Structures 35 (2022) 577–590

Contents lists available at ScienceDirect

Structures

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

The twenty-first century of structural engineering research: A topic modeling approach

Yazhou Xie [*], Chunxiao Ning, Lijun Sun

Department of Civil Engineering, McGill University, QC H3A0C3, Canada

ARTICLE INFO

Keywords:
Topic modeling
Latent Dirichlet allocation
Structural engineering research
Cold and hot topics
Journal topic distribution
Country/region topic distribution

ABSTRACT

Aiming at disclosing a general research landscape of structural engineering in the twenty-first century, this study applies the latent Dirichlet allocation (LDA), a topic modeling approach, to analyze 51,346 article abstracts from 23 prestigious journals in structural engineering with a publication period from 2000 to 2020. The LDA analyzes the literature inventory by extracting 50 distinguishable wordclouds, each centered around one distinct research theme and assigned a unique topic name. Subsequently, various measures have been proposed to integrate the posterior distributions of these research topics with article information such as publication year, journal name, and correspondence address. The increase index identifies five cold and hot topics, which reflect the shift of research interests in the community. Emerging research topics such as seismic risk assessment and composite material have received much more attention in recent years. Moreover, advanced metrics have been proposed to analyze the research similarity and evolution across different journals and countries/regions. As discussed in the paper, analysis findings would enable community stakeholders (e.g., students, engineers, researchers, conference organizers, journal editors, funding agencies) to explore the state of the research and develop viable strategies to further foster the healthy growth of the community. Such strategies can be (1) researchers submitting a paper to the most appropriate journal; (2) journal editors adjusting the journal focus to enhance its impact; and (3) funding agencies prioritizing research supports that best fit regional needs and circumstances, among others.

1. Introduction

As one of the oldest engineering disciplines, structural engineering deals with the analysis and design of buildings, bridges, and other constructed facilities that support self-weight and resist other imposed loads. The evolution of structural engineering features an incremental process that began with empirical approaches by observing actual behaviors, followed by more scientific methods as testing of materials and elements became possible, and led to the development of standards and regulations over time to recognize practice and research findings [1]. The development of numerical modeling and structural health monitoring in the twenty-first century further fostered a hierarchy of research that addressed various emerging topics at different scales. As expected to become master builders, stewards of the environment, innovators, managers of risk, and leaders in public policy [2], the next generation of civil engineers should be made aware of the established solid research base in structural engineering [1]. Motivated by this need, this study aims to leverage the recent advances in statistical and machine learning [3,4] to explore the expansive body of knowledge embedded in the numerous peer-reviewed publications in structural engineering. Moreover, this exploration will enable scientific experts to have a solid grasp on which topics are relevant to their interests, which research areas are rising or falling in popularity, and how different topics are distributed in different journals and countries/regions. Analysis outcomes from this study will also help journal editors and funding agencies to identify and prioritize novel research topics that bear a strong promise to impact the discipline.

As a popular statistical tool for text analysis, topic modeling extracts latent variables from a large collection of documents [5]. In general, approaches for topic modeling can be categorized as latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA), latent Dirichlet allocation (LDA), and correlated topic model (CTM), among others. LSA represents the text as a document term matrix and applies singular value decomposition (SVD) to reduce its dimensionality and encode it using latent features (i.e., topics) [6], while PLSA expands the LSA with a foundation of statistics – instead of relying on the SVD, the PLSA is based on the likelihood principle and defines a proper generative model of the data [7,8]. As a step further, LDA develops a generative

[*] Corresponding author.
E-mail address: tim.xie@mcgill.ca (Y. Xie).

https://doi.org/10.1016/j.istruc.2021.11.018
Received 9 September 2021; Received in revised form 8 November 2021; Accepted 9 November 2021
Available online 25 November 2021

probabilistic model of a corpus by representing documents as random mixtures over latent topics and topics as probabilistic distributions over words [9]. By using the same inference framework, researchers have also made extensions to the LDA, where improved models include the CTM that captures the correlations across different topics [10], and the dynamic topic model [11] and topic over time model [12] that analyze the evolution of latent topics over time.

LDA is arguably the most popular approach in topic modeling; it has been widely applied to natural language processing, text mining, social media analysis, and information retrieval, etc. [13]. For dealing with research articles, one of the first attempts has been made by Griffiths and Steyvers [14], who used LDA to analyze article abstracts from the *Proceedings of the National Academy of Science* (PNAS). Their study has shown the effectiveness and consistency of the extracted research topics in capturing the meaningful latent structure in the documents. Following this study, LDA has been further leveraged to (1) develop the author-topic model where authorship information is included as a multinomial distribution over topics [15]; (2) identify research topics and top-cited papers in computer-based sentiment analysis [16]; (3) understand topic evolution by incorporating the citation network [17]; (4) infer key research topics in transportation research [18]; and (5) explore the sustainability literature in maritime studies [19], etc. In addition, Yalcinkaya and Singh [20] have also applied LSA to identify principle research areas in the field of building information modeling (BIM). Recently, Ezzeldin and El-Dakhakhni [21] have used LDA to analyze research articles from two structural engineering journals, namely *Journal of Structural Engineering* and *Engineering Structures*.

Motivated by the need to disclose a general research landscape for the discipline of structural engineering in the twenty-first century, this study extracts 51,346 articles from 23 related prestigious journals with a publication period from 2000 to 2020. In this regard, previous studies have proven that consistent topic coherence and topic ranking can be achieved by using abstract versus full-text data when the document collection is large [22]. Therefore, the LDA analysis framework developed by Sun and Yin for transportation research [18] is adapted herein to analyze the collected large inventory of article abstracts in structural engineering. This study first discusses the theoretical background of implementing LDA in topic modeling, and then introduces the extracted abstract data and the methods used for data processing. The applied LDA framework successfully identifies 50 distinguishable research topics, which further enables an extensive topic analysis that involves a variety of measures to quantify the topic distributions against time, journal, and country/region. Finally, a discussion of analysis findings and potential applications concludes the article in anticipation of stimulating more relevant discussions within stakeholders toward the healthy growth of the research community.

2. Latent Dirichlet allocation (LDA) for topic modeling

Topic modeling aims to automatically uncover the hidden thematic structure from a collection of documents. As a generative probabilistic model introduced by Blei et al. (2003) [9], the LDA utilizes a three-level hierarchical Bayesian model to postulate a topic structure that can most likely generate the observed document-word data. LDA is an unsupervised model where the hidden topic structure is captured by obtaining the posterior distribution given the observed documents. The generative process of LDA is described below:

1) Define K topics and determine the word distribution for every topic k as β_k $\text{Dirichlet}_V(\eta)$, where the subscript V is the size of vocabulary bank and β_k is the parameter for the multinomial word distribution for topic $k = 1, 2, \ldots, K$.
2) Determine the topic distribution for every document d as θ_d $\text{Dirichlet}_K(\alpha)$, where θ_d is the parameter for the multinomial topic distribution for document $d = 1, 2, \ldots, D$.
3) The n_{th} word in each document d is first assigned with a topic $z_{d,n}$ $\text{Multinomial}_K(\theta_d)$ based on the topic distribution in this document, where $z_{d,n}$ represents the n_{th} word topic assignment for document d.
4) The n_{th} word in each document d is then determined as $w_{d,n}$ $\text{Multinomial}_V(\beta_{z_{d,n}})$ according to its topic assignment ($z_{d,n}$) and the per-topic word distribution (β_k). This step generates N words $w_{d,n}$ for document d where $n = 1, 2, \ldots, N$.

This generative process is illustrated in Fig. 1 using a probabilistic graphical model in plate notation [23]. In Fig. 1, the unshaded nodes represent the hidden random variables, the shaded nodes the observed random variables, and the edges the conditional dependencies between them. The rectangles are called plates that represent replication. As is depicted, each topic β_k is considered as a Dirichlet distribution, $\beta_k \sim \text{Dirichlet}_V(\eta)$, over the vocabulary V, while every document is represented as a separate Dirichlet distribution, $\theta_d \sim \text{Dirichlet}_V(\alpha)$, over K topics. As such, each word in document d is generated by assigning a topic ($z_{d,n}$) and choosing a word based on β_k under the given topic k. The sampled words are then compared with the observed words, and the joint distribution of all the hidden variables β_K (topics), θ_D (per-document topic proportions), z_D (word topic assignments), and observed variables w_D (words in documents) is expressed by:

$$p(\beta_K, \theta_D, z_D, w_D|\alpha, \eta) = \prod\nolimits_{k=1}^{K} p(\beta_k|\eta) \prod\nolimits_{d=1}^{D} p(\theta_d|\alpha) \prod\nolimits_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_k) \tag{1}$$

With the aid of chain rule in probability [24], the joint distribution shown in Eq. (1) provides a viable solution to guide both the training and inference processes for LDA. In particular, the dependencies shown in Eq. (1) are utilized to compute the posterior distribution of LDA's topic structure, defined by the per-topic word distribution parameter β_K, per-document topic distribution parameter θ_D, and per-word topic assignment z_D. The posterior of these concurring parameters can be expressed as shown in Eq. (2). It is worth mentioning that Eq. (1) can also be utilized to infer other posterior distributions such as the per-document topic distribution $p(\theta_D|w_D)$ and per-topic word distribution $p(\beta_K|w_D)$ by marginalizing other relevant variables.

$$p(\beta_K, \theta_D, z_D|w_D) = \frac{p(\beta_K, \theta_D, z_D, w_D)}{p(w_D)} \tag{2}$$

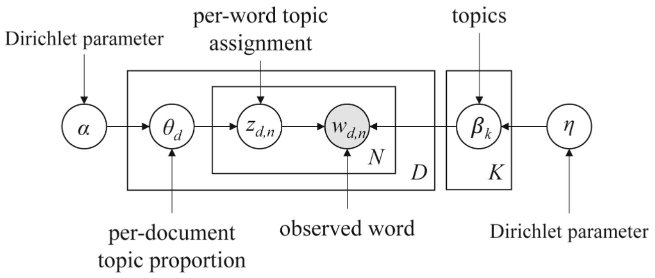


Fig. 1. Graphical model representation of LDA.

Computation of the posterior is intractable because of the denominator, which can be dealt with by applying the variational expectation–maximization algorithm [9] or through Gibbs sampling [14]. These techniques can provide a close approximation to the true posterior through statistical inference. To this end, the identified topics and their per-document distributions are coupled with the article information (i.e., journal name, publishing year, and country of the corresponding author's affiliation) to discover the temporal and regional trends in structural engineering research.

3. Article-abstract data in structural engineering research

Structural engineering is defined as a discipline that deals with the analysis and design of buildings, bridges, and other constructed facilities that support self-weight and resist other imposed loads. Based on this definition, 23 prestigious journals listed in Table 1 are selected from the *Web of Science Core Collection* under the search category "Engineering, Civil". The journal list chosen excludes several top-tier interdisciplinary journals that also publish new research findings in other fields. For instance, *Computer-aided Civil and Infrastructure Engineering* also covers transportation, water resources engineering, and management of infrastructure systems; building energy, maintenance, and management are prevalent research topics in the *Journal of Building Engineering*; the *International Journal of Structural Stability and Dynamics* welcomes research articles that deal with aerospace structures, marine structures, biostructures, and nano-structures; and *Computers & Structures* includes papers in all areas of mechanics. Likewise, this study also excludes some top-notch journals that mainly focus on construction materials (e.g., *Construction and Building Materials*, *Journal of Composites for Construction*, *Structure Concrete*, *Journal of Materials in Civil Engineering*, etc.) and geotechnical engineering (e.g., *Journal of Geotechnical and Geoenvironmental Engineering*, *Soil Dynamics and Earthquake Engineering*, etc.).

Table 1
Journal data considered in structural engineering research.

Journal	Abbreviation	Articles	Year
ACI Structural Journal	ACI STRUCT J	1936	2000–2020
Bulletin of Earthquake Engineering	B EARTHQ ENG	1893	2003–2020
Earthquake Engineering & Structural Dynamics	EARTHQ ENG STRUCT D	2267	2000–2020
Earthquake Spectra	EARTHQ SPECTRA	1517	2002–2020
Earthquakes and Structures	EARTHQ STRUCT	940	2010–2020
Engineering Structures	ENG STRUCT	10,059	2000–2020
Journal of Bridge Engineering	J BRIDGE ENG	1930	2003–2020
Journal of Constructional Steel Research	J CONSTR STEEL RES	4257	2000–2020
Journal of Earthquake Engineering	J EARTHQ ENG	1189	2000–2020
Journal of Performance of Constructed Facilities	J PERFORM CONSTR FAC	1617	2002–2020
Journal of Structural Engineering	J STRUCT ENG	3083	2000–2020
Journal of Wind Engineering and Industrial Aerodynamics	J WIND ENG IND AEROD	2872	2000–2020
Smart Structures and Systems	SMART STRUCT SYST	1277	2005–2020
Steel and Composite Structures	STEEL COMPOS STRUCT	1693	2002–2020
Structural Control & Health Monitoring	STRUCT CONTROL HLTH	1458	2005–2020
Structural Design of Tall and Special Buildings	STRUCT DES TALL SPEC	1107	2003–2020
Structural Engineering and Mechanics	STRUCT ENG MECH	3520	2000–2020
Structural Safety	STRUCT SAF	816	2000–2020
Structure and Infrastructure Engineering	STRUCT INFRASTRUCT E	1230	2005–2020
Structures	STRUCTURES	1203	2015–2020
Sustainable and Resilient Infrastructure	SUS RES INFRASTRUCT	76	2016–2020
Thin-Walled Structures	THIN WALL STRUCT	4444	2000–2020
Wind and Structures	WIND STRUCT	962	2000–2020

The abstract data of each selected journal is extracted from the *Web of Science* (https://webofknowledge.com/). A preliminary data analysis indicates a significant temporal oscillation of research topics for abstracts published before year 2000, which results from that (1) more than half of the journals were newly launched in this century; and (2) much fewer articles were published in each existing journal in the last century. Therefore, this study only considers the articles published since 2000, framing a general study scope of exploring themes and trends in structural engineering research in the twenty-first century. Eventually, 51,346 article abstracts have been collected as the document inventory that spans 21 years from 2000 to 2020. Table 1 also lists the total number of articles obtained from each journal, where a significant journal variability can be observed: *Engineering Structures* owns the largest body of article data (10,059 articles), which is 130 times more than those collected from the recently created journal of *Sustainable and Resilient Infrastructure*. Moreover, Fig. 2 provides a journal temporal disaggregation of the article data, showing a generally increasing number of published articles for most of the journals.

The collected article abstracts are further utilized to extract a word corpus for topic modeling. In this respect, the following steps have been carried out to preprocess the abstract data such that it can be conveniently analyzed through LDA. First, a full abstract is split into a series of words using delimiters such as space, comma, and colon. Two types of words are then eliminated from the corpus: those that appear less than 10 times or belong to the standard *stop list* recommended by the Natural Language Toolkit (http://www.nltk.org/). This study also removes a list of common words (i.e., those appearing more than 1500 times or more than 80% of all the abstracts) that bear trivial contextual meaning, such as *loading*, *model*, *result*, *use*, *effect*, *structure*, *apply*, *reduce*, *enhance*, *require*, etc. Moreover, a translation table is established for lemmatization to ensure that the same words in different forms are interchangeable. For instance, both *isolator* and *isolate* are considered the same as *isolation*, while *optimization* is equivalent to *optimal*, *optimum*, *optimize*, *optimisation*, etc. As an essential step, n-grams analysis has been conducted to automatically identify bigrams and trigrams by combining two and three words in a contiguous sequence. This analysis focuses on those word combinations that appear at least 200 times in the corpus. To this end, the total vocabulary has been reduced from 39,067 words to 9,557 words that occur 3,012,970 times in total in the entire collection.

4. Discovering research topics

The posterior inference of LDA is obtained through an efficient Gibbs sampler provided by the MALLET package [25]. A sensitivity analysis has been conducted to determine the input parameters for LDA. It is found that the number of topics $K = 50$ is able to achieve a converged group of research topics. The hyperparameter α on the Dirichlet distribution controls the mean shape and sparsity of per-document topic distributions. Namely, a larger α favors more uniform topic distributions. This study considers a small value $\alpha = 5/K = 0.1$ towards sparse topic distributions for every document, given the relatively narrow definition of structural engineering research. Besides, the hyperparameter η on topic word distribution β_k is considered as $\eta = 0.01$. Using these parameters, the LDA is carried out through 20 random runs for initialization and 8000 iterations for sampling. To this end, the LDA model provides two types of posterior distributions, namely the posterior per-document topic distribution, θ_d, and the posterior word distribution of each topic, β_k.

The posterior word distribution, β_k, for each of the 50 research topics is illustrated as a wordcloud in Figs. 3 and 4. Note that only the top words with the highest posterior probability are shown in the figures, and the size of each word is in proportion to its probability. A topic name is further assigned on each wordcloud by examining the most relevant

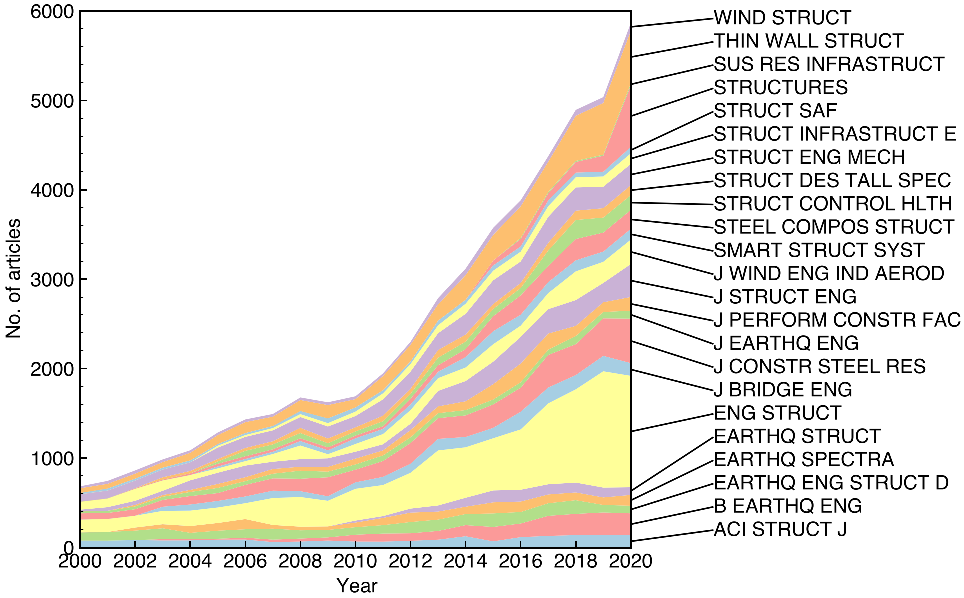


Fig. 2. A journal temporal disaggregation of the article data collected in this study.

words and their top ten contributing journal articles. The assigned topic name explicitly captures the intrinsic meaning associated with each wordcloud. In particular, expert judgement is relied upon in this process to capture the subtle difference between wordcoulds that share the same words. For instance, both Topics #6 and #21 share the word *FEM* with the highest posterior probability. However, a closer look at other relevant narratives and contributing journal articles indicate that Topic #6 is more about implementing FEM as a technique for numerical simulation, while Topic #21 focuses on developing the FEM itself (e.g., finite element formulation). Such expert judgement also deals with some topics that cover related but distinct aspects, issues, or structures (e.g., *seismic fragility/risk* in Topic #2 versus *risk and resilience* in Topic #19 and *regional seismic risk* in Topic #22). As a result, 49 out of 50 research topics bear evident contextual meaning that can be equivalently treated as a research area, yet Topic #29 turns out to be a general topic that is frequently used in academic writing – its relevant words are *construction*, *research*, *material*, *building*, *review*, *engineering*, *performance*, *issue*, *project*, *module*, *engineer*, etc. In general, the discovered 50 research topics provide a thorough landscape for the structural engineering community to classify research fields in the literature – researchers can easily identify one or more research topics that belong to their areas of expertise. It is worth noting that the research topics are defined by extracting a common research theme from each wordcloud. In this regard, some wordclouds also include keywords that belong to a specific method or tool. For example, research Topic #38 has a keyword of *ANN* (*artificial neural network*), which indicates the popularity of machine learning as a viable tool [26] to deal with reliability-related problems in structural engineering.

The LDA model for the discovered 50 research topics is further visualized through PyLDAvis, a python library for interactive topic model visualization (https://pyldavis.readthedocs.io/en/latest/). Fig. 5 (a) illustrates the inter-topic distance map where each research topic is denoted as a numbered circle – its area also represents the frequency of the topic over the entire corpus. As shown in the figure, these 50 topics are projected into a two-dimensional plane using principal coordinates analysis with the distance matrix created through the Jensen-Shannon divergence (JSD) [27]. As such, the distance between the circles turns out to be a direct measure of the similarity between topics. In particular, the overlapped circles in Fig. 5(a) represent inter-related research topics that share a common research theme, including (1) *cold-formed steel* (Topic #46) for *steel joints* (Topic #41); (2) *shear behavior* (Topic #16) for *beam-column joints* (Topic #28); (3) *wind turbine* (Topic #25) and *torsion* (Topic #34) under *wind flow & turbulence* (Topic #49); and (4) *structural control* (Topic #4) under *wind load* (Topic #48), among others. Other than the inter-topic distance map, Fig. 5(b) also shows the worldcloud frequency for Topic #35, *sensor monitoring*. The histogram lists the estimated term frequency for each word within Topic #35 (red color) out of its total frequency over the entire documents (blue color). As listed, *sensor*, *image*, *wireless*, etc., are somewhat unique words that mainly belong to Topic #35, while terms like *database*, *detection*, and *technique* are shared by other research topics.

5. Topic distribution over time

The per-document topic distribution θ_d can be further coupled with the publishing year of each article to analyze the temporal evolution of the discovered 50 research topics. The topic rising and falling represents the scientific interest it generates in the research community, which is probably a result of social forces, emerging techniques, extreme events, and scientific preferences. Moreover, the temporal variation of these topics provides a straightforward means to understand the dynamics of structural engineering research, which is particularly useful for determining potential targets for scientific funding. In this respect, more advanced topic-time joint models (e.g., the dynamic topic model by Blei and Lafferty [11] and the topic over time model by Wang and McCallum [12]) should be explored to quantify the temporal evolution of a specific research topic explicitly. Developing such joint models requires fitting a separate statistical model with a continuous distribution over timesteps in the generation process. Such efforts are considered outside the scope of the current study; instead, this study adopts a basic analysis method introduced by Griffiths and Steyvers [14], who examined the linear trend of θ_d by year in a post hoc manner. In particular, the temporal variation of research topics is measured using $\theta_k^{[t]}$, the proportion of topic k within the topic distribution at time t for all articles:

$$\theta_k^{[t]} = \frac{\sum_{d=1}^{D}\theta_{dk} \times \mathbf{I}(t_d = t)}{\sum_{d=1}^{D}\mathbf{I}(t_d = t)} \tag{3}$$

where $\mathbf{I}(e) = 1$ if e is true and 0 otherwise. As shown in Fig. 6(a), $\theta_k^{[t]}$ offers a quantitative measure to explore the temporal dynamics of all research topics, where the topics are shown in order (i.e., Topic #1 to #50) from the bottom to the top. Several general trends can be observed

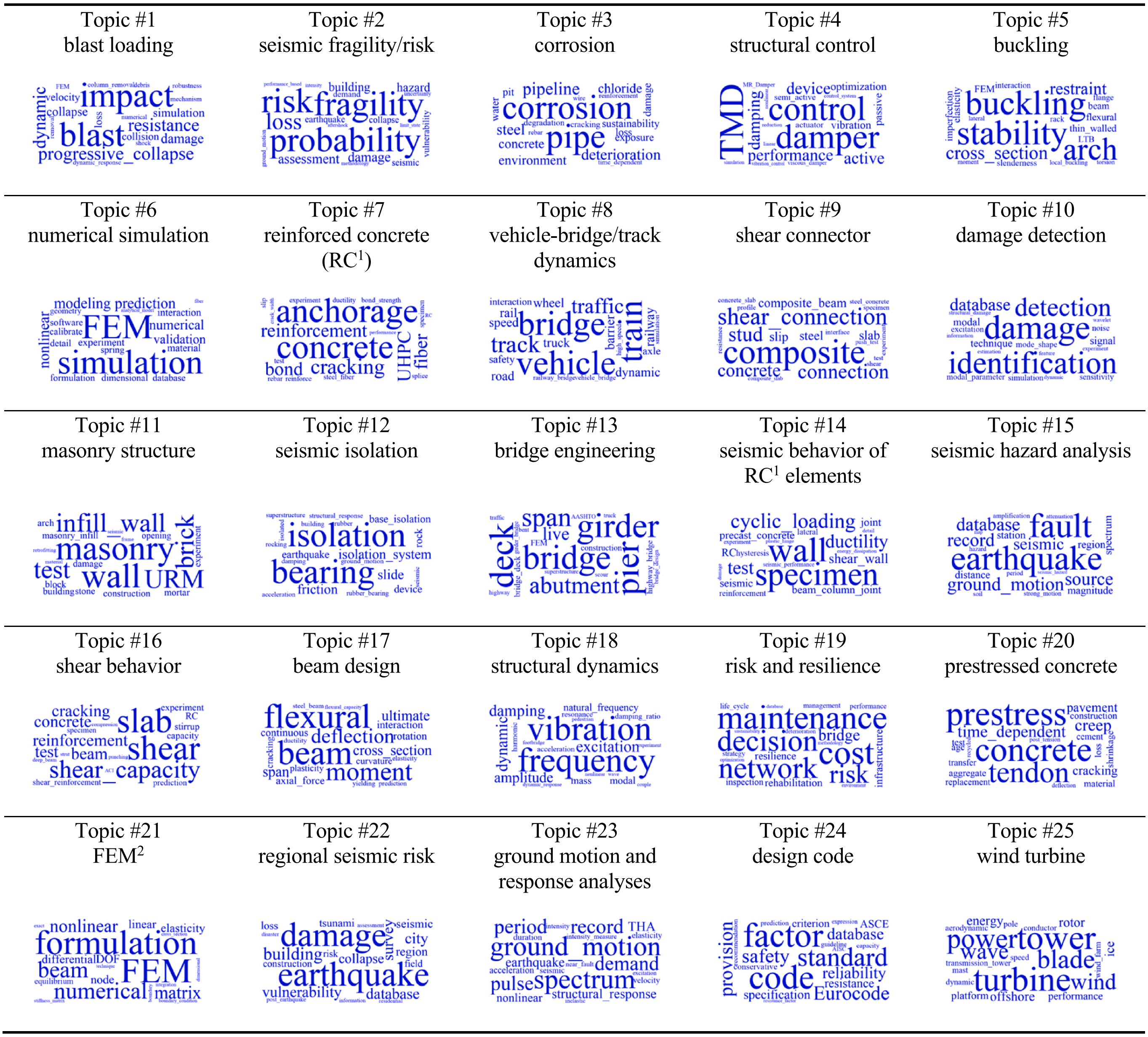


Fig. 3. Wordcloud of Topic #1 – Topic #25 (^{1}RC = reinforced concrete; ^{2}FEM = finite element method.)

from the figure. For instance, the most popular five topics are Topics #21 – *FEM*, #49 – *wind flow & turbulence*, #32 – *seismic evaluation of buildings*, #4 – *structural control*, and #14 – *seismic behavior of RC elements*. In contrast, certain topics have received relatively limited scientific interest, including Topics #41 – *steel joint*, #34 – *torsion*, and #43 – *hybrid simulation*. Fig. 6(b) also presents a comparison of the temporal trends among the most popular five topics. Despite their overall prevalence in the corpus, some topics indeed exhibit a decreasing trend in popularity over time, such as Topics #21 – *FEM*, #49 – *wind flow & turbulence*, and #4 – *structural control*.

Fig. 6(b) leads to a subsequent question to explore hot and cold topics. This study adopts the increase index, r_k, developed by Sun and Yin [18] to measure the popularity change of topic k at two distinct time windows:

$$r_k = \frac{\sum_{t=2015}^{2020} \theta_k^{[t]}}{\sum_{t=2000}^{2005} \theta_k^{[t]}} \tag{4}$$

where $r_k < 1$ means that topic k became less popular in 2015–2020 than 2000–2005, and vice versa. Therefore, the hottest and coldest topics can be obtained by pinpointing those with the largest and smallest r_k values, respectively. As shown in Fig. 7(a), the five coldest topics can be generally classified into two types. First, Topics #21 – *FEM*, #4 – *structural control*, and #49 – *wind flow & turbulence* have gained significant research interest at the beginning of the 21st century, yet the research momentum on these topics decreased substantially in recent years. Unlike these once-popular topics, Topics #34 – *torsion* and #24 – *design code* have remained cold throughout the past two decades. Fig. 7 (b) exhibits a different research view: the five hottest topics are Topics #1 – *blast loading*, #26 – *SCB & BRB*, #2 – *seismic fragility/risk*, #40 – *thin-walled tube*, and #9 – *shear connector*. By further checking the relevant top words within each topic, it can be concluded that these five hottest topics represent a shift of research interest towards rigorous numerical simulations, integrated seismic risk assessment, innovative design and protective devices, and the use of composite materials.

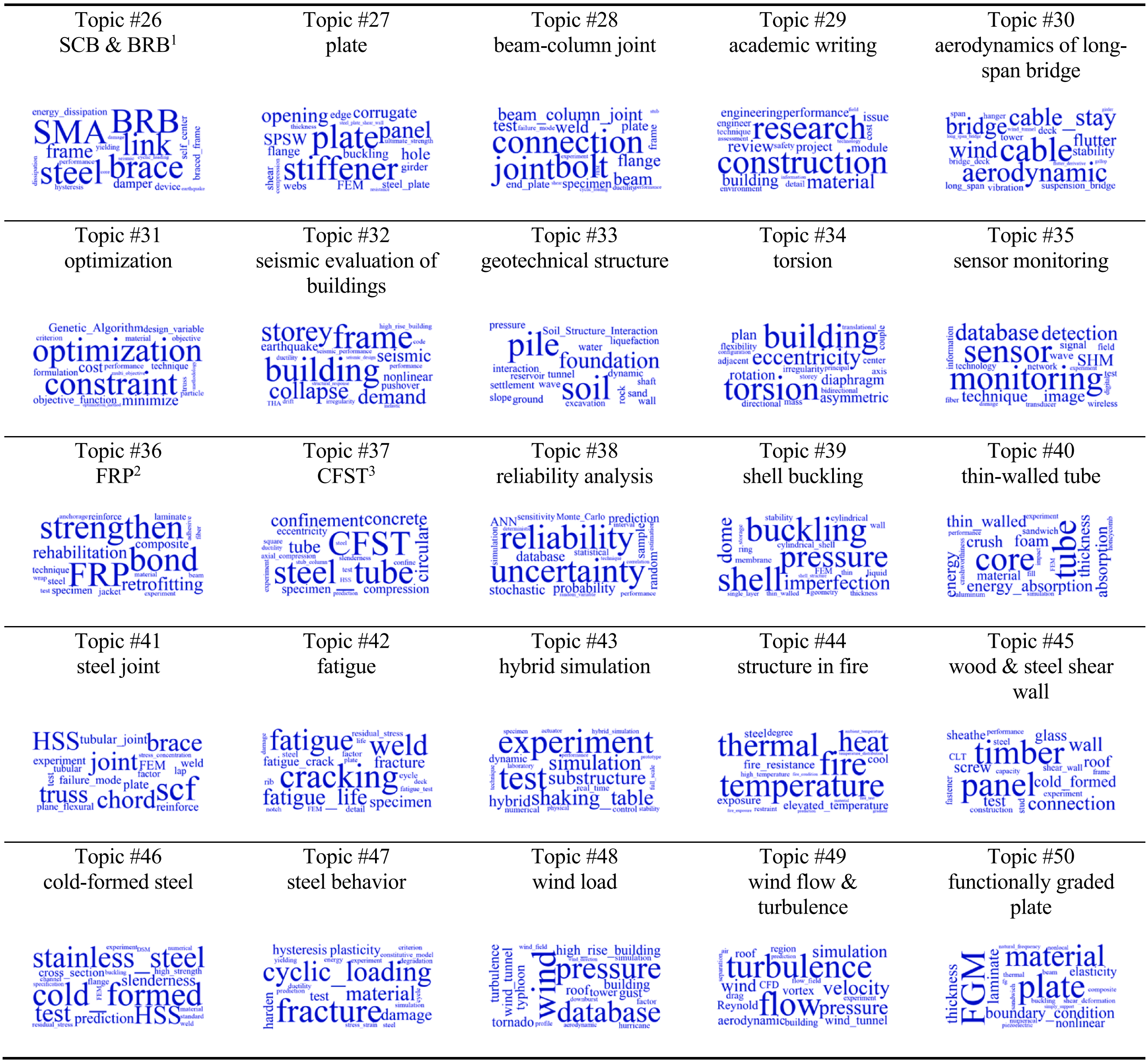

Topic #26
SCB & BRB¹
SMA BRB link frame steel brace damper device energy_dissipation self_center braced_frame
Topic #27
plate
opening corrugate SPSW plate panel stiffener hole girder flange buckling webs FEM steel_plate shear
Topic #28
beam-column joint
beam_column_joint test weld plate connection joint bolt flange beam end_plate specimen frame
Topic #29
academic writing
engineering performance research issue review project module construction building detail material
Topic #30
aerodynamics of long-span bridge
bridge cable_stay wind cable flutter stability tower deck aerodynamic long_span vibration suspension_bridge
Topic #31
optimization
Genetic_Algorithm design_variable optimization cost constraint objective_function minimize particle
Topic #32
seismic evaluation of buildings
storey frame earthquake seismic building collapse demand nonlinear pushover
Topic #33
geotechnical structure
pressure Soil_Structure_Interaction pile foundation liquefaction soil tunnel slope ground sand wall
Topic #34
torsion
plan building eccentricity rotation torsion diaphragm asymmetric
Topic #35
sensor monitoring
database detection sensor SHM monitoring technique image wireless
Topic #36
FRP²
strengthen rehabilitation composite bond FRP retrofitting specimen jacket
Topic #37
CFST³
confinement concrete eccentricity tube CFST steel tube circular compression
Topic #38
reliability analysis
ANN Monte_Carlo prediction reliability database uncertainty stochastic probability random
Topic #39
shell buckling
dome buckling cylindrical_shell pressure shell imperfection
Topic #40
thin-walled tube
thin_walled crush foam core tube energy_absorption thickness absorption
Topic #41
steel joint
HSS tubular_joint brace joint FEM weld truss chord SCF reinforce
Topic #42
fatigue
fatigue weld fracture cracking fatigue_life specimen
Topic #43
hybrid simulation
experiment simulation test substructure hybrid shaking_table
Topic #44
structure in fire
thermal heat fire temperature exposure elevated_temperature
Topic #45
wood & steel shear wall
sheathe glass timber wall panel cold_formed connection screw roof
Topic #46
cold-formed steel
stainless_steel slenderness cold_formed test prediction HSS
Topic #47
steel behavior
hysteresis plasticity cyclic_loading material fracture damage harden
Topic #48
wind load
wind high_rise_building pressure roof tower gust database tornado
Topic #49
wind flow & turbulence
simulation turbulence wind velocity flow pressure aerodynamic wind_tunnel
Topic #50
functionally graded plate
FGM material plate elasticity boundary_condition nonlinear thickness laminate

Fig. 4. Wordcloud of Topic #26 – Topic #50 (¹SCB = self-centering brace, BRB = buckling restrained brace; ²FRP = fiber-reinforced polymer; ³CFST = concrete filled steel tube).

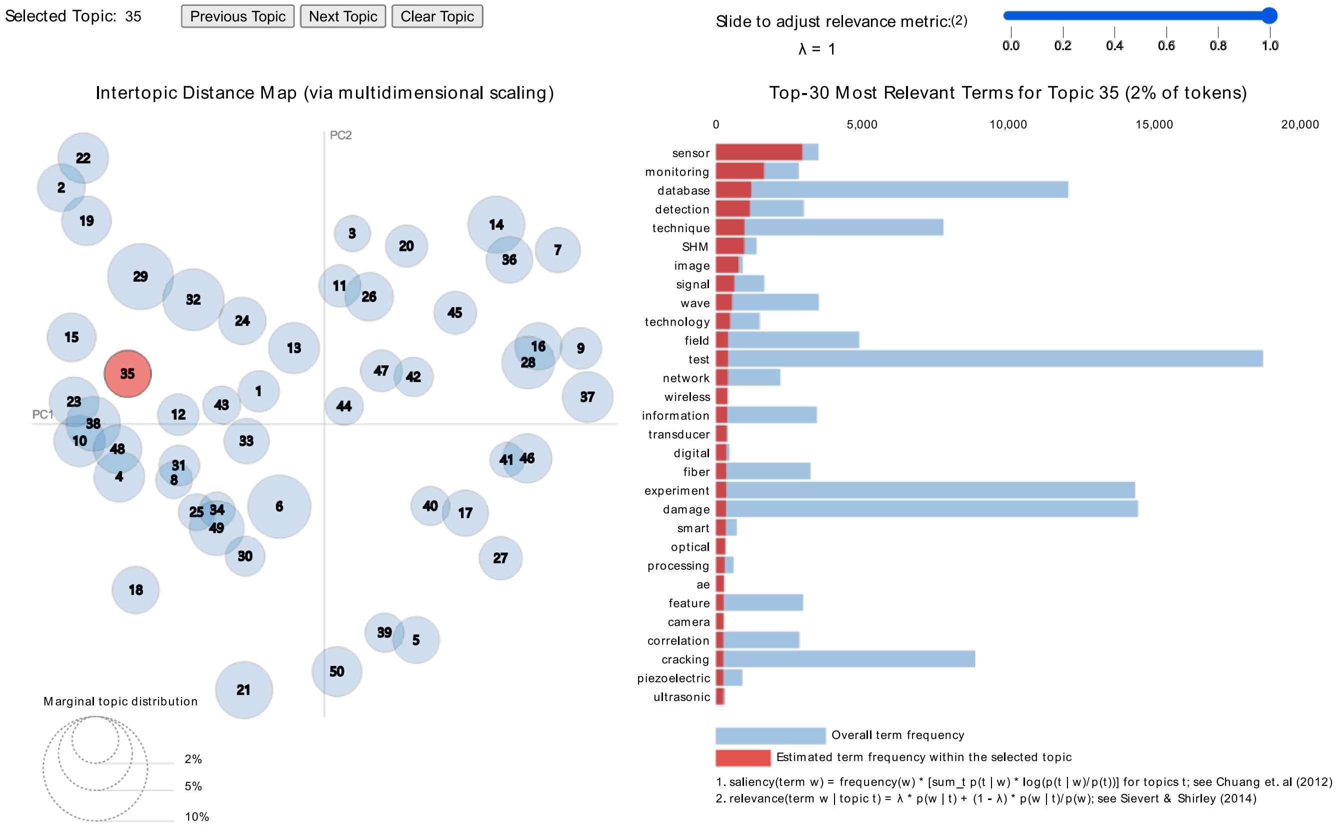
6. Journal topic distribution

6.1. Journal topic distribution and similarity

Using the same temporal analysis concept, the topic distribution across different journals can be measured through the metric, θ_k^j, the proportion of topic k within the topic distribution in journal j:

$$\theta_k^j = \frac{\sum_{d=1}^{D} \theta_{dk} \times \mathbf{I}(j_d = j)}{\sum_{d=1}^{D} \mathbf{I}(j_d = j)} \tag{5}$$

The metric θ_k^j turns out to be a two-dimensional matrix whose values vary per topic and journal. θ_k^j is illustrated in Fig. 8 with each row representing the topic distribution of a specific journal. The existence of red colors denotes a sparse distribution of research topics in a journal. The figure shows that widely distributed research topics exist in three comprehensive journals – *Structures*, *Engineering Structures*, and *Journal of Structural Engineering*. By contrast, as also can be inferred from their names, a few journals in structural engineering have their signatures – they focus on a certain set of research topics. For instance, both *Structural Control & Health Monitoring* and *Smart Structures and Systems* have a large body of research on Topics #10 – *damage detection* and #35 – *sensor monitoring*; the two wind-related journals, *Wind and Structures* and *Journal of Wind Engineering and Industrial Aerodynamics*, primarily deal with wind hazard (i.e., Topics #48 – *wind load* and #49 – *wind flow & turbulence*); *ACI Structural Journal* focuses on reinforced concrete behavior with three dominant research topics: #7 – *reinforced concrete*, #14 – *seismic behavior of RC elements*, and #16 – *shear behavior*; *Structural Safety* somewhat spells Topic #38 – *reliability analysis*, etc. Moreover, Fig. 8 captures some subtle differences among the five seismic-related journals: the three journals of *Journal of Earthquake Engineering*, *Bulletin of Earthquake Engineering*, and *Earthquake Spectra* somewhat share a common focus on Topic #15 – *seismic hazard analysis*, while the remaining two journals, *Earthquakes and Structures* and *Earthquake Engineering & Structural Dynamics*, are more balanced in covering several earthquake-related research topics (e.g., Topics #2, #4, #12, #23,



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

(a) (b)

Fig. 5. Topic model visualization through PyLDAvis: (a) the inter-topic distance map and (b) the top 30 most relevant terms for each topic.

#32).

By taking into account all topics, θ^j can be considered as the averaged topic distribution of all articles in journal j. This overall topic distribution is further used to quantify the similarity between journals. First, the difference between the overall topic distributions (θ^u and θ^v) of two journals, u and v, can be computed using the JSD [27]:

$$JSD(\theta^u,\theta^v) = \frac{1}{2}KLD(\theta^u,\overline{\theta}) + \frac{1}{2}KLD(\theta^v,\overline{\theta}) \qquad (6)$$

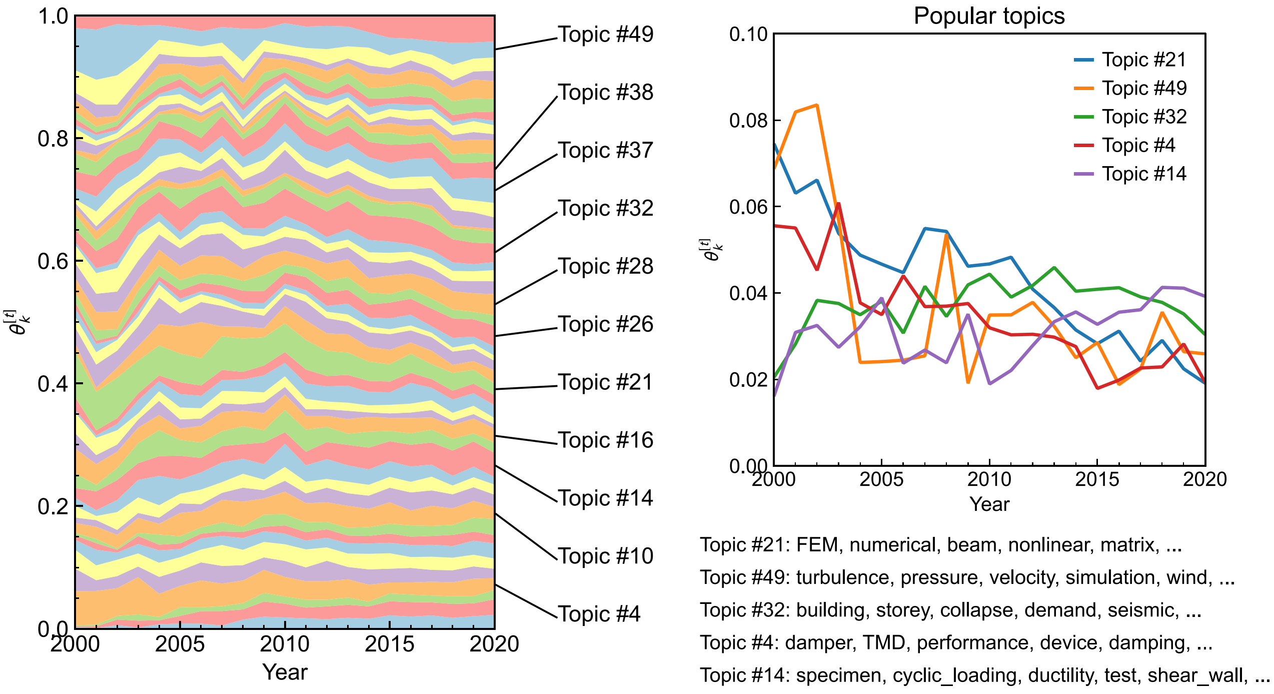
where $\overline{\theta} = \frac{1}{2}(\theta^u + \theta^v)$ and $KLD(\theta,\theta') = \sum_{k=1}^{K}\theta_k \log\frac{\theta_k}{\theta'_k}$ is the Kullback-Leibler divergence between two topic distribution θ and θ'. Second, the Jensen-Shannon distance, which is the square root of JSD, is adopted to measure the distance between two journals. This measured distance is then used to perform hierarchical clustering across all journals, where the complete linkage method is utilized to compute distances between paired clusters. The result of hierarchical clustering is shown as the dendrogram on the left panel of Fig. 8, where a smaller distance represents a higher degree of similarity. As can be seen from the figure, the three closest pairs of journals that share the shortest distance are the pair of *Journal of Earthquake Engineering* and *Bulletin of Earthquake Engineering* in the purple cluster, the pair of *Wind and Structures* and *Journal of Wind Engineering and Industrial Aerodynamics* in the green cluster, and the pair of *Structures* and *Engineering Structures* in the red cluster. The dendrogram also captures several distinct research fields in structural engineering, including structural health monitoring in the pink cluster, earthquake engineering in the orange and purple clusters, wind engineering in the green cluster, and steel structures in one of the grey clusters. Besides, there exist stand-alone journals that deal with specific research themes, such as *ACI Structural Journal* for reinforced concrete structures, *Structural Safety* for reliability analysis, and *Journal of Bridge Engineering* for bridges, etc.

6.2. Journal topic distribution over time

By combining Eq. (3) and Eq. (5), the temporal topic variation within each journal can be measured through $\theta_k^{j[t]}$, the proportion of topic k for the topic distribution in journal j at time t:

$$\theta_k^{j[t]} = \frac{\sum_{d=1}^{D}\theta_{dk} \times \mathbf{I}(t_d = t, j_d = j)}{\sum_{d=1}^{D}\mathbf{I}(t_d = t, j_d = j)} \qquad (7)$$

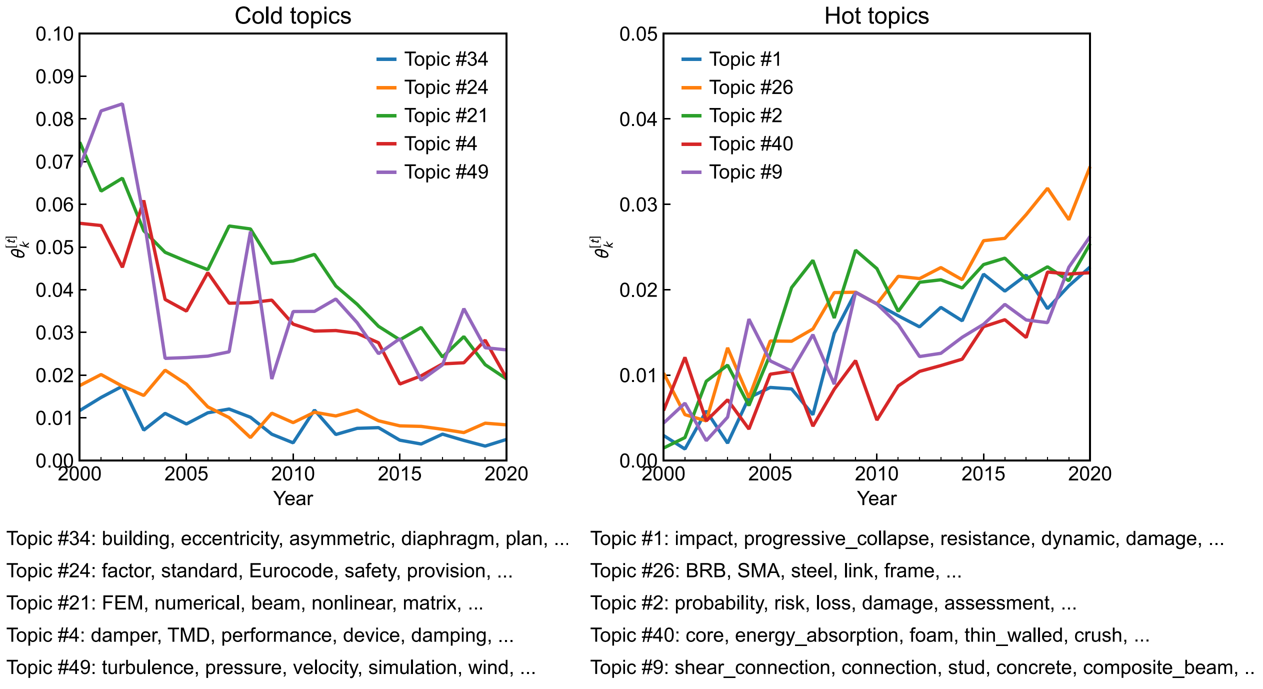
$\theta_k^{j[t]}$ offers a new metric that examines the temporal evolution of research topics for each journal. Fig. 9 presents the $\theta_k^{j[t]}$ results for the selected 23 journals, where the topics are listed in order (i.e., Topic #1 to #50) from the bottom to the top. The overall topic distribution shown in Fig. 9 indicates a consistent trend observed from Fig. 8 – other than the three comprehensive journals, *Structures*, *Engineering Structures*, and *Journal of Structural Engineering*, which handle widely distributed research topics, the remaining journals possess distinct research scopes where certain topics take significant proportions. Moreover, a comparison between Figs. 7 and 9 discloses two major mechanisms of producing cold and hot topics. First, the rising and falling of some topics represent the changing scientific interests they generate within each journal. For instance, as one of the coldest topics, Topic #4 – *structural control* was indeed a once-popular topic for the journals of *Earthquake Engineering & Structural Dynamics* and *Structural Control & Health Monitoring*. However, both



(a)

(b)

Fig. 6. Topic distribution over time: (a) all topics, and (b) the most popular five topics.



(a)

(b)

Fig. 7. (a) Five coldest and (b) five hottest topics identified from increase ratio, r_k.

journals have exhibited a decreased interest in publishing related articles after 2015. The same observation also applies to the Cold Topic #21 – *FEM* in the journals of *Engineering Structures*, *Structural Engineering and Mechanics*, and *Thin-Walled Structures*, showing a decreasing proportion, as well as the Hot Topic #2 – *seismic fragility/risk* in all five seismic-related journals with growing popularity. By contrast, despite their

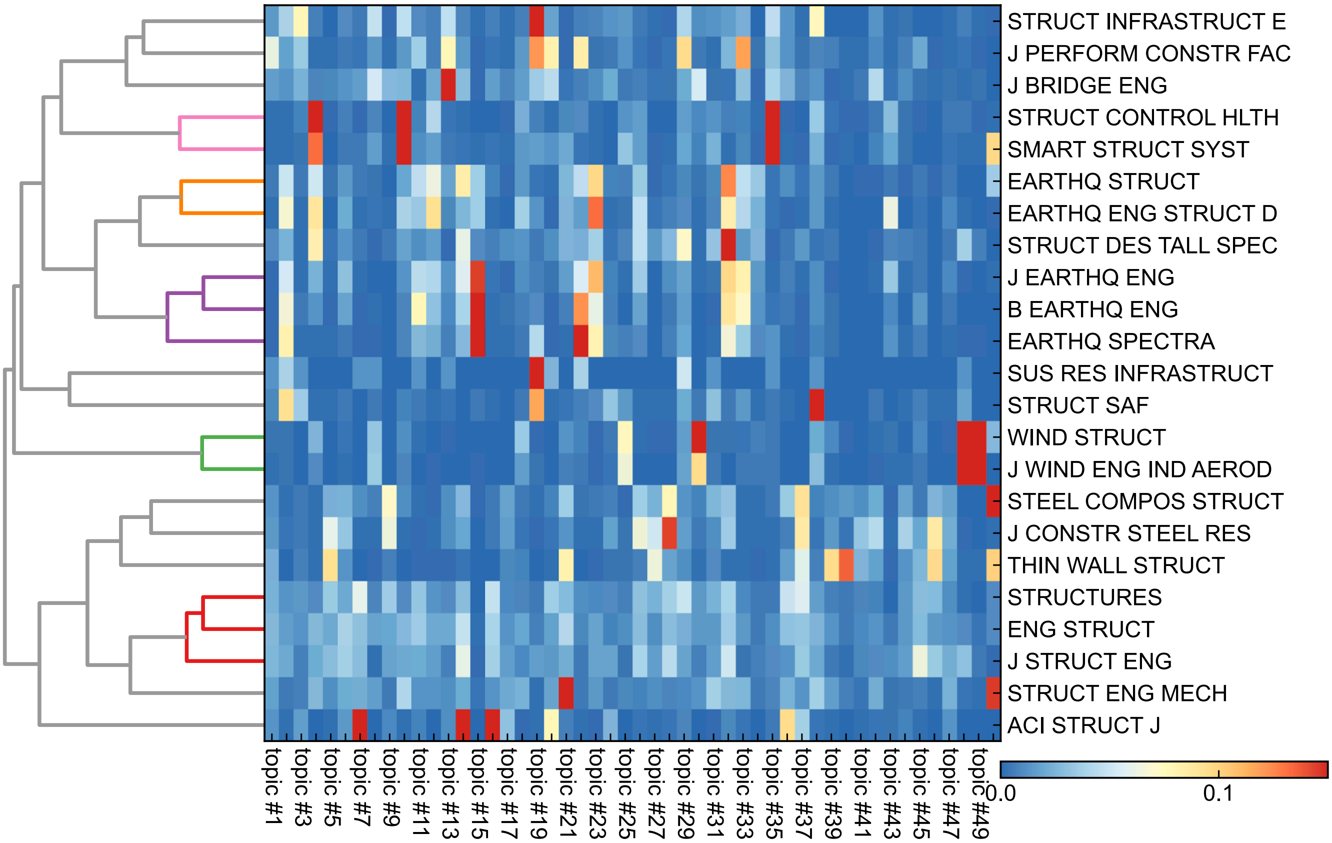


Fig. 8. Journal topic distribution and journal similarity.

overall changing popularities over time, some topics remain constantly favored in a particular journal. For example, Cold Topic #49 – *wind flow & turbulence* stays the most dominant topic in the *Journal of Wind Engineering and Industrial Aerodynamics* during the past two decades. Topic #49 became a cold topic in Fig. 7(a) due to its reduced proportion in other journals, as well as the emergence of new topics and new journals.

The journal-level temporal evolution of research topics also reveals new trends that have not been observed before. As shown in Fig. 9, some topics have grown substantially in some specific journals. For instance, Topic #50 – *functionally graded plate* has become a central topic for three journals: *Smart Structures and Systems*, *Steel and Composite Structures*, and *Structural Engineering and Mechanics*. Besides, two topics, Topics #10 – *damage detection* and #35 – *sensor monitoring*, have experienced reversed trends of interests between journals of *Smart Structures and Systems* and *Structural Control & Health Monitoring*. The former shows reduced interest in these two topics, while the latter exhibits growing preference. Fig. 9 can also help detect some anomalies in the history of a journal. Taking *Earthquake Spectra* as an example, research topics in this journal experienced a significant fluctuation in 2008, when Topic #15 – *seismic hazard analysis* suddenly became predominant. A retrospect of the journal indicates that in this year, 13 research articles were published through a special issue that summarized the principal results of a five-year research program, the Next Generation of Attenuation (NGA) Relations Project. Research outcomes presented in these 13 articles all belong to Topic #15, leading to its significantly increased proportion when compared with other research topics, such as Topic #22 – *regional seismic risk*.

7. Country/region topic distribution

7.1. Country/region topic distribution and similarity

Similar to the topic distribution analysis at the journal level, distributions of different research topics are further correlated with articles' correspondence addresses using $\theta_k^{(c)}$, which measures the proportion of topic k within country/region c:

$$\theta_k^{(c)} = \frac{\sum_{d=1}^{D} \theta_{dk} \times \mathbf{I}(c_d = c)}{\sum_{d=1}^{D} \mathbf{I}(c_d = c)} \tag{8}$$

The metric $\theta_k^{(c)}$ varies per topic and country/region. Fig. 10 presents the $\theta_k^{(c)}$ results where each row shows the topic distribution of a specific country/region. Due to large datasets, only countries/regions that have more than 200 articles are shown in the figure. It can be observed that research topics in most of the countries/regions are widely distributed, and only three countries exhibit noted research preferences in orange or red colors. First, Topic #23 – *ground motion and response analyses* turns out to be a popular topic for researchers from Mexico, a country that is prone to strong earthquakes. Moreover, Topic #50 – *functionally graded plate*, a topic that involves composite materials, has gained significant research interests in Iran and Algeria.

An overview of the topic distributions in Fig. 10 also indicates a great regional diversity, where no similar distributions can be visualized between any pairs of countries/regions. To this end, the similarity analysis shown in Eq. (6) is congruously applied to cluster countries/regions with similar research interests, whereas the results are provided as the dendrogram on the left panel of Fig. 10. The similarity analysis successfully identifies a few pairs of countries displaying strong similarity: Canada and United States, Greece and Italy, Korea and China, India and Turkey, Germany and France, Belgium and Brazil, and Spain and United Kingdom. As can be observed, graphical locations, development stages, and geological characteristics play crucial roles in determining research similarities across the world. Fig. 10 in general exhibits five different clusters – the yellow cluster that consists of Canada, United States, Switzerland, and New Zealand; the cluster in grey including Greece, Italy, Mexico, and Chile, where earthquake hazard is of significant

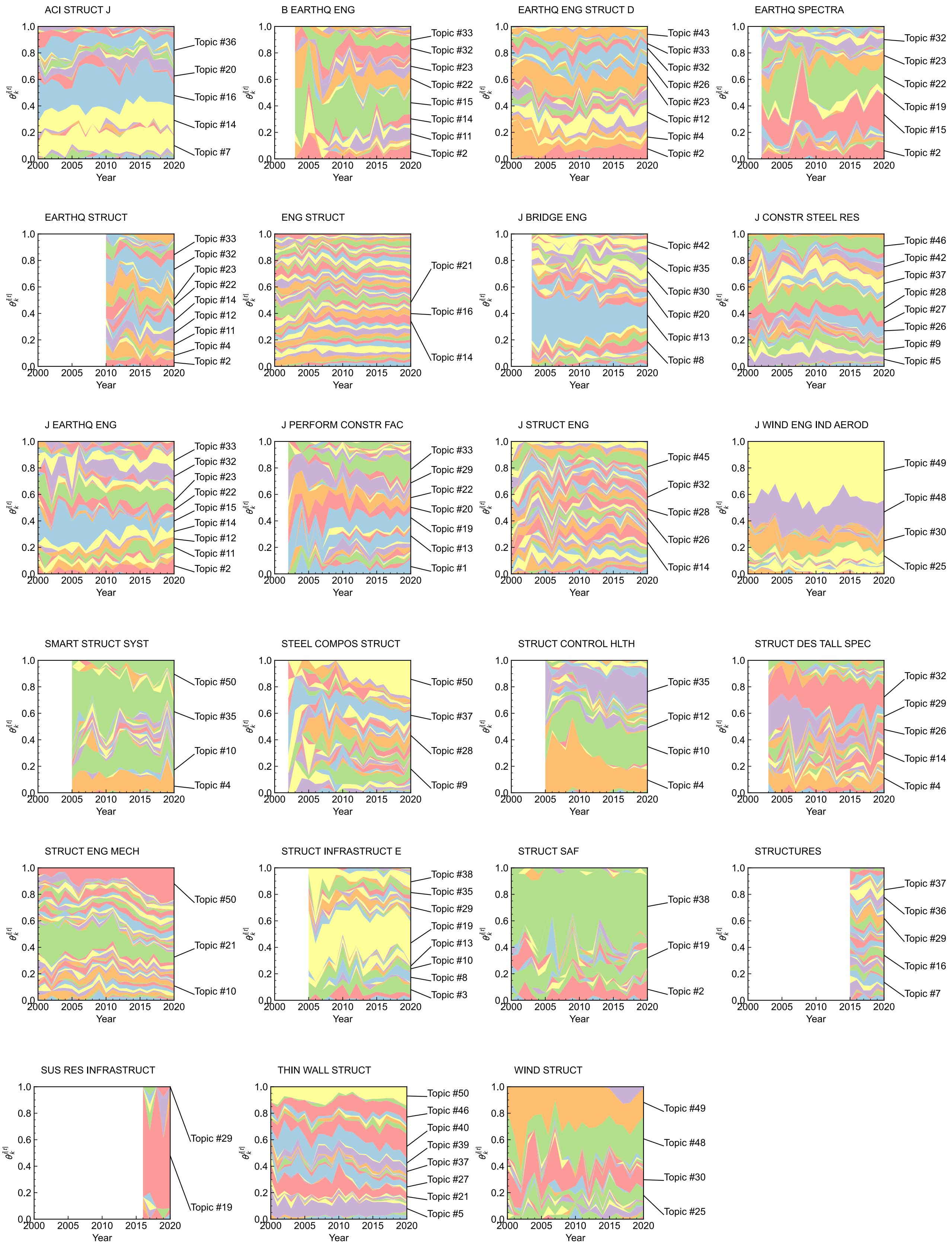


Fig. 9. Topic distribution over time for each individual journal.

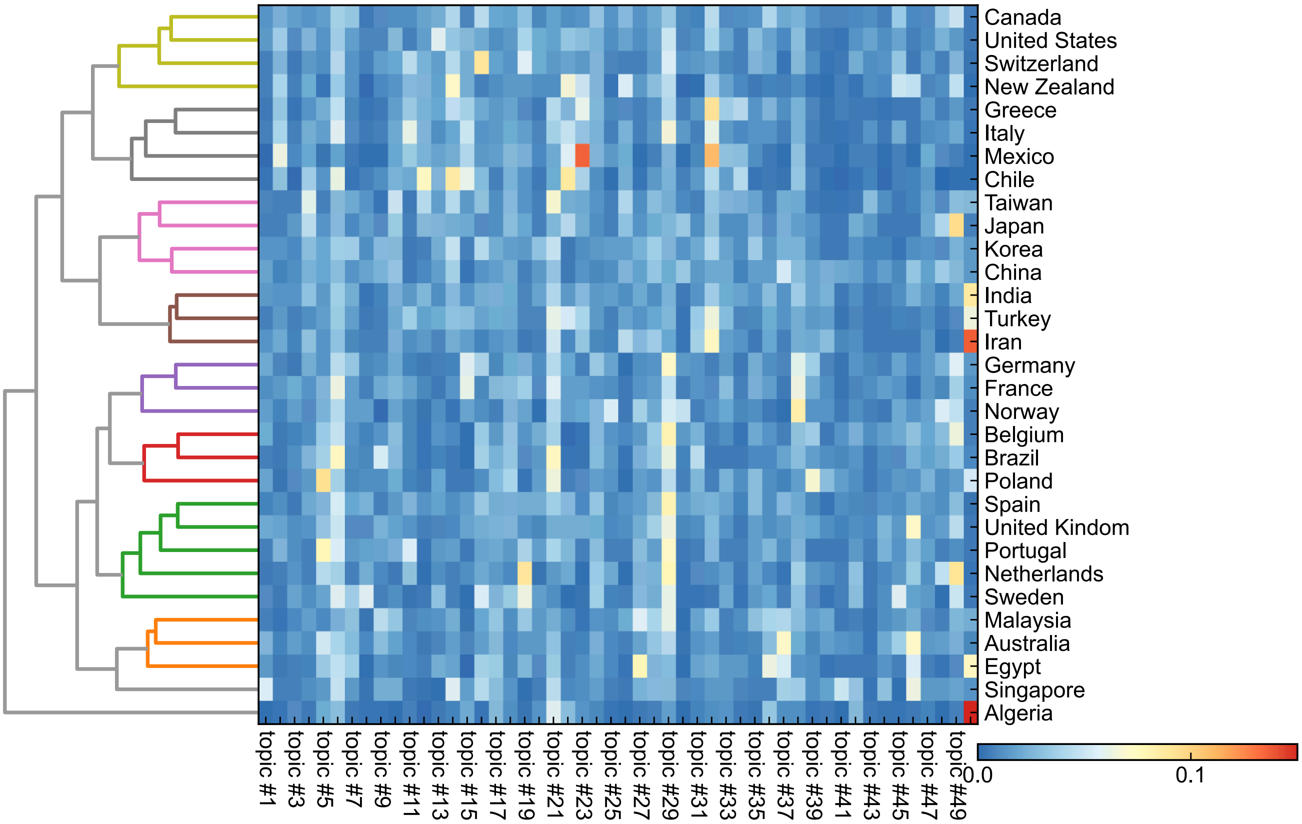


Fig. 10. Country/region topic distribution and similarity.

concern (e.g., Topics #23 – *ground motion and response analyses* and #32 *seismic evaluation of buildings* are popular topics in these countries); the Asian cluster consisting of Taiwan, Japan, Korea, China, India, Turkey, and Iran in pink and brown; the European cluster including Germany, France, Norway, Belgium, United Kingdom, Sweden, etc., colored in purple, red, and green; and the remaining countries that feature relatively unique research signatures.

7.2. Country topic distribution over time

A natural step forward is to explore the temporal topic variation within each country/region. This study computes $\theta_k^{(c)[t]}$ to measure the

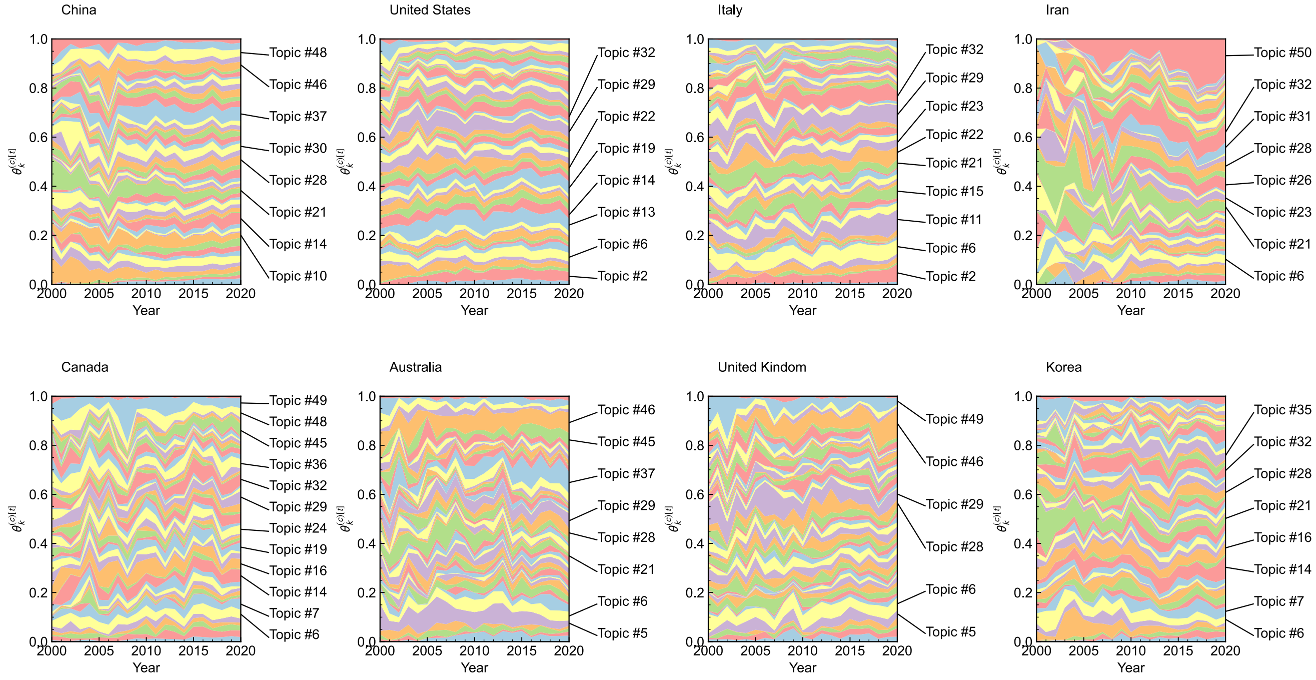


Fig. 11. Topic distribution over time for the top 8 countries.

proportion of topic k for the topic distribution in country/region c at time t:

$$\theta_k^{(c)[t]} = \frac{\sum_{d=1}^{D}\theta_{dk} \times \mathbf{I}(t_d = t, c_d = c)}{\sum_{d=1}^{D}\mathbf{I}(t_d = t, c_d = c)} \tag{9}$$

Fig. 11 presents the $\theta_k^{(c)[t]}$ results for the top 8 countries, where the topics are provided in order (i.e., Topic #1 to #50) from the bottom to the top. Popular topics have also been pinpointed for each country and listed on the right side of each figure. Fig. 11 illustrates the temporal evolutions of research focus across different countries, where similar research interests can also be identified. For instance, Topic #6 – *numerical simulation* turns out to be a universally popular topic among these 8 countries, representing the continued interest around the world in relying upon computational simulations to conduct research in structural engineering. Other than Topic #6, frequently emerged popular topics include (1) Topic #28 – *beam-column joint* in China, Iran, Australia, United Kingdom, and Korea; (2) Topic #14 – *seismic behavior of RC elements* in China, United States, Canada, and Korea; (3) Topic #21 – *FEM* in China, Italy, Iran, Australia, and Korea; and (4) Topic #32 – *seismic evaluation of buildings* in United States, Italy, Iran, Canada, and Korea. In addition to these shared popular topics, Fig. 10 also exhibits research similarities in the time dimension. For example, research interest on Topic #21 – *FEM* has constantly been falling in China, Italy, Iran, Australia, and Korea, while Topic #46 – *cold-formed steel* has shown growing popularity in both Australia and United Kingdom.

In contrast, Fig. 11 also demonstrates research diversities among these 8 countries. Taking Topic #50 – *functionally graded plate* as an example, it remains a non-central topic in all countries except Iran, where it has become the most dominant research theme in recent years. This is probably because a certain group of active researchers in Iran have produced considerable research outcomes in this new area. Another notable difference can be observed regarding three research topics related to seismic risk: Topic #2 – *seismic fragility/risk*, Topic #19 – *risk and resilience*, and Topic #22 – *regional seismic risk*. Topics #2 and #22 are popular in the United States and Italy but not in other countries, which is somewhat reasonable due to the high seismicity in these two countries. However, as the third topic, Topic #19 involves the new resilience concept, which makes it one of the most popular topics in the United States, but not in Italy. To this end, it can be concluded that the topic distribution of a country/region not only depends on the scientific interest in the community, but also is practice-oriented and may be influenced by government policies, funding mechanisms, and the initiation of new research concepts.

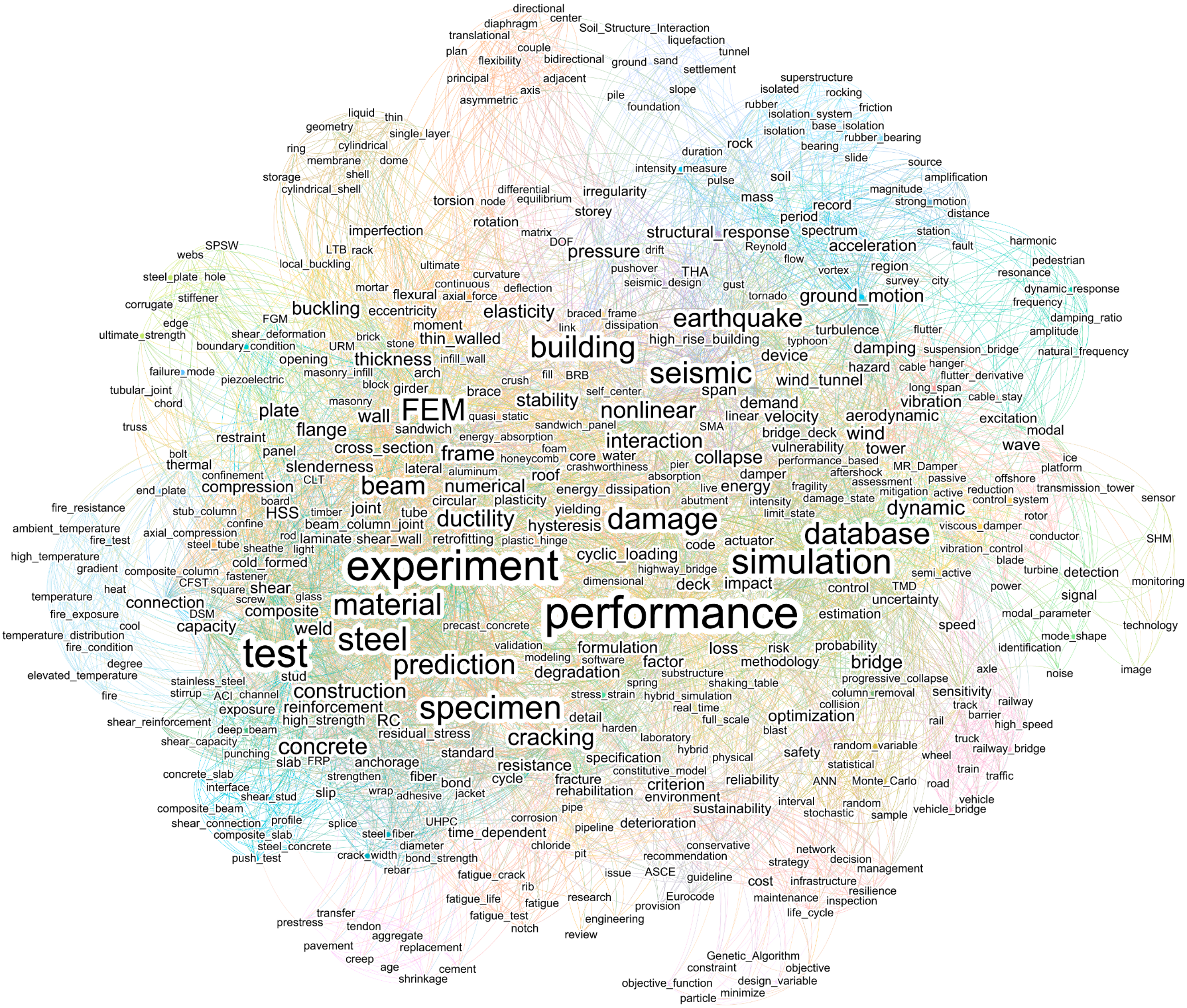


Fig. 12. Co-presence structure of words across topics.

8. Network of word co-presence

The co-presence of words in different topics is further explored. To achieve this, a binary matrix is defined as $P = [\beta_k^v \geq 0.075]$ of size $V \times K$, where V is the number of words and K is the number of topics. In matrix P, each element $p_{vk} = 1$ if $\beta_k^v \geq 0.075$ and 0 otherwise, capturing whether word v is a substantial component of topic k. Moreover, an adjacency matrix of words is defined as $Q = PP^T$, which quantifies the number of topics that have both $\beta_k^u \geq 0.075$ and $\beta_k^v \geq 0.075$ (i.e., the co-presence of two words in a topic). The word co-presence network defined by matrix Q is visualized in Fig. 12, where only the largest connected components (i.e., 482 vertices and 6,588 edges) are illustrated. As shown from the figure, the size of each word is proportional to its occurrence frequency, whereas the wordcloud from the same research topic is provided with the same cluster color. The word co-presence network shown in Fig. 12 provides a graphical means to discover linkages among research topics in structural engineering. First, the words with the largest font sizes indicate those that have the most frequent co-occurrence. It can be observed that numerical simulation (i.e., words include *FEM*, *simulation*, *database*) and experimental testing (i.e., words such as *test*, *experiment*, *specimen*) are the two primary tools that have been widely utilized across the research community. The relevant words also stay close to the center of the network, representing the strong connections among numerical simulation, experimental testing, and the remaining research topics. In addition, other frequently occurred words include *performance*, *earthquake*, *building*, *seismic*, *damage*, *material*, *steel*, *prediction*, etc., which denote a significant interest in seismic-related research. In general, the word co-presence network captures the latent research topics as different clusters, as well as their interconnections shared by connecting edges and common words. For instance, the word *soil* connects Topic #23 - *ground motion and response analyses*, Topic #12 - *seismic isolation*, and Topic #33 - *geotechnical structure* (i.e., see the three clusters in blue at the top right of Fig. 12). One more example lies in the word *speed*: although it bears different contextual meanings in different research areas, it serves as the connecting word between Topic #8 - *vehicle-bridge/track dynamics* and Topic #25 - *wind turbine*, as shown on the bottom right of Fig. 12. To this end, the word network shows a general research landscape towards a global understanding of how different research topics (i.e., word clusters) are allocated and interconnected with each other.

9. Discussions and conclusions

This study applies LDA to analyze 51,346 article abstracts from 23 peer-reviewed journals in structural engineering with a publication period from 2000 to 2020. The LDA successfully identifies 50 research topics that define the current state of research in the community. Posterior distributions of document-topic and topic-word are further combined with the publication year, journal, and correspondence address for a series of analyses to explore the context of each topic, the associated research trends, and the topic similarity/variance across journals and regions. These analyses provide a viable strategy to probe the core content of structural engineering research in the twenty-first century, which is expected to benefit all community stakeholders (e.g., students, engineers, researchers, conference organizers, journal editors, funding agencies) in multiple ways. This section discusses the analysis findings, potential applications, and future research needs to further promote such benefits.

First, the identified cold and hot topics reflect the shift of research interests in the structural engineering community. The research momentum on once-popular topics, such as *FEM*, *structural control*, and *wind flow & turbulence*, has been decreased in recent years. By contrast, *blast loading*, *SCB & BRB*, *seismic fragility/risk*, *thin-walled tube*, and *shear connector* (i.e., engaging composite materials) seem to attract increased research attention over time. The emergence of cold and hot topics can help researchers understand the research trend, capture the embedded research need, and switch their research focus if necessary. To this end, the temporal evolution of a specific research topic can be analyzed in more depth through the use of more advanced topic-time joint models (e.g., the dynamic topic model [11] and the topic over time model [12]), whereas additional analyses are also required to uncover the underlying mechanisms that cause the popularity change of these research topics.

Furthermore, this study analyzes the topic distribution and evolution at the journal level. In structural engineering, scientific journals can be classified into comprehensive ones (i.e., *Structures*, *Engineering Structures*, and *Journal of Structural Engineering*), which cover a broad range of research topics, general ones that focus on certain research areas (e.g., seismic versus wind, steel versus concrete, etc.), and more unique ones that mainly deal with specific topics, such as *Structural Safety* for reliability analysis. The findings on journal-level topic similarity and evolution can help researchers identify target journals for manuscript submission. For instance, the journal of *Smart Structures and Systems* has recently changed its research focus from the two once-dominant topics, namely Topics #10 – *damage detection* and #35 – *sensor monitoring*, to a novel theme of Topic #50 - *functionally graded plate*. However, growing interest can still be observed in the journal of *Structural Control & Health Monitoring* on both Topics #10 and #35. Moreover, journal editors and publishers can utilize the journal-topic distribution information to (1) re-evaluate the appropriateness of the journal scope and focus and (2) make efforts to identify and prioritize research themes that would enhance the journal impact. In this regard, the citation network [17] can be further incorporated into the LDA model to better explore the interconnections among different journals, as well as discover the most impactful research topics for each journal.

By linking to the correspondence address of each article, the LDA model is further utilized to analyze the topic distribution and evolution at the country/region level. In general, distributions of research topics across the globe are affected by factors such as the graphical location, development stage, and geological characteristics. In addition, the initiation of a new research concept, such as *risk and resilience*, can be country/region-specific due to the distinctions in government policy and funding mechanism, etc. In this respect, the identified topic distributions can help funding agencies to (1) better understand the research needs of their regions; (2) abandon those research themes that are not compatible with the regional development; and (3) prioritize specific topics that are either practical to solve urgent problems, or more fundamental to bear long-standing research and practical impacts.

In summary, the research findings from the current study reflect a general landscape of the state of research in structural engineering in the twenty-first century. The discovered research topics, as well as their distributions over time, journal, and region, are expected to stimulate more relevant discussions within the community, from which new insights can be generated toward concrete actions for different stakeholders to foster the healthy growth of the structural engineering research community.

Declaration of Competing Interest

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

References

[1] Clarke B, Middleton C, Rogers C. The Future of Geotechnical and Structural Engineering Research. Proc Inst Civ Eng - Civ Eng 2016;169:1–18.

[2] ASCE. Achieving the Vision for Civil Engineering in 2025: A Roadmap for the Profession. 2007. https://doi.org/10.1061/9780784478868.002.

[3] Salehi H, Burgueño R. Emerging artificial intelligence methods in structural engineering. Eng Struct 2018;171:170–89. https://doi.org/10.1016/j.engstruct.2018.05.084.

[4] Xie Y, Ebad Sichani M, Padgett JE, DesRoches R. The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. Earthq Spectra 2020;36:1769–801. https://doi.org/10.1177/8755293020919419.

[5] Blei D, Carin L, Dunson D. Probabilistic topic models. IEEE Signal Process Mag 2010;27:55–65. https://doi.org/10.1109/MSP.2010.938079.

[6] Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Process 1998;25:259–84. https://doi.org/10.1080/01638539809545028.

[7] Hofmann T. Probabilistic latent semantic analysis. ArXiv 2013;1301.6705:289–96.

[8] Hofmann T. Probabilistic latent semantic indexing. Proc 22nd Annu Int ACM SIGIR Conf Res Dev Inf Retrieval. SIGIR 1999;1999(51):50–7. https://doi.org/10.1145/312624.312649.

[9] Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res 2003;3:993–1022. https://doi.org/10.1016/B978-0-12-411519-4.00006-9.

[10] Blei DM, Lafferty JD. Correlated topic models. Adv Neural Inf Process Syst 2005:147–54.

[11] Blei DM, Lafferty JD. Dynamic topic models. Proc 23rd Int Conf. Mach Learn 2006:1–8.

[12] Wang X, McCallum A. Topics over Time: A non-markov continuous-time model of topical trends. In: Proc ACM SIGKDD Int Conf Knowl Discov Data Min 2006; 2006. p. 424–33.

[13] Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 2019;78:15169–211. https://doi.org/10.1007/s11042-018-6894-4.

[14] Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A 2004;101:5228–35. https://doi.org/10.1073/pnas.0307752101.

[15] Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. ArXiv Prepr 2012;1207(4169):487–94. https://doi.org/10.1016/s0030-5898(20)32328-2.

[16] Mäntylä MV, Graziotin D, Kuutila M. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Comput Sci Rev 2018;27:16–32. https://doi.org/10.1016/j.cosrev.2017.10.002.

[17] He Q, Chen B, Pei J, Qiu B, Mitra P, Giles L. Detecting topic evolution in scientific literature: How can citations help? Int Conf Inf Knowl Manag Proc 2009:957–66. https://doi.org/10.1145/1645953.1646076.

[18] Sun L, Yin Y. Discovering themes and trends in transportation research using topic modeling. Transp Res Part C Emerg Technol 2017;77:49–66. https://doi.org/10.1016/j.trc.2017.01.013.

[19] Shin SH, Kwon OK, Ruan X, Chhetri P, Lee PTW, Shahparvari S. Analyzing sustainability literature in maritime studies with text mining. Sustain 2018;10. https://doi.org/10.3390/su10103522.

[20] Yalcinkaya M, Singh V. Patterns and trends in Building Information Modeling (BIM) research: A Latent Semantic Analysis. Autom Constr 2015;59:68–80. https://doi.org/10.1016/j.autcon.2015.07.012.

[21] Ezzeldin M, El-Dakhakhni W. Metaresearching Structural Engineering Using Text Mining: Trend Identifications and Knowledge Gap Discoveries. J Struct Eng 2020;146:04020061. https://doi.org/10.1061/(asce)st.1943-541x.0002523.

[22] Syed S, Spruit M. Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation. Proc - 2017 Int Conf Data Sci Adv Anal DSAA 2017 2017;2018-January:165–74. https://doi.org/10.1109/DSAA.2017.61.

[23] Buntine WL. Operations for Learning with Graphical Models. J Artif Intell Res 1994;2:159–225. https://doi.org/10.1613/jair.62.

[24] Schum DA. The Evidential Foundations of Probabilistic Reasoning. Northwestern University Press; 1994.

[25] McCallum AK. MALLET: A machine learning for language toolkit; 2002.

[26] Reich Y. Machine Learning Techniques for Civil Engineering Problems. Comput Civ Infrastruct Eng 1997;12:295–310. https://doi.org/10.1111/0885-9507.00065.

[27] Lin J. Divergence Measures Based on the Shannon Entropy. IEEE Trans Inf Theory 1991;37:145–51. https://doi.org/10.1109/18.61115.