Taylor & Francis
Taylor & Francis Group

Science and Technology for the Built Environment

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uhvc21

What we talk about when we talk about EEMs: using text mining and topic modeling to understand building energy efficiency measures (1836-RP)

Apoorv Khanuja & Amanda L. Webb

To cite this article: Apoorv Khanuja & Amanda L. Webb (2022): What we talk about when we talk about EEMs: using text mining and topic modeling to understand building energy efficiency measures (1836-RP), Science and Technology for the Built Environment, DOI: 10.1080/23744731.2022.2133329

To link to this article: https://doi.org/10.1080/23744731.2022.2133329

Science and Technology for the Built Environment, (2022) **0**, 1–15

ISSN: 2374-4731 print / 2374-474X online
DOI: 10.1080/23744731.2022.2133329

Taylor & Francis
Taylor & Francis Group

What we talk about when we talk about EEMs: using text mining and topic modeling to understand building energy efficiency measures (1836-RP)

APOORV KHANUJA AND AMANDA L. WEBB*

Department of Civil and Architectural Engineering and Construction Management, University of Cincinnati, Cincinnati, OH, USA

Energy Efficiency Measures (EEMs) play a central role throughout the building energy efficiency industry, and lists of EEMs therefore exist in a variety of resources. However, each of these use different conventions for describing and organizing measures, which presents a major challenge for aggregating information across these resources. The goal of this study is to discover trends in how existing resources describe and organize EEMs using topic modeling and other text mining methods. A unique dataset of 3,490 EEMs from 16 different documents was compiled and analyzed using frequency analysis, part of speech tagging, and topic modeling. The results showed three major trends. First, a typical EEM contains six words and is phrased in verb-noun format, although these characteristics varied widely. Second, there are words and bigrams commonly used across many EEMs, and these include action words, specific building components, broader building systems, and descriptor terms. Third, there are thematic similarities between the EEM lists, which in some cases highlight the ways in which these lists are derived from one another. These findings provide insight into the nature of EEMs and can be used as the basis for developing a standardized system for organizing and describing EEMs.

Introduction

Energy efficiency measures (EEMs) are the fundamental mechanism for improving energy performance in buildings. An EEM is defined as "an action taken in the operation or equipment in a building that reduces energy use of the building while maintaining or enhancing the building's safety, comfort, and functionality" (ASHRAE. 2018a). This broad definition underscores the foundational nature of EEMs throughout the building energy efficiency industry. There are many parties that may be involved in a building efficiency project—including energy modelers, energy auditors, energy managers, building owners and utilities, among other stakeholders—and each party works with EEMs in various ways. In energy modeling, for example, EEMs define design alternatives, allowing modelers and designers to explore potential what-if scenarios for improving the building. In energy auditing, EEMs are the basis for the energy auditor's recommended list of actions for the building owner, and are a key component of an audit report. EEMs are also the basis for awarding financial incentives in utility-sponsored efficiency programs.

True to their widespread role, lists and descriptions of EEMs exist in a variety of different resources. Statewide Technical Reference Manuals (TRMs), which provide information about EEMs for use in many utility-sponsored energy efficiency programs, list a variety of EEMs along with transparent methods for calculating energy savings for each measure (Illinois Energy Efficiency Stakeholder Advisory Group 2019; New York State Joint Utilities 2019). Reference books intended to aid practicing energy auditors or energy managers (variously called handbooks, sourcebooks, or manuals, among other terms) typically contain descriptions of common EEMs, along with methods for calculating resulting energy savings (Wulfinghoff 1999; Thumann 1992). Several standards and guidelines addressing energy efficiency in existing buildings also contain lists of EEMs, including ASHRAE Standard 100-2018, which enumerates over 200 EEMs for use in existing buildings (ASHRAE. 2018b). Energy modeling and building data exchange tools, such as the Commercial Building Energy Saver (Hong et al. 2015), and BuildingSync (Long et al. 2021) also contain lists of EEMs, and users can select from these lists to add EEMs to a given project.

While these various resources provide lists of EEMs, they use a variety of different conventions for naming,

Received March 31, 2022; accepted September 23, 2022
Apoorv Khanuja, Student Member ASHRAE, is a Graduate Student. **Amanda L. Webb**, PhD, Full Member ASHRAE, is an Assistant Professor.
*Corresponding author e-mail: amanda.webb@uc.edu

organizing, and describing measures. This presents a major challenge for exchanging and analyzing EEM-related data. It limits the ability to aggregate information across EEM datasets and compare EEMs (and EEM savings and cost effectiveness) from one dataset to another. It also makes it difficult to leverage these existing resources to develop new, comprehensive lists of measures for use in energy modeling and data exchange tools, in guidelines and standards, and in building efficiency programs and policies. To begin to address this barrier, a better understanding of current methods of organizing and describing EEMs is needed, as well as insight into how these existing systems relate to one another.

Text mining and related natural language processing (NLP) techniques, such as topic modeling, present a promising strategy for analyzing EEM names and descriptions. Text mining, broadly, is the process of automatically extracting previously unknown information and insights from unstructured text within any written resource (Hearst 1999). Topic modeling is an unsupervised text mining technique that can be used to uncover hidden themes (i.e., topics) across a collection of documents, as well as within individual documents (Blei 2012). Text mining and topic modeling have been used to analyze textual data in a variety of applications, like examining newspaper articles related to government funding of artists and arts organizations (DiMaggio, Nag, and Blei 2013), uncovering themes in educational leadership research literature over time (Wang, Bowers, and Fikis 2017), and evaluating Consumer Financial Protection Bureau complaints (Bastani, Namavari, and Shaffer 2019). Research has also been conducted testing the effectiveness of topic models in analyzing twitter data (L. Hong and Davison 2010). Overall, these studies show that topic modeling is a valuable technique to analyze large collections of texts where manual review would be unfeasible, and that it works well in uncovering the thematic makeup of documents across a variety of different fields.

Text mining and topic modeling have also been successfully applied in a buildings context to understand textual data. Abdelrahman et al. (2021) used text mining to capture the relationship between data science techniques and building energy efficiency applications in research literature. S. Hong, Kim, and Yang (2022) and Bouabdallaoui et al. (2020) both used text mining and machine learning to classify building maintenance request data to improve facility management. Lai and Kontokosta (2019) used topic modeling to discover themes in construction activities across major U.S. cities by examining text data from building permits. Specific to EEMs, Lai et al. (2022) used NLP to extract information about recommended EEMs from energy audit reports and match them to post-audit building permit descriptions to estimate the likelihood of EEM adoption. They found data quality—including inconsistent EEM naming—to be a significant concern, further highlighting the need for this current study.

The goal of this study is to discover trends in how existing resources describe and organize EEMs using topic modeling and other text mining methods. Existing lists of EEMs were identified and collected through a comprehensive literature review, and then compiled into a dataset. This resulted in a total of 3,490 EEMs from 16 different documents which were used as the basis of this analysis. A variety of text mining techniques were then used to analyze the data. First, frequency analysis was used to quantify variation in EEM length and to identify commonly occurring and co-occurring terms. Then, part of speech tagging was performed to find typical EEM formats. Finally, topic modeling and cosine similarity were applied to reveal underlying themes and find similar documents.

This study makes a novel contribution to the research literature by systematically analyzing the structure of EEMs and providing deeper insight into the nature of EEMs and the ways in which they are used and described across the building energy efficiency industry. The use of text mining techniques to obtain these insights is especially important, as it represents a replicable and scalable process that could be applied to understand other sets of EEMs. The large dataset of 3,490 EEMs assembled for this study also provides a valuable source of data for other researchers working on this topic. More broadly, the insights gained from this study can be used as the basis for developing a standardized system for organizing and describing EEMs. In this respect, this study represents a foundational step toward greater standardization of EEMs and EEM-related data.

Methodology

Data

A comprehensive literature review was conducted from September 2019 through July 2020 to identify existing lists of EEMs. An initial list of suggested documents was collected from members of a Project Advisory Board of industry professionals, and additional documents were added through the literature review process. For a document to be included in the analysis, it needed to contain a list of EEMs. A few documents—including ASHRAE's Procedures for Commercial Building Energy Audits (ASHRAE. 2011), colloquially known as the "green book," and ASHRAE's Advanced Energy Design Guides (ASHRAE. 2019)—that discuss EEMs were given initial review but ultimately not included as part of the analysis because they lack a well-defined list of EEMs. ASHRAE's Procedures for Commercial Building Energy Audits, for example, describes various types of measures (e.g., low-cost vs. capital investment), but does not include a list of EEMs. The Advanced Energy Design Guides contain recommended design criteria for various building components for different building types in each U.S. climate zone (e.g., for office buildings in climate zone 4, Maximum solar heat gain coefficient (SHGC), Fixed: 0.34), but do not contain a list of specific actions.

A total of 16 sources were included in the analysis, and these are broadly representative of EEM lists commonly used across industry. These sources are treated as documents for this analysis, and these terms are used interchangeably throughout this article. Table 1 lists the full title of each source included in the analysis, along with its citation and

Table 1. List of documents analyzed.

Abbrev.	Title	Type	Reference
1651RP	ASHRAE 1651-RP, Development of Maximum Technically Achievable Energy Targets for Commercial Buildings: Ultra-Low Energy Use Building Set	Other	(Glazer 2015)
ATT	Audit Template, Release 2020.2.0	Tool	(Pacific Northwest National Laboratory 2020)
BCL	Building Component Library	Tool	(National Renewable Energy Laboratory 2020a)
BEQ	ASHRAE Building EQ	Tool	(ASHRAE. 2020)
BSYNC	BuildingSync, Version 2.0	Tool	(National Renewable Energy Laboratory 2020b)
CBES	Commercial Building Energy Saver	Tool	(Lawrence Berkeley National Laboratory 2020a)
DOTY	Commercial Energy Auditing Reference Handbook	Handbook	(Doty 2011)
IEA11	Source Book for Energy Auditors, Vol. 1	Handbook	(Lyberg 1987)
IEA46	Energy Efficient Technologies and Measures for Building Renovation: Sourcebook	Handbook	(Zhivov and Nasseri 2014)
ILTRM	Illinois Statewide Technical Reference Manual for Energy Efficiency, Version 8.0	TRM	(Illinois Energy Efficiency Stakeholder Advisory Group 2019)
NYTRM	New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs – Residential, Multi-Family, and Commercial/Industrial Measures, Version 7	TRM	(New York State Joint Utilities 2019)
REMDB	National Residential Efficiency Measures Database, Version 3.1.0	Other	(National Renewable Energy Laboratory 2018)
STD100	ASHRAE Standard 100-2018, Energy Efficiency in Existing Buildings	Standard	(ASHRAE. 2018b)
THUM	Energy Conservation in Existing Buildings Deskbook	Handbook	(Thumann 1992)
WSU	Energy Audit Workbook	Handbook	(Washington State University Cooperative Extension and Energy Program 2003)
WULF	Energy Efficiency Manual	Handbook	(Wulfinghoff 1999)

an abbreviation assigned to each source that is used to refer to the source throughout this study. Table 1 also groups each source into one of five types: (1) tools, which include software or Web-based tools; (2) Technical Reference Manuals (TRMs); (3) handbooks, which include textbooks and instructional manuals; (4) standards; (5) other documents that did not fit under the other categories. These distinct types of documents are evidence of the wide range of uses for EEM lists in practice. Notably, while some of these EEM lists were developed by individuals (especially the handbooks), most of them represent the result of collaborative projects or processes. Collectively, the documents span over 30 years and represent decades of assembling and organizing EEMs. The only documents that were known a priori to be similar were BSYNC and ATT. The only difference between these is that the categories for some EEMs differ between the documents. While these lists are very similar, they represent two distinct tools and applications of EEM lists, and were therefore both included in the study.

The EEM lists were manually extracted from each source and stored in a comma separated value (CSV) file that contained: a unique identifier for each EEM, the name of each EEM, the name of its corresponding category (and subcategory, if present), and the name of the source. The categories were not used in this text mining analysis, but were analyzed as part of a separate qualitative analysis of EEM

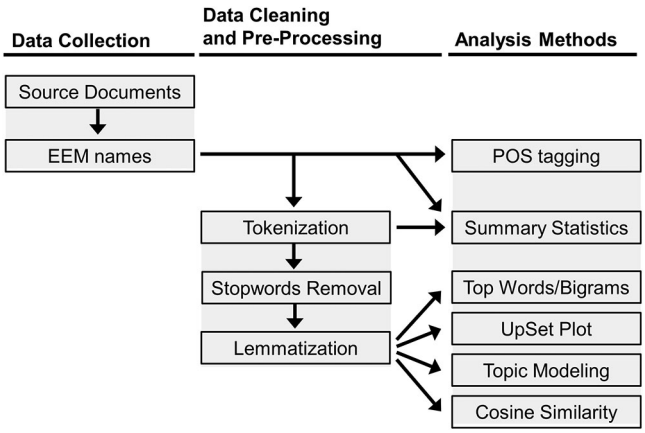

Fig. 1. Schematic of data processing workflow.

categorization systems (Webb and Khanuja 2022). To maintain fidelity with the original source documents, the text of each EEM was extracted exactly as it was written, preserving typos in the rare cases in which they occurred. Note that for WULF and WSU, some EEMs were subsidiary (i.e., more specific) versions of other EEMs, and in these cases the subsidiary EEM was appended to the less specific EEM descriptions to form a single measure. As a result, these EEMs became longer than they appear in the original source. For example, in the EEM, "Minimize the duration of boiler plant operation. For applications with regular schedules, install clock controls to start and stop boilers" the second half was appended to the first half to form a single EEM. The EEM lists from each document were then combined into a main list containing a total of 3,490 EEMs from across the 16 documents (Khanuja and Webb 2022).

Data cleaning and pre-processing

Since the EEM names were derived from a variety of documents with different string lengths and linguistic styles, they were homogenized using several pre-processing techniques prior to analysis. A schematic of the data processing workflow is shown in Figure 1. The data cleaning and pre-processing workflow followed in this study is similar to many of the previous text mining papers reviewed, and is an important step to reduce noise from the data (Wang, Bowers, and Fikis 2017; Lai and Kontokosta 2019; Bastani, Namavari, and Shaffer 2019; Abdelrahman et al. 2021). The statistical computing software R was used for all pre-processing and analysis (R Core Team 2014).

First, the EEM names were tokenized into individual words. Tokenization is the process of breaking up the sequence of strings into smaller pieces called tokens. These tokens can be single words, n-grams (a contiguous sequence of n words), or even complete sentences. This was done using the tidytext R package (Silge and Robinson 2016). Tokenization using tidytext also removes all punctuation and whitespace and converts all words to lower case. For terms containing a hyphen (e.g., low-e) or slash (e.g., and/or), punctuation is removed and these terms are converted into multiple tokens. This process strips away all context of the sentences describing the EEMs and essentially transforms them into a collection of standalone words called a "bag of words" (Aldous 1985).

After tokenization, the stop words were removed from this bag of words using the R package stopwords (Benoit, Muhr, and Watanabe 2021). Stop words are frequently occurring but un-informative words (e.g., and, or, to, the) and are often removed from textual data prior to text mining. The list of stop words used for this analysis came from the snowball lexicon within the stopwords package, which was selected because its relatively short list of stop words would retain most of the EEM text. In addition to removing stop words, the tokens in which the first character was a number were also removed. This was because these tokens generally provided unnecessary level of detail (e.g., specific temperature setpoints, COP values, or the name of a standard such as ASHRAE 62.1) that was not essential to describing the EEM. However, tokens that contained numbers but started with an alphanumeric letter (e.g., T8, T12, CO2, etc.) were not removed since they provided useful information regarding the specific type of building component affected by an EEM.

Finally, the remaining tokens were lemmatized into their root form using the textstem R package (Rinker 2018). Lemmatization removes the inflection from the words and converts them into their root form (called lemma). This prevents the analysis from counting the different forms of the same word as different words. For example, the words "reduce," "reduced," and "reduces" have the same lemma "reduce". This resulted in a cleaned-up bag of words, which was then used for much of the text mining analysis.

Analysis methods

First, frequency analysis was used to quantify variation in EEM names across different documents and to identify commonly occurring terms. Summary statistics for each source were computed, including the number of EEMs per source, number of duplicate EEMs per source, and the minimum, median, average, and maximum number of words (i.e., tokens) per EEM. Statistics for the number of total and duplicate EEMs were computed using the original (i.e., pre-cleaned) text, and statistics for words per EEM were computed using the tokenized text before removing stopwords. Using the lemmatized text, the 20 most frequent words and bigrams (i.e., n-grams for $n = 2$) in the corpus were found, along with their frequency of occurrence in individual documents.

The co-occurrence of words within EEMs was then explored using the lemmatized text, to understand how commonly occurring terms combine with one another. A subset of five commonly occurring terms was selected, and a script was developed in R to identify the EEMs containing each of these terms. To visualize the number of EEMs containing each term, as well as the number of EEMs in which the terms co-occur, UpSet plots were created using the UpSetR R package (Conway, Lex, and Gehlenborg 2017). The UpSet plots are better than traditional Venn diagrams at representing set interactions for more than three sets. Like Venn diagrams, UpSet plots visualize the relationships

Table 2. Variation in EEMs across documents.

	Number of EEMs		Words per EEM			
Source	Total	Duplicates	Min.	Median	Avg.	Max.
1651RP	398	0	1	5.0	5.2	17
ATT	223	82	1	4.0	4.2	14
BCL	302	0	1	3.0	3.9	14
BEQ	295	1	2	12.0	11.9	41
BSYNC	223	82	1	4.0	4.2	14
CBES	102	0	2	7.0	7.5	19
DOTY	69	0	1	4.0	4.8	11
IEA11	232	0	2	5.0	5.3	13
IEA46	420	4	1	12.5	16.7	109
ILTRM	193	4	2	4.0	4.5	12
NYTRM	108	20	1	4.0	4.2	13
REMDB	136	3	1	4.0	4.5	14
STD100	241	1	2	15.0	18.3	103
THUM	52	0	2	6.0	5.8	15
WSU	130	0	2	6.0	6.5	17
WULF	366	13	2	11.0	12.6	41
TOTAL	3490	511	1	6.0	8.6	109

between sets, however, unlike Venn diagrams, UpSet plots visualize set intersections in a matrix layout (Conway, Lex, and Gehlenborg 2017). The sets are visualized as rows, with the total size of each set represented using a barplot at the left of the figure. Every possible intersection is represented by a bottom plot (dots and lines), and their frequency of occurrence is shown on a barplot at the top of the figure.

Second, part of speech (POS) tagging was used to uncover the syntactical structure of the EEMs. The POS tagging was performed using the RDRPOSTagger R package (Nguyen et al. 2014). The tagger annotates each word in the EEM name with a POS tag based on its definition and the context in which it is used. The result of the analysis is a list of words from each EEM automatically tagged with their corresponding part of speech (e.g., verb, noun, adjective). Note that this analysis was performed using the original (i.e., pre-cleaned) text, since tokenization and removing stop words strips the text of the context, thereby making it difficult for the tagger to determine the POS of the remaining words.

Third, to understand how the words and documents in the corpus relate to one another, topic models were developed using the topicmodels R package (Grün and Hornik 2011). This analysis was performed using the lemmatized text. Latent Dirichlet Allocation (LDA) was used for this study, which assumes that each document is made up of an underlying, unknown collection of topics, and each topic is made up of an underlying collection of words (Blei, Ng, and Jordan 2003; Blei 2012). In order to generate topic models, a document term matrix (DTM) was first created. A DTM is a large matrix with the document names as rows, the terms occurring within those documents as columns and the counts of those terms in those documents as the values of the matrix. This converts each document of arbitrary length in the corpus into a fixed length vector of real numbers. For this analysis, each source with their list of EEMs was treated as a document for the DTM. The values in the matrix were predominantly zeroes since most of the terms were not common to all documents. This DTM was then used to uncover the hidden topics across the documents.

Even though topic modeling is an unsupervised algorithm, the expected number of topics (k) still needs to be specified. If the value of k is too low, the LDA model will be too coarse to differentiate between topics. However, if the value of k is too high, it will make the model too complex and granular. For this analysis, the perplexity values for topic models from $k = 2$ to $k = 12$ topics were calculated using the topicmodels R package (Grün and Hornik 2011). Perplexity is a statistical measure of how well a probability model predicts a sample, with low values meaning that the model is a better predictor (Blei, Ng, and Jordan 2003; Zhao et al. 2015). Six topics were selected for this analysis based on a combination of diminishing returns in the perplexity analysis curve and keeping the number of topics relatively small.

The LDA topic model created using the topicmodels package returns two matrices relevant to the analysis: the beta matrix, which contains the probability distribution of words within topics, and the gamma matrix, which contains the probability distribution of topics within each document. The topic model detects the words most likely to occur in each topic, however, it is up to the analyst to interpret what the topics could mean using their domain-specific knowledge. For the beta matrix, a threshold of 1% was used, and only the terms with a probability higher than that were considered while interpreting the topics.

Cosine similarity was then used to find similar documents within the corpus. To compute this, the cosine distance between the documents was calculated using the term frequency values in the DTM. This served as the measure of similarity between the documents. The cosine distance gives a value between zero and one, where 0 signifies that the documents are completely dissimilar and 1 signifies completely identical documents.

Results

Summary counts of the number of EEMs per document indicate wide variation across the documents. Table 2 shows the total number of EEMs and duplicate EEMs in each document, and across all documents. The number of duplicate EEMs was calculated by subtracting the number of unique EEMs from the total number of EEMs in the document. The results show a wide spread in the number of EEMs within each source, ranging from a low of 52 EEMs in THUM to a high of 420 EEMs in IEA46. While the majority of the documents have few or no duplicate EEMs, several of the documents repeat the same measure name across multiple categories, resulting in a high number of duplicate EEMs for those documents. For example, BSYNC repeats the EEM "Clean and/or repair" 18 times, once in each category. BSYNC and ATT have the highest number of duplicate EEMs, with duplicate EEMs representing over one-third of their total EEMs. The total number of duplicate EEMs across all documents is 511. Note that this number is greater

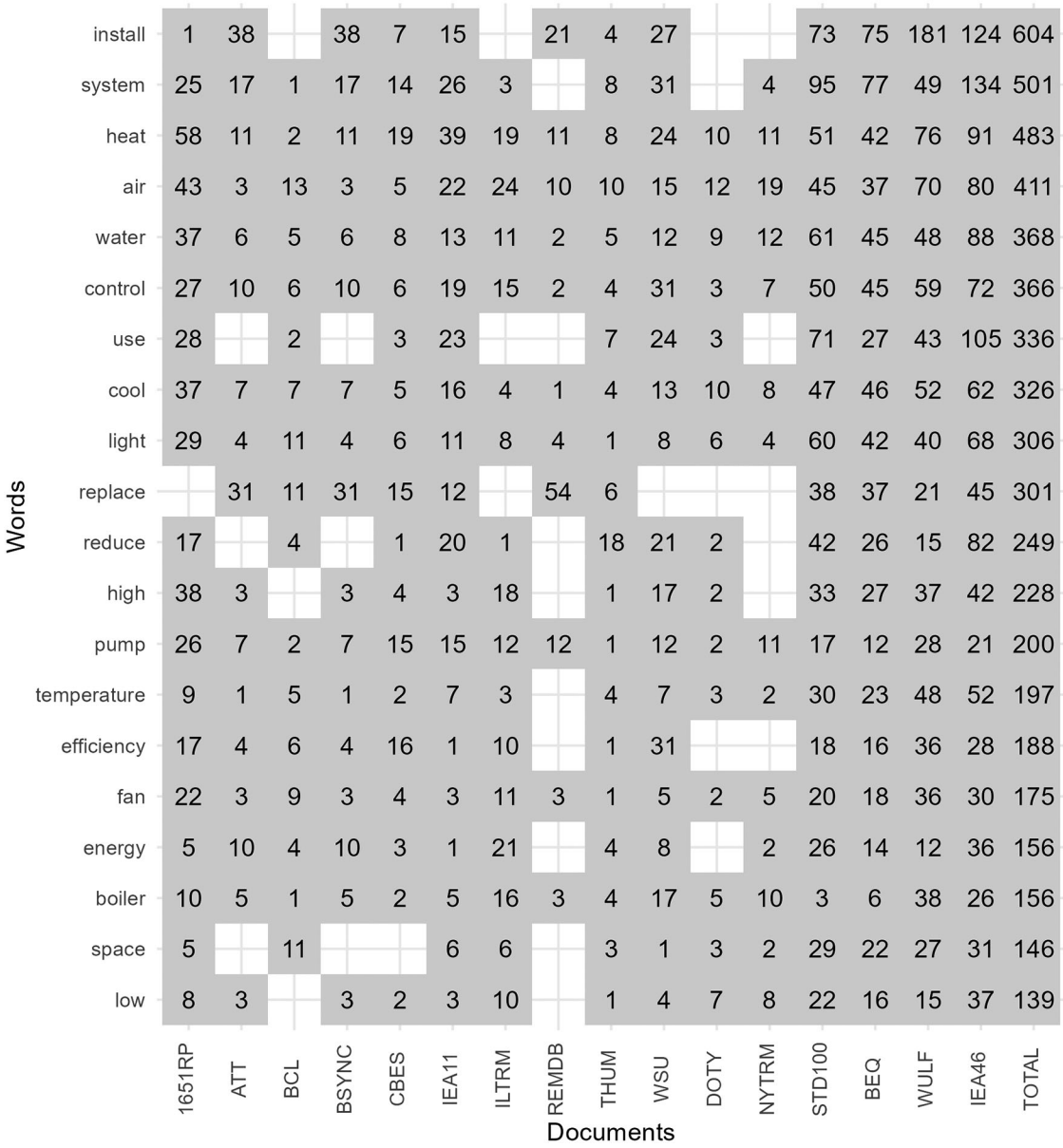

Fig. 2. Frequency of top 20 words by document.

than the sum of duplicate EEMs within each document, since it accounts for EEMs duplicated across different documents, in addition to those duplicated within a document. Note also that this only accounts for exact duplicates and does not account for duplicate EEMs that describe the same action but are phrased differently.

Counts of the number of words per EEM also indicate wide variation across the documents. For each source, the number of words per EEM was counted and summary statistics (minimum, maximum, median, and average number of words) were computed for each document and are displayed in Table 2. The results show that, across all documents, EEMs can be as short as a single word and as long as 109 words, with the median EEM containing six words. Four documents—STD100, WULF, IEA46 and BEQ—contain particularly long EEMs, with high median, average and maximum word counts. The rest of the documents stay within the range of 4–7 words per EEM on average. The mean number of words per EEM across all documents is 8.6, which is a bit higher than the median of 6.0 words per EEM and implies a slight positive skew in the data.

Word frequency counts show that the most frequently occurring words across the corpus are a mix of verbs and nouns. Figure 2 shows the frequency distribution of the top 20 words across the corpus of documents. The marginal total for each word is shown in the right-most column. Note that these represent the top 20 words across all documents, not the top 20 words for each document. Four of the top 20 words are verbs—install, use, replace, and reduce—and a verb (install) is also the most frequently occurring word across the corpus. These verbs describe the action performed in the implementation of the EEM, and their presence among the most common

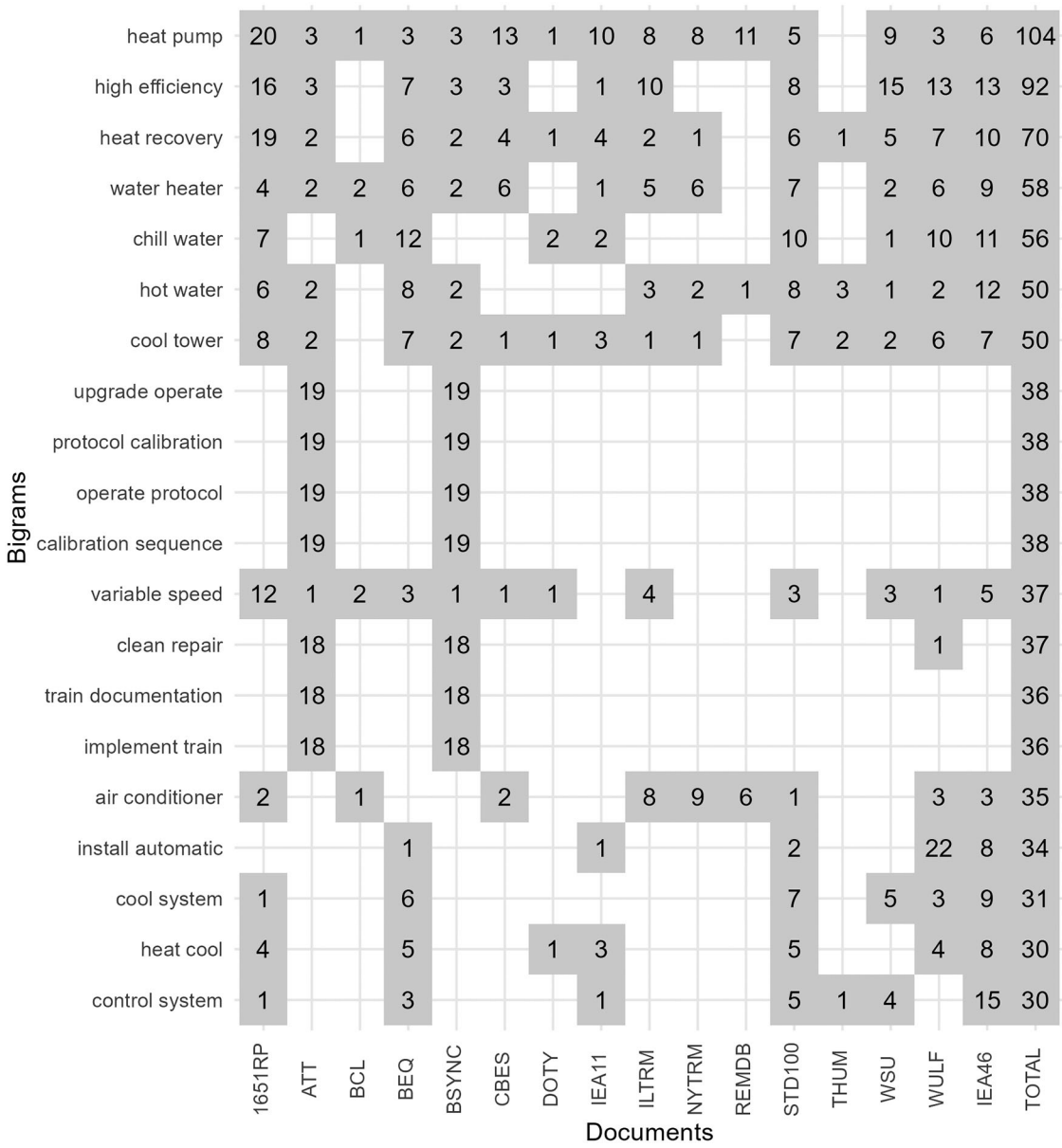

Fig. 3. Frequency of top 20 bigrams by document.

words suggests that an action term is an important component of an EEM. These verbs also suggest that synonymous terms may be common, as "install" and "use" have potentially similar meanings in a building energy efficiency context. Most of the remining top 20 words are nouns and represent a specific component (e.g., pump, fan, boiler) or building system (e.g., heating, cooling, air, water, lighting, control) affected by the EEM. The presence of words like "high" and "efficiency" among the top 20 words suggests that EEMs commonly contain descriptor terms such as "high efficiency" to characterize the desired performance of an EEM.

Figure 2 also shows wide variation in the occurrence of top 20 words by document. The frequency of occurrence of a word in an individual document ranges from zero (empty cells) to 181 instances of the same word (for install in WULF). Five documents—IEA11, WULF, STD100, BEQ and IEA46—contain all of the top 20 words. In the latter four of these documents the top 20 words occur with high frequency, which matches the observation from Table 2 that these documents contain long, wordy EEMs. In contrast, REMDB is missing nine of the top 20 words in the corpus.

In contrast to the word counts, the most frequently occurring bigrams are primarily nouns. Figure 3 shows the frequency distribution of the top 20 bigrams in the corpus and their distribution across documents. Note that these represent the top 20 bigrams across all documents, rather than the top 20 bigrams for each document. Note also that some bigrams may seem confusing due to the removal of stop words from between the bigram (e.g., the bigram "clean repair" originally had the term "and/or" between the words). Figure 3 shows that the top bigrams consist of specific retrofit technologies (e.g., heat recovery, heat pump, water heater, cooling tower) with only a

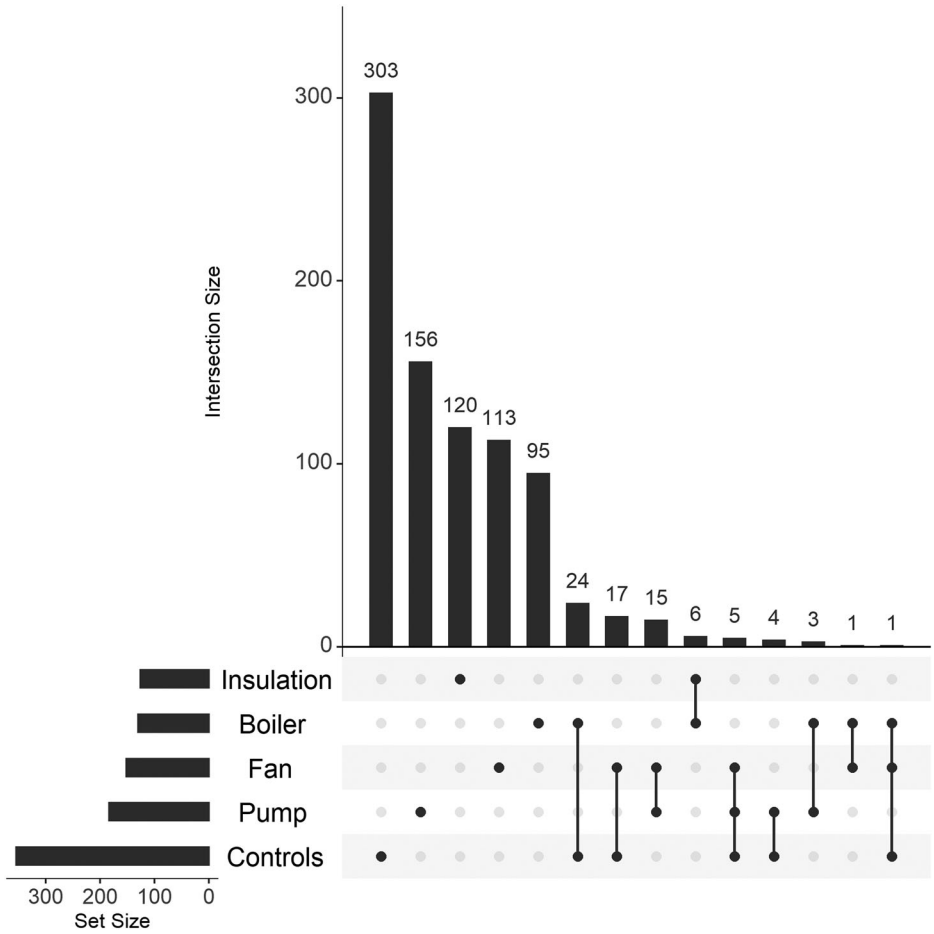

Fig. 4. UpSet plot illustrating frequency of words in EEM names.

few instances of bigrams containing a verb (e.g., upgrade operate, clean repair, install automatic). The bigram "high efficiency" is the second most frequent bigram in the list, and highlights the occurrence of this common descriptor term across most of the documents analyzed.

Figure 3 also illustrates the uneven distribution of the bigrams across the documents. The most frequent bigrams (e.g., heat pump, high efficiency, heat recovery, water heater) occur in nearly every document. However, several of the top 20 bigrams only occur in a small subset of documents. For example, ATT and BSYNC are the only documents containing many of the top 20 bigrams; this is because the EEMs "Clean and/or repair," "Implement training and/or documentation," "Upgrade operating protocols, calibration, and/or sequencing" are repeated 18–19 times in these two documents, once in each category.

Figure 3 also shows that some documents contain only a few of the top 20 bigrams. Only three of the top 20 bigrams are present in REMDB, and only five of the top 20 bigrams are present in the BCL. In the case of the REMDB, this result can be explained by two features of the REMDB. First, unlike most of the other documents, it is focused on residential buildings, and some bigrams describing components common in commercial buildings (e.g., cool tower, chill water) may not be relevant to the scope of the REMDB. Second, the structure and terminology in the REMDB measure list excludes some of the top 20 bigrams that would be relevant to residential buildings. For example, "water heater" is a sub-category in REMDB and the word "water heater" is therefore implied in the EEM name and never shows up in the EEM name itself (an example EEM name from the Water Heater sub-category: "Replace Electric Tank with Heat Pump"). As another example, the word "heat recovery" does not appear in an EEM in REMDB because the abbreviation HRV is used instead (an example EEM from the category Airflow: "Install HRV/ERV"). In the case of the BCL, this source is an energy modeling measure database to which contributors have added measures on an ad hoc basis, and may not have complete coverage of all building systems or components. Moreover, the BCL has few rules about EEM naming and allows contributors to invent their own names for measures, resulting in a wide variety of naming conventions (e.g., an example measure from the BCL is a single run-on word "AedgK12InteriorLighting").

The analysis of word co-occurrence indicates that words generally occur more frequently on their own than in combination with other terms. Figure 4 shows an UpSet plot for five words of interest: controls, pump, fan, boiler, and insulation (four of these words are shown among the top 20 words in Figure 2). Each row represents one of the words of

Table 3. Sample of EEMs with their naming formats.

#	EEM	Format	Variation	Source
1	Insulation	Noun only	[N]	DOTY
2	De-lamping	Verb only	[V]	DOTY
3	Replace glazing	Verb-Noun	[V-N]	ATT
4	Jockey boilers	Noun only	[N-N]	DOTY
5	Cool roof	Noun only	[Adj-N]	NYTRM
6	Add heat recovery	Verb-Noun	[V-(N-N)]	ATT
7	Boiler combustion fan control	Noun only	[N(x4)]	DOTY
8	Lower chilled water condensing temperature	Verb-Noun	[V-(Adj-N)-(Adj-N)]	DOTY
9	Angled filters instead of flat filters	Existing-Proposed	–	DOTY
10	Convert system from steam to hot water	Existing-Proposed	–	ATT
11	Double layers of gypsum board as a way of getting increased thermal storage capacity.	Complex	–	1651RP
12	In any spaces with fenestration, evaluate opportunities for daylight harvesting by determining the spatial daylight autonomy (sDA) in accordance with IES LM-83. In spaces where $sDA_{300,50\%}$ is greater than 55%, consider installing daylight switching or daylight dimming controls (and appropriate ballasts if the lighting system is fluorescent or high-intensity discharge [HID]) to reduce use of electric lighting.	Complex	–	STD100

interest, and the left barplot represents the total number of EEMs containing that word. The bottom part of the plot represents every possible combination of words, and the top barplot represents the number of EEMs containing that combination of words. Note that the counts in Figure 4 are based on the number of EEMs in which a word appears, and these differ from the counts in Figure 2, which are based on the number of times the word itself occurs. Figure 4 shows that EEMs that contain only one of these words occur far more frequently than those with combinations of these words. Figure 4 also shows that some words occur in combination more frequently than others. The words pump, controls, and fan all have multiple intersections with each other, whereas the word "insulation" only co-occurs in boiler EEMs.

The results of the POS tagging revealed that each EEM in the corpus can generally be grouped into one of five typical EEM formats: verb-noun, verb only, noun only, existing-proposed, and complex. Note that the first three formats (verb-noun, verb only, noun only) were a direct result from the POS tagger, whereas the last two formats were identified using manual interpretation of the POS tagger results. Table 3 shows the results for the POS tagging analysis for 12 example EEMs from the overall list, along with their source of origin, arranged in the ascending order of their length. The various parts of speech are represented in Table 3 using the following abbreviations: verb (V), noun (N), adjective (Adj). The 12 example EEMs were selected to illustrate the full range of typical formats and their variations.

Most of the EEMs in the corpus are in the verb-noun format, and variations in this format are shown by EEMs #3, #6, and #8 in Table 3. The EEMs with this syntax can also be described as having an action-component format, using one action word and one or more building components to describe the EEM. EEM #2 illustrates verb only format, in which the EEM contains only a verb. EEMs #1, #4, #5, and #7 are all noun only format, in which the EEM contains only a noun representing the building component affected by the EEM. EEMs #9 and #10 are variations on the existing-proposed format, in which both the existing condition and proposed condition are specified in the EEM name. Finally, EEMs #11 and #12 are full sentences and represent the complex format. EEMs with the complex format were mostly found in BEQ, IEA46, STD100 and WULF, documents already identified in Table 2 as containing longer, wordier EEMs.

Topic models were employed to uncover six hidden themes (or topics) within the documents using a probabilistic framework based on the frequency and co-occurrence of words. The results of the topic modeling are shown in Table 4. Each panel in the table represents a different topic and shows the top 15 words in that topic along with their corresponding beta probabilities. The words are displayed in decreasing order of their beta probabilities (represented as a percentage), which is the probability of that word belonging to that topic. Note that the topic model only determines which words belong to each topic, and the modeler is then left to interpret the results and determine how to describe

Table 4. Topic modeling distribution of words across topics.

Topic 1: CONTROLS/ REDUCE		Topic 2: SYSTEMS/ LIGHTING/WATER		Topic 3: HVAC/METRICS	
Word	Beta	Word	Beta	Word	Beta
heat	4.1%	install	3.5%	add	3.1%
cool	3.7%	system	2.6%	zone	1.8%
control	3.6%	light	2.0%	set	1.8%
air	3.2%	water	1.8%	build	1.5%
system	3.1%	use	1.6%	cop	1.4%
use	2.7%	replace	1.6%	eer	1.2%
water	2.3%	reduce	1.5%	doas	1.1%
high	2.2%	energy	1.1%	story	1.1%
temperature	2.1%	hour	0.9%	area	1.0%
reduce	1.8%	consider	0.9%	demand	1.0%
efficiency	1.7%	lamp	0.9%	economizer	1.0%
light	1.6%	sensor	0.8%	hvac	1.0%
chill	1.4%	build	0.8%	value	1.0%
fan	1.4%	zone	0.8%	type	1.0%
motor	1.2%	space	0.8%	efficiency	0.9%
Topic 4: HEATING		**Topic 5: INSTALL/ REPLACE**		**Topic 6: ACTIONS**	
Word	Beta	Word	Beta	Word	Beta
air	3.9%	install	9.3%	upgrade	5.7%
heat	2.6%	replace	3.4%	install	5.5%
pump	2.3%	unit	2.3%	replace	4.4%
boiler	2.0%	insulate	1.8%	add	3.4%
heater	1.9%	remove	1.5%	repair	3.3%
water	1.8%	pump	1.4%	system	3.1%
insulation	1.4%	fixture	1.4%	implement	2.8%
energy	1.2%	operation	1.4%	clean	2.7%
conditioner	1.2%	minimize	1.3%	operate	2.3%
fan	1.2%	automatic	1.3%	sequence	2.3%
high	1.2%	spray	1.2%	calibration	1.8%
light	1.2%	seal	1.1%	protocol	1.7%
low	1.1%	turn	1.1%	documentation	1.6%
recovery	1.1%	plant	1.0%	train	1.5%
furnace	1.1%	tank	1.0%	insulation	1.3%

each topic. The words with a probability of less than 1% of belonging to that topic were disregarded when coming up with the topic labels. The words used to describe the topics are shown in bold text at the top of each panel in Table 4.

Table 4 shows that Topic 1 (CONTROLS/REDUCE) is about adding controls to air-side and water-side heating and cooling systems. Some lower probability words in Topic 1 reveal that the topic could also encompass reducing the lighting system usage or adding more efficient lighting. The top words in Topic 2 (SYSTEMS/LIGHTING/WATER) suggest a fairly broad theme. It addresses systems, broadly speaking, and contains several verbs (install, use, replace, reduce). Light and water both have relatively high beta probabilities in this topic, so it may address lighting fixtures, reducing the lighting usage or water system usage. Topic 3 (HVAC/METRICS) appears be about adding components to air distribution systems and air-side HVAC (add, DOAS, demand, economizer, HVAC), and also includes performance metrics (COP, EER) and zone-related terms (zone, set). Topic 4 (HEATING) mostly consists of words related to the heating system (heater, heat pump, boiler, furnace) including both air-side (air, fan) and water-side (pump, boiler) systems. Topic 4 is also unique among the other topics in that it contains no verbs. Topic 5 (INSTALL/REPLACE) is about installing or replacing equipment (unit, pump, fixture, tank). There is a considerable difference in the probability of occurrence of the word install and the remaining words in this topic. Some of the lower probability words in Topic 5 reveal that the topic could also include weatherization measures (insulate, spray, seal). Topic 6 (ACTIONS) consists largely of a variety of verbs, describing all the actions that could be performed on various building systems, most prominently upgrade, install, replace, add, repair, implement, and clean.

The distribution of the above topics across the documents can be used to make inferences about similarities and differences between the documents. The breakdown of topics by

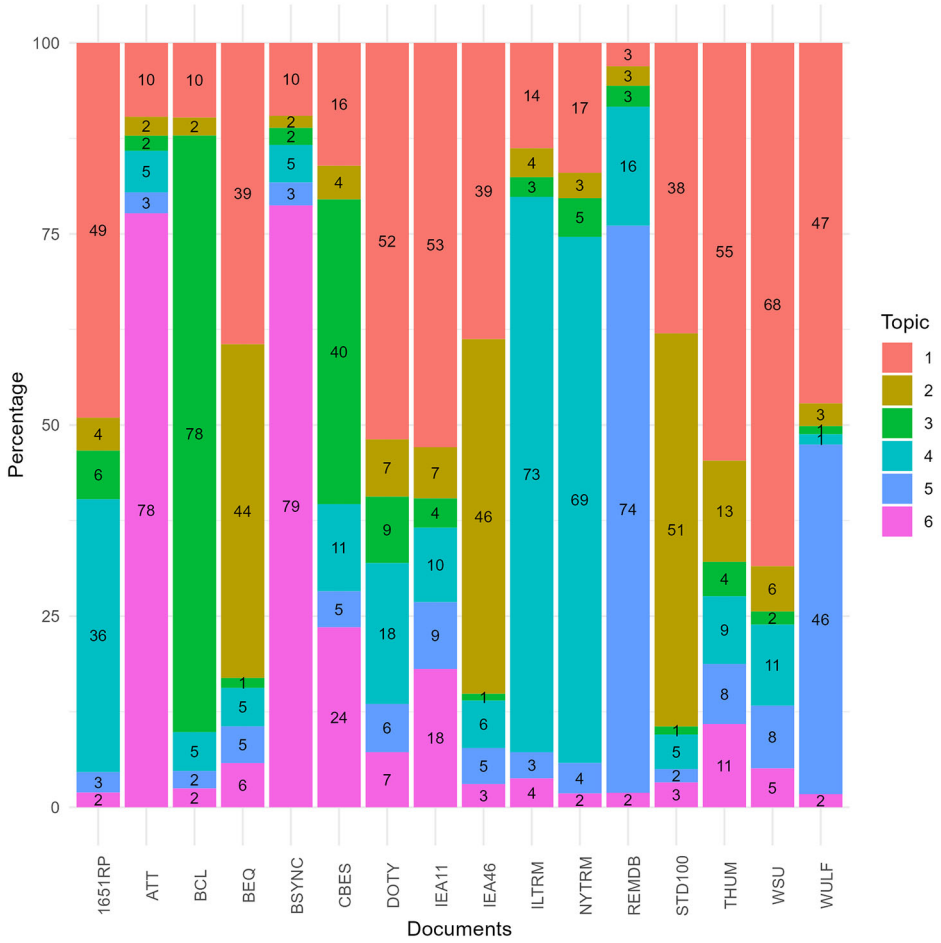

Fig. 5. Topic modeling distribution of topics across documents.

document is shown in Figure 5. The results indicate that Topic 1 (CONTROLS/REDUCE) accounts for a relatively high proportion in almost all the documents, reflecting control and conservation as core principles in many EEMs. Topic 1 comprises the majority of WSU, THUM, IEA11, and DOTY. Topic 2 (SYSTEMS/LIGHTING/WATER) is the majority topic in STD100, IEA46, and BEQ. The similar topic distributions for these documents match their historical development and dependence: BEQ was based on STD100, which was based on IEA46. Sizeable proportions of Topic 3 (HVAC/METRICS) occur in the BCL (78% Topic 3), as well as in CBES (40% Topic 3). Both of these documents are heavily linked to energy modeling, and Topic 3 could be considered the modeling-related topic. Both NYTRM and ILTRM contain a majority of words from Topic 4 (HEATING), which shows a prevalence of heating EEMs in these TRMs and suggests that there are similarities in the TRM structures in general. Topic 5 (INSTALL/REPLACE) accounts for 74% of REMDB and almost half of WULF. This is because the verbs replace, install, and insulate describe the action taken in the majority of the REMDB EEMs and also matches the high prevalence of the term "install" in WULF. BSYNC and ATT are largely composed of Topic 6 (ACTIONS) and have an almost equal breakdown across all six topics. This captures the prevalence of a variety of verbs in both of these documents, as well as the fact that BSYNC is the basis for ATT and the two documents contain identical lists of EEMs. Note that while the ATT and BSYNC lists are exactly the same, the gamma distribution shows a slight difference. This is due to the fact that topic modeling algorithm begins by randomly assigning words to topics and then iteratively improves those assignments.

The similarities and differences between documents observed in the topic models are also illustrated in the cosine similarity analysis. Table 5 shows cosine similarity (in percentage) between each pair of documents as a pairwise matrix. These scores range from 14% to 100% (i.e., identical) similarity. The cells for document pairs that are over 60% similar are shown in bold text with grey shading. The results show that BSYNC and ATT are identical, again reflecting the fact that BSYNC is the basis for ATT. STD100 is 94% identical to BEQ and 96% identical to IEA46; whereas IEA46 is 91% identical to BEQ. This again reflects the fact that BEQ was based on STD100, which was based on IEA46. The most dissimilar documents are DOTY and REMDB with a cosine similarity score of only 14%.

Discussion and conclusions

This study used a range of text mining techniques to discover trends in how existing resources describe and organize

Table 5. Cosine similarity matrix.

	WULF	WSU	THUM	STD100	REMDB	NYTRM	ILTRM	IEA46	IEA11	DOTY	CBES	BSYNC	BEQ	BCL	ATT	1651RP
1651RP	60	**68**	57	**73**	20	59	**62**	**72**	**71**	**63**	44	27	**71**	35	27	**100**
ATT	49	38	32	46	36	20	24	47	47	19	49	**100**	52	24	**100**	
BCL	28	24	24	36	18	24	25	33	32	28	33	24	37	**100**		
BEQ	**80**	**73**	**62**	**94**	34	50	50	**91**	**76**	52	46	52	**100**			
BSYNC	49	38	32	46	36	20	24	47	47	19	49	**100**				
CBES	38	44	31	44	26	32	34	43	49	30	**100**					
DOTY	46	44	46	54	14	56	48	55	56	**100**						
IEA11	**68**	**71**	**69**	**77**	30	52	49	**79**	**100**							
IEA46	**80**	**80**	**73**	**96**	30	51	53	**100**								
ILTRM	47	54	41	50	18	**70**	**100**									
NYTRM	44	46	41	49	22	**100**										
REMDB	30	19	26	30	**100**											
STD100	**76**	**76**	**67**	**100**												
THUM	55	**66**	**100**													
WSU	**71**	**100**														
WULF	**100**															

EEMs. A unique list of 3,490 EEMs from 16 different documents was compiled through a literature review and analyzed using several text mining methods: frequency analysis, POS tagging, and topic modeling. The results of this analysis revealed three major trends about the nature of EEMs and the ways they are currently described.

First, the summary word counts and POS tagging identified typical EEM length and structure, as well as their variations. A typical EEM is around six words in length and is phrased in verb-noun format. However, there is enormous variety in these characteristics across the documents, with EEM length differing by two orders of magnitude, from as low as one word to more than 100 words. EEM syntax varied widely, with some EEMs consisting of only a verb or only a noun, while others detailed both the existing and proposed condition relative to the EEM.

Second, the frequency counts of words and bigrams, as well as the co-occurrence analysis using UpSet plots, established that there are words and bigrams that are commonly used across EEMs. These common terms include both verbs and nouns, and the nouns include specific building components and technologies, broader building systems, and descriptor terms. The frequency counts also suggested that synonymous terms and abbreviations are common in EEM names. Although it only included five terms, the UpSet plot showed that these terms occur in an EEM more frequently in isolation, rather than in combination, providing preliminary evidence that each EEM can potentially be characterized by a single primary noun. However, the results also showed that not all of these common terms and bigrams occurred in all of the documents, indicating that while common EEM terminology exists, each document ultimately has its own unique vocabulary.

This wide array of terminology has important impacts on working with EEM data. Terms such as “high efficiency” and “high performance” are common in the building energy efficiency industry but are vague in meaning. Similarly, synonymous terms and abbreviations (e.g., heat recovery and HRV) are also commonly used, but can be confusing, especially to non-experts. The use of alternative terminology across different documents also impacts the data cleaning and pre-processing phase in text mining. Because tokenization removes symbols and punctuation, alternative spellings of terms are treated differently. For example, the term “T-8” would be treated as two separate tokens “T” and “8,” while “T8” would be treated as a single token. This should be considered when cleaning EEM-related text data, as punctuated words are common (e.g., “low-e,” “A/C”). Overall, these challenges suggest that when naming EEMs, terminology should be carefully selected and vague and synonymous terminology avoided. Issues related to synonymous terminology and abbreviations in EEM names are discussed further in Webb and Khanuja (2022).

Third, the topic modeling uncovered six underlying themes, which along with the cosine similarity results highlight similarities and differences between the documents. The topic modeling yielded several unique insights. First, the use and selection of verbs in EEM names are a key

differentiator between documents. The topic modeling suggests that documents can use a wide variety of verbs (documents with large percentages of TOPIC 6: ACTIONS), be limited to just a few verbs (documents with large percentages of TOPIC 5: INSTALL/REPLACE), or use minimal verbs (TOPIC 4: HEATING). Given the importance of an action to the definition of an EEM, this differential treatment of verbs across documents was somewhat surprising. Second, dependencies between documents are important and not always explicit a priori. Documents that were not necessarily expected to be similar were revealed to be similar through topic modeling and cosine similarity. For example, IEA46, Standard 100, and Building EQ have a similar topic distribution and high cosine similarity, but there is nothing in these documents that indicates that they are related. The similarities between these documents observed in the topic modeling were confirmed by the authors via personal communication with the authors of these documents. These dependencies also suggest that the development and evolution of EEM lists have often been ad hoc rather than systematic, relying extensively on borrowing from previous lists. Conversely, documents that were expected to be similar were revealed to be dissimilar. One might reasonably expect the types of document identified in Table 1 (e.g., tool, TRM, handbook, standard, other) to be reflected in the topic modeling, but that was not necessarily the case. For example, WULF is a handbook but has a very different topic distribution than DOTY and THUM, which are also handbooks.

Previous studies have demonstrated the ability of text mining methods to provide insights on large collections of unstructured data from a variety of different fields (Bastani, Namavari, and Shaffer 2019; Bouabdallaoui et al. 2020), including building energy-related data (Lai and Kontokosta 2019; Lai et al. 2022). The results from this study provide further evidence of text mining's utility, and further expands its application to EEMs. Lai et al. (2022) observed inconsistent EEM naming to be a significant concern in working with EEM-related data, and this study systematically demonstrates the significant variation in EEM names. It also identifies trends that can inform standardization efforts and help resolve these data quality concerns. This study suggests that there are several key features of an EEM name that, if standardized, might improve data quality: length (how succinct is it?), terminology (does it use synonyms, abbreviations, or vague terms?), and format (does it use verbs, and which verbs?). Using text mining, these features can be used to analyze the characteristics and consistency in a set of EEM names. Ultimately, this study has shown how EEM names are not only important for communicating the intent of a measure, but also serve as a powerful organizing and analytic principle.

There are several notable limitations to this study. First, this study uses a relatively small number of documents for analysis. Studies that apply text mining to analyze documents often do so on large corpora containing thousands of documents (Lai et al. 2022; Abdelrahman et al. 2021; Bastani, Namavari, and Shaffer 2019; Wang, Bowers, and Fikis 2017). Using a small number of documents leads to a sparse DTM, which can sometimes reveal strong associations between documents or topics that might not be true. Reducing the sparsity of the DTM by ignoring the terms below a certain threshold term frequency value could potentially reveal more accurate insights (Wang, Bowers, and Fikis 2017). Second, the POS tagger was a useful aid in uncovering the syntactical structure of EEM names, but was not trained on EEM data prior to use, resulting in many incorrectly tagged terms. EEM names present a special context for determining parts of speech, as many EEM names are simply phrases instead of full sentences, and as common terminology is often used in specific ways (e.g., the words "heat" and "pump" are almost exclusively used as nouns in the corpus but were frequently tagged as verbs). The POS tagger would need to be better trained before the POS analysis could be fully automated. Third, the topic modeling was completely unsupervised. Several prior studies recommend starting with a dictionary of common terms related to the domain prior to text mining (Abdelrahman et al. 2021; Lai et al. 2022). This extra step may provide additional insight into the data, as a topic model with pre-defined seed words would search for words with high beta probabilities around these initial words.

More broadly, the insights gained from this study can be used as the basis for developing a standardized system for organizing and describing EEMs. The findings here have two larger implications for standardizing EEMs and EEM-related data: first, the need for a common EEM format, second, the need for a standardized EEM vocabulary.

The wide variation in EEM length and format points to the need for a common format in a standardized system. The variation in format also suggests that some EEMs are far more explicit than others in describing the intended action, and that the specifics of an EEM are often implied. As just one example, the EEM "Replace boiler" (from BSYNC) does not explicitly state what the boiler is being replaced with, and it is left implied that it is being replaced with another, more efficient boiler. Such implicitness is not desirable in a standardized system, and there is therefore a need to develop an explicit EEM format.

The existence of common words and bigrams in the corpus suggests that these could be a promising basis for tagging and sorting EEMs. A standard, preset vocabulary of verbs and nouns could be constructed to support this. These terms would effectively act as the basic building blocks for EEM names. This would expand on existing efforts to standardize building energy efficiency terminology, such as the Building Energy Data Exchange Specification (BEDES) (Lawrence Berkeley National Laboratory 2020b), which does not directly address the standardization of EEMs. However, the finding in this study that not all of the common words occurred in all of the documents suggests that EEM terminology is highly diverse, and a common vocabulary may therefore be challenging to develop.

Future work should continue to explore and improve the application of text mining methods to building energy-related data. While much of the data related to building energy performance is quantitative in nature, mining available unstructured, qualitative data, such as EEM names and

descriptions, can provide important insight about the building stock. As the amount of textual data continues to grow with the expansion of mandatory energy audit ordinances—15 U.S. cities currently have such a policy (Institute for Market Transformation 2021)—and other ambitious energy policies such as building performance standards targeting existing buildings, the ability to make use of this data to improve building energy performance is becoming increasingly important.

Most importantly, future work should leverage the results from this study to develop a standardized system for categorizing EEMs. In addition to a categorization hierarchy, this system should define a common format for EEMs, as well as a standard vocabulary of terms that can be used to tag and sort EEMs. The development of such a system would enable key stakeholders—including energy auditors, incentive program managers, and policymakers—to more clearly communicate the intent of an EEM and share EEM-related data across building projects, portfolios, and programs.

Acknowledgments

The authors gratefully acknowledge the members of the Project Monitoring Subcommittee (PMS) for their guidance and feedback: Chris Balbach (Chair), Rob Hitchcock, and Adam Hinge. Special thanks to the members of the Project Advisory Board (PAB) for generously sharing their time, technical expertise, and insight: Marco Ascazubi, Honey Berk, David Hodgins, Jim Kelsey, Nicholas Long, Paul Mathew, Ben O'Donnell, and David Sachs. Additional thanks to others who contributed their expertise at PMS and PAS meetings, especially Travis Walter and Harry Bergmann.

Funding

This work was funded by ASHRAE through 1836-RP, Developing a Standardized Categorization System for Energy Efficiency Measures. The project was sponsored by TC 7.6 Building Energy Performance, and co-sponsored by SSPC 100 Energy Efficiency in Existing Buildings and the Building EQ Committee.

ORCID

Apoorv Khanuja http://orcid.org/0000-0001-6287-5484
Amanda L. Webb http://orcid.org/0000-0002-1941-6087

Data availability

The data that support the findings of this study are openly available in Zenodo at http://doi.org/10.5281/zenodo.6726629. The code used to produce this analysis is available at: https://github.com/retrofit-lab/ashrae-1836-rp-text-mining.

References

Abdelrahman, M. M., S. Zhan, C. Miller, and A. Chong. 2021. Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature. *Energy and Buildings* 242:110885. doi:10.1016/j.enbuild.2021.110885

Aldous, D. 1985. *Exchangeability and related topics*. Berlin: Springer.

ASHRAE. 2011. *Procedures for commercial building energy audits*. 2nd ed. Atlanta: ASHRAE.

ASHRAE. 2018a. *ASHRAE standard 211-2018, Standard for commercial building energy audits*. Atlanta: ASHRAE.

ASHRAE. 2018b. *ASHRAE standard 100-2018, Energy efficiency in existing buildings*. Atlanta: ASHRAE.

ASHRAE. 2019. *Achieving zero energy: Advanced energy design guide for small to medium office buildings*. Atlanta: ASHRAE. https://aedg.ashrae.org/.

ASHRAE. 2020. "Building EQ." https://buildingeq.ashrae.org/.

Bastani, K., H. Namavari, and J. Shaffer. 2019. Latent Dirichlet Allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications* 127:256–71. doi:10.1016/j.eswa.2019.03.001

Benoit, K., D. Muhr, and K. Watanabe. 2021. "Stopwords: Multilingual Stopword Lists." https://CRAN.R-project.org/package=stopwords.

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55 (4):77–84. doi:10.1145/2133806.2133826

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Bouabdallaoui, Y., Z. Lafhaj, P. Yim, L. Ducoulombier, and B. Bennadji. 2020. Natural language processing model for managing maintenance requests in buildings. *Buildings* 10 (9): 160. doi:10.3390/buildings10090160

Conway, J. R., A. Lex, and N. Gehlenborg. 2017. UpSetR: An R Package for the visualization of intersecting sets and their properties. *Bioinformatics* 33 (18):2938–40. doi:10.1093/bioinformatics/btx364.

DiMaggio, P., M. Nag, and D. Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. Government arts funding. *Poetics, Topic Models and the Cultural Sciences* 41 (6): 570–606. doi:10.1016/j.poetic.2013.08.004

Doty, S. 2011. *Commercial energy auditing reference handbook*. 2nd ed. Boca Raton: Fairmont Press.

Glazer, J. 2015. *Development of maximum technically achievable energy targets for commercial buildings: Ultra-low energy use building set. ASHRAE Research Project 1651-RP Final Report*. Arlington Heights, IL: Gard Analytics.

Grün, B, and K. Hornik. 2011. Topicmodels: An R package for fitting topic models. *Journal of Statistical Software* 40 (13):1–30. doi:10.18637/jss.v040.i13

Hearst, M. A. 1999. Untangling text data mining. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, 3–10. USA: Association for Computational Linguistics. doi:10.3115/1034678.1034679

Hong, L, and B. D. Davison. 2010. Empirical study of topic modeling in Twitter. In Proceedings of the First Workshop on Social Media Analytics, 80–88. New York, NY, USA: Association for Computing Machinery. doi:10.1145/1964858.1964870

Hong, S., J. Kim, and E. Yang. 2022. Automated text classification of maintenance data of higher education buildings using text mining and machine learning techniques. *Journal of Architectural Engineering* 28 (1):4021045. doi:10.1061/(ASCE)AE.1943-5568.0000522

Hong, T., M. A. Piette, Y. Chen, S. H. Lee, S. C. Taylor-Lange, R. Zhang, K. Sun, and P. Price. 2015. Commercial building energy saver: An energy retrofit analysis toolkit. *Applied Energy* 159: 298–309. doi:10.1016/j.apenergy.2015.09.002

Illinois Energy Efficiency Stakeholder Advisory Group. 2019. *2020 Illinois Statewide Technical Reference Manual for Energy Efficiency Version 8.0.* https://www.ilsag.info/technical-reference-manual/il_trm_version_8/.

Institute for Market Transformation. 2021. "Comparison of U.S. Commercial Building Energy Benchmarking and Transparency Policies." https://www.imt.org/resources/comparison-of-commercial-building-benchmarking-policies/.

Khanuja, A, and A. Webb. 2022. ASHRAE 1836-RP Main List of Energy Efficiency Measures (v1.0) [Data Set]. *Zotero*. doi:10.5281/zenodo.6726629

Lai, Y, and C. E. Kontokosta. 2019. Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. *Computers, Environment and Urban Systems* 78:101383. doi:10.1016/j.compenvurbsys.2019.101383

Lai, Y., S. Papadopoulos, F. Fuerst, G. Pivo, J. Sagi, and C. E. Kontokosta. 2022. Building retrofit hurdle rates and risk aversion in energy efficiency investments. *Applied Energy* 306:118048. doi:10.1016/j.apenergy.2021.118048

Lawrence Berkeley National Laboratory. 2020a. "Commercial Building Energy Saver". http://cbes.lbl.gov/.

Lawrence Berkeley National Laboratory. 2020b. "Building Energy Data Exchange Specification (BEDES)". https://bedes.lbl.gov/.

Long, N., K. Fleming, C. CaraDonna, and C. Mosiman. 2021. BuildingSync: A schema for commercial building energy audit data exchange. *Developments in the Built Environment* 7 (July): 100054. doi:10.1016/j.dibe.2021.100054

Lyberg, Mats Douglas, Ed. 1987. *Source book for energy auditors* (Vol. 1). Stockholm, Sweden: Swedish Council for Building Research. https://www.iea-ebc.org/projects/project?AnnexID=11.

National Renewable Energy Laboratory. 2018. "National Residential Energy Efficiency Measures Database, Version 3.1.0". https://remdb.nrel.gov/.

National Renewable Energy Laboratory. 2020a. "Building Component Library". https://bcl.nrel.gov/.

National Renewable Energy Laboratory. 2020b. "BuildingSync, Version 2.0". https://buildingsync.net/.

New York State Joint Utilities. 2019. *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs – Residential, Multi-Family, and Commercial/Industrial Measures Version 7.* http://www3.dps.ny.gov/W/PSCWeb.nsf/All/72C23DECFF52920A85257F1100671BDD.

Nguyen, D. Q., D. Q. Nguyen, D. D. Pham, and S. B. Pham. 2014. RDRPOSTagger: A ripple down rules-based part-of-speech tagger. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, 17–20. Gothenburg, Sweden: Association for Computational Linguistics. doi:10.3115/v1/E14-2005

Pacific Northwest National Laboratory. 2020. "Audit Template, Release 2020.2.0." https://buildingenergyscore.energy.gov/.

R Core Team. 2014. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rinker, T. W. 2018. *Textstem: Tools for stemming and lemmatizing text* (version 0.1.4). Buffalo, New York. http://github.com/trinker/textstem.

Silge, J, and D. Robinson. 2016. Tidytext: Text mining and analysis using tidy data principles in R. *The Open Journal* 1 (3):37. doi:10.21105/joss.00037

Thumann, Albert, Ed. 1992. *Energy conservation in existing buildings deskbook*. Lilburn, GA: Fairmont Press.

Wang, Y., A. J. Bowers, and D. J. Fikis. 2017. Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of EAQ articles from 1965 to 2014. *Educational Administration Quarterly* 53 (2):289–323. doi:10.1177/0013161X16660585

Washington State University Cooperative Extension and Energy Program. 2003. *Washington State University Energy Program Energy Audit Workbook* WSUCEEP2003-043. http://www.energy.wsu.edu/PublicationsandTools.aspx.

Webb, A, and A. Khanuja. 2022. Developing a Standardized Categorization System for Energy Efficiency Measures. Final Report RP-1836. ASHRAE.

Wulfinghoff, D. 1999. *Energy efficiency manual: For everyone who uses energy, pays for utilities, controls energy usage, designs and builds, is interested in energy and environmental preservation*. Wheaton, MD: Energy Institute Press.

Zhao, W., J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou. 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics* 16 (S13):S8. doi:10.1186/1471-2105-16-S13-S8

Zhivov, Alexander, and Cyrus Nasseri, eds. 2014. *Energy efficient technologies and measures for building renovation: Sourcebook*. IEA ECBS Annex 46. https://www.iea-ebc.org/Data/publications/EBC_Annex_46_Technologies_and_Measures_Sourcebook.pdf.