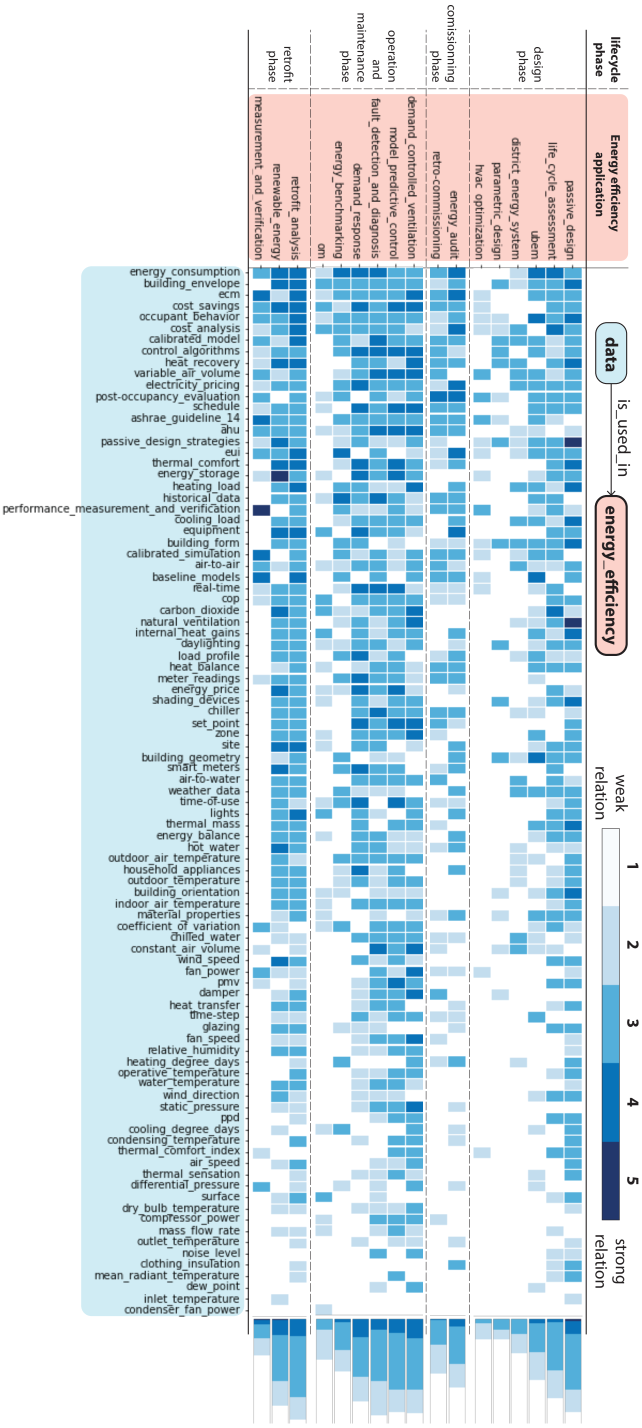


Figure 8: The relation between data points and different energy-efficiency applications. The energy efficiency applications (shown on the Y-axis with red highlight) are chronologically grouped based on the building life-cycle phases. The data-points (shown on the X-axis blue highlight) are sorted based on their appearance frequency

Figure 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. Figure 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 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 strong relation to data sources. These relations have different 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.

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 histogram (Figures 10 and 11). The relations across each two distinct categories are extracted using the cosine similarity between pairs of words from each category .

3.3.1. Hierarchical agglomerative clustering of concepts

The HAC has been applied for words in each distinct category using the Ward's method [107]. On the one hand, Figure 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 commissioning phase. However, the data_science category has been clustered into 5 different groups/subgroups. These are, Machine Learning **(ML)**, Deep Learning **(DL)**, Data pre/post-processing **(PP)**, Optimization **(OP)**,

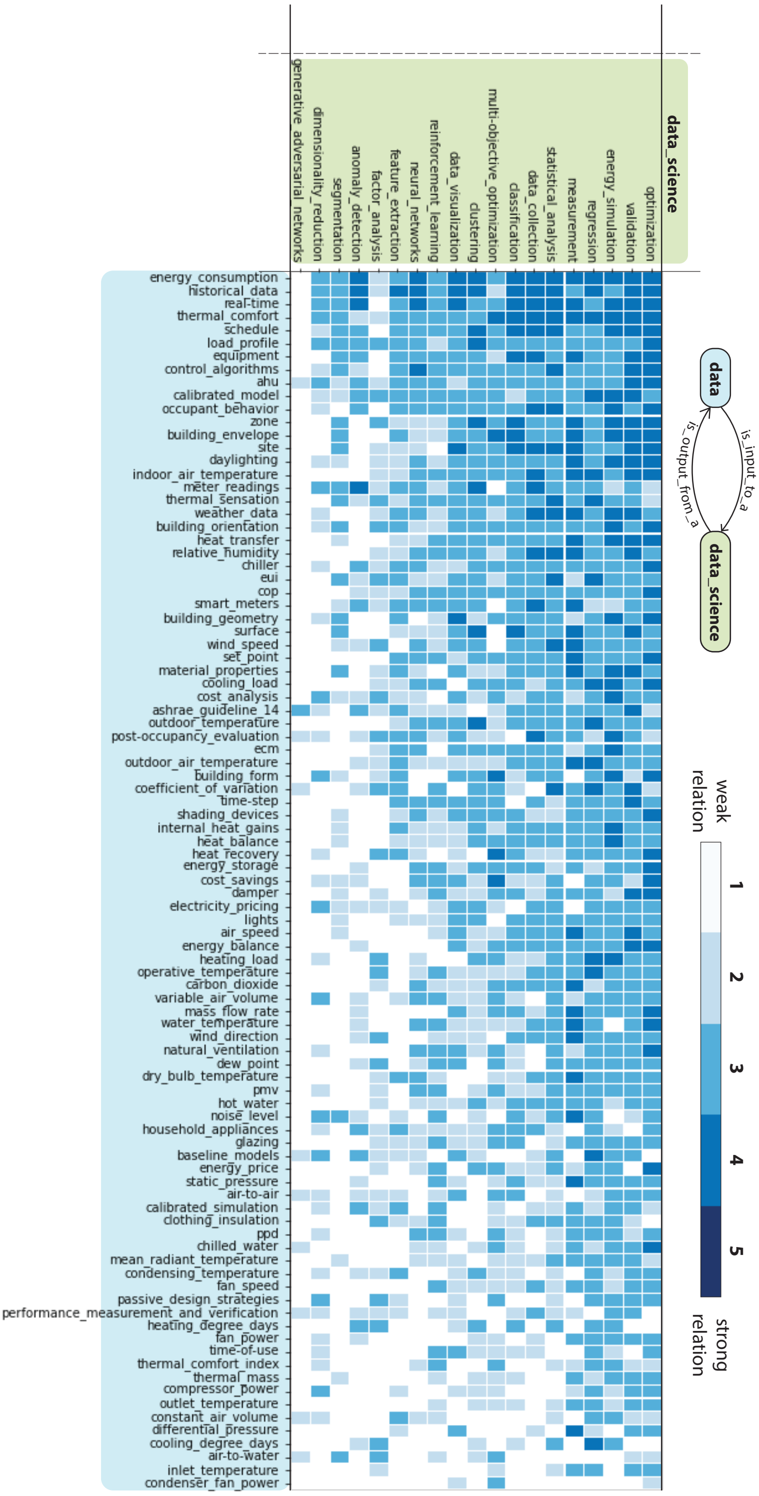


Figure 9: The relation between data points (shown on the X-axis with blue highlight) and data-science algorithms (shown on the Y-axis with green highlight). Both of them are sorted based on the sum of the similarity per each row/column

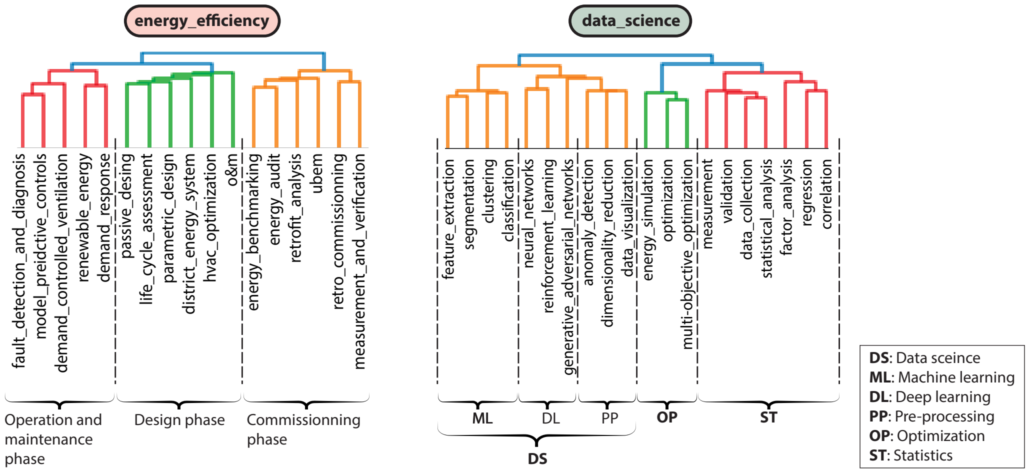


Figure 10: The hierarchical agglomerative clustering (HAC) of the energy efficiency and data science categories using Ward's method.

and Statistical methods **(St)**. On the other hand, Figure 11 the HAC of the data category (on the top) and the correlation matrix between words (the heat map on the bottom). There are ten different groups can be observed in from this matrix:

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.

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.

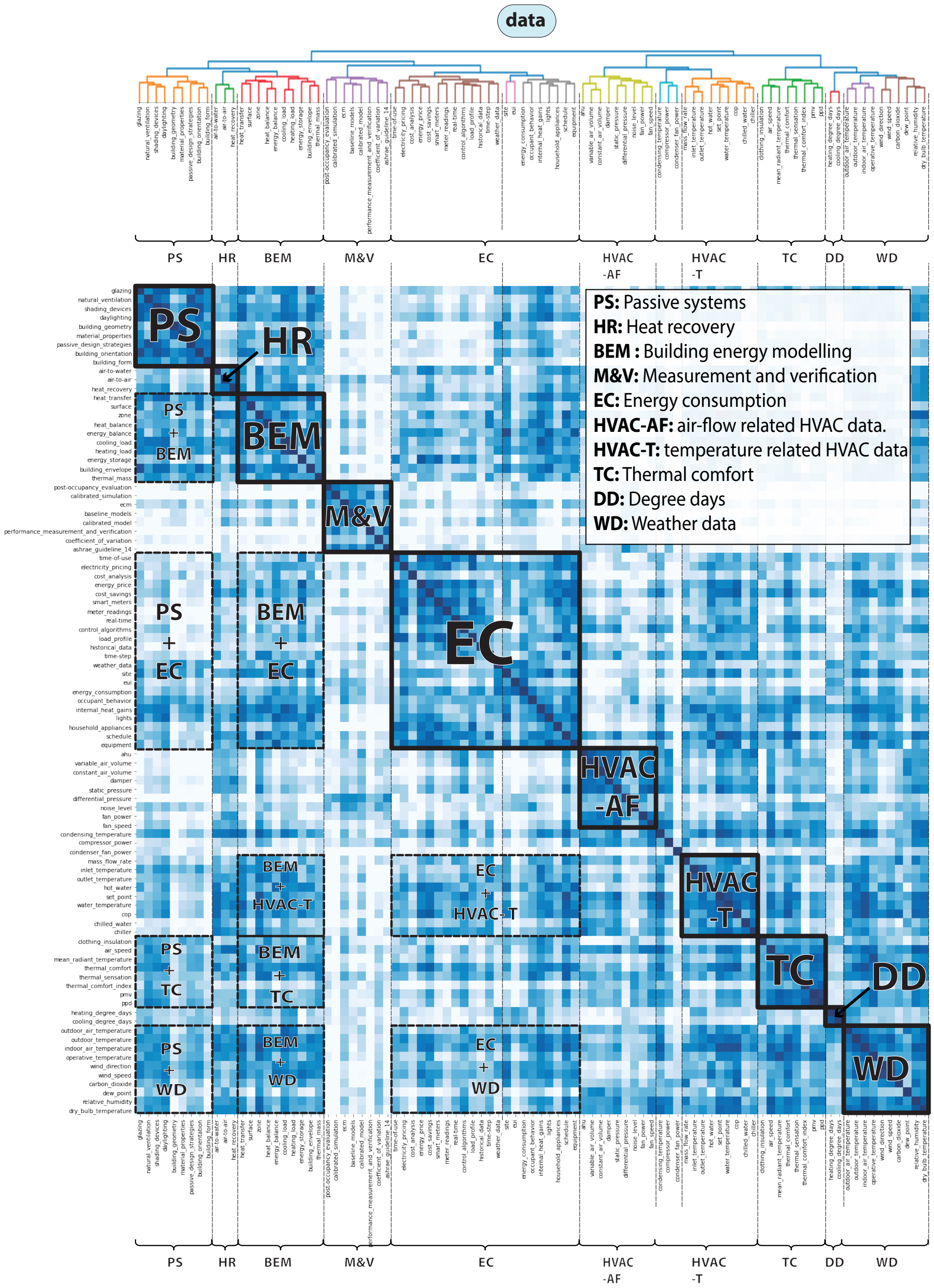


Figure 11: A hierarchically clustered correlation matrix is used to group similar data sources together and to extract the correlation between words

9. **Weather data (WD)** such as wind speed and direction, CO2, dew point, temperature and relative humidity.

From this correlation matrix, areas of intersection between different data can be identified. For example, the Energy consumption (EC) group has a high relationship with most of the other groups such as PS, BEM, HVAC, and WD. Also BEM data has high relationship with PS, EC, HVAC, TC, and WD. Similarly, the HR group exhibits a high relationship with most of the other groups. However, the M&V group shows the lowest relationship with other groups except with the HVAC airflow data as well as some EC-related data-points. The intra-relationships among data-points can possibly reveal useful patterns specially when compared to the data-science, energy efficiency, and phases categories. The intra-correlation matrix of the energy-efficiency category also shows interesting patterns.

3.3.2. Comparison of clustering across the four categories

The final visualization in this section focuses on converging the four categories with the clustering techniques (Figure 12).

It can be observed that the data-science category is grouped into five clusters. These clusters are: 1) Machine Learning Algorithms **(ML)** including classification, clustering, segmentation, and feature extraction. 2) Deep learning algorithms **(DL)** such as neural networks, reinforcement learning **(RL)**, and generative adversarial networks (GANs). 3) Data Pre/Post processing **(PP)** including anomaly detection, dimensionality reduction, and data visualization. Those three clusters can fall under Data Science **(DS)** category. 4) Optimization **(OP)** including single-objective and multi-objective optimization as well as energy simulation. Finally, 5) Statistics **(ST)** to which statistical-related methods belong. For example, data collection, factor analysis, correlation, and regression. Although these clusters are distinct from each other, there is still overlapping among them.

Passive systems **(PS)** data exhibit high relationships in the design phase energy applications such as passive design, lifecycle assessment, and parametric design. It is also used for retrofit analysis, and urban energy modelling (UBEM). Heat recovery **(HR)** data are also highly used during the operation phase specially for demand-controlled ventilation, demand response, and fault detection and diagnosis. It is also used for the passive design, as well as district energy systems. Furthermore, BEM data are extensively used in demand response, renewable energy, passive design, and retrofit analysis applications. Unsurprisingly, Energy consumption **(EC)** data showed high correlation with operation and maintenance lifecycle phase energy-efficiency applications. For example, demand-response is highly correlated with time of use, electricity prices, cost saving, control algorithms, load profile, and equipment; renewable energy resources is also highly correlated with energy prices, smart meters, energy consumption data, and equipment; furthermore, model predictive controls are highly correlated with real-time data, control algorithms, and schedules. During the commissioning phase, energy benchmarking exhibits high correlation with energy use intensity (EUI) and energy consumption data. However, in the design phase, the EC data seems not to be adequately utilized in HVAC optimization, district energy systems, and O&M. This could be attributed to the fact that these data are generated from a building during its operation. The figure also shows that **HVAC** data are used intensively during the operation phase, more specifically, in demand controlled ventilation, model predictive controls, and fault detection and diagnosis. However, in the design phase, it is used in HVAC optimization, and district energy systems. Moreover, in the commissioning phase, HVAC data are mainly used in energy audit, and retrofit analysis. On the other hand, thermal comfort **(TC)** data are highly

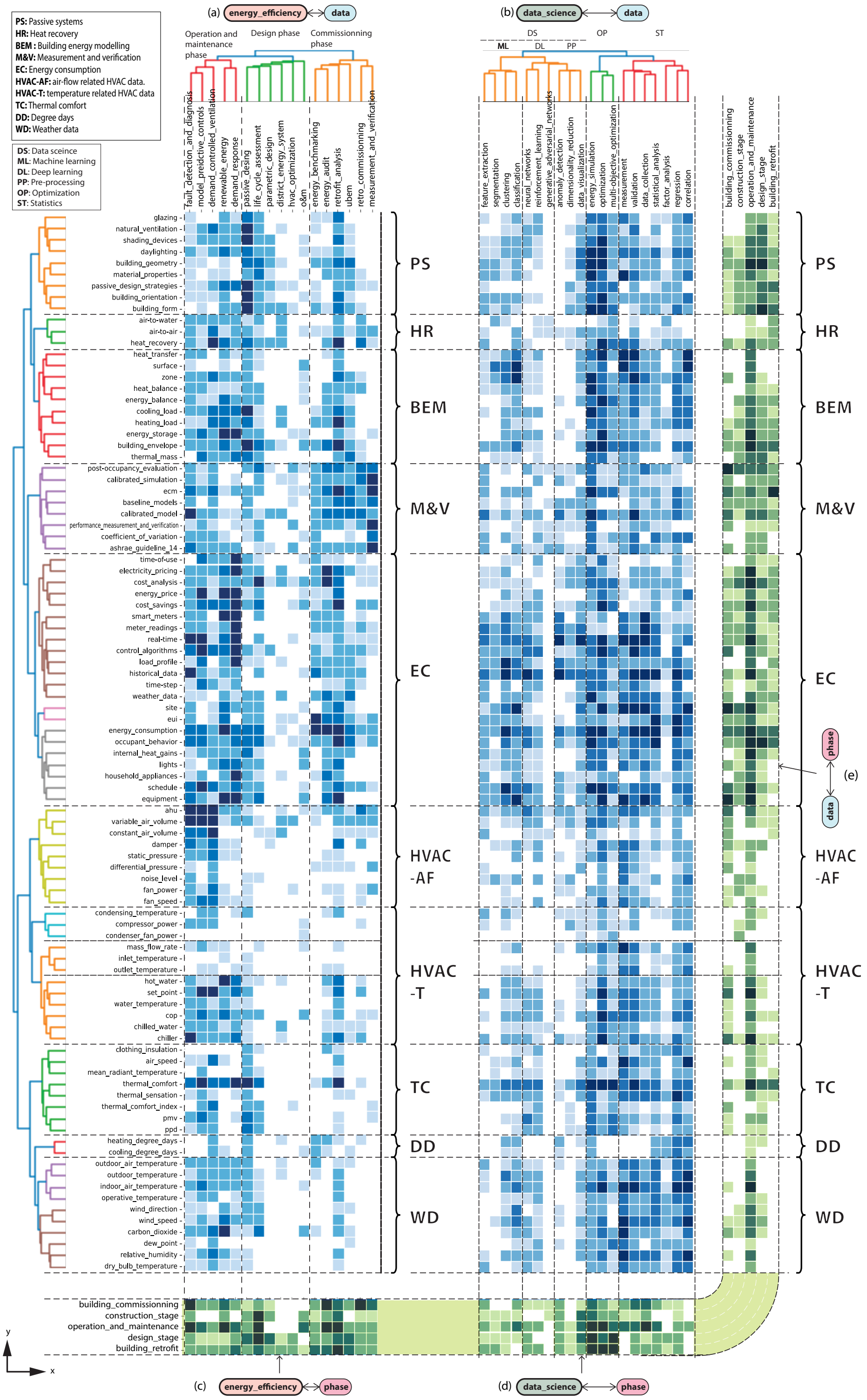


Figure 12: This figure shows the relation between different pairs of categories after performing HAC within each category. The blue heat-maps show relations between a) data and energy_efficiency; and b) data and data_science. The green heat-maps show relations between c) phase and energy_efficiency; d) phase and data_science; and e) phase and data

correlated with different energy efficiency applications during the operation phase. However, it is used in parametric design, and lifecycle assessment in the design phase.

It can be observed that optimization category **(OP)** is widely used in most of the energy-efficiency applications. This OP includes single objective and multi-objective optimization, as well as energy simulation. Data pre/post-processing **(PP)** is also widely used in most of the energy-efficiency applications. On the other hand, energy efficiency applications in the operation-and-maintenance and the commissioning phases show high relation with most of the data-science methods. However, the design phase energy-efficiency application don't make much use of the data-science methods.

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.

Figure 13 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 Figure 13 are intuitively those related to the traditional building energy domain techniques of `simulation`, `optimization`, `neural networks`, `reinforcement learning`, and `statistical analysis` (See Figure 14 - the black bold keywords).

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 [34] and optimization engines such as BEopt [28], GenOpt [150], and jEplus [158]. 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 [133] and R [72]. As illustrated in Figure 13 and Figure 14, 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 [68]. Crawley et al. provides an overview into the capabilities of twenty major building performance simulation programs [33]. Despite its long standing history and developments, challenges remain leading to opportunities in research and development. Hong, Langevin and Sun lists the ten BPS challenges [64]. Table 1 lists each of these ten challenges. Additionally, we include the relevant publications and existing open-source repositories and data-sources that forms the foundation in addressing these challenges.

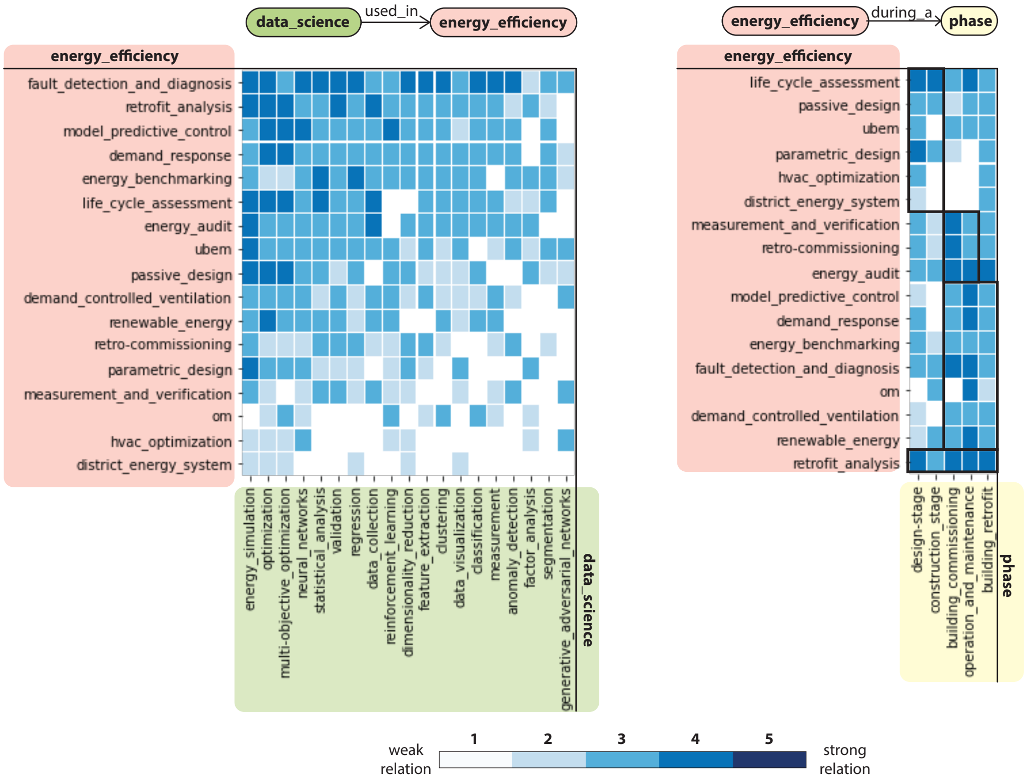


Figure 13: The figure on the left shows the relation between data-science algorithms and energy-efficiency applications sorted based on usability. On the right, the relation between energy-efficiency applications and life-cycle phases are illustrated. The heatmaps' axes are all sorted based on strength correlation except life-cycle phase which is in a chronological order

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 1990's as a mean 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 [79]. Although matured, there have been recent developments in AFDD as a result of advancements in Artificial Intelligence techniques [160] and anomaly detection [117, 118]. 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 [155] and chiller [154] AFDD methods. However, amongst the data science techniques listed in Figure 13, 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

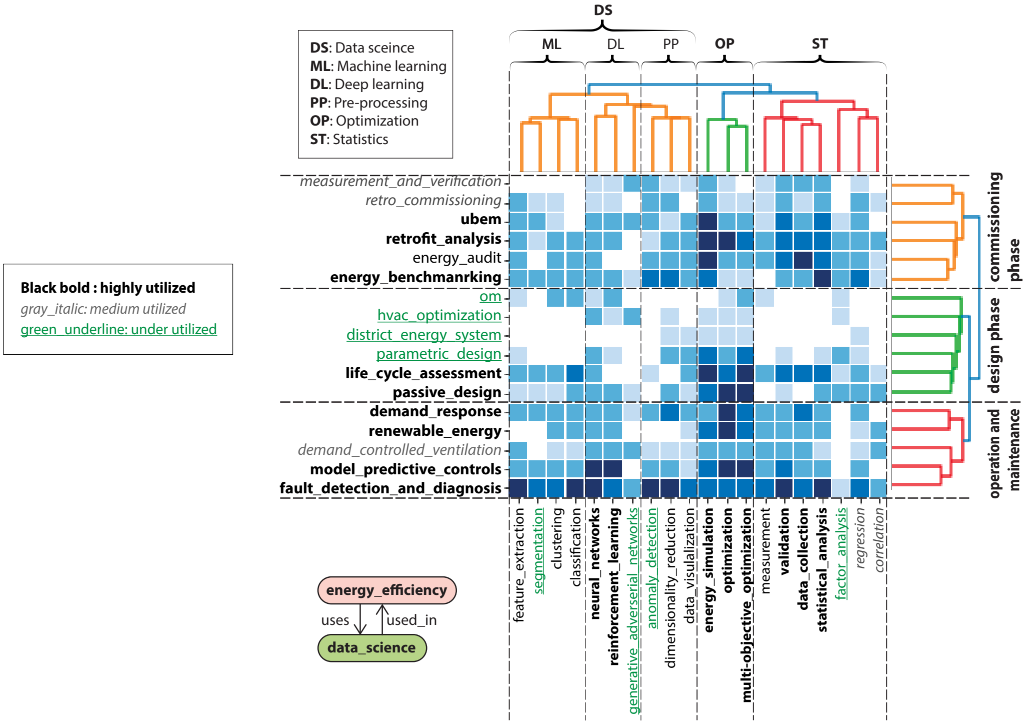


Figure 14: Finding relationships between application and data science techniques: the HAC relation between data science (X-Axis) and energy efficiency applications (Y-Axis). The black bold text refers to keywords that are highly utilized applications, gray italic text refers to medium utilized applications, green underlined text refers to underutilized applications.

for generating balanced dataset [119, 120]. Additionally, GANs recently has shown promising results in semi-real-time simulation of urban solar radiation simulation as well as urban wind simulation using Pix2Pix [30].

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 [9]. As shown in Figure 13, 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 [59]. 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 [60]. 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 [26, 149].

Model predictive control (MPC) has gained traction in the last two decades through

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 [37]; Causes [142]; Credibility gap [15]	ObepME [78]; WinProGen (Occupant-behaviour gap) [19]	[69, 104, 19]
2. Modeling human-building interactions	Occupant modeling methodology in BPS [153] Challenges and opportunities [112]	Buildings.Occupants [148]	Occupant behavior [67]
3. Model calibration	Calibration methods and techniques [32, 122, 43]; Sensitivity analysis [140]	Bayesian calibration [25, 121]; Optimization [21]	OpenStudio Calibration examples [109]
4. Modeling operation, controls and retrofits	Retrofit toolkits [86, 87]; Model based commissioning [146]	Crowd-sourced ML for buildings [102]	Large, open meter data [103]
5. Modeling operational faults	Energy performance optimization [14]	Openstudio Fault Models Gem [24]	Open fault detection data [52]
6. Zero-net-energy and grid-responsive buildings	Gaps and needs [12, 83]; Grid-responsive buildings [110]	BeOpt [29]	NZEB occupant behaviour [82]; Watts per person[113]
7. Urban-scale building energy modeling	Modeling methodology and workflows [123]; Challenges and future opportunities [63]	PyCity [134, 138, 1]	City buildings dataset [23] SynCity[130], NYC-UBEM [128]
8. Evaluating the energy-saving potential of building technologies at national or regional scales	E3 [94], Building stock energy prediction [88]	INTERDYME [8], PortableDyme [54], Scout [111]	Open data use for city-wide benchmarking [129]
9. Modeling energy efficient technology adoption	[42, 70, 50]	N.A.	Air-Conditioning Heating and Refrigeration Institute (AHRI) open data [58]
10. Integrated modeling and simulation	Progress, prospects, and requirements [31]	IFC[18], GBXML[40], OpenFOAM[71], EnergyPlus[34],Co-simulation e.g. obFMU[65]	Physics-based and data-driven modelling for NYC [130]

numerous case study-based implementations [4]. 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 [6]. Neural network models are becoming more popular due to its stronger modeling capability [5]. 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 [41]. Over the few past years, RL is becoming a major competitor of MPC with its advantages of lower requirements on predictive model and better adaptability [159]. 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.

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 [125]. 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 [77]. 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 [27]. Also, any 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 [89]. Besides, at 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 [73].

`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 [11] and even redefining the way buildings are categorized for benchmarking [114, 157]. The large increase in open data sets available has created opportunities to target specific strategies to cities based on their specific needs [156] and using a combination physics-based and data-driven methods [130]. Data-driven improvements have been suggested related to generalizability [100] and interpretability [101].

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. [75, 76] 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 on both outdoor weather as well as control and social behaviour of consumers [55].

`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. [135] 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 [56, 16, 44]. 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 [61]. 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) [53]. 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 [2] Lunchbox and OWL [81]; Optimization and multi-objective optimization such as OPOSSUM [151], octopus, Galapagos, and Optimus [35]; Energy Modelling such as Ladybug tools, [131], BuildFit [3]; Data visualization and deep learning using Gh_CPython [93]. These tools have grasped the attention of a large body of researchers and end-use programmers recently and may have a great potentials 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. The aim of this process is to review the datascience methods that are used in different building energy efficiency applications through 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 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, 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). Secondly, 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 a energy efficiency applications. Future directions include improving the model so that it results in a directional graph (di-garph). This digraph can be drawn out using relation extraction models based part-of-speech and stop-words. Also, adding high performance pre-trained models that are based on transformers such as BERT [139] and Generative Pre-trained Transformer 3 (GPT-3) [17] can be useful 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

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 M. Abdelrahman and Ahmed Mohamed Yousef Toutou. ANT: A Machine Learning Approach for Building Performance Simulation: Methods and Development. *The Academic Research Community publication*, 3(1):205, 2019. doi:10.21625/archive.v3i1.442.

[3] 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.

[4] Abdul Afram and Farrokh Janabi-Sharifi. Theory and applications of hvac control systems–a review of model predictive control (mpc). *Building and Environment*, 72:343–355, 2014. doi:https://doi.org/10.1016/j.buildenv.2013.11.016.

[5] Abdul Afram, Farrokh Janabi-Sharifi, Alan S Fung, and 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 and Buildings*, 141:96–113, 2017. doi:10.1016/j.enbuild.2017.02.012.

[6] Zakia Afroz, GM Shafiullah, Tania Urmee, and Gary Higgins. Modeling techniques used in building hvac control systems: A review. *Renewable and Sustainable Energy Reviews*, 83:64–84, 2018. doi:10.1016/j.rser.2017.10.044.

[7] 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. doi:https://doi.org/10.1177/1094428120943281.

[8] Clopper Almon Jr. Interdyme: A package of programs for building interindustry dynamic macroeconomic models, version 3.10. *Department of Economics, University of Maryland*, 2000.

[9] 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 policy*, 35(3):1819–1829, 2007. doi:https://doi.org/10.1016/j.enpol.2006.05.018.

[10] Yu Qian Ang, Zachary Michael Berzolla, and Christoph F Reinhart. From concept to application: A review of use cases in urban building energy modeling. *Applied Energy*, 279:115738, 2020. doi:https://doi.org/10.1016/j.apenergy.2020.115738.

[11] Pandarasamy Arjunan, Kameshwar Poolla, and Clayton Miller. EnergyStar++: Towards more accurate and explanatory building energy benchmarking. *Appl. Energy*, 276:115413, October 2020. doi:https://doi.org/10.1016/j.apenergy.2020.115413.

[12] Shady Attia, Mohamed Hamdy, William O'Brien, and Salvatore Carlucci. Assessing gaps and needs for integrating building performance optimization tools in net zero energy buildings design. *Energy and Buildings*, 60:110–124, 2013. doi:10.1016/j.enbuild.2013.01.016.

[13] Muhammad Rizwan Bashir and Asif Qumer Gill. Towards an IoT big data analytics framework: Smart buildings systems. *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. doi:10.1109/HPCC-SmartCity-DSS.2016.0188.

[14] 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. doi:10.1063/1.5053110.

[15] Bill Bordass, William Bordass Associates, and Robert Cohen. Energy Performance of Non-Domestic Buildings: Closing the Credibility Gap. *8th International Conference on Improving Energy Efficiency in Commercial Buildings*, pages 1–10, 2004. URL: https://www.buildup.eu/en/node/1900.

[16] Silvio Brandi, Marco Savino Piscitelli, Marco Martellacci, and Alfonso Capozzoli. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy Build.*, 224:110225, October 2020. doi:https://doi.org/10.1016/j.enbuild.2020.110225.

[17] 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.

[18] BuildingSMART. Industry Foundation Classes (IFC) - buildingSMART Technical. 2020. URL: https://technical.buildingsmart.org/standards/ifc.

[19] Davide Calì, Mark Thomas Wesseling, and Dirk Müller. WinProGen: A Markov-Chain-based stochastic window status profile generator for the simulation of realistic energy performance in buildings. *Building and Environment*, 136:240–258, 2018. doi:10.1016/j.buildenv.2018.03.048.

[20] Debaditya Chakraborty and Hazem Elzarka. Advanced machine learning techniques for building performance simulation: a comparative analysis. *Journal of Building Performance Simulation*, 12(2):193–207, 2019. doi:https://doi.org/10.1080/19401493.2018.1498538.

[21] Gaurav Chaudhary, Joshua New, Jibonananda Sanyal, Piljae Im, Zheng O'Neill, and Vishal Garg. Evaluation of "autotune" calibration against manual calibration of building energy models. *Applied energy*, 182:115–134, 2016. doi:https://doi.org/10.1016/j.apenergy.2016.08.073.

[22] 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. doi:https://doi.org/10.1016/j.autcon.2018.05.028.

[23] Yixing Chen, Tianzhen Hong, Xuan Luo, and Barry Hooper. Development of city buildings dataset for urban building energy modeling. *Energy and Buildings*, 183:252–265, 2019. doi:10.1016/j.enbuild.2018.11.008.

[24] 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.

[25] Adrian Chong and Kathrin Menberg. Guidelines for the bayesian calibration of building energy models. *Energy and Buildings*, 174:527–547, 2018. doi:https://doi.org/10.1016/j.enbuild.2018.06.028.

[26] Adrian Chong, Weili Xu, Song Chao, and Ngoc-Tri Ngo. Continuous-time bayesian calibration of energy models using bim and energy data. *Energy and Buildings*, 194:177–190, 2019. doi:https://doi.org/10.1016/j.enbuild.2019.04.017.

[27] Despoina Christantoni, Simeon Oxizidis, Damian Flynn, and Donal P Finn. Implementation of demand response strategies in a multi-purpose commercial building using a whole-building simulation model approach. *Energy and Buildings*, 131:76–86, 2016. doi:10.1016/j.enbuild.2016.09.017.

[28] C Christensen, R Anderson, S Horowitz, A Courtney, and J Spencer. Beopt(tm) software for building energy optimization: Features and capabilities. 8 2006. doi:10.2172/891598.

[29] Craig Christensen, Adam Courtney, Scott Horowitz, Todd Givler, and Greg Barker. Beopt: Software for identifying optimal building designs on the path to zero net energy. *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.

[30] 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.

[31] J. A. Clarke and J. L.M. Hensen. Integrated building performance simulation: Progress, prospects and requirements. *Building and Environment*, 91:294–306, 2015. doi:10.1016/j.buildenv.2015.04.002.

[32] Daniel Coakley, Paul Raftery, and Marcus Keane. A review of methods to match building energy simulation models to measured data. *Renewable and sustainable energy reviews*, 37:123–141, 2014. doi:https://doi.org/10.1016/j.rser.2014.05.007.

[33] Drury B Crawley, Jon W Hand, Michaël Kummert, and Brent T Griffith. Contrasting the capabilities of building energy performance simulation programs. *Building and environment*, 43(4):661–673, 2008. doi:https://doi.org/10.1016/j.buildenv.2006.10.027.

[34] 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, and Jason Glazer. EnergyPlus: Creating a new-generation building energy simulation program. *Energy and Buildings*, 33(4):319–331, 2001. doi:10.1016/S0378-7788(00)00114-6.

[35] Cemre Cubukcuoglu, Berk Ekici, Mehmet Fatih Tasgetiren, and Sevil Sariyildiz. OPTIMUS: Self-Adaptive Differential Evolution with Ensemble of Mutation Strategies for Grasshopper Algorithmic Modeling. *Algorithms*, 12(7):141, 2019. doi:10.3390/a12070141.

[36] Hanna De Vries, Victor Bekkers, and Lars Tummers. Innovation in the public sector: A systematic review and future research agenda. *Public administration*, 94(1):146–166, 2016. doi:https://doi.org/10.1111/padm.12209.

[37] Pieter De Wilde. The gap between predicted and measured energy performance of buildings: A framework for investigation. *Automation in construction*, 41:40–49, 2014. doi:https://doi.org/10.1016/j.autcon.2014.02.009.

[38] 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, doi:https://doi.org/10.1016/j.eswa.2007.01.035.

[39] Krisztina Demeter, Levente Szász, and Andrea Kő. A text mining based overview of inventory research in the isir special issues 1994–2016. *International Journal of Production Economics*, 209:134–146, 2019. doi:https://doi.org/10.1016/j.ijpe.2018.06.006.

[40] Vanda Dimitriou, Steven K Firth, Tarek M Hassan, and Farid Fouchal. BIM enabled building energy modelling: development and verification of a GBXML to IDF conversion method. *Ibpsa*, pages 12–14, 2016. URL: https://dspace.lboro.ac.uk/2134/22818.

[41] 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. doi:10.1016/j.arcontrol.2020.09.001.

[42] 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.

[43] Enrico Fabrizio and Valentina Monetti. Methodologies and advancements in the calibration of building energy models. *Energies*, 8(4):2548–2574, 2015. doi:https://doi.org/10.3390/en8042548.

[44] 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. doi:https://doi.org/10.1016/j.apenergy.2018.04.118.

[45] Cheng Fan, Yongjun Sun, Yang Zhao, Mengjie Song, and Jiayuan Wang. Deep learning-based feature engineering methods for improved building energy prediction. *Appl. Energy*, 240:35–45, April 2019. doi:https://doi.org/10.1016/j.apenergy.2019.02.052.

[46] Cheng Fan, Fu Xiao, Zhengdao Li, and Jiayuan Wang. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings*, 159:296–308, 2018. doi:https://doi.org/10.1016/j.enbuild.2017.11.008.

[47] Cheng Fan, Da Yan, Fu Xiao, Ao Li, Jingjing An, and Xuyuan Kang. Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches. *Build. Simul.*, October 2020. doi:https://doi.org/10.1007/s12273-020-0723-1.

[48] Hannah Fontenot and Bing Dong. Modeling and control of building-integrated microgrids for optimal energy management–a review. *Applied Energy*, 254:113689, 2019. doi:https://doi.org/10.1016/j.apenergy.2019.113689.

[49] 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, doi:https://doi.org/10.1016/j.enbuild.2018.02.023.

[50] 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.

[51] Yoav Goldberg and Omer Levy. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. 2014. arXiv:1402.3722.

[52] Jessica Granderson, Guanjing Lin, Ari Harding, Piljae Im, and Yan Chen. Building fault detection data to aid diagnostic algorithm creation and performance testing. *Sci Data*, 7(1):65, February 2020. doi:https://doi.org/10.6084/m9.figshare.11743074.

[53] 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.

[54] A GroBmann, F Hohmann, and K Wiebe. Portabledyme-a simplified software package for econometric model building. *Macroeconomic Modelling For Policy Evaluation*, 120:33, 2013.

[55] 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.

[56] Burak Gunay and Weiming Shen. Connected and Distributed Sensing in Buildings: Improving Operation and Maintenance. *IEEE Systems, Man, and Cybernetics Magazine*, 3(4):27–34, 2017. doi:10.1109/msmc.2017.2702386.

[57] H Burak Gunay, Weiming Shen, and Guy Newsham. Data analytics to improve building performance: A critical review. *Automation in Construction*, 97:96–109, 2019. doi:https://doi.org/10.1016/j.autcon.2018.10.020.

[58] American Heating and Refrigeration Institute (AHRI). Historical Data: Statistical information on HVACR equipment shipments, 2020. URL: http://www.ahrinet.org/resources/statistics/historical-data.

[59] Jan LM Hensen and Roberto Lamberts. *Building performance simulation for design and operation*. Routledge, 2 edition, 2019. doi:https://doi.org/10.1201/9780429402296.

[60] Yeonsook Heo, Ruchi Choudhary, and GA Augenbroe. Calibration of building energy models for retrofit analysis under uncertainty. *Energy and Buildings*, 47:550–560, 2012. doi:https://doi.org/10.1016/j.enbuild.2011.12.029.

[61] Carlos Roberto Barrios Hernandez. Thinking parametric design: Introducing parametric Gaudi. *Design Studies*, 27(3):309–324, 2006. doi:10.1016/j.destud.2005.11.006.

[62] Tianzhen Hong, Yixing Chen, Xuan Luo, Na Luo, and Sang Hoon Lee. Ten questions on urban building energy modeling. *Building and Environment*, 168:106508, 2020. doi:https://doi.org/10.1016/j.buildenv.2019.106508.

[63] Tianzhen Hong, Yixing Chen, Xuan Luo, Na Luo, and Sang Hoon Lee. Ten questions on urban building energy modeling. *Building and Environment*, 168:106508, 2020. doi:https://doi.org/10.1016/j.buildenv.2019.106508.

[64] Tianzhen Hong, Jared Langevin, and Kaiyu Sun. Building simulation: Ten challenges. *Building Simulation*, 11(5):871–898, 2018. doi:10.1007/s12273-018-0444-x.

[65] Tianzhen Hong, Hongsan Sun, Yixing Chen, Sarah C. Taylor-Lange, and Da Yan. An occupant behavior modeling tool for co-simulation. *Energy and Buildings*, 117:272–281, 2016. doi:10.1016/j.enbuild.2015.10.033.

[66] Tianzhen Hong, Zhe Wang, Xuan Luo, and Wanni Zhang. State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, 212:109831, 2020. doi:https://doi.org/10.1016/j.enbuild.2020.109831.

[67] Gesche Margarethe Huebner and Ardeshir Mahdavi. A structured open data collection on occupant behaviour in buildings. *Scientific data*, 6(1):1–4, 2019. doi:https://doi.org/10.1038/s41597-019-0276-2.

[68] 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.

[69] Salah Imam, David A. Coley, and Ian Walker. The building performance gap: Are modellers literate? *Building Services Engineering Research and Technology*, 38(3):351–375, 2017. doi:10.1177/0143624416684641.

[70] Mark Jaccard and Margo Dennis. Estimating home energy decision parameters for a hybrid energy-economy policy model. *Environmental Modeling and Assessment*, 11(2):91–100, 2006. doi:10.1007/s10666-005-9036-0.

[71] Hrvoje Jasak. OpenFOAM: Open source CFD in research and industry. *International Journal of Naval Architecture and Ocean Engineering*, 1(2):89–94, 2009. doi:10.2478/ijnaoe-2013-0011.

[72] Hongyuan Jia and Adrian Chong. eplusr: A framework for integrating building energy simulation and data-driven analytics. **In Review**, 2020. URL: https://CRAN.R-project.org/package=eplusr, doi:10.13140/RG.2.2.34326.16966.

[73] Anish Jindal, Neeraj Kumar, and Mukesh Singh. A unified framework for big data acquisition, storage, and analytics for demand response management in smart cities. *Future Generation Computer Systems*, 108:921–934, 2020. doi:10.1016/j.future.2018.02.039.

[74] Arif E Jinha. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010. doi:https://doi.org/10.1087/20100308.

[75] Christian Johansson, Markus Bergkvist, Davy Geysen, Oscar De Somer, Niklas Lavesson, and Dirk Vanhoudt. Operational Demand Forecasting in District Heating Systems Using Ensembles of Online Machine Learning Algorithms. *Energy Procedia*, 116:208–216, 2017. doi:10.1016/j.egypro.2017.05.068.

[76] Christian Johansson and Blekinge Tekniska Hogskola. On intelligent district heating. 2014. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:bth-00587.

[77] A Rezaee Jordehi. Optimisation of demand response in electric power systems, a review. *Renewable and sustainable energy reviews*, 103:308–319, 2019. doi:10.1016/j.rser.2018.12.054.

[78] M. Jradi, K. Arendt, F. C. Sangogboye, C. G. Mattera, E. Markoska, M. B. Kjærgaard, C. T. Veje, and B. N. Jørgensen. ObepME: An online building energy performance monitoring and evaluation tool to reduce energy

performance gaps. *Energy and Buildings*, 166:196–209, 2018. doi:10.1016/j.enbuild.2018.02.005.

[79] 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. doi:doi.org/10.1080/10789669.2005.10391123.

[80] 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. doi:10.1145/2025528.2025537.

[81] 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.

[82] Mika Yagi Kim. " Watts Per Person " Paradigm To Design Net Zero Energy Buildings : Examining Technology Interventions and Integrating Occupant. doi:10.25549/usctheses-c3-281325.

[83] D. Kolokotsa, D. Rovas, E. Kosmatopoulos, and K. Kalaitzakis. A roadmap towards intelligent net zero- and positive-energy buildings. *Solar Energy*, 85(12):3067–3084, 2011. doi:10.1016/j.solener.2010.09.001.

[84] 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. doi:10.1007/11575832_13.

[85] Andrew Lake, Behnaz Rezaie, and Steven Beyerlein. Review of district heating and cooling systems for a sustainable future. *Renewable and Sustainable Energy Reviews*, 67:417–425, 2017. doi:https://doi.org/10.1016/j.rser.2016.09.061.

[86] 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.

[87] Sang Hoon Lee, Tianzhen Hong, Mary Ann Piette, and Sarah C. Taylor-Lange. Energy retrofit analysis toolkits for commercial buildings: A review. *Energy*, 89:1087–1100, 2015. doi:10.1016/j.energy.2015.06.112.

[88] Hyunwoo Lim and Zhiqiang John Zhai. Review on stochastic modeling methods for building stock energy prediction. *Building Simulation*, 10(5):607–624, 2017. doi:10.1007/s12273-017-0383-y.

[89] Shunfu Lin, Fangxing Li, Erwei Tian, Yang Fu, and Dongdong Li. Clustering load profiles for demand response applications. *IEEE Transactions on Smart Grid*, 10(2):1599–1607, 2017. doi:10.1109/TSG.2017.2773573.

[90] 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. doi:https://doi.org/10.1177/0312896219877678.

[91] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. 2002. URL: http://arxiv.org/abs/cs/0205028, doi:10.3115/1118108.1118117.

[92] Emilio T Maddalena, Yingzhao Lian, and Colin N Jones. Data-driven methods for building control—a review and promising future directions. *Control Engineering Practice*, 95:104211, 2020. doi:https://doi.org/10.1016/j.conengprac.2019.104211.

[93] Mahmoud Mohamed Abdelrahman. Gh_CPython: CPython plugin for grasshopper. 2017. doi:10.5281/zenodo.888148.

[94] 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. doi:10.3390/en11082002.

[95] Paul A. Mathew, Laurel N. Dunn, Michael D. Sohn, Andrea Mercado, Claudine Custudio, and Travis Walter. Big-data for building energy performance: Lessons from assembling a very large national database of building energy use. *Applied Energy*, 140:85–93, 2015. doi:10.1016/j.apenergy.2014.11.042.

[96] 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.

[97] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 52–55, 2019. arXiv:1712.09405.

[98] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26:3111–3119, 2013. arXiv:1310.4546.

[99] Tomáš Mikolov, Wen-tau Yih, and 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*, pages 746–751, 2013.

[100] Clayton Miller. More buildings make more generalizable Models—Benchmarking prediction methods on open electrical meter data. *Machine Learning and Knowledge Extraction*, 1(3):974–993, August 2019. doi:https://doi.org/10.3390/make1030056.

[101] Clayton Miller. What's in the box?! towards explainable machine learning applied to non-residential building smart meter classification. *Energy Build.*, 199:523–536, September 2019. doi:https://doi.org/10.1016/j.enbuild.2019.07.019.

[102] 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. doi:https://doi.org/10.1080/23744731.2020.1795514.

[103] 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. doi:https://doi.org/10.1038/s41597-020-00712-x.

[104] 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. doi:10.17632/7CVGWS3MX3.1.

[105] Miguel Molina-Solana, María Ros, M Dolores Ruiz, Juan Gómez-Romero, and María J Martín-Bautista. Data science for building energy management: A review. *Renewable and Sustainable Energy Reviews*, 70:598–609, 2017. doi:https://doi.org/10.1016/j.rser.2016.11.132.

[106] José A Moral-Muñoz, Antonio G López-Herrera, Enrique Herrera-Viedma, and Manuel J Cobo. Science mapping analysis software tools: A review. In *Springer Handbook of Science and Technology Indicators*, pages 159–185. Springer, 2019. doi:https://doi.org/10.1007/978-3-030-02511-3_7.

[107] Fionn Murtagh and Pierre Legendre. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3):274–295, 2014. doi:10.1007/s00357-014-9161-z.

[108] Cristina Nichiforov, Grigore Stamatescu, Iulia Stamatescu, and Ioana Făgărăşan. Evaluation of Sequence-Learning models for Large-Commercial-Building load forecasting. *Information*, 10(6):189, June 2019. doi:https://doi.org/10.3390/info10060189.

[109] 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.

[110] Monica Nuekomm, Valerie Nubbe, and Robert Fares. Grid-interactive Efficient Buildings: Overview. (April):1–36, 2019. doi:doi.org/10.2172/1508212.

[111] US. Department of Energy Building Technology Office (BTO). Scout: Github repository, 2020. URL: https://github.com/trynthink/scout.

[112] 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. doi:https://doi.org/10.1016/j.buildenv.2020.106738.

[113] 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. doi:10.1038/s41597-019-0275-3.

[114] 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. doi:https://doi.org/10.1016/j.apenergy.2018.12.025.

[115] 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. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[116] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. doi:10.3115/v1/D14-1162.

[117] 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. doi:https://doi.org/10.1016/j.apenergy.2019.113727.

[118] 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. doi:https://doi.org/10.1007/s12273-020-0650-1.

[119] Matias Quintana and Clayton Miller. Towards class-balancing human comfort datasets with GANs. *BuildSys 2019 - Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 391–392, 2019. doi:10.1145/3360322.3361016.

[120] Matias Quintana, Stefano Schiavon, Kwok Wai Tham, and 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*, pages 120–129, 2020. doi:https://doi.org/10.1145/3408308.3427612.

[121] Loıc Raillon, Simon Rouchier, and Sarah Juricic. pysip: an open-source tool for bayesian inference and prediction of heat transfer in buildings.

[122] T Agami Reddy. Literature review on calibration of building energy simulation programs: uses, problems, procedures, uncertainty, and tools. *ASHRAE transactions*, 112:226, 2006.

[123] Christoph F Reinhart and Carlos Cerezo Davila. Urban building energy modeling–a review of a nascent field. *Building and Environment*, 97:196–202, 2016. doi:https://doi.org/10.1016/j.buildenv.2015.12.001.

[124] Behnaz Rezaie and Marc A Rosen. District heating and cooling: Review of technology and potential enhancements. *Applied energy*, 93:2–10, 2012. doi:https://doi.org/10.1016/j.apenergy.2011.04.020.

[125] 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. doi:10.1016/j.scs.2018.04.008.

[126] Danny Rodrigues Alves, Giovanni Colavizza, and Frédéric Kaplan. Deep reference mining from scholarly literature in the arts and humanities. *Frontiers in Research Metrics and Analytics*, 3:21, 2018. doi:https://doi.org/10.3389/frma.2018.00021.

[127] Adam Rohloff. Data analytics from cradle to grave. *ASHRAE Journal*, 58(2):34, 2016.

[128] Jonathan Roth. Github: Nyc - urban building energy model for new york city. URL: https://github.com/jmr385/UBEM_NYC.

[129] 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.

[130] Jonathan Roth, Amory Martin, Clayton Miller, and 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. *Applied Energy*, 280, 2020. doi:10.1016/j.apenergy.2020.115981.

[131] Mostapha Sadeghipour Roudsari and Michelle Pak. Ladybug: A parametric environmental plugin for grasshopper to help designers create an environmentally-conscious design. *Proceedings of BS 2013: 13th Conference of the International Building Performance Simulation Association*, pages 3128–3135, 2013.

[132] Pattarin Sanguankaew and Vichita Vathanophas Ractham. Bibliometric review of research on knowledge management and sustainability, 1994–2018. *Sustainability*, 11(16):4388, 2019. doi:https://doi.org/10.3390/su11164388.

[133] Philip Santosh. Eppy: Scripting language for e+, energyplus. https://github.com/santoshphilip/eppy, 2014. Accessed: 2020-12-03.

[134] Jan Schiefelbein, Jana Rudnick, Anna Scholl, Peter Remmen, Marcus Fuchs, and Dirk Müller. Automated urban energy system modeling and thermal building simulation based on OpenStreetMap data sets. *Building and Environment*, 149:630–639, feb 2019. doi:10.1016/j.buildenv.2018.12.025.

[135] 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.

[136] 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. doi:https://doi.org/10.1016/j.joi.2016.03.008.

[137] 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. doi:https://doi.org/10.1186/s12859-019-2607-x.

[138] 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. doi:10.1109/ISGTEurope.2017.8260206.

[139] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 4593–4601, 2020. arXiv:1905.05950, doi:10.18653/v1/p19-1452.

[140] Wei Tian. A review of sensitivity analysis methods in building energy analysis. *Renewable and Sustainable Energy Reviews*, 20:411–419, 2013. doi:https://doi.org/10.1016/j.rser.2012.12.014.

[141] 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. doi:https://doi.org/10.1007/s10032-015-0249-8.

[142] 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. doi:10.3389/fmech.2015.00017.

[143] Nees Jan Van Eck and Ludo Waltman. Text mining and visualization using vosviewer. *arXiv preprint arXiv:1109.2058*, 2011. arXiv:1109.2058.

[144] Nees Jan Van Eck and Ludo Waltman. Citnetexplorer: A new software tool for analyzing and visualizing citation networks. *Journal of informetrics*, 8(4):802–823, 2014. doi:https://doi.org/10.1016/j.joi.2014.07.006.

[145] Richard Van Noorden. Elsevier opens its papers to text-mining., 2 2014. doi:10.1038/506017a.

[146] J. Verhelst, G. Van Ham, D. Saelens, and L. Helsen. Model selection for continuous commissioning of HVAC-systems in office buildings: A review. *Renewable and Sustainable Energy Reviews*, 76:673–686, 2017. doi:10.1016/j.rser.2017.01.119.

[147] 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. doi:https://doi.org/10.1016/j.rser.2016.10.079.

[148] Zhe Wang, Tianzhen Hong, and Ruoxi Jia. Buildings.Occupants: a Modelica package for modelling occupant behaviour in buildings. *Journal of Building Performance Simulation*, 12(4):433–444, 2019. doi:10.1080/19401493.2018.1543352.

[149] Rebecca Ward, Ruchi Choudhary, Alastair Gregory, and Mark Girolami. Continuous calibration of a digital twin: comparison of particle filter and bayesian calibration approaches. *arXiv preprint arXiv:2011.09810*, 2020. arXiv:2011.09810.

[150] 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.

[151] Thomas Wortmann. OPOSSUM: Introducing and Evaluating a Model-based Optimization Tool for Grasshopper. *Proceedings of the CAADRIA 17*, (April):283 – 292, 2017. URL: http://papers.cumincad.org/data/works/att/caadria2017_124.pdf.

[152] Ibrahim Y Wuni, Geoffrey QP 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. URL: https://doi.org/10.1016/j.enbuild.2019.02.010.

[153] 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 and Buildings*, 107:264–278, 2015. doi:https://doi.org/10.1016/j.enbuild.2015.08.032.

[154] Ke Yan, Adrian Chong, and Yuchang Mo. Generative adversarial network for fault detection diagnosis of chillers. *Building and Environment*, 172:106698, 2020. doi:https://doi.org/10.1016/j.buildenv.2020.106698.

[155] Ke Yan, Jing Huang, Wen Shen, and Zhiwei Ji. Unsupervised learning for fault detection and diagnosis of air handling units. *Energy and Buildings*, 210:109689, 2020. doi:https://doi.org/10.1016/j.enbuild.2019.109689.

[156] Zheng Yang, Jonathan Roth, and 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. doi:https://doi.org/10.1016/j.enbuild.2017.12.040.

[157] Sicheng Zhan, Zhaoru Liu, Adrian Chong, and Da Yan. Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking. *Applied Energy*, 269:114920, 2020. doi:10.1016/j.apenergy.2020.114920.

[158] Yi Zhang and Ivan Korolija. Performing complex parametric simulations with jeplus. In *SET2010-9th International Conference on Sustainable Energy Technologies*, pages 24–27, 2010.

[159] Zhiang Zhang, Adrian Chong, Yuqi Pan, Chenlu Zhang, and Khee Poh Lam. Whole building energy model for hvac optimal control: A practical framework based on deep reinforcement learning. *Energy and Buildings*, 199:472–490, 2019. doi:10.1016/j.enbuild.2019.07.029.

[160] Yang Zhao, Tingting Li, Xuejun Zhang, and Chaobo Zhang. Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future. *Renewable and Sustainable Energy Reviews*, 109:85–101, 2019. doi:https://doi.org/10.1016/j.rser.2019.04.021.