Trends in Transportation Research

Exploring Content Analysis in Topics

Subasish Das, Karen Dixon, Xiaoduan Sun, Anandi Dutta,
and Michelle Zupancich

Proceedings of journal and conference papers are good sources of big textual data to examine research trends in various branches of science. The contents, usually unstructured in nature, require fast machine-learning algorithms to be deciphered. Exploratory analysis through text mining usually provides the descriptive nature of the contents but lacks quantification of the topics and their correlations. Topic models are algorithms designed to discover the main theme or trend in massive collections of unstructured documents. Through the use of a structural topic model, an extension of latent Dirichlet allocation, this study introduced distinct topic models on the basis of the relative frequencies of the words used in the abstracts of 15,357 TRB compendium papers. With data from 7 years (2008 through 2014) of TRB annual meeting compendium papers, the 20 most dominant topics emerged from a bag of 4 million words. The findings of this study contributed to the understanding of topical trends in the complex and evolving field of transportation engineering research.

In any branch of science, it is always difficult to predict the stringent issues that will dominate in decades to come. In the complex and evolving field of transportation engineering, it is fair to say that topical trends in upcoming years will be too complex to acquire valuable insights through prediction. Research in statistical models of topical co-occurrence, however, has led to the development of a variety of useful topic models. Researchers use these techniques to discover hidden trends inside unstructured, larger textual contents.

TRB organizes the largest and most comprehensive annual transportation conference in the world. Established in 1920 as the National Advisory Board on Highway Research, TRB provides a platform for the exchange of information and research results about every aspect of transportation engineering. TRB's multifarious activities involve large numbers of engineers, scientists, researchers, and practitioners from the public and private sectors and academia. State transportation departments and federal agencies, including the component administrations of the U.S. Department of Transportation, support this conference program. More than 12,000 policy makers, administrators, practitioners, researchers, and representatives of government, industry, and academic institutions attended the 95th TRB annual meeting in January 2016. More than 5,000 presentations in nearly 750 sessions and workshops were made, which covered a broad area of transportation science and engineering.

To understand the research trends in the domain of complex transportation engineering, the exploration of the TRB compendium papers could be a good point to start. Das et al. used the titles and abstracts of 7 years (2008 through 2014) of TRB compendium papers to perform text mining and latent Dirichlet allocation (LDA) topic modeling (*1*). The study provided an exploratory view of the data set and uncovered clusters of topics in unstructured form in the document groups through the use of bag-of-pattern representations of LDA. However, the Das et al. study did not use supporting metadata to determine more document-specific topics and to identify a correlation between topics. The current research aims to mitigate the limitations of this previously conducted research.

LITERATURE REVIEW

To understand large amounts of textual content, probabilistic topic models, such as LDA, have become commonly used tools in the present day (*2*). Although the principal purpose of this algorithm is to conduct exploratory analysis, researchers consider the importance of topic models as a tool to measure latent linguistic significance (*3*). Most of the text-mining tasks in text corpora (large and structured set of texts) employ statistical topic models such as probabilistic latent semantic analysis (*4*) and LDA (*5*). However, these unsupervised models (i.e., models with no clear definition of explanatory and response variables) can result in topical trends that are not interpretable (*6, 7*).

In recent years, researchers have proposed many knowledge-based topic models (*8–15*) and dynamic topic models (*16–20*) to overcome the issues associated with conventional LDA. Researchers also have investigated the performance of automatic coherence measurement of topic models (*21*). Moreover, research has been conducted on the development of an unsupervised method that improves the coherence score by considering the co-occurrence of words in a corpus. Some dynamic topic models have been proposed to mine dynamic patterns [e.g., topic over time (*18*) and dynamic mixture model (*5, 16, 18, 19*)]. Time is a significant consideration in these models.

The framework behind consideration of additional information, or metadata about the structure of the corpus in modeling the framework, uses the altercation of prior distributions to partially shape information across similar documents. Researchers have explored the incorporation of metadata into models from the vantage point of various aspects [e.g., author (*22*), topical content and ideology (*23*), and geography (*24*)]. Approaches that target corpus structure reflect by making inferences about observed covariates rather than predict

S. Das, K. Dixon, and M. Zupancich, Texas A&M Transportation Institute, 3135 TAMU, College Station, TX 77843-3135. X. Sun, Civil Engineering Department, University of Louisiana, Lafayette, LA 70504. A. Dutta, Computer Science and Engineering Department, Texas A&M University System, 3112 TAMU, College Station, TX 77843-3112. Corresponding author: S. Das, s-das@tti.tamu.edu.

Transportation Research Record: Journal of the Transportation Research Board,
No. 2614, 2017, pp. 27–38.
http://dx.doi.org/10.3141/2614-04

covariate values in data. Supervised LDA targets a prediction task and assumes a generative model for document-level variables (*25*). This approach finds a low-dimensional representation that predicts words and the covariate. Partially labeled LDA allows the inclusion of prior information from particular documents that are somewhat pertinent (*26*). Finally, factorial LDA has a mathematical setup similar to that of the structural topic model but focuses on latent covariates with an emphasis on interpretation (*27*). For details on current development of various topic models, see Yao et al. (*28*). In the current study, a web portal was developed to provide interested readers with a detailed bibliography on text mining and topic modeling (*29*).

Limited work in topic modeling has been done in the field of transportation with the exception of work done by Das et al., who applied LDA to TRB compendium papers (*1*). The current study extended LDA with structural topic modeling (STM) through the use of the titles and abstracts of 15,357 TRB compendium papers, along with metadata from 7 years (2008 through 2014) of TRB annual meeting compendium papers.

TOPIC MODELING

STM Approach

LDA lacks additional document-level information, in which variation can be seen in different theoretical interests. The use of LDA, and then the performance of a post hoc evaluation of variation with a certain covariate of interest, could be a reasonable solution. STM accommodates corpus structure through document-level covariates. The key idea behind STM is to specify the priors as generalized linear models through which it is possible to condition on arbitrary observed data. This approach directly allows an estimation of the quantities of interest in the unstructured textual contents.

The model (Figure 1) combines and extends three existing models: the correlated topic model (*30*), the Dirichlet multinomial regression (*31*) topic model, and the sparse additive generative topic model (*21*).

The notations used for the theoretical part are as follows:

$d \in \{1 \dots D\}$ = index of the documents,
$n \in \{1 \dots N\}$ = index of the tokens in the documents,
$v \in \{1 \dots V\}$ = index of a vocabulary of words,
$\omega_{d,n}$ = observed token (a conditionally independent drawn from $v \in \{1 \dots V\}$),
$k \in \{1 \dots K\}$ = index of topics,
X = topic prevalence matrix with dimension D by P,
Y = topic content matrix with dimension D by A,
m_v = baseline log frequency of each word in the vocabulary, and
s, r, σ^2, ρ = hyperparameters.

Core Language Model

The core language model builds on the correlated topic model that models correlations in the document-topic proportions with the use of the logistic normal distribution. For a model with K topics, it follows that

$$\eta \sim N\left(\mu, \sum\right) \tag{1}$$

$$\theta_k = \frac{\exp(\eta_k)}{\left(\sum_i \exp(\eta_i)\right)} \tag{2}$$

where η_k is fixed to 0 for identification.

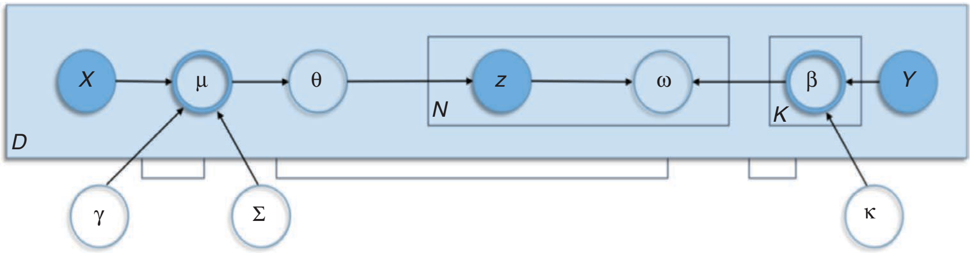
For each token within a document, a topic is sampled from a multinomial distribution $z \sim M(\theta)$ and, conditional on that sampled topic, a word is chosen from the distribution over words β_z. Here μ and β are specific to the document covariates.

Topic Prevalence

The topic prevalence component permits the expected document-topic proportions to vary by covariates (X), rather than arise from a single shared prior. It models the mean vector of the logistic normal as a simple linear model such that $\mu_d = X_{d\gamma_d}$, where γ is a regularizing prior to avoid overfitting. It takes the form of a normal multivariate linear model with shared covariance parameters, which will reduce to the standard correlated topic model formulation with an unpenalized intercept and no covariates.

Topical Content

The idea of topical content depends on the parameterization of the distribution over words as deviations in log space from a corpuswide



Topic Prevalence
$\mu_{d,k} = X_{d\gamma_k}$
$\gamma_k \sim N(0,\sigma_k^2)$
$\sigma_k^2 \sim \text{gamma}(s^\gamma, r^\gamma)$

Language Model
$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$
$z_{d,n} \sim \text{mult}(\theta_d)$
$\omega_{d,n} \sim \text{mult}(\beta_d^{k=z_{d,n}})$

Topical Content
$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$
$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$
$\tau_v^{y,k} \sim \text{gamma}(s^k, r^k)$

FIGURE 1 Structural topic model.

baseline *m*. Thus, for the simple case with a single covariate (y) that denotes a mutually exclusive and exhaustive group of documents, the distribution over words is

$$\beta_{d,k,v} \propto \exp\left(m_v + \kappa_v^{\cdot,k} + \kappa_v^{Y,\cdot} + \kappa_v^{Y,k}\right) \quad (3)$$

where

m_v = baseline log frequency of words,
κ_k = deviations due to each topic,
κ_c = nontopic-specific deviations due to covariates Y, and
κ_I = topic-specific covariate deviations.

The effect of topic and covariate represents sparse deviation from the corpuswide empirical word frequency. It replaces the multinomial likelihood on words with a multinomial logistic regression where the covariates are the token-level topic latent variables z, the user-supplied covariates Y, and their interaction. In principle, it does not require the restriction of models with single categorical covariates. In practice, computation considers the number of levels of topical content covariates to be relatively small. If no covariates are included and the parameters are not regularized, the model reduces to the standard maximum likelihood estimation of the topical content parameters given in the simpler LDA and correlated topic models.

With the infrastructure developed as described, customization of a topic model to a particular data set involves only the specification of a model for the linear predictors of topic prevalence and topical content. To fit the model, STM uses a semicollapsed, variational expectation-maximization algorithm that gives an estimate of the model parameters upon convergence. Regularized prior distributions are used to enhance interpretation and prevent overfitting. The model is estimated by using semicollapsed, variational expectation maximization. In the E step, the joint optimum of the document's topic proportions and the token-level assignments are solved. In the M step, the global parameters are inferred that control the priors on topical prevalence and content. For details on STM, see Roberts et al. (*32*).

METHODOLOGY

The titles and abstracts from 15,357 compendium papers, along with metadata collected from TRB annual meetings from 2008 through 2014, were used in the analysis. Figure 2 shows the yearly frequency of the compendium papers. Although the number of accepted compendium papers increased over the years, a sudden jump (186% increase) occurred from 2008 to 2009.

TRB provides the following information for all compendium papers:

- Publication year,
- Title,
- Abstract,
- Author's name,
- First author's institutional affiliation,
- Review committee's code, and
- Review committee's name.

A crucial contribution of this study is that it incorporates information about each document [e.g., publication year, author name, author affiliation, review committee (code or name)]. The paper abstracts were used to estimate the topic modeling. Other items

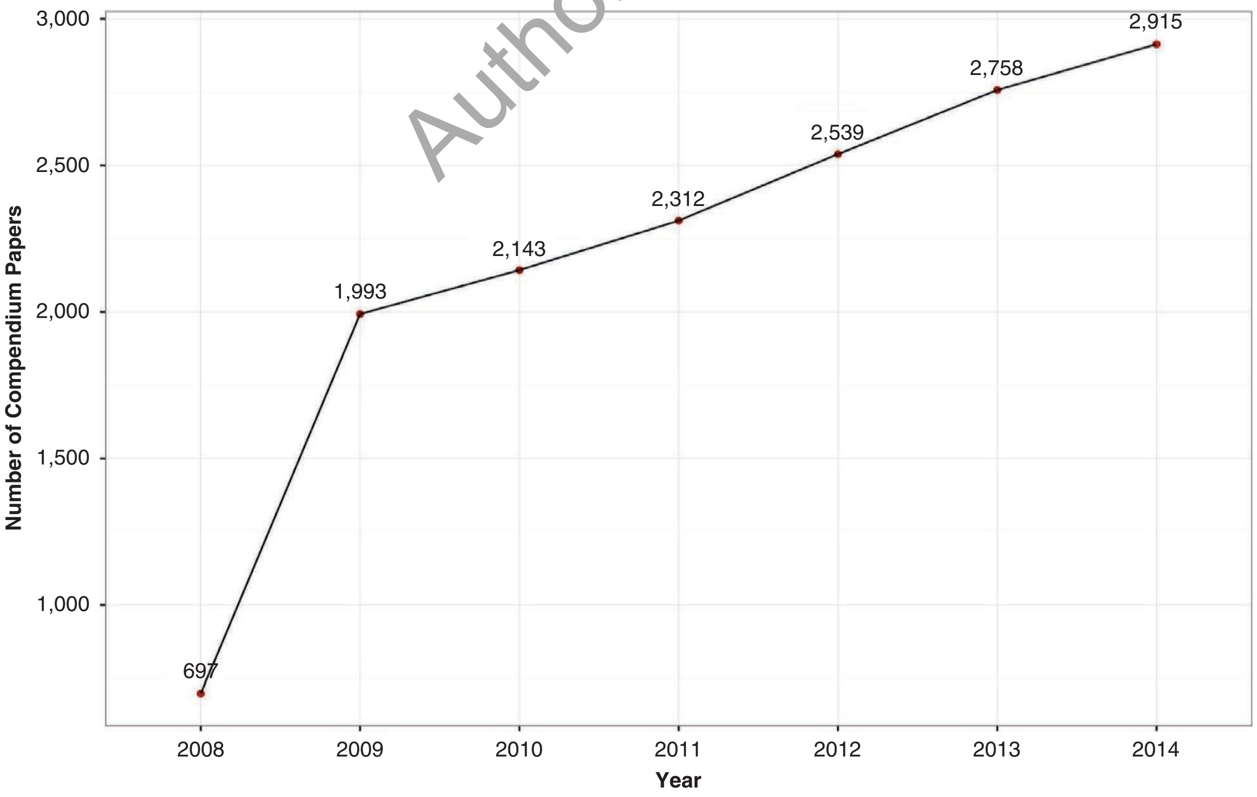


FIGURE 2 Frequency of TRB compendium papers.

(e.g., publication year, review committee, author name, author affiliation) were considered metadata. The faster machine-learning algorithm of STM provides fast, transparent, replicable analyses that can develop topic models from nearly 4 million words generated from 15,357 compendium papers. To determine the impact of different items in the metadata, the open-source R package tm was used (*33*). To perform this analysis, text corpora were created by combining all topics for 7 years filtered by the code of the review committee. Figure 3 illustrates the heat map of the most frequent items in the top 10 review committees.

Figure 4 illustrates the heat maps of the most frequent items in the top 10 review committees in each year (2008 was excluded because of its unrepresentative number of papers). The figure clearly shows that the year and the committee name had significance in generating distribution of high frequency of words. The findings from the abstracts provided support for the use of the year and the committee name as metadata to perform topic prevalence from the paper abstracts. The committee codes and committee names follow.

- ABE90: Transportation in the Developing Countries,
- ADB10: Traveler Behavior and Values,
- ADB30: Transportation Network Modeling,
- ADB40: Transportation Demand Forecasting,
- ADC20: Transportation and Air Quality,
- ADC80: Alternative Transportation Fuels and Technologies,
- AFK30: Characteristics of Nonasphalt Components of Asphalt Paving Mixtures,
- AFK50: Characteristics of Asphalt Paving Mixtures to Meet Structural Requirements,
- AHB15: Intelligent Transportation Systems,
- AHB20: Operations and Traffic Management,
- AHB25: Traffic Signal Systems,
- AHB40: Highway Capacity and Quality of Service,
- AHB45: Traffic Flow Theory and Characteristics,
- ANB20: Safety Data, Analysis and Evaluation,
- ANF10: Pedestrians, and
- ANF20: Bicycle Transportation.

In the STM framework, researchers can choose covariates to incorporate into the model. These covariates inform either the topic prevalence or the topical content latent variables with observed information about the respondent. In this study the publication year and the names of the review committees were used as the covariates in the topical prevalence portion of the model (*X*). These observed covariates will affect how much the respondent is to discuss a particular topic.

Often it is useful to engage in some processing of the text data before it is modeled. The most common processing steps are stemming [reducing words to their root form; exceptions were made in several cases (e.g., crash and crashes, policy and policies)] to consider the weightage of the frequencies and stop word removal (e.g., the, is, are, at). The study used open source R package stm to perform this analysis (*34*). The structural topic model package provides many useful features, including rich ways to explore topics, as well as appropriate uncertainty estimation and extensive visualization options. This package also is helpful to remove noncharacter text and html code, as well as to pass custom stop word lists. Through a preliminary analysis, the study developed a list of redundant and insignificant words to consider "stop words." The functions used in this package properly associate metadata with text data and reindex this relationship when text data fields are blank or become blank after preprocessing (e.g., with stop word removal).

The data import process will output documents, vocabulary, and metadata that can be used for analysis. STM incorporates metadata into the topic modeling framework. In STM, metadata can be entered in the topic model in two ways: topical prevalence and topical content. Metadata covariates for topical prevalence allow the observed metadata to affect the frequency with which a topic is discussed. Covariates in topical content allow the observed metadata

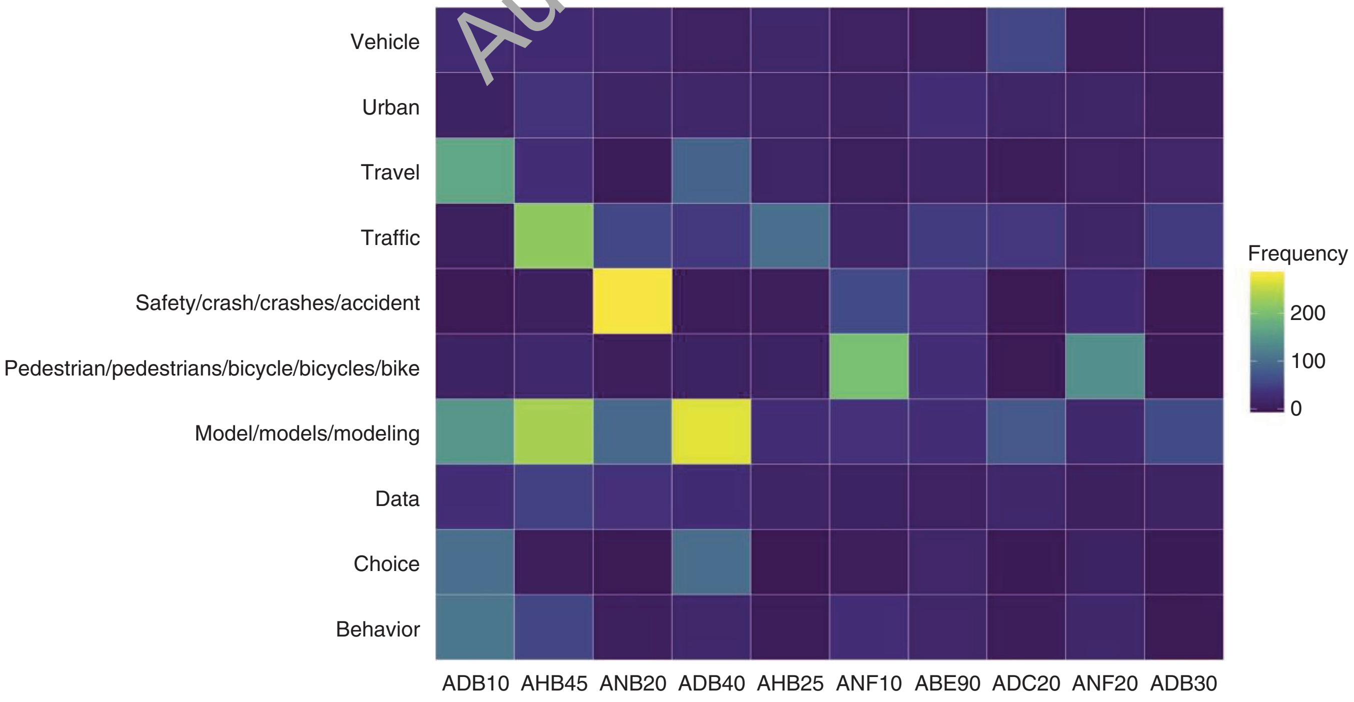


FIGURE 3 Heat map of most frequent words in different review committees.

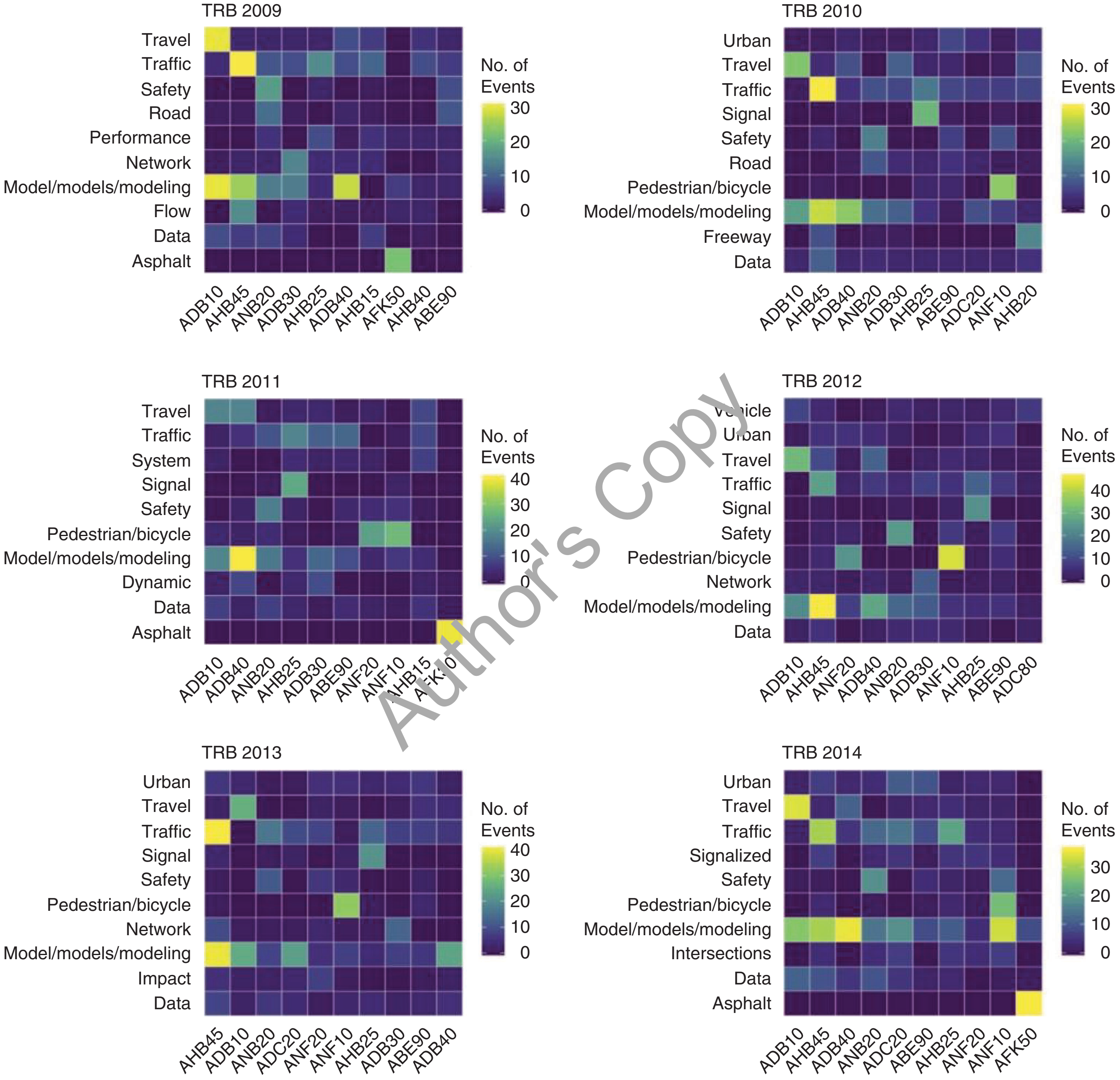


FIGURE 4 Heat maps of most frequent words used in different review committees (2009–2014).

to affect the word rate use within a given topic. Topical prevalence is a vector that sums up to 1 for each individual text or response: for example, in a four-topic model, one response may be deemed by the model to belong 70% to Topic 1 and 10% each to Topics 2, 3, and 4. Topical prevalence captures how much each topic contributes to a document. Because different documents come from different sources, it is scientific to allow this prevalence to vary with metadata about document sources. The model is set to run for a maximum of 100 expectation-maximization iterations. Typically, convergence of the model is monitored by the change in the approximate bound between expectation-maximization iterations. The current model is converged after 17 iterations.

Figure 5 illustrates the corpus-level visualization of the top topics from a 20-topic model and the frequency of words in each of these topics. It shows the expected proportion of the corpus that belongs to each topic. High-frequency topics include project, management, data, information, travel, and choice. Figure 6 shows 20 top topics (identified by the top three words in each topic) on the basis of highest probability and frequency–exclusivity.

Exclusivity is defined as the rank by distribution of topic given word. Semantic coherence has its basis in co-occurrence statistics for the top *n* words in a topic. Table 1 lists the average semantic coherence and exclusivity scores for each model (represented by topic numbers).

Figure 7 plots the semantic coherence and exclusivity scores of the topics in STM with 20 topic models. On the exclusivity dimension, the differences are small, which indicates that the top words for the topics can appear within top words of other topics. On the semantic coherence dimension, the differences are larger. It indicates that the words that are most associated with the corresponding themes do not occur equally within the documents. For example, the difference in the exclusivity of Topic 17 Speeding and Topic 3 Roadway Design is much less. It indicates that the top words in Topic 17 can reappear in Topic 3 and vice versa. The distinctness of the topics can be measured by the relative distance of these points. For example, Topic 6 Air Quality is relatively far away from Topic 18 Project Management, which means that these two topics are distinct in nature.

In addition, STM permits correlations between topics. Positive correlations between topics indicate that both topics are likely to be discussed within a document. Figure 8 shows Cluster 1: Topics 4, 8, 11, 17; Cluster 2: Topics 1, 9, 12, 13, 15, 18; and Cluster 3: Topics 2, 3, 5, 10, 16, 19, 20. The figure also shows that Topic 7 is associated with only one topic (Topic 8). Topic 7 contains information about travel estimation that has a closer relation with traffic flow contents in Topic 8. In Cluster 1 (Topics 4, 8, 11, and 17), the key words are "crash," "flow," "pedestrian," and "speed," respectively. These key words represent associations between traffic safety research methods. In Cluster 2 (Topics 1, 9, 12, 13, 15, and 18), the key words represent relationships between traffic flow and transit-related research. In Cluster 3, which shows networks between Topics 2, 3, 5, 10, 16, 19, and 20, the top key words are "mixture," "performance," "bridge," "concrete," "pavement," "test," and "fatigue," respectively. These key words represent associations between pavement-based research methods.

This study analyzed the frequency distribution (on the basis of the relative distance and size of the text) for the correlated topic models. Three groups of topic models were considered to illustrate the relationship between the correlated topic groups. The first group contained Topic 4 Crash Risk and Topic 8 Traffic Flow. Figure 9*a* indicates the higher presence of words like "crash," "factors," "flow," and "capacity." The relative closeness between the words in Topic 8 is more closely compared with the words in Topic 4. The second group contains Topic 11 and Topic 17. Figure 9*b* indicates a higher presence of words like "pedestrian," "drivers," and "speed." The relative closeness between the words in Topic 11 Pedestrian Safety is more closely compared with the words in Topic 17 Speeding. The third group contains Topic 16 Surface Treatment and Topic 19 Soil Test. Figure 9*c* indicates the higher presence of words like "pavements," "tests," and "results." The relative closeness between the words in Topic 19 is more closely compared with the words in Topic 16.

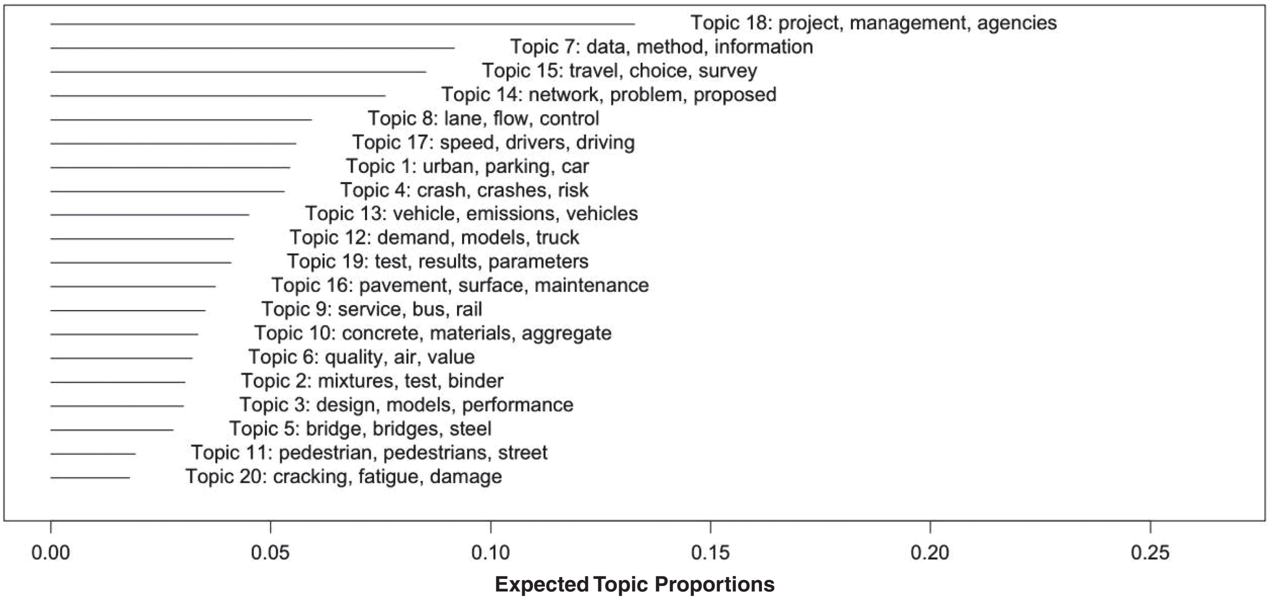


FIGURE 5 Expected topic proportions in top 20 topics.

Topic 1 [Urban Accessibility] Top Words
Highest Probability: urban, parking, car, congestion, areas, public, policy
FREX: parking, car, accessibility, pricing, urban, policy, policies

Topic 2 [Mix Binder Performance] Top Words
Highest Probability: mixtures, test, binder, mix, mixture, performance, temperature
FREX: binder, mixtures, mix, mixture, binders, HMA, temperature

Topic 3 [Roadway Design] Top Words
Highest Probability: design, models, performance, pavement, data, developed, used
FREX: design, MEPDG, calibration, guide, procedure, predicted, input

Topic 4 [Crash Risk] Top Words
Highest Probability: crash, crashes, risk, factors, weather, data, severity
FREX: crashes, crash, injury, severity, risk, weather, accident

Topic 5 [Bridge Design] Top Words
Highest Probability: bridge, bridges, steel, design, structures, system, concrete
FREX: bridge, bridges, steel, structures, reinforced, structural, specifications

Topic 6 [Air Quality] Top Words
Highest Probability: quality, air, value, impact, values, noise, airport
FREX: air, quality, noise, airport, value, track, exposure

Topic 7 [Trip Estimation] Top Words
Highest Probability: data, method, information, used, travel, estimation, proposed
FREX: estimation, collection, data, accuracy, GPS, incident, real

Topic 8 [Traffic Flow] Top Words
Highest Probability: lane, flow, control, capacity, lanes, signal, delay
FREX: lanes, signal, delay, lane, turn, flow, freeway

Topic 9 [Public Transport] Top Words
Highest Probability: service, bus, rail, passenger, system, stations, services
FREX: bus, passengers, train, station, buses, stations, passenger

Topic 10 [Pavement Materials] Top Words
Highest Probability: concrete, materials, aggregate, strength, material, base, content
FREX: materials, strength, cement, aggregates, concrete, water, aggregate

Topic 11 [Pedestrian Safety] Top Words
Highest Probability: pedestrian, pedestrians, street, crossing, walking, intersections, streets
FREX: pedestrian, pedestrians, crossing, street, streets, walking, facilities

Topic 12 [Travel Demand] Top Words
Highest Probability: demand, models, truck, freight, spatial, modeling, regional
FREX: truck, freight, region, regional, spatial, demand, modeling

Topic 13 [Vehicle Emissions] Top Words
Highest Probability: vehicle, emissions, vehicles, fuel, costs, cost, energy
FREX: emissions, fuel, consumption, emission, gas, costs, energy

Topic 14 [Network Optimization] Top Words
Highest Probability: network, problem, proposed, route, algorithm, networks, optimal
FREX: problem, network, toll, optimization, optimal, networks, assignment

Topic 15 [Travel Choice] Top Words
Highest Probability: travel, choice, survey, behavior, mode, activity, trip
FREX: bicycle, choice, household, survey, individuals, activity, travel

Topic 16 [Surface Treatment] Top Words
Highest Probability: pavement, surface, maintenance, pavements, sections, condition, performance
FREX: surface, sections, pavement, maintenance, friction, pavements, treatment

Topic 17 [Speeding] Top Words
Highest Probability: speed, drivers, driving, driver, vehicle, speeds, zone
FREX: drivers, speed, driver, driving, speeds, signs, zone

Topic 18 [Project Management] Top Words
Highest Probability: project, management, agencies, system, projects, process, infrastructure
FREX: agencies, management, projects, infrastructure, project, challenges, practices

Topic 19 [Soil Test] Top Words
Highest Probability: test, results, parameters, tests, soil, load, measured
FREX: soil, parameters, element, finite, pressure, tests, measured

Topic 20 [Pavement Condition] Top Words
Highest Probability: cracking, fatigue, damage, loading, failure, strain, stress
FREX: fatigue, damage, cracking, strain, failure, loading, stress

FIGURE 6 Top 20 topics on basis of highest probability and frequency–exclusivity. [Highest probability = group of words within each topic with highest probability (inferred directly from topic-word distribution parameter); FREX = group of words both frequent and exclusive, identifying words that distinguish topics, and calculated by taking the harmonic mean of rank by probability within the topic (frequency) and rank by distribution of topic given word (exclusivity); HMA = hot-mix asphalt; MEPDG = *Mechanistic–Empirical Pavement Design Guide*.]

TABLE 1 Average Semantic Coherence and Exclusivity of Top 20 Topics

Topic	Semantic Coherence	Exclusivity
Topic 1 [Urban Accessibility]: urban, parking, car	−101.93	9.92
Topic 2 [Mix Binder Performance]: mixture, test, binder	−54.31	9.87
Topic 3 [Roadway Design]: design, models, performance	−95.54	9.71
Topic 4 [Crash Risk]: crash, crashes, risk	−94.09	9.74
Topic 5 [Bridge Design]: bridge, bridges, steel	−90.70	9.69
Topic 6 [Air Quality]: quality, air, value	−128.43	9.92
Topic 7 [Trip Estimation]: data, method, information	−82.60	9.56
Topic 8 [Traffic Flow]: lane, flow, control	−84.99	9.92
Topic 9 [Public Transport]: service, bus, rail	−105.42	9.93
Topic 10 [Pavement Materials]: concrete, materials, aggregate	−87.20	9.96
Topic 11 [Pedestrian Safety]: pedestrian, pedestrians, street	−105.52	9.86
Topic 12 [Travel Demand]: demand, models, truck	−105.89	9.81
Topic 13 [Vehicle Emissions]: vehicle, emissions, vehicles	−93.13	9.92
Topic 14 [Network Optimization]: network, problem, proposed	−85.99	9.89
Topic 15 [Travel Choice]: travel, choice, survey	−96.54	9.74
Topic 16 [Surface Treatment]: pavement, surface, maintenance	−105.08	9.93
Topic 17 [Speeding]: speed, drivers, driving	−95.04	9.72
Topic 18 [Project Management]: project, management, agencies	−85.47	9.37
Topic 19 [Soil Test]: test, results, parameters	−101.77	9.66
Topic 20 [Pavement Condition]: cracking, fatigue, damage	−89.95	9.90

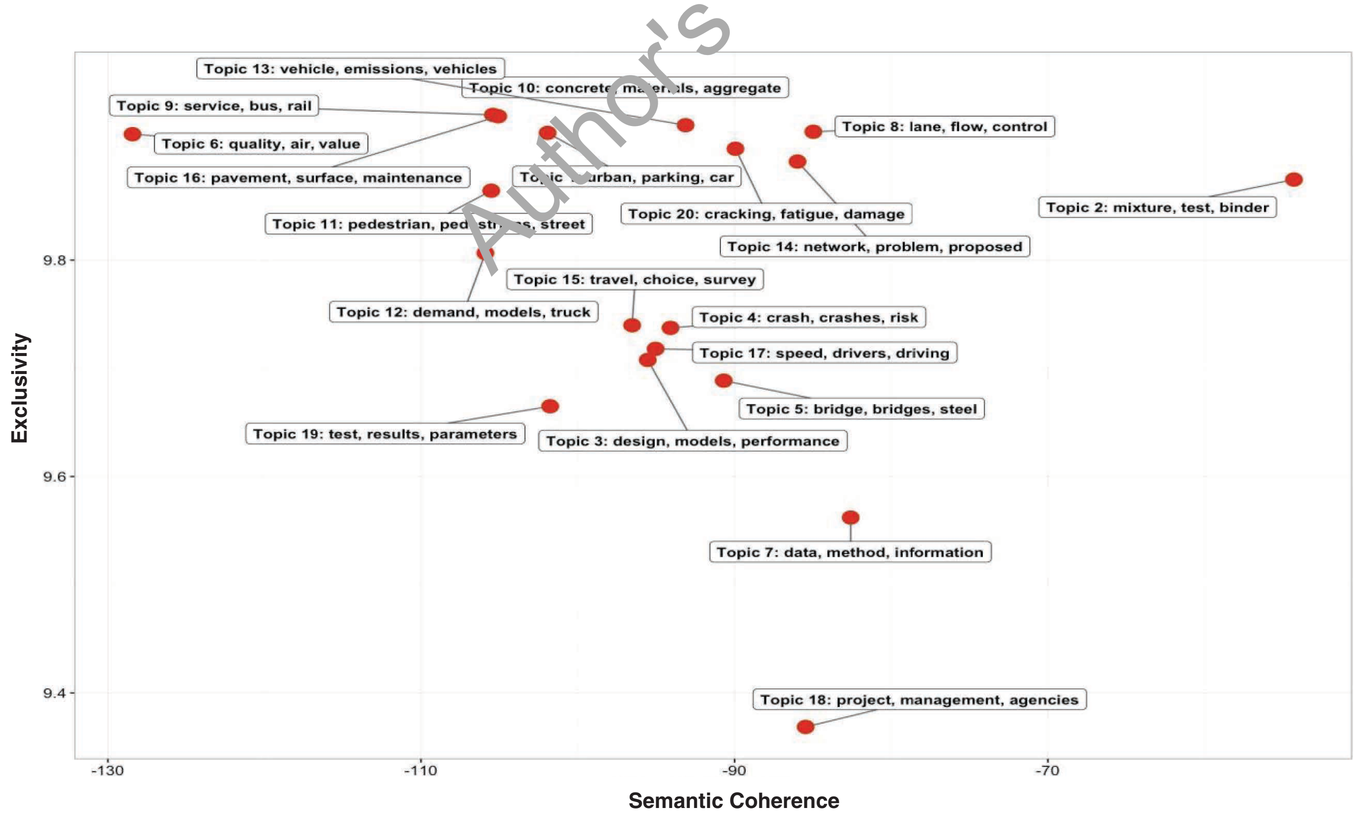


FIGURE 7 Topic quality plot on basis of average semantic coherence and exclusivity.

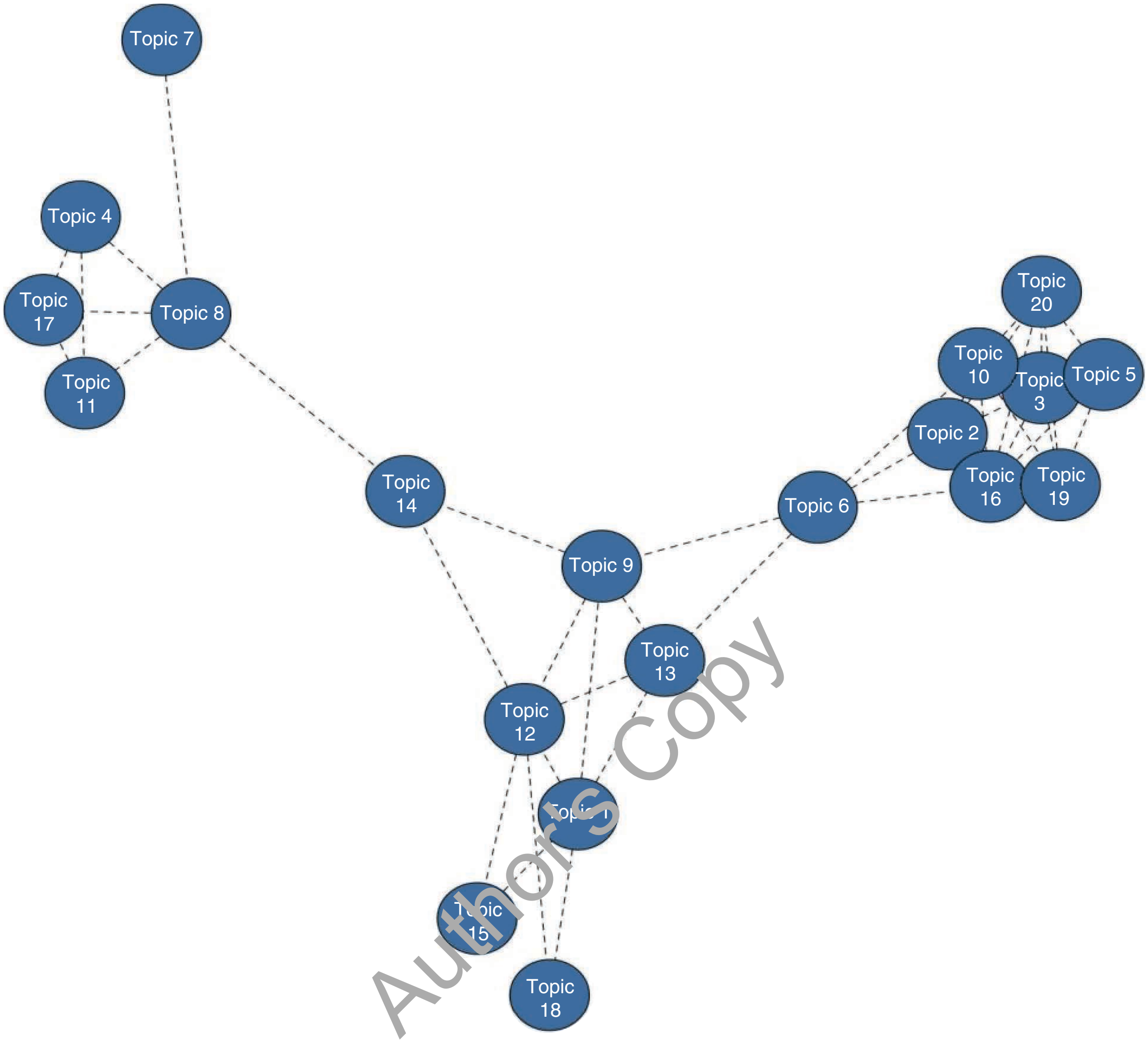


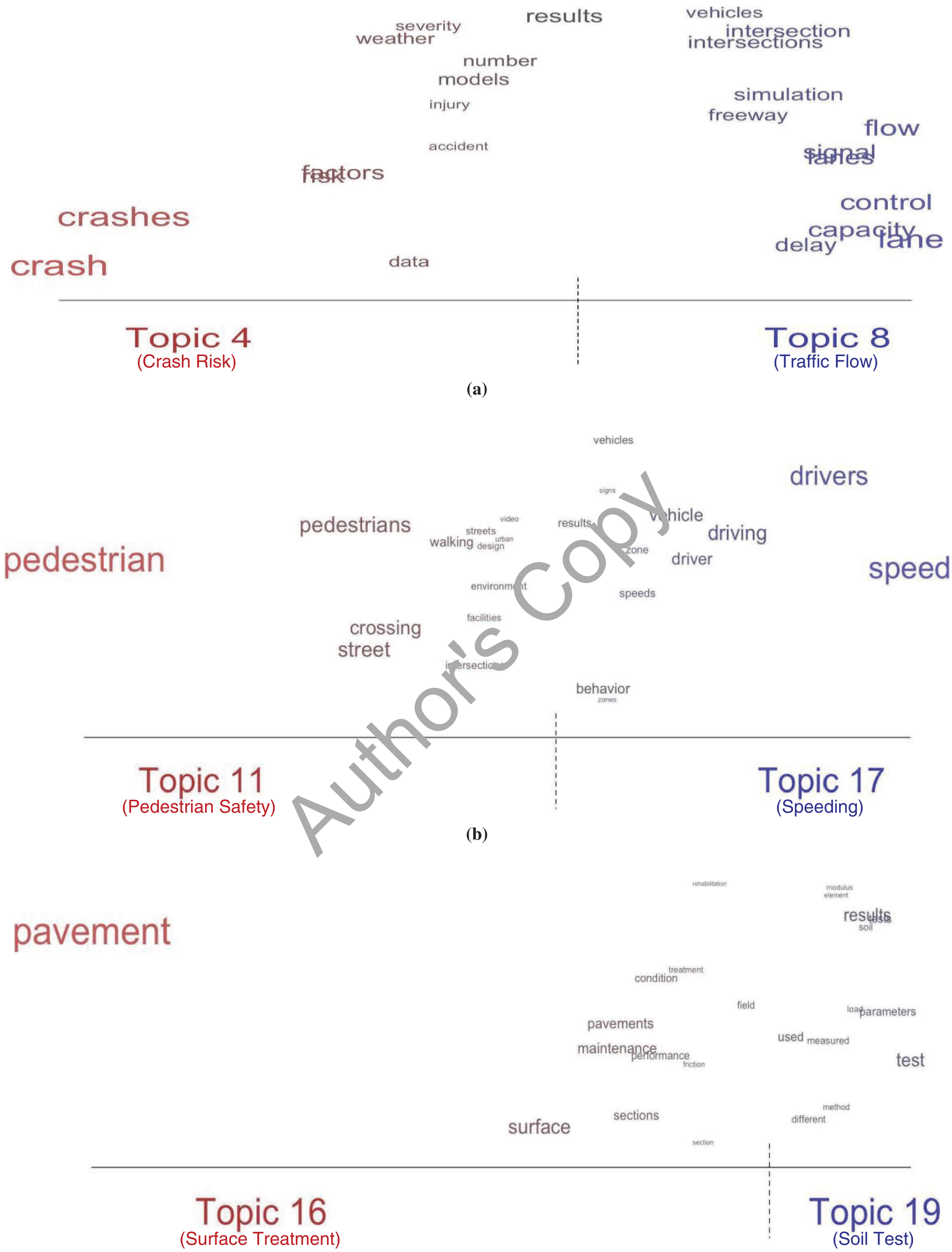
FIGURE 8 Correlation of 20 topic models.

The study findings were as follows:

- In topic proportions, traffic safety and traffic operational research were more often discussed than pavement-related research.
- The top three topics on the basis of expected higher topic proportions were Topic 18 Project Management, Topic 7 Trip Estimation, and Topic 15 Travel Choice.
- Top 20 topic groups provided high-frequency words on the basis of two scores. The cluster of words in each topic group implied the higher presence of these key words in each group.
- The semantic coherence and exclusivity scores provided the distinctness of each topic. The relative closeness of the topics in the exclusivity dimension was less than the relative closeness of the topics in the semantic coherence dimension.
- Correlation plots showed correlations through networks. These networks indicated the most related topic models of the topic 20 models. Three specific clusters were visible in the top 20 topic models.

CONCLUSION

Topic trend extraction from big text corpus is fundamental in most of the topic models. The mining of knowledge hidden behind big data is a popular research topic around the world, but to date limited research on the topic has been conducted in the field of transportation engineering. This study performed text mining on 4 million words from the titles and abstracts of 15,357 compendium papers from seven TRB annual meetings. The study used the metadata for each of the abstracts to determine more document-specific topics and identified the top 20 topics by providing high-frequency words on the basis of two scores: high probability and frequency–exclusivity. The results showed that the conduct of traffic safety and traffic operational research was higher proportionally than that of pavement-related research. In addition, semantic coherence and exclusivity scores were explored to provide the distinctness of each topic. The study provided a unique tool to explore topical prevalence and content and to iden-



(a)

(b)

(c)

FIGURE 9 Trends of topics in three correlated topic groups.

tify more relevant trends in the expansive fields of transportation research to generate disaggregate level correlation. A practical use of this research could be to implement topic prevalence in the assignment of papers to the appropriate committees. Future application might include the use of additional topic models for different sets of transportation research papers and reports, (e.g., published papers in the *Transportation Research Record* and NCHRP research reports).

ACKNOWLEDGMENTS

Mark R. Norman, TRB Technical Activities Division Director, and Michael DeCarmine, TRB Program Officer, provided the data that contain information on TRB annual meeting compendium papers. The authors appreciate their interest in, and support of, this study.

REFERENCES

1. Das, S., X. Sun, and A. Dutta. Text Mining and Topic Modeling of Compendiums of Papers from Transportation Research Board Annual Meetings. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2552, 2016, pp. 48–56. https://dx.doi.org/10.3141/2552-07.
2. Blei, D. Probabilistic Topic Models. *Communications of the ACM,* Vol. 55, No. 4, 2012, pp. 77–84. https://doi.org/10.1145/2133806.2133826.
3. Grimmer, J., and B. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis. *Political Analysis,* Vol. 21, No. 3, 2013, pp. 267–297. https://doi.org/10.1093/pan/mps028.
4. Hofmann, T. Probabilistic Latent Semantic Indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM, New York, 1999, pp. 50–57. https://doi.org/10.1145/312624.312649.
5. Blei, D., A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research,* Vol. 3, 2003, pp. 993–1022.
6. Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. *Proceedings of Advances in Neural Information Processing Systems 22* (Y. Bengio, ed.), Curran Associates, Inc., Red Hook, N.Y., 2009, pp. 288–296.
7. Mimno, D., H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing Semantic Coherence in Topic Models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* Association for Computational Linguistics, Stroudsburg, Pa., 2011, pp. 262–272.
8. Andrzejewski, D., and X. Zhu. Latent Dirichlet Allocation with Topic-In-Set Knowledge. *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing,* Association for Computational Linguistics, Stroudsburg, Pa., 2009, pp. 43–48.
9. Andrzejewski, D., X. Zhu, and M. Craven. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. *Proceedings of the 26th Annual International Conference on Machine Learning,* ACM, New York, 2009, pp. 25–32.
10. Andrzejewski, D., X. Zhu, M. Craven, and B. Recht. Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation Using First-Order Logic. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence,* Vol. 2, AAAI Press, Palo Alto, Calif., 2011, pp. 1171–1177.
11. Chemudugunta, C., A. Holloway, P. Smyth, and M. Steyvers. Modeling Documents by Combining Semantic Concepts with Unsupervised Statistical Learning. *Proceedings of the Semantic Web: 7th International Semantic Web Conference* (A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, and K. Thirunarayan, eds.), Springer International, Switzerland, 2008, pp. 229–244.
12. Chen, Z., A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Exploiting Domain Knowledge in Aspect Extraction. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Association for Computational Linguistics, Stroudsburg, Pa., 2013, pp. 1655–1667.
13. Chen, Z., A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Leveraging Multi-Domain Prior Knowledge in Topic Models. *Proceedings of the 23rd International Joint Conference on Artificial Intelligence,* AAAI Press, Palo Alto, Calif., 2013, pp. 2071–2077.
14. Doshi-Velez, F., B. Wallace, and R. Adams. Graph-Sparse LDA: Topic Model with Structured Sparsity. *Proceedings of the 29th AAAI Conference on Artificial Intelligence,* 2015, AAAI Press, Palo Alto, Calif., pp. 2575–2581.
15. Yao, L., Y. Zhang, B. Wei, H. Qian, and Y. Wang. Incorporating Probabilistic Knowledge into Topic Models. *Proceedings of the 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining,* Springer International, Switzerland, 2015, pp. 586–597.
16. Blei, D., and J. Lafferty. Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning,* ACM, New York, 2006, pp. 113–120. https://doi.org/10.1145/1143844.1143859.
17. Kalyanam, J., A. Mantrach, D. Saez-Trumper, H. Vahabi, and G. Lanckriet. Leveraging Social Context for Modeling Topic Evolution. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, New York, 2015, pp. 517–526. https://doi.org/10.1145/2783258.2783319.
18. Wang, X., and A. McCallum. Topics over Time: Non-Markov Continuous-Time Model of Topical Trends. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, New York, 2006, pp. 424–433. https://doi.org/10.1145/1150402.1150450.
19. Wei, X., J. Sun, and X. Wang. Dynamic Mixture Models for Multiple Time-Series. *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Morgan Kaufmann Publishers, Inc., San Francisco, Calif., 2007, pp. 2909–2914.
20. Yan, X., J. Guo, Y. Lan, J. Xu, and X. Cheng. Probabilistic Model for Bursty Topic Discovery in Microblogs. *Proceedings of the 29th AAAI Conference on Artificial Intelligence,* AAAI Press, Palo Alto, Calif., 2015, pp. 353–359.
21. Eisenstein, J., A. Ahmed, and E. Xing. Sparse Additive Generative Models of Text. *Proceedings of the 29th International Conference on Machine Learning,* ICML, Bellevue, Wash., 2011, pp. 1041–1048.
22. Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smyth. The Author-Topic Model for Authors and Documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence,* AUAI, Arlington, Va., 2004, pp. 422–429.
23. Ahmed, A., and E. P. Xing. Staying Informed: Supervised and Semi-Supervised Multi-View Topical Analysis. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* Association for Computational Linguistics, Stroudsburg, Pa., 2010, pp. 1140–1150.
24. Eisenstein, J., B. O'Connor, N. Smith, and E. Xing. Latent Variable Model for Geographic Lexical Variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* Association for Computational Linguistics, Stroudsburg, Pa., 2010, pp. 1277–1287.
25. Blei, D., and J. McAuliffe. Supervised Topic Models. *Proceedings of Advances in Neural Information Processing Systems* 20 (J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds.), Neural Information Processing Systems Foundation, Inc., La Jolla, Calif., 2007.
26. Ramage, D., C. Manning, and S. Dumais. Partially Labeled Topic Models for Interpretable Text Mining. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, New York, 2011, pp. 457–465. https://doi.org/10.1145/2020408.2020481.
27. Paul, M., and M. Dredze. Factorial LDA: Sparse Multi-Dimensional Text Models. *Proceedings of Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), Neural Information Processing Systems Foundation, Inc., La Jolla, Calif., 2012, pp. 2591–2599.
28. Yao, L., Y. Zhang, B. Wei, L. Li, F. Wu, P. Zhang, and Y. Bian. Concept over Time: The Combination of Probabilistic Topic Model with Wikipedia Knowledge. In *Expert Systems with Applications,* Vol. 60, Elsevier B.V., Amsterdam, Netherlands, 2016, pp. 27–38. https://doi.org/10.1016/j.eswa.2016.04.014.

29. Bibliography on Topic Modeling: Research Papers and Abstracts. http://subasish.github.io/pages/TRB2016/topicm.html. Accessed July 2016.
30. Blei, D., and J. Lafferty. Correlated Topic Model of Science. *Annals of Applied Statistics,* Vol. 1, No. 1, 2007, pp. 17–35.
31. Mimno, D., and A. McCallum. Topic Models Conditioned on Arbitrary Features with Dirichlet–Multinomial Regression. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence,* AUAI Press, Arlington, Va., 2008, pp. 411–418.
32. Roberts, M., B. Stewart, D. Tingley, and E. Airoldi. *Structural Topic Model and Applied Social Science.* Presented at Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation, 2013.
33. Feinerer, I., and K. Hornik. Tm: Text Mining Package. R Package Version 0.6-2, 2015. http://CRAN.R-project.org/package=tm. Accessed July 2016.
34. Roberts, M., B. Stewart, and D. Tingley. Stm: R Package for Structural Topic Models. 2016. http://www.structuraltopicmodel.com. Accessed July 2016.

The Standing Committee on Library and Information Science for Transportation peer-reviewed this paper.