Transportation Research Record Articles: A Case Study of Trend Mining

Subasish Das, Ph.D.
(Corresponding author)
Associate Transportation Researcher, Texas A&M Transportation Institute
1111 RELLIS Parkway, Room 4414, Bryan, TX 77807
Email: s-das@tti.tamu.edu
ORCID: 0000-0002-1671-2753

Anandi Dutta, Ph.D.
Senior Lecturer, Computer Science and Engineering
Ohio State University, 2015 Neil Ave, Columbus, OH 43210
Email: dutta.34@osu.edu

Marcus A. Brewer, P.E., PMP
Research Engineer, Texas A&M Transportation Institute
1111 RELLIS Parkway, Room 4426, Bryan, TX 77807
Email: m-brewer@tti.tamu.edu
ORCID: 0000-0002-1996-3259

Submitted to Transportation Research Record
Date: December 20, 2019

ABSTRACT

This study employs two topic models to perform trend mining on an abundance of textual data to determine trends in research topics from massive collections of unstructured documents over the years. This study collected data from the titles and abstracts of the papers published in *Transportation Research Record: Journal of the Transportation Research Board*, since 1978. The content of these papers was ideal for examining research trends in various fields of research because it contains large textual data. In previous studies, exploratory analysis tools such as text mining were used to provide descriptive information about the data. However, this method does not provide researchers with quantifications of the topics and their correlations. Furthermore, the contents examined in this study are largely unstructured, and therefore they require faster machine learning algorithms to decipher them. For these reasons, the research team chose to employ two topic modeling tools, Latent Dirichlet Allocation (LDA) and Structural Topic Model (STM), to perform trend mining. This analysis succeeded in extracting twenty main topics, identified by keywords, from the data. The research team also developed two interactive topic model visualizations that can be used to extract topics from journal titles and abstracts, respectively. The findings from this study provide researchers with a further understanding of trends within the complex and ever-evolving field of transportation engineering research.

Keywords: trend mining, latent Dirichlet allocation, topic model, natural language processing.

INTRODUCTION

Transportation is an increasingly important factor in people's quality of life. Rapid growth in population, miles traveled, urbanization, and emerging technologies like connected and autonomous vehicles have had a substantial impact on modern living. Transportation research has produced numerous benefits in both engineering and science, providing better transportation services and systems. The rising application of emerging technologies; the increasing number of peer reviewed journals and conference proceedings; and the significant growth in interdisciplinary collaborations reflects the significance of the size and scope of transportation research. Transportation challenges and problems, however, have changed over time, and the transportation research scope has also become more diverse. The domain of transportation research includes a broad inter-disciplinary coverage of topics, ranging from classic topics such as signal control and traffic congestion to societal problems such as environmental justice and sustainability to new technologies such as big data analytics, connected vehicles, autonomous vehicles, and artificial intelligence. Due to the consistent evolution from the advances in solutions/technologies developed and the specific questions raised, transportation research has experienced an upsurge of research publications in recent decades.

Predicting future salient issues in any field of science that will dominate research is always a challenge, but as transportation research becomes more complex and cross-cutting, this challenge will increase. Research regarding statistical models of co-occurrence of trending topics has led to the growth of different useful topic models. This efficient machine learning technique helps researchers find concealed trends inside unstructured larger textual contents.

The Transportation Research Board (TRB) coordinates the most comprehensive and largest annual transportation conference in the world. With their establishment in 1920 as the National Advisory Board on Highway Research, TRB has provided a platform to convert research results into applicable information about every facet pertaining to transportation engineering. Thousands of scientists, engineers, and other transportation practitioners and researchers from the private and public sectors and academia are all included in the TRB's various activities. This program has gained the approval of the state transportation departments and federal agencies, including the component administrations of the U.S. Department of Transportation.

The Transportation Research Record (TRR) series is the official Journal of the Transportation Research Board and has technical papers that have been accepted for publication through a rigorous peer-review process refereed by TRB technical committees. These papers provide extensive documentation of the research activities undertaken by the transportation research community, and they provide a unique insight into the research topics that have remained active over the long term as well as topics that have recently emerged into the forefront.

To comprehend the research trends in the realm of complex transportation engineering, an analysis of TRR journal articles would be beneficial. By applying latent Dirichlet allocation (LDA) model, this study presents an empirical analysis of 30,784 articles published in the Transportation Research Record from 1978 to 2019 to identify trends in topics, keywords, and authors over time.

LITERATURE REVIEW

In recent years, probabilistic topic models, such as Latent Dirichlet Allocation (LDA) *(1)*, have become a popular research tool to interpret large amounts of textual data. Researchers have noted the importance of topic models *(2)* in measuring latent linguistic significance. Most studies involving text mining analysis employ statistical topic models such as Probabilistic Latent Semantic Analysis (PLSA) *(3)*, in addition to LDA *(4)*. However, these models are unsupervised, meaning the explanatory variables and response variables are not clearly defined; this can result in topics that are not interpretable *(5, 6)*.

To overcome the challenges associated with LDA and other conventional methods, researchers have proposed many knowledge-based topic models *(7-14)* and dynamic topic models (DTMs) *(15-19)*. Furthermore, researchers examined the performance of the automatic coherence measure of topic models and developed an unsupervised method to improve the coherence score by considering the word "co-occurrence" within a collection of texts *(20)*. Researchers have also proposed DTMs in which time is a significant consideration, such as Topic over Time (TOT) *(17)* and Dynamic Mixture Model (DMM), to mine dynamic patterns *(4, 21, 17, 18)*. Additionally, several researchers have recently suggested nonparametric Bayesian models, based on Dirichlet Process (DP), to model topics over space and time *(16-20)*.

McLaurin et al. (*21*) applied topic modeling to driving data and distinguished key associations between drivers with obstructive sleep apnea and normal drivers. Sun et al. (*22*) used the temporal doubly stochastic Dirichlet process (TDSDP) mixture model and presented an unsupervised tracking algorithm for human and car trajectory detection. In their study, Sun and Yin (*23*) used a latent Dirichlet allocation (LDA) model on article abstracts to deduce 50 key topics. Their results indicated that the characterized topics are representative and meaningful, specifically in regard to the established sub-fields of transportation research.

Venkatraman et al. (*24*) investigated "differences between drivers" lateral responses in various events utilizing probabilistic topic modeling. Another worthwhile study of topic modeling in the transportation field was conducted by Das et al. (*25*) in which they studied topic changes of abstracts from papers in the Transportation Research Board (TRB) Annual Meeting from 2008 to 2014. Das et al. used text mining and topic modeling in several other transportation studies *(26-29)*. Two recent studies used text mining and topic modeling TRB compendium papers and TRR papers (*30-31*). In a recent study, Biehl (*32*) used both text mining and topic modeling techniques to investigate the promotion of walking and cycling adoption utilizing several focus groups of the local residents in two geographic communities: Chicago's Humboldt Park neighborhood and the suburb of Evanston. They combined traditional qualitative discourse analysis with popular natural language tools such as topic modeling and sentiment analysis.

The framework behind considering additional information or meta-data about the structure of the corpus in the modeling framework uses the altercation of the prior distributions to partially pool information amongst similar documents. Researchers have explored incorporating meta-data into models from various aspects: author-topic model (*33, 34*), topical content/ideology (*35*), geography (*36*), trend analysis (*26*), attitudes on self-driving cars (*37*), and aviation incident reports (*38*).

TOPIC MODELING

Latent Dirichlet Allocation

In 2003, Blei et al. developed the Latent Dirichlet Allocation (LDA) model to address the issues found in the probabilistic latent semantic analysis (PLSI) model (*4, 39-40*). Improving upon the

PLSI model, the LDA model uses a K-dimensional latent random variable. This variable follows the Dirichlet distribution to show the topic mixture ratio of the document. The LDA model has been proven as one of the most widely used topic models (*41*).

Having a stronger descriptive power than others, the LDA model is more capable of matching the semantic conditions than other models. The parameter space of the LDA model is simpler than the PLSI model. Additionally, this hierarchical model, with a more stable structure, avoids any overfitting condition because its parameter space is not relevant to the number of training documents in LDA (*41*). This model is generally considered as a complete probability generative model (*41, 42*).

The authors mostly followed a study conducted by Kim and Shim (*43*) for a brief overview of LDA. Consider U and D_u denote the set of users and the 'bag of words' generated by a user $u \in U$ respectively. Consider V be the set of distinct words appearing in a bag of words D_u at least once for a user $u \in U$. To represent the set of latent topics where the number of topics is given as a parameter, Z is utilized by the user. In the generative process of LDA, each user u has his own preference over the topics represented by a probabilistic distribution $\overrightarrow{\theta_u}$, which is a multinomial distribution over Z. Also, each topic z has a multinomial distribution over V, denoted by $\overrightarrow{\emptyset_z}$.

Figure 1 illustrates the graphical representation of the LDA model. The generative process of this method can be defined as following (*43*):

- For each topic $z \in Z$, consider a multinomial distribution $\emptyset_z \sim Dir(\vec{\beta})$.
- For each user $u \in U$, consider a multinomial distribution $\emptyset_u \sim Dir(\vec{\alpha})$.
- For each word $w \in D_u$,
    - consider a topic $z \sim Multinomial(\overrightarrow{\theta_u})$.
    - consider a word w$\sim Multinomial(\overrightarrow{\emptyset_z})$.

The LDA model considers that the multinomial distributions $\overrightarrow{\theta_u}$ and $\overrightarrow{\emptyset_z}$ are calculated from conjugate prior distributions, known as Dirichlet distribution, whose parameters are given as $\vec{\alpha}$ and $\vec{\beta}$ respectively. Each word w in D_u is supposed to be designated by first drawing a topic z with following the topic preference distribution $\overrightarrow{\theta_u}$ and next choosing a word w from the corresponding distribution $\overrightarrow{\emptyset_z}$ of the selected topic z. According to this modeling approach, the probability that a word w is generated by a user u can be determined by:

$$\int Dir(\theta_u; \alpha) \left(\sum_{z=1}^{|Z|} \theta_{uz} \emptyset_{zw} \right) d\theta_u.$$

Structural Topic Model (STM)

In political science and linguistics, the STM has been used for text data analysis (*44-49*). Both LDA and STM are Bayesian generative topic models that assume that each topic is a distribution over words and each document is a mixture of corpus-wide topics (*2, 5*). The algorithm of STM identifies document-level structure information to influence topical prevalence (for example, proportion of topics by document frequency) and topic content (distribution of the keywords in topics). It emphasizes the suitability determination of investigating how covariates affect the content of text documents. A brief introduction on STM is described below, which is based on theories described in other studies (2, 5).

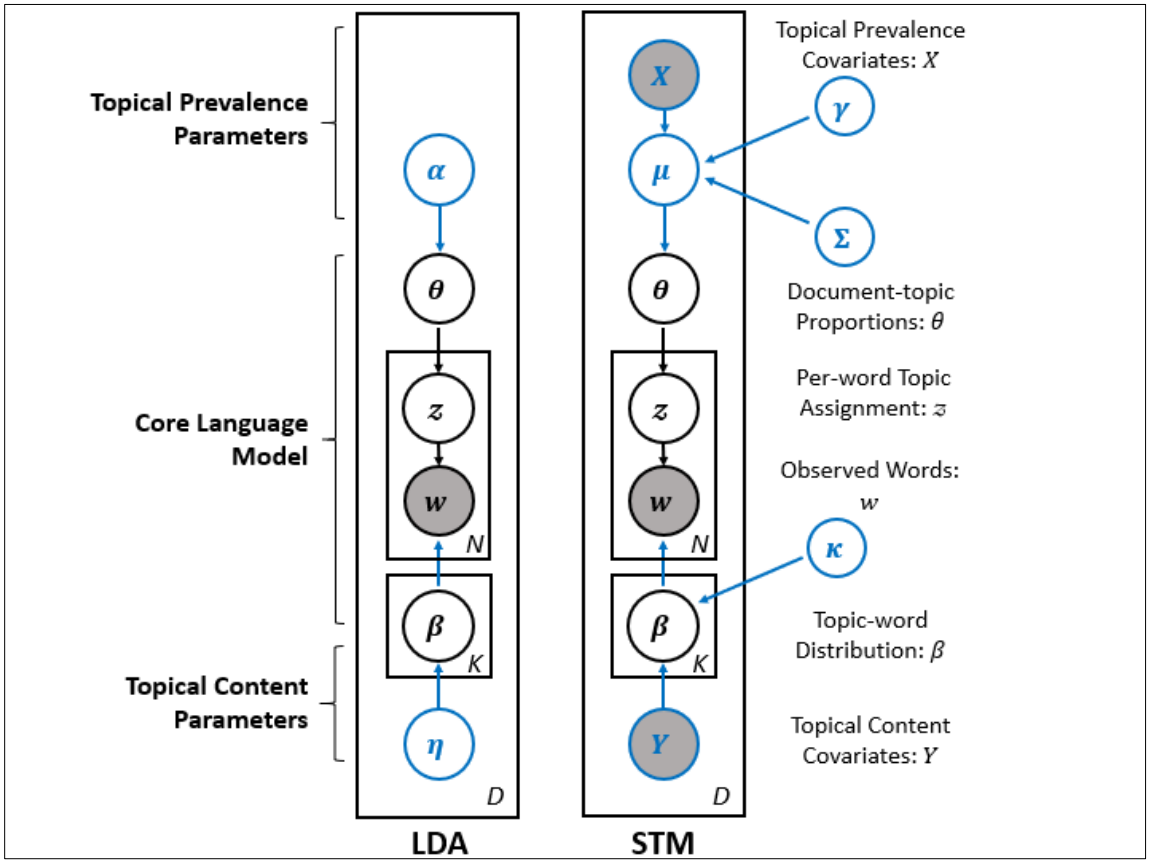


Figure 1 LDA and STM frameworks.

Figure 1 presents the technical differences between the frameworks of STM and LDA models. A variable labeled with its role in the data generating process represents each node. The shaded nodes are the observed or real variables and the unshaded nodes are hidden variables. The rectangles indicate replication: $n \in \{1, 2, \dots, N\}$ indexes words within a document; $k \in \{1, 2, \dots, K\}$ indexes each topic assuming the user-specified number of topics is K; and $d \in \{1, 2, \dots, D\}$ represents the document indexes. Figure 2 also illustrates that only node w (i.e., words in documents) is seen in LDA and STM. As such, the objective of these two models is to conclude the hidden topic information from the observed words, W and output two critical matrices, per-document topic proportions, θ and topic-word distributions, β. Additionally, Figure 2 indicates that STM and LDA have a similar framework with three components: topical prevalence parameters, the core language model, and consider topical content parameters. The elements of the core language model of LDA and STM are the same, where θ_d and $\beta_{d,k,v}$ denote the hidden per-document topic proportions and per-corpus topic-word distributions, respectively; $z_{d,n}$ implies the hidden topic assignment of each stated term; and $w_{d,n}$ denotes the stated term, which is drawn from words indexed by $v \in \{1, 2, \dots, V\}$. The core language model of STM and LDA follows the two-step generative process for each document d in the corpus (*2*, *5*).

- Step 1. Perform random choice of a distribution over topics θ_d for document d.
- Step 2. For each word w_n in the document d, (a) conduct random choice of a topic $z_{d,n}$ from the distribution over topics θ_d in Step 1. (b) conduct random choice of a word w_n from the corresponding distribution over the vocabulary $\beta_{d,k,v}$, where $k = z_{d,n}$.

STM can be differentiated from LDA by the topical prevalence parameters and topical content parameters. Particularly, the topical prevalence (content) parameters in LDA are certain shared prior Dirichlet parameters $\alpha(\eta)$, while those of STM are replaced with prior structures specified in the form of generalized linear models parameterized by document specific covariates $X(Y)$ (*44*).

The algorithm differences uncover prominent distinctions in STM when compared with LDA. First, with the help of STM, the researchers can introduce document-level covariates to the topical prevalence parameters that affect document-topic proportions *(44).* Second, STM also encourages scholars to introduce document-level covariates to the topic content parameters that affect the topic-word distributions *(44).* Third, STM is an extension of the correlated topic model, in which topics can be correlated with one another (*50*), thereby enabling the user to clearly examine the correlation structures among topics.

METHODOLOGY

Data Collection

The research team desired a robust, long-standing journal to conduct an analysis of trends for topics, authors, and keywords. They selected the TRR series based on its long history and its inclusion of a wide variety of subject matter. The TRR series was also attractive because of its rigorous review process and widespread use among both academia and practitioners. The research team used the Transport Research International Documentation (TRID) website to develop the databases for this study. All TRR articles are first saved in the research information system (RIS) format. Later, the database is converted into spreadsheet format. The columns in the database include the title of the paper, keywords, abstract, authors, and publication year. This analysis included 30,784 articles (see Table 1) published between 1974 and 2019. Publication years of the articles were extracted from TRID metadata.

Table 1 Number of Journal Articles and Word Counts in Titles and Abstracts by Year

Year	Number of Articles	Total Words in Titles	Total Words in Abstracts
1974	368	3,002	52,494
1975	222	1,741	31,397
1976	623	5,256	96,121
1977	446	3,686	70,583
1978	479	4,080	80,370
1979	456	3,930	73,217
1980	474	4,006	71,791
1981	526	4,527	80,871
1982	589	4,947	98,103
1983	613	5,507	104,646
1984	607	5,306	93,568
1985	505	4,650	84,221
1986	542	4,934	88,476
1987	591	5,590	102,616
1988	543	5,146	92,709
1989	471	4,467	81,260
1990	587	5,579	102,819
1991	797	7,507	140,139
1992	615	5,888	110,179
1993	638	6,129	112,400
1994	605	5,935	114,134

1995	614	6,026	114,238
1996	727	6,952	133,387
1997	595	6,046	113,570
1998	613	6,313	115,072
1999	729	7,516	142,228
2000	703	7,104	136,395
2001	676	7,163	127,888
2002	662	6,942	127,733
2003	759	8,096	150,420
2004	689	7,395	134,013
2005	834	9,274	163,373
2006	816	9,070	160,726
2007	825	9,277	166,216
2008	703	7,932	142,645
2009	779	8,973	156,838
2010	951	11,160	189,703
2011	995	11,754	202,750
2012	939	11,177	191,928
2013	931	11,115	193,258
2014	932	11,377	195,497
2015	971	11,837	204,261
2016	875	10,847	186,506
2017	866	10,812	182,574
2018	719	9,300	153,102
2019 (partial)	584	7,461	124,637
Grand Total	**30,784**	**322,732**	**5,791,072**

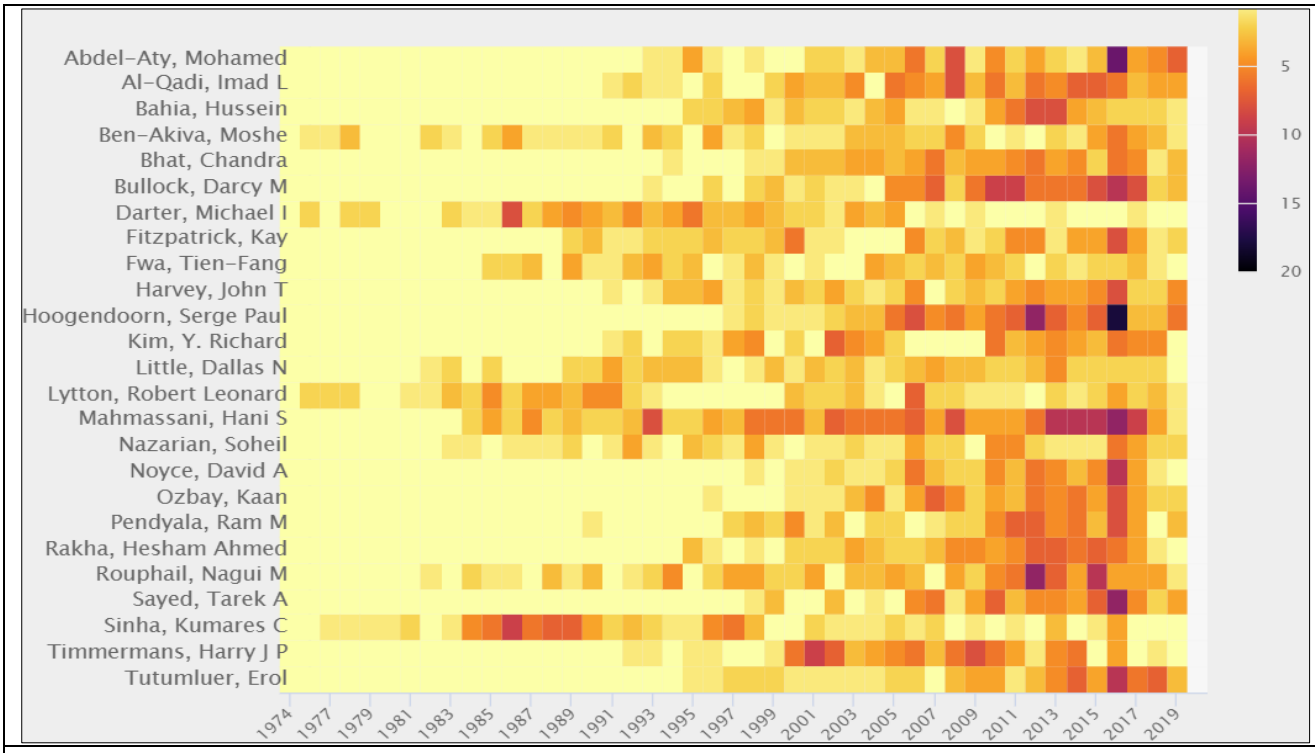
Exploratory Text Mining

Prolific Authors and Co-Author Networks

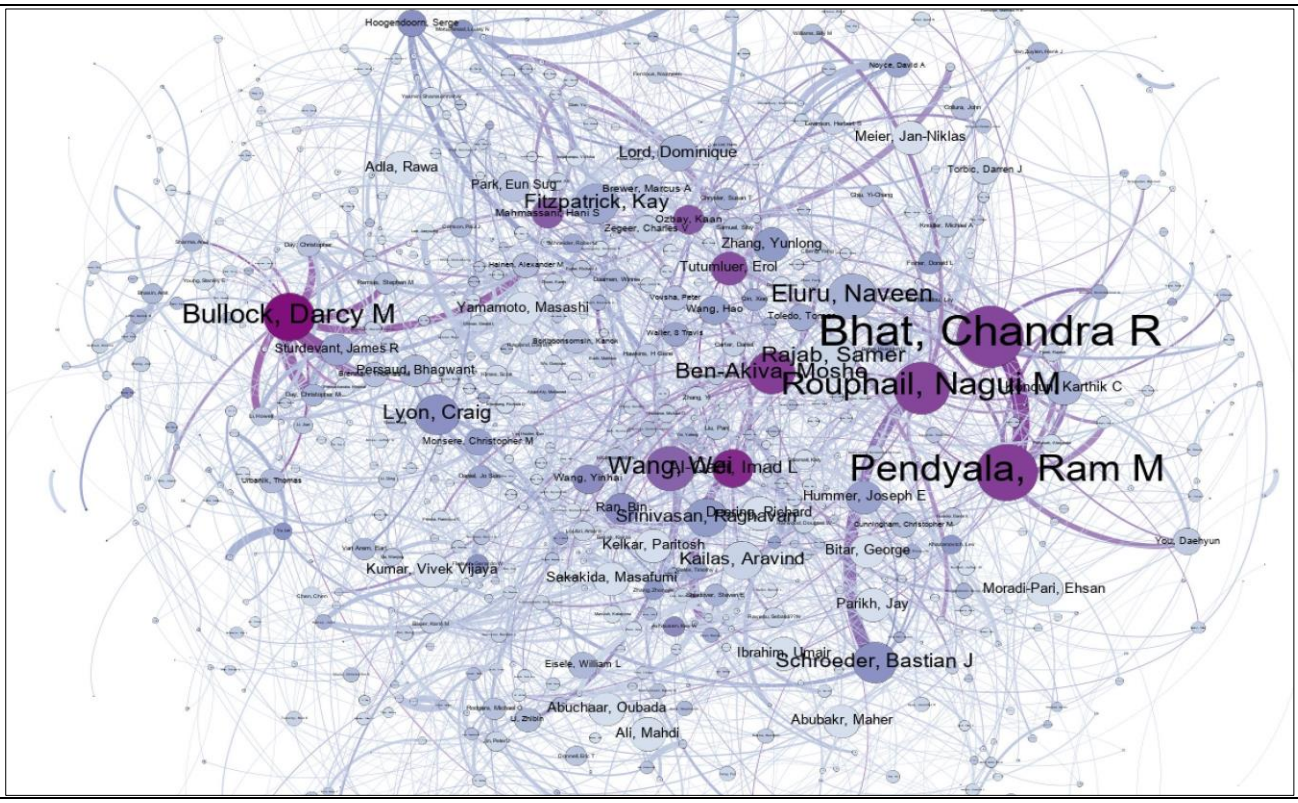
The authors developed a web-based interactive visualization of the heatmap shown in Figure 2a to list the top 25 most prolific authors of the TRR articles and illustrate the frequency of publication for each author *(51).* The authors are listed in alphabetical order by last name. The colors indicate the number of TRR articles published by each author by year. The light-yellow color, primarily shown toward the left side of the heat map, indicates the beginning of the scientific careers for these prolific authors. The darker colors indicate a greater number of articles published by the author in the given year. As shown in Figure 2a, Serge Hoogendoorn published 18 TRR articles in 2016, which is the maximum number of TRR articles by research as an author or co-author in one year. Hani Mahmassani has published 183 TRR articles since 1984, the highest total in that period. Four of the listed researchers started their publication in the early days of the TRR journals. Another four researchers started their careers in the ‘80s. Most of the other authors started publishing papers in the TRR after 1990.

Co-authorship networks can be used to investigate the structure of scientific collaborations. As transportation research has become increasingly cross-disciplinary, it is important to investigate the patterns and trends of the authors. The current study is only limited to develop a co-author network plot for a quick understanding of the complex interdisciplinary networks between the authors. Future studies can explore the development of advanced analysis like author-topic model development. The network plot, shown in Figure 2b, shows the network patterns of the authors that have at least one TRR article as author or co-author. The complexity

of this network indicates the massive number of nodes and links between the authors. The research team used Gephi 0.9.2 to create network plots to explore the network of co-authors. First, the research team used the R software to create a GDF file, with link-in and link-out counts as an attribute. Then the GDF file was imported to Gephi. The research team then prepared a network visualization using the ForceAtlas algorithm, which will make the nodes with similar connection grouped. The node sizes are proportional to link-in counts and color by different nodes. Imad Al-Qadi and Darcy Bullock are the two authors with the highest number of co-author connections (127 and 123, respectively). The research team also created an interactive web tool that allows the users to explore the co-authorship network interactively *(52)*.



(a) Heatmap of prolific TRR authors (https://rpubs.com/subasish/507543)



(b) Coauthorship network (http://subasish.github.io/pages/gephi_html/TRR_C/network/)

Figure 2 Prolific authors and co-authorship network.

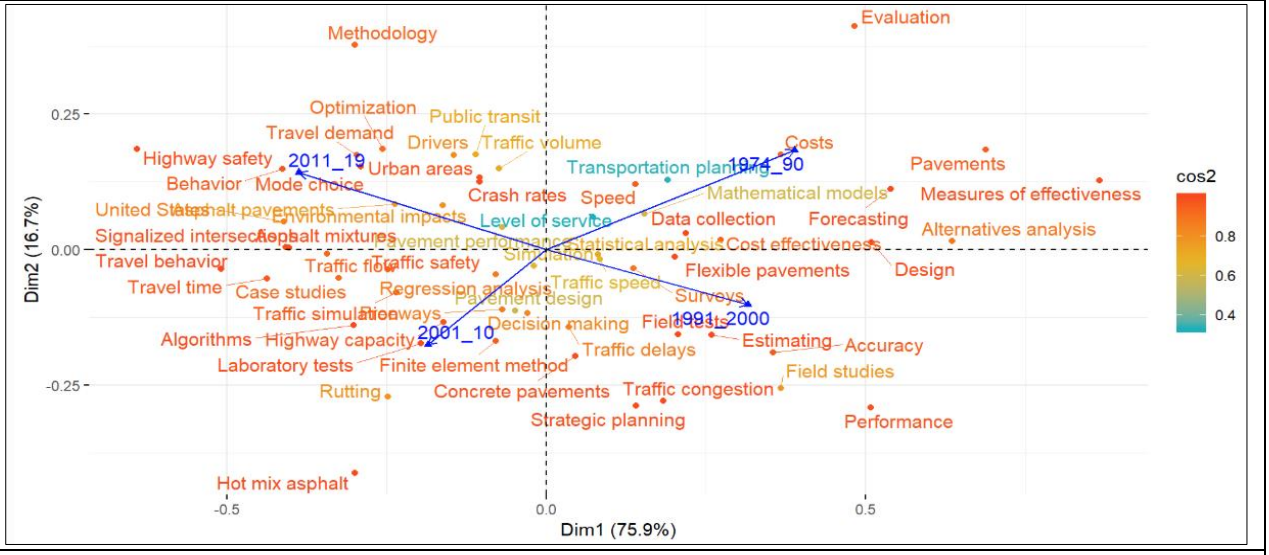
Multiple Correspondence Analysis (MCA) of Top Keywords

Multiple Correspondence Analysis, a data mining method known for dimension reduction, does not distinguish between explanatory variables and the response variable but requires the construction of a matrix based on pairwise cross-tabulation of each variable *(53).* By considering *P* to be the number of attributes (in this case: 'keywords') and *I* as the number of transactions (as rows). This will generate a matrix of *'I x P.'* If L_p is the number of categories for variable *p*, the total number of categories for all variables can be defined as $L = \sum_{p=1}^{p} L_p$. In the new matrix *'I x L,'* each of the keywords will contain several columns to show all of their possible values. The cloud of categories is considered as a weighted combination of *J* points. Category *j* is represented by a point denoted by C^j with the weight of n_j. For each of the variables, the sum of the weights of category points is *n*. In this way, for the whole set *J* the sum is nP. The relative weight w_j for point C^j is $w_j = n_j/(nP) = f_j/P$. Interested readers can consult other studies for a comprehensive theory of MCA *(54, 55).*

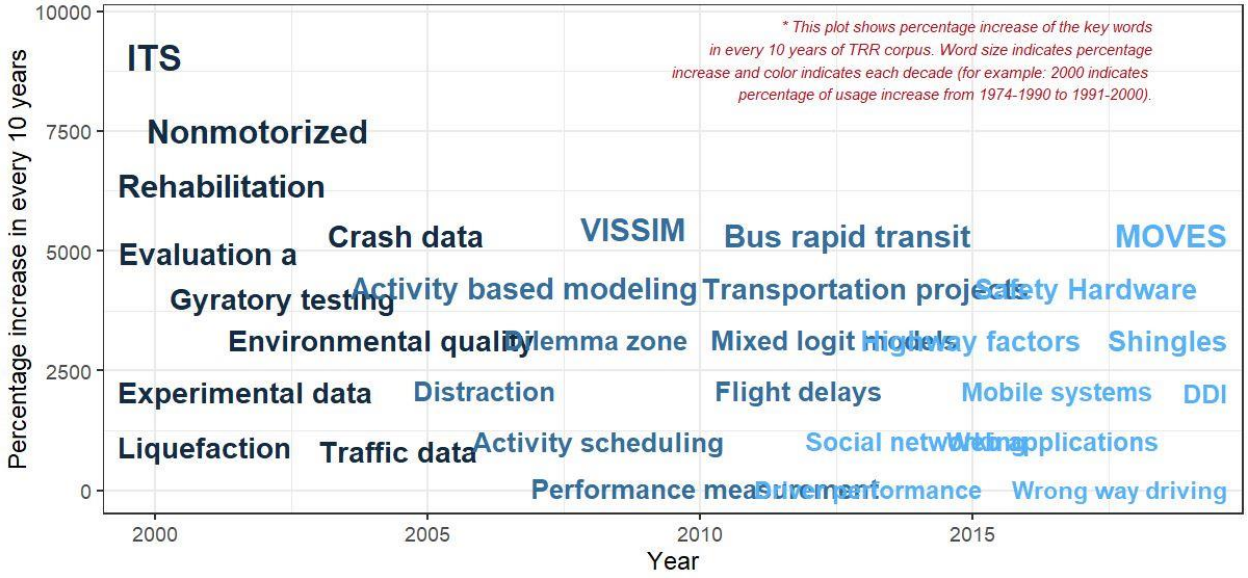
To define different clusters by the attributes, MCA generates several parameters. An example of parameters is the two-dimensional coordinates of the attributes, which indicate the clustering patterns of the attributes. Due to the overlapping in coordinates with the use of a large set of attributes, biplot is sometimes limited in visualization capacity. The parameter squared cosine or cos2 indicates the quality of the representation that measures the degree of association between attributes and an axis. If the attribute is well represented by two dimensions, the sum of the cos2 will be approximately one. Attributes with larger cos2 values contribute the most to the

definition of the dimensions. The publication years are grouped into four categories (by decade) for easy interpretation. The location of the years indicates a clockwise rotation (Figure 3a). Four different clusters of keywords have been developed based on their coordinates (cluster 1: upper right, cluster 2: upper left, cluster 3: lower left, cluster 4: lower right).

Table 2 lists the parameters developed by the top 100 keywords. Figure 3b shows the percentage increase of the keywords in the four-decade groups. Word size indicates the percentage over the decade groups, and color represents the decade groups. The higher percentages of some of the terms indicate that exponential growth occurred during these time periods. For example, the keywords 'social network' shows around a 1000% increase during 2010-2019 compared to 2001-2009.



(a) Distribution of keywords by year (MCA plot)



(b) Variations of used words from 1974 to 2019

Figure 3 Word distribution over the years.

Table 2 MCA Measures for the Top Hundred Keywords

Key Words	Rank	Coordinates		Contributions		cos^2		Cluster
		Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2	
Evaluation	57	0.4837	0.4117	2.3821	7.8198	0.5538	0.4013	Cluster 1
Pavements	11	0.6888	0.1840	9.5413	3.0833	0.8963	0.0639	Cluster 1
Costs	16	0.3676	0.1748	2.5501	2.6123	0.7517	0.1700	Cluster 1
Transportation planning	9	0.1905	0.1279	0.7646	1.5613	0.2301	0.1037	Cluster 1
Measures of effectiveness	59	0.8672	0.1272	7.2738	0.7090	0.9712	0.0209	Cluster 1
Speed	54	0.1401	0.1203	0.2046	0.6828	0.5503	0.4054	Cluster 1
Forecasting	4	0.5395	0.1113	7.4870	1.4431	0.9581	0.0408	Cluster 1
Mathematical models	1	0.1543	0.0654	0.9596	0.7812	0.5132	0.0922	Cluster 1
Level of service	43	0.0730	0.0598	0.0592	0.1799	0.1966	0.1319	Cluster 1
Data collection	3	0.2195	0.0298	1.3648	0.1141	0.9819	0.0181	Cluster 1
Cost effectiveness	40	0.2738	0.0172	0.8562	0.0154	0.9322	0.0037	Cluster 1
Alternatives analysis	12	0.6357	0.0150	8.0645	0.0202	0.8881	0.0005	Cluster 1
Design	29	0.5094	0.0135	3.9053	0.0124	0.9967	0.0007	Cluster 1
Methodology	38	-0.2995	0.3766	1.0367	7.4241	0.3626	0.5732	Cluster 2
Optimization	21	-0.2560	0.1848	1.1240	2.6528	0.6034	0.3144	Cluster 2
Highway safety	15	-0.6401	0.1848	7.8393	2.9586	0.9180	0.0765	Cluster 2
Public transit	6	-0.1094	0.1748	0.2946	3.4049	0.2018	0.5150	Cluster 2
Drivers	24	-0.1443	0.1738	0.3458	2.2752	0.3293	0.4783	Cluster 2
Travel demand	19	-0.2963	0.1732	1.5809	2.4489	0.6988	0.2390	Cluster 2
Mode choice	47	-0.2899	0.1519	0.9231	1.1492	0.7624	0.2095	Cluster 2
Traffic volume	41	-0.0732	0.1494	0.0610	1.1501	0.1391	0.5785	Cluster 2
Behavior	31	-0.4127	0.1480	2.4630	1.4348	0.8548	0.1099	Cluster 2
Urban areas	23	-0.1033	0.1328	0.1785	1.3347	0.3707	0.6117	Cluster 2
Crash rates	60	-0.1036	0.1249	0.1035	0.6813	0.4071	0.5917	Cluster 2
Asphalt pavements	35	-0.2368	0.0829	0.6820	0.3788	0.6774	0.0831	Cluster 2
Environmental impacts	58	-0.1616	0.0809	0.2588	0.2935	0.6343	0.1588	Cluster 2
United States	51	-0.4106	0.0505	1.7794	0.1217	0.8472	0.0128	Cluster 2
Pavement performance	8	-0.0686	0.0407	0.1022	0.1631	0.4506	0.1587	Cluster 2
Signalized intersections	32	-0.4079	0.0035	2.2652	0.0008	0.9846	0.0001	Cluster 2
Asphalt mixtures	36	-0.4024	0.0032	1.9222	0.0005	0.9899	0.0001	Cluster 2
Traffic flow	20	-0.3425	-0.0077	2.0272	0.0047	0.9351	0.0005	Cluster 3
Simulation	7	-0.0188	-0.0303	0.0079	0.0931	0.1810	0.4715	Cluster 3
Travel behavior	22	-0.5087	-0.0364	4.3608	0.1010	0.9949	0.0051	Cluster 3
Traffic safety	39	-0.2472	-0.0375	0.7001	0.0729	0.9531	0.0219	Cluster 3
Regression analysis	49	-0.0783	-0.0460	0.0672	0.1049	0.6459	0.2228	Cluster 3
Case studies	2	-0.3244	-0.0526	3.7134	0.4426	0.8960	0.0236	Cluster 3
Travel time	5	-0.4370	-0.0540	4.7353	0.3275	0.9601	0.0147	Cluster 3
Traffic simulation	37	-0.2344	-0.0799	0.6393	0.3361	0.8652	0.1004	Cluster 3

Freeways	25	-0.0688	-0.1112	0.0760	0.9010	0.2401	0.6279	Cluster 3
Pavement design	45	-0.0480	-0.1130	0.0255	0.6389	0.0910	0.5041	Cluster 3
Decision making	18	-0.0292	-0.1179	0.0155	1.1482	0.0480	0.7845	Cluster 3
Highway capacity	44	-0.1603	-0.1345	0.2849	0.9087	0.5718	0.4026	Cluster 3
Algorithms	17	-0.3012	-0.1403	1.7033	1.6734	0.8211	0.1781	Cluster 3
Finite element method	48	-0.0780	-0.1687	0.0668	1.4146	0.1687	0.7887	Cluster 3
Laboratory tests	13	-0.1957	-0.1730	0.7334	2.5963	0.5577	0.4358	Cluster 3
Rutting	55	-0.2470	-0.2710	0.6332	3.4539	0.3582	0.4312	Cluster 3
Hot mix asphalt	53	-0.2995	-0.4117	0.9381	8.0284	0.3459	0.6535	Cluster 3
Statistical analysis	46	0.0816	-0.0090	0.0735	0.0041	0.7133	0.0087	Cluster 4
Flexible pavements	50	0.2018	-0.0134	0.4379	0.0088	0.9954	0.0044	Cluster 4
Traffic speed	56	0.0847	-0.0184	0.0732	0.0156	0.6179	0.0291	Cluster 4
Surveys	10	0.1371	-0.0353	0.3921	0.1177	0.8461	0.0561	Cluster 4
Traffic delays	33	0.0357	-0.1429	0.0166	1.2114	0.0509	0.8178	Cluster 4
Field tests	30	0.2070	-0.1562	0.6414	1.6550	0.6371	0.3629	Cluster 4
Estimating	42	0.2598	-0.1581	0.7653	1.2846	0.7276	0.2696	Cluster 4
Accuracy	28	0.3553	-0.1902	1.9050	2.4720	0.7526	0.2156	Cluster 4
Concrete pavements	52	0.0457	-0.1965	0.0219	1.8367	0.0507	0.9382	Cluster 4
Field studies	34	0.3684	-0.2564	1.7067	3.7463	0.5282	0.2559	Cluster 4
Traffic congestion	14	0.1839	-0.2793	0.6475	6.7676	0.3011	0.6945	Cluster 4
Strategic planning	26	0.1416	-0.2882	0.3152	5.9140	0.1932	0.7999	Cluster 4
Performance	27	0.5082	-0.2914	3.9024	5.8118	0.7405	0.2434	Cluster 4

TOPIC MODELING RESULTS AND DISCUSSIONS

Structural Topic Model

The research team performed the analysis by using open source 'R' package structural topic model, 'stm,' *(44)* and topic model, 'tm' *(56)*. The data import process will produce documents, vocabulary, and metadata that STM incorporates into the topic modeling framework.

Metadata covariates for topical prevalence allow the observed metadata to affect the frequency with which a topic is discussed. The model is set to run for a maximum of 150 iterations. The convergence of the model will typically be monitored by the change in the approximate bound between EM iterations. The current model is converged after 46 iterations.

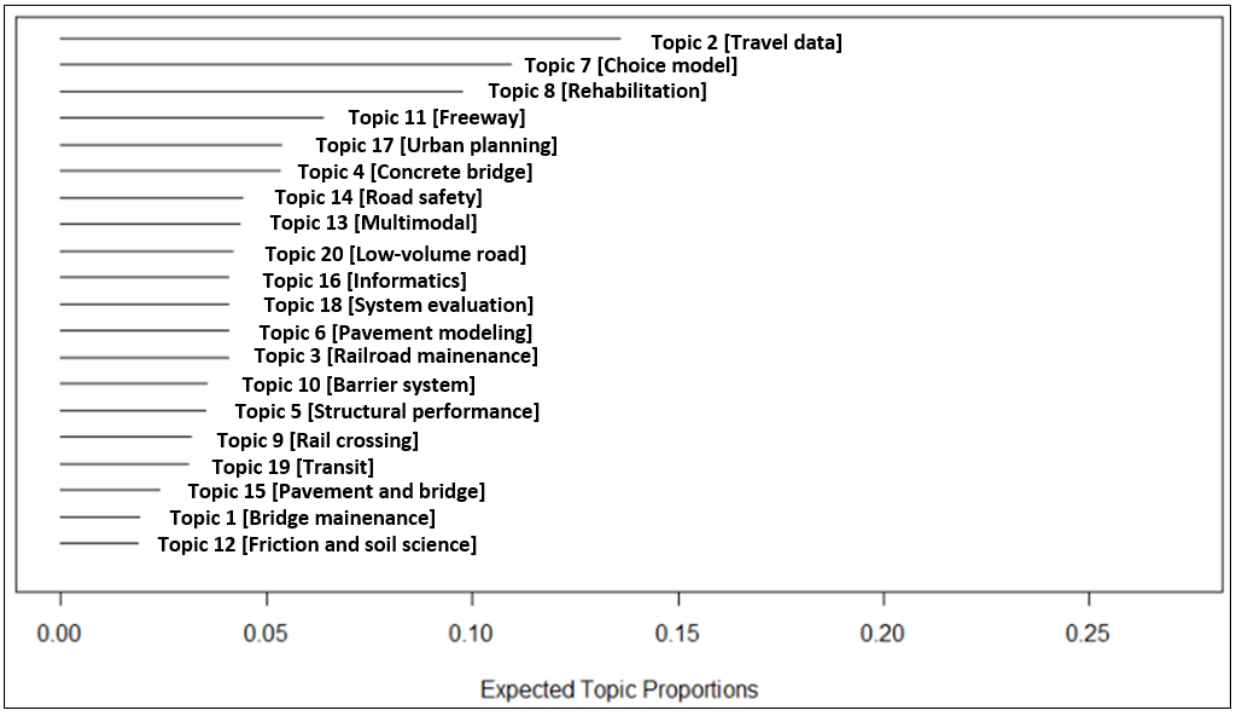


Figure 4 illustrates the corpus level visualization of the top topics from a 20- topic model. It shows the expected proportion of the corpus that belongs to each topic. High-frequency topics include Topic 2 (travel data), Topic 7 (choice model), Topic 8 (rehabilitation), Topic 11 (freeway), and Topic 17 (urban planning).

Table 3 lists the top 20 topics (with the top four words in each topic) based on the highest probability (Prob) and frequency-exclusivity (FREX) measures. These measures are developed to identify terms that define a topic. 'Prob' infers the probability that a term occurs in the topic. The other measure 'FREX' considers two criteria: 1) determining how often a term occurs in each topic, and 2) developing adjustment based on the degree to which the term is exclusive to that topic. Table 3 shows words identified by their 'Prob' and 'FREX' values as important for each topic.

Table 3 Top 20 Topics Based on Highest Probability and FREX

Topic 1 [Bridge maintenance] Top Words	**Topic 11 [Freeway] Top Words**
Highest Probability[1]: barrier, bridge, concrete, system, highways, maintenance, performance	Highest Prob: system, evaluation, performance, information, freeway, management
FREX[2]: barrier, lateral, needs, experimental, hazardous, improving, high-speed	FREX: terminal, congestion, capacity, aggregate, freeway, speed
Topic 2 [Travel data] Top Words	**Topic 12 [Friction and soil science] Top Words**
Highest Probability: data, travel, performance, modeling, vehicle, driving, evaluation	Highest Prob: utility, performance, soil, track, rail, soils, areas
FREX: driving, reliability, estimation, pedestrian, modeling, data, bicycle	FREX: utility, track, soil, soils, rapid, areas, clay, microcomputer, friction
Topic 3 [Railroad maintenance] Top Words	**Topic 13 [Multimodal] Top Words**
Highest Probability: maintenance, evaluation, recent, bus, rail, railroad, program	Highest Prob: data, modeling, vehicle, performance, effects, design, management
FREX: recent, railroad, maintenance, computer, effectiveness, noise, bus	FREX: longitudinal, modeling, multimodal, speed, prediction, cracking, adaptive
Topic 4 [Concrete bridge] Top Words	**Topic 14 [Road safety] Top Words**
Highest Prob: evaluation, urban, design, bridge, system, stresses, concrete	Highest Prob: evaluation, car, performance, urban, vehicle, crash, system
FREX: stresses, small, urban, bicycle, public, energy, accidents	FREX: road, experience, car, comparative, needs, predicting, safety
Topic 5 [Structura performance] Top Words	**Topic 15 [Pavement and bridge] Top Words**
Highest Prob: system, vehicle, evaluation, concrete, stiffness, performance, vehicles	Highest Prob: concrete, related, design, bridge, vehicle, management, properties
FREX: stiffness, automated, vehicles, crash, vehicle, railway, factors	FREX: related, accident, deformation, properties, fatigue, procedures, strategic
Topic 6 [Pavement modeling] Top Words	**Topic 16 [Informatics] Top Words**
Highest Prob: design, concrete, modeling, information, evaluation, system, management	Highest Prob: system, performance, information, data, concrete, specification, vehicle
FREX: guardrail, area, efficiency, modeling, intermodal, information, binders	FREX: specification, real, information, patterns, dynamic, validation
Topic 7 [Choice model] Top Words	**Topic 17 [Urban planning] Top Words**
Highest Prob: travel, models, data, design, system, network, choice	Highest Prob: automobile, travel, design, evaluation, urban, quality, models
FREX: hot-mix, choice, network, toll, pricing, motor, intersections	FREX: automobile, quality, considerations, demand, air, project, freeway
Topic 8 [Rehabilitation] Top Words	**Topic 18 [System evaluation] Top Words**
Highest Prob: evaluation, concrete, design, closure, management, discussion, pavements	Highest Prob: concrete, design, bridge, system, rail, used, evaluation
FREX: closure, discussion, flexible, rehabilitation, reinforced, testing, operations	FREX: used, light, cement, weight, deflectometer, statistical, conditions
Topic 9 [Rail crossing] Top Words	**Topic 19 [Transit] Top Words**
Highest Prob: behavior, rail, data, transition, application, signalized, evaluation	Highest Prob: design, concept, performance, system, evaluation, bus, management
FREX: transition, signalized, generation, accuracy, routing, tests	FREX: concept, trucks, large, structures, pressure, bridges, change
Topic 10 [Barrier system] Top Words	**Topic 20 [Low-volume road] Top Words**
Highest Prob: barriers, system, evaluation, travel, potential, service, design	Highest Prob: roads, evaluation, low-volume, models, sign, system, performance
FREX: barriers, zones, potential, empirical, proposed, economic, improved	FREX: roads, low-volume, sign, traffic, measurement, models, impacts

Note: [1]Highest Probability is the group of words within each topic with the highest probability.
[2]FREX determines the frequency (harmonic mean of rank by probability within the topic) and exclusivity (rank by the distribution of topic given word) of the words by identifying words that distinguish topics.

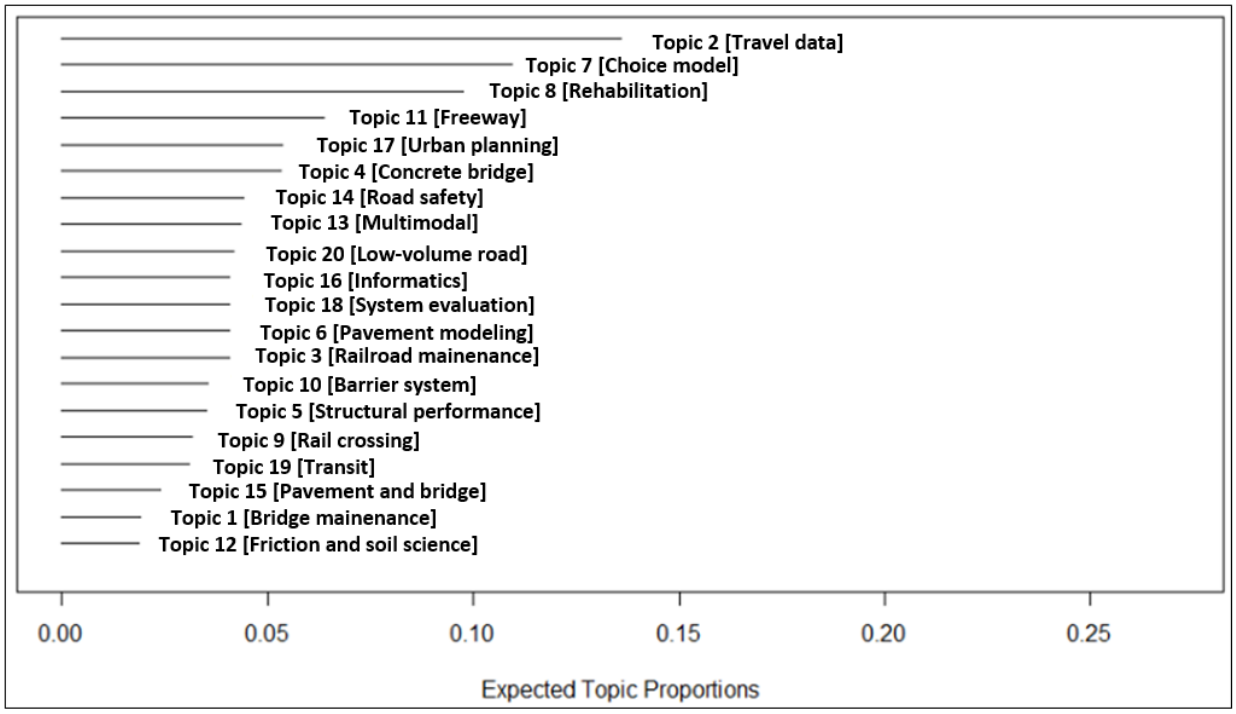


Figure 4 Top 20 topics.

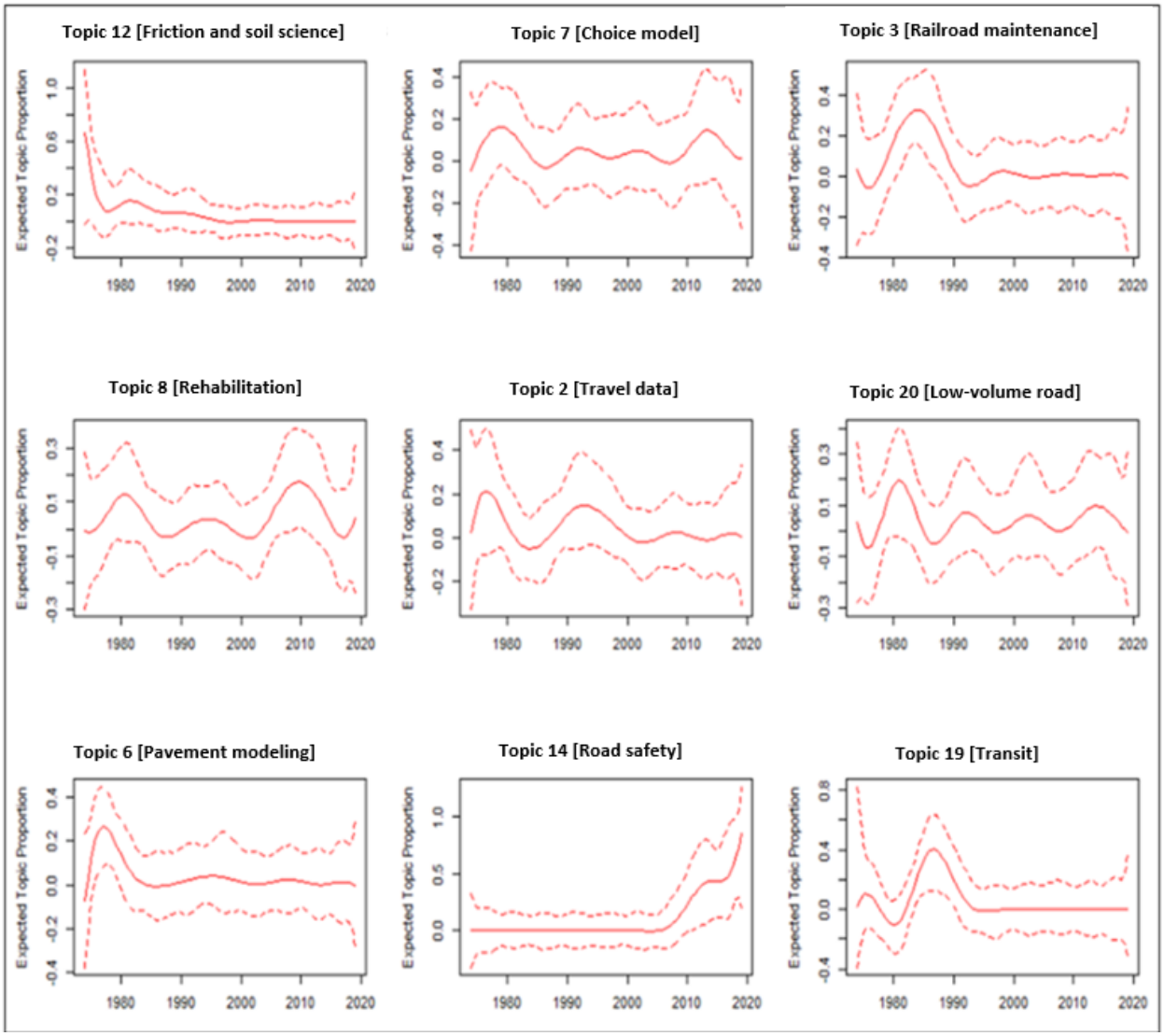


Figure 5 shows the distribution of expected topic proportions by years. Nine topics have been randomly selected to show the trend over the years. From 1970 to 2019, one topic showed an overall upward trend; the topic included the keywords in Topic 14 (*"crash," "based," "driving," and "empirical").* Another topic (Topic 12), with the words *"clay," "microcomputer," "simulation," and "discuss,"* showed a sharp decline in expected topic proportion after 1970 and then remained at consistently low value from 1980 to 2019. Two topics showed a peak increase from about 1980 to 1990, before decreasing to a value of approximately zero. One of these topics contained the keywords *"rehabilitation," "truck," "closure," and "road"* (words in Topic 19); the other topic contained the keywords *"railroad," "closure," "space," and "discuss"* (words in Topic 13). Another topic showed a sharp increase from about 1970 to 1975 before decreasing to an approximate value of zero; this topic contained the keywords *"aggregate," "air," and "small"* (words in Topic 6). The other remaining topics generally remained consistent throughout the years, with minor fluctuations over time.

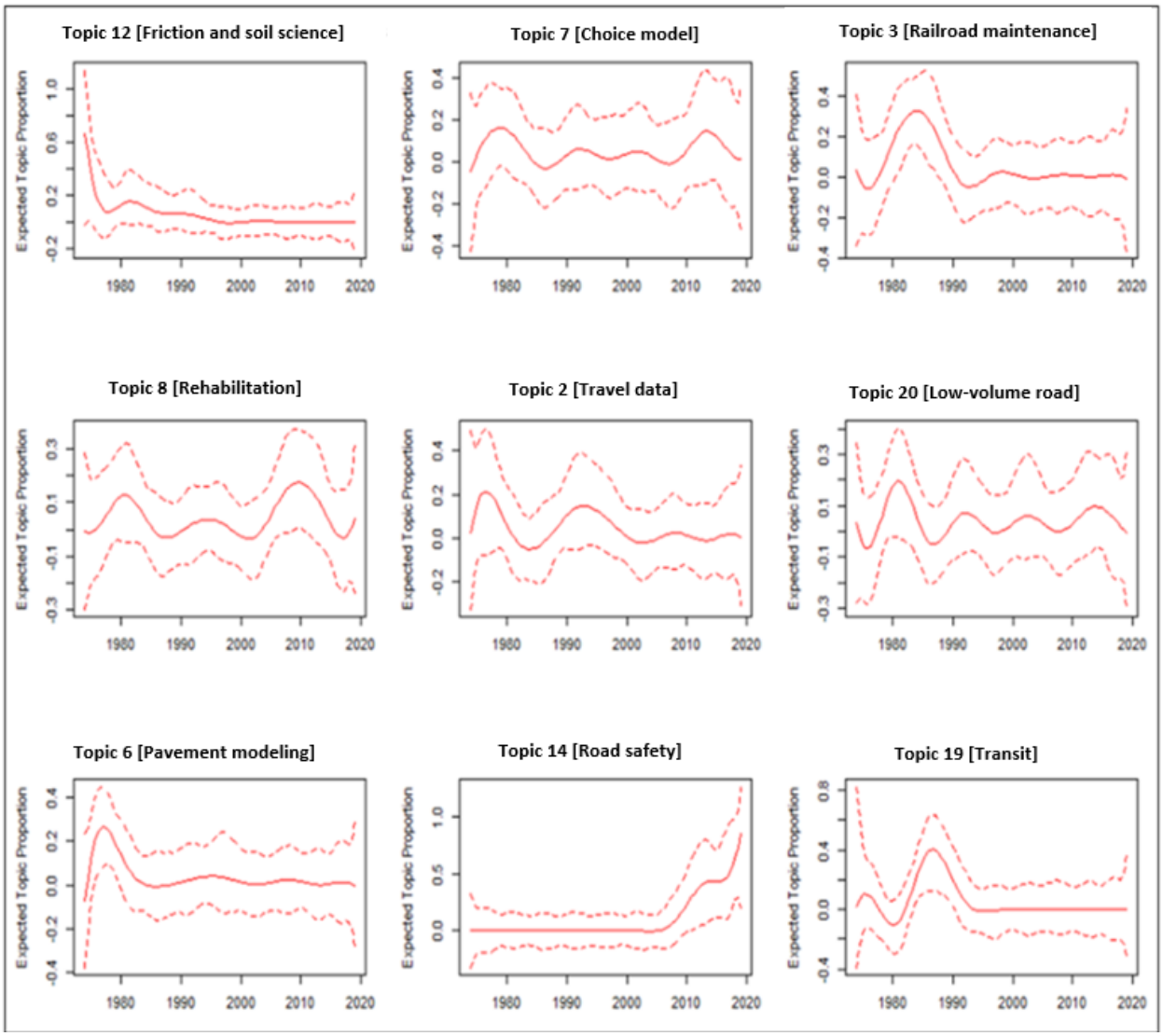


Figure 5 Expected topic proportions by year (dotted lines indicate 95% confidence intervale values).

Visualizations of LDA Models

By using metadata, STM functions explain the trends over the years. As the current study is based on large textual contents (i.e., the titles have a bag of approximately 323,000 words, and the abstracts have a bag of approximately six million words), there is a need to develop an interactive and comprehensive topic model. Recently, the LDA model has been used primarily to visualize the output of topic models fit. However, the high dimensionality of the fitted model produces challenges in creating these visualizations. LDA is normally applied to thousands of documents, representing combinations of dozens to hundreds of topics, which are modeled as distributions across thousands of terms. To mitigate these challenges, interactivity is the best technique to create LDA visualizations. Interactivity is a basic technique that is both compact and thorough. In this study, the LDAvis package was employed to develop interactive LDA models (*57*). Figure 6(a and b) presents an interactive visualization of LDA topic models. The research team developed two web tools to demonstrate these interactive plots *(58, 59).* The plots are comprised of two sections:

- The left section of the graphics represents a global perspective on the topic model. The topics are plotted as circles in a two-dimensional biplot. The locations of the topics are based on the measures of principal component analysis (PCA). By measuring the distance between topics and projecting the inter-topic distances onto two dimensions using multidimensional scaling, the centers of these circles are placed in the visualization. The overall prevalence of each topic is then encoded using the areas of the circles to allow sorting the topics in decreasing order of prevalence *(56).*
- The right section of the graphics displays a bar chart (keywords are shown horizontally). The bars represent the individual terms that are the most useful for interpreting the topics on the left, based on which topic is currently selected. This allows users to comprehend the meaning of each topic. The overlaid bars in the plot represent both the corpus-wide frequency of a given term and the topic-specific frequency of the term.
- Both sections of this visualization are connected. When the user selects a topic (on the left), the bar plot on the right highlights the most useful terms (on the right) for interpreting the selected topic. Additionally, selecting a term from the bar plot reveals the conditional distribution over topics in the biplot for the selected term. This allows users to efficiently examine many topic-term relationships.

The findings of the paper are as follows:

- The co-author network is very complex; it indicates that transportation research co-authorship is multi-disciplinary and broad.
- In topic proportions, research topics are diversified; however, travel demand-related studies showed higher topic proportions than other topics.
- Top 20 topic groups provide high-frequency words based on two scores. The cluster of words in each topic group infers the higher presence of these keywords in each group.
- The interactive visualization of the LDA topic models indicates that the topics developed from the journal titles are distinct when compared with topic models developed from journal abstracts.

CONCLUSIONS

In the fields of engineering and science, transportation is a key research area. Throughout the world, mining big data for potential trends and patterns has become an increasingly popular research topic. However, there has been a lack of research conducted in the field of transportation engineering to mine the data. The challenges and problems encountered in transportation research have constantly changed over time. Additionally, the transportation research scope has become more diverse with expanding and inter-disciplinary coverage of topics. As a result, there has been an outbreak of transportation research publications within the last decade. In the present study, the research team performed topic modeling on text containing approximately six million words from the peer-reviewed abstracts and titles of 30,784 published TRR articles, dating back to 1974. To identify the research patterns from that period, this study applied two popular topic models, STM and LDA. This study also identified the top 20 topics that produced the highest word frequency measured by two scores: FREX and high probability. To explore more relevant patterns in the broad fields of transportation research, this study presents a unique tool to probe present content and prevalence to develop a disaggregated level correlation. In addition, this study produced two topic model interactive tools cultivated separately for TRR paper abstracts and titles. These specific methods have not yet been applied

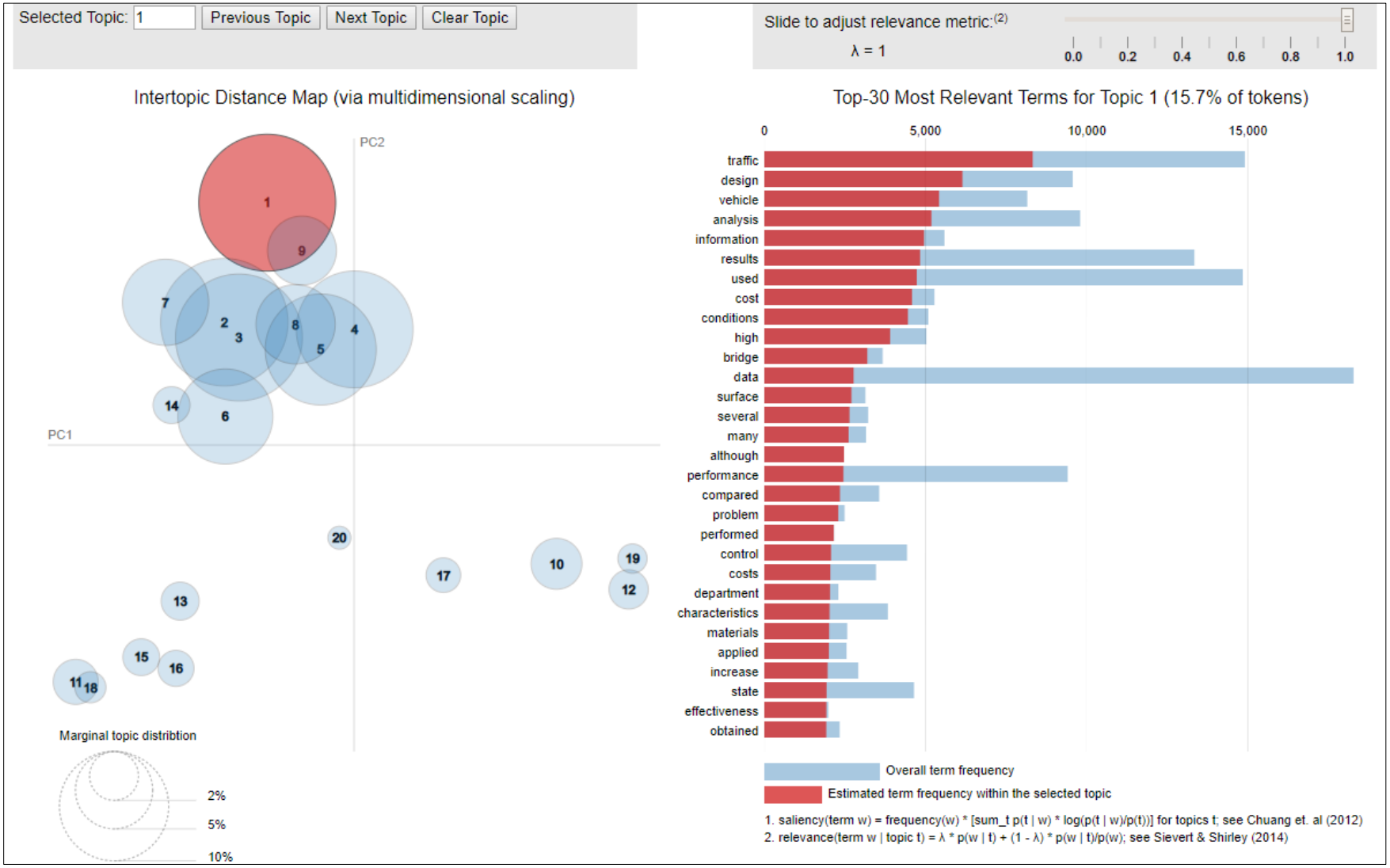


Figure 6a Interactive LDAvis tool for TRR abstracts (http://subasish.github.io/pages/trr_abstract/).

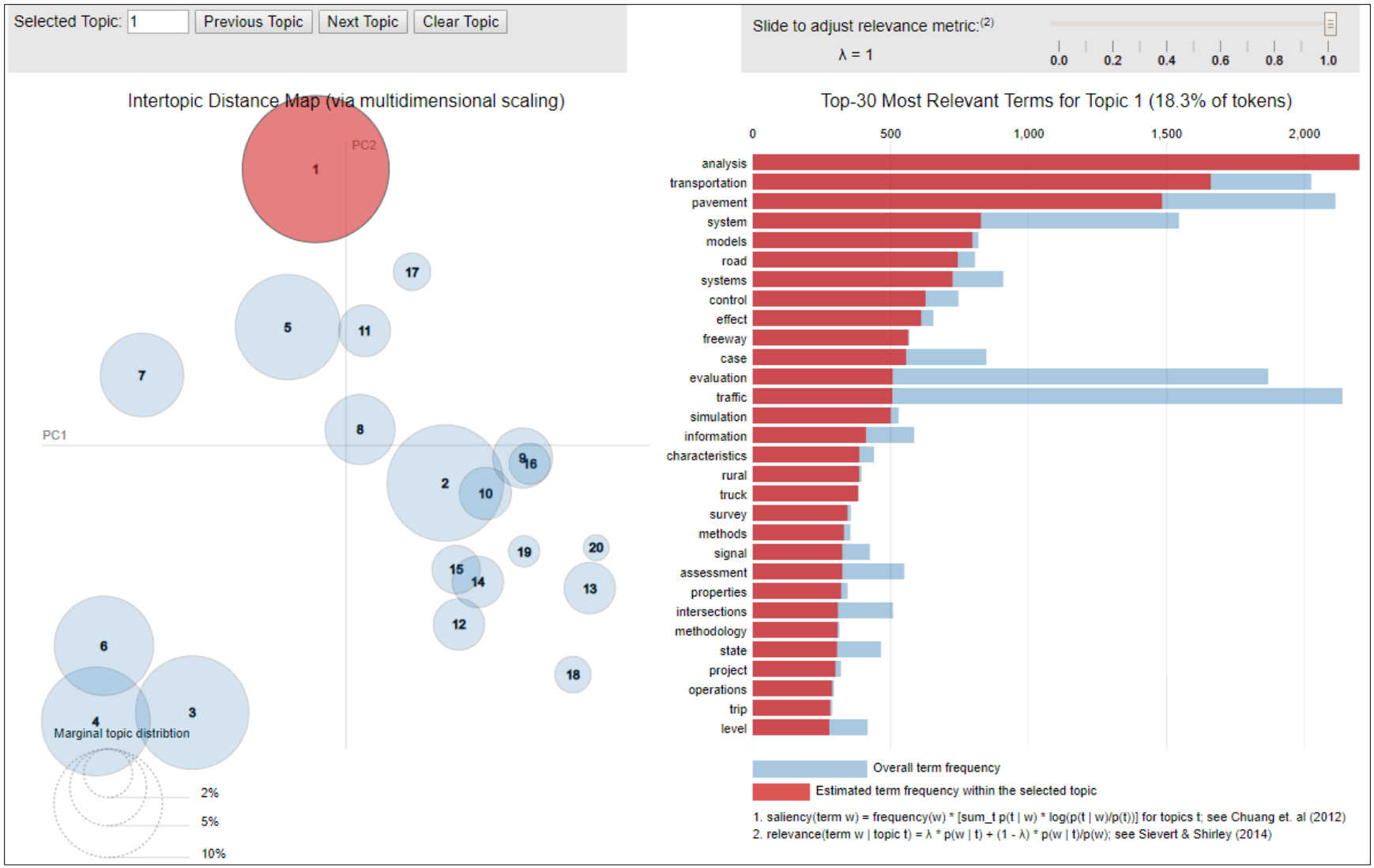


Figure 6b Interactive LDAvis tool for TRR titles (http://subasish.github.io/pages/trr_title/).

to the identification of the research trends from TRR articles. However, the present study demonstrates how STM, LDA, and other similar methods could be utilized to offer the potential of natural language processing works in transportation research. Future research can improve the natural language processing methods used in the study by incorporating additional databases, such as state DOT reports and other national reports from top transportation journals.

This study identified topics that were both meaningful and representative; they mostly corresponded to established sub-fields in the transportation research field. The identified fields unearth a landscape for further transportation research. This methodology is also suitable for other areas related to transportation engineering. The framework established in this study can be used in other studies within the domain of natural language processing.

ACKNOWLEDGMENT

The authors appreciate the assistance provided by the students on this project: Bita Maraghehpour, and Ly-Na Tran.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm the contribution to the paper as follows: study conception and design: Subasish Das; data collection: Subasish Das and Anandi Dutta; analysis and interpretation of results: Subasish Das and Anandi Dutta; draft manuscript preparation: Subasish Das, Anandi Dutta, and Marcus Brewer. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Blei, D. Probabilistic topic models, *Communications of the ACM*, Vol. 55, No. 4, 2012, pp. 77-84.
2. 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.
3. Hofmann, T. Probabilistic Latent Semantic Indexing, *ACM SIGIR*, 1999.
4. Blei, D., A. Ng and M. Jordan. Latent Dirichlet Allocation, *The Journal of machine Learning research*, Vol. 3, 2003, pp. 993–1022.
5. Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang and D. M. Blei. Reading tea leaves: How humans interpret topic models, *NIPS*, Vol. 22, 2009, pp. 288–296.
6. Mimno, D., H. Wallach, E. Talley, M. Leenders and A. McCallum. Optimizing Semantic Coherence in Topic Models, *EMNLP*, 2011.
7. Andrzejewski D. and X. Zhu. Latent Dirichlet allocation with topic-in-set knowledge, *NAACL HLT*, 2009.
8. Andrzejewski, D., X. Zhu and M. Craven. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors, *ACM*, 2009.
9. Andrzejewski, D., X. Zhu, M. Craven and B. Recht. A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation Using First-Order Logic, 2011.
10. Chemudugunta, C., A. Holloway, P. Smyth and M. Steyvers. Modeling Documents by Combining Semantic Concepts with Unsupervised Statistical Learning, *The semantic Web-ISWC*, 2008, pp. 229–244.
11. Chen, Z., A. Mukherjee, B. Liu, M. Hsu, M. Castellanos and R. Ghosh. Exploiting Domain Knowledge in Aspect Extraction, 2013.

12. Chen, Z., A. Mukherjee, B. Liu, M. Hsu, M. Castellanos and R. Ghosh. Leveraging Multi-Domain Prior Knowledge in Topic Models, 2013.
13. Doshi-Velez F., B. Wallace and R. Adams. Graph-Sparse lda: A Topic Model with Structured Sparsity, 2015.
14. Yao, L., Y. Zhang, B. Wei, H. Qian and Y. Wang. Incorporating Probabilistic Knowledge into Topic Models, *PAKDD*, 2015, pp. 586–597.
15. Blei, D. and J. Lafferty. Dynamic Topic Models, *ACM*, 2006.
16. Kalyanam, J., A. Mantrach, D. Saez-Trumper, H. Vahabi and G. Lanckriet. Leveraging Social Context for Modeling Topic Evolution, *ACM SIGKDD*, 2015.
17. Wang, X. and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends, *ACM SIGKDD*, 2006.
18. Wei, X., J. Sun and X. Wang. Dynamic mixture models for multiple time-series, *Ijcai*, Vol. 7, 2007, pp. 2909–2914.
19. Yan, X., J. Guo, Y. Lan, J. Xu and X. Cheng. A probabilistic model for bursty topic discovery in microblogs, *AAAI*, 2015.
20. Eisenstein, J., A. Ahmed and E. Xing. Sparse additive generative models of text, *ICML*, 2011, pp. 1041–1048.
21. McLaurin, E., A. D. McDonald, J. D. Lee, N. Aksan, J. Dawson, J. Tippin, M. Rizzo. Variations on a Theme: Topic Modeling of Naturalistic Driving Data, 2014. https://journals-sagepub-com.srv-proxy2.library.tamu.edu/doi/abs/10.1177/1541931214581443. Accessed Jul. 20, 2019.
22. Sun, X., N. H. C. Yung, and E. Y. Lam. Unsupervised Tracking with the Doubly Stochastic Dirichlet Process Mixture Model. IEEE Transactions on Intelligent Transportation Systems, Vol. 17, No. 9, 2016, pp. 2594–2599. https://doi.org/10.1109/TITS.2016.2518212.
23. Sun, L., and Y. Yin. Discovering Themes and Trends in Transportation Research Using Topic Modeling. Transportation Research Part C: Emerging Technologies, Vol. 77, 2017, pp. 49–66.
24. Venkatraman, V., Y. Liang, E. McLaurin, W. Horrey, and M. Lesch. Exploring Driver Responses to Unexpected and Expected Events Using Probabilistic Topic Models. *Driving Assessment Conference*, 2017, pp. 375–381.
25. 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*, Vol. 2552, No. 1, 2016, pp. 48–56.
26. Das, S., K. Dixon, X. Sun, A. Dutta, and M. Zupancich. Trends in Transportation Research: Exploring Content Analysis in Topics. Transportation Research Record: Journal of the Transportation Research Board, No. 2614, 2017, pp 27-38.
27. Das, S. #TRBAM: Social Media Interactions from the Largest Transportation Conference. TR News, 2019, November-December Issue, pp. 18-23.
28. Das, S., A. Dutta, T. Lindheimer, M. Jalayer, and Z. Elgart. YouTube as a Source of Information in Understanding Autonomous Vehicle Consumers: Natural Language Processing Study. Transportation Research Record: Journal of the Transportation Research Board, 2019, p. 12p.
29. Das, S., A. Dutta, G. Medina, L. Minjares-Kyle, and Z. Elgart. Extracting Patterns from Twitter to Promote Biking. IATSS Research, Vol. 43, No. 1, 2019, p. pp 51-59.

30. Boyer, R., W. Scherer, and M. Smith. Trends Over Two Decades of Transportation Research: A Machine Learning Approach. Transportation Research Record: Journal of the Transportation Research Board. Issue: 2614, pp. 1-7, 2017.
31. Hong, J., R. Tamakloe, G. Lee, and D. Park. Insight from Scientific Study in Logistics using Text Mining. Transportation Research Record: Journal of the Transportation Research Board. Issue: 4, pp. 97-107, 2019.
32. Biehl, A., Y. Chen, K. Sanabria-Veaz, D. Uttal, and A. Stathopoulos. Where Does Active Travel Fit within Local Community Narratives of Mobility Space and Place? Transportation Research Part A: Policy and Practice, Vol. 123, 2019, pp. 269–287.
33. Rosen-Zvi, M., T. Griffiths, M. Steyvers and P. Smyth. The author-topic model for authors and documents, UAI, 2004.
34. Qi, W., M. Quing, B. Xia, and N. An. Discovering regulatory concerns on bridge management: An author-topic model based approach. Transport Policy. Vol 75, 2019.
35. Ahmed and E. Xing. Staying informed: supervised and semi-supervised multi-view topical analysis, EMNLP, 2010, pp. 1140-1150.
36. Eisenstein, J., B. O'Connor, N. Smith and E. Xing. A latent variable model for geographic lexical variation, EMNLP, 2010, pp. 1277–1287.
37. Lee, J. D. and K. Kolodge. Understanding Attitudes Towards Self-Driving Vehicles: Quantitative Analysis of Qualitative Data, 2018. Accessed July 2019.
38. Kuhn, K. D. Using Structural Topic Modeling to Identify Latent Topics and Trends in Aviation Incident Reports. Transportation Research Part C: Emerging Technologies, Vol. 87, 2018, pp. 105–122.
39. Blei, D. M. Probabilistic models of text and images. California: University of California, 2004.
40. Blei, D. M. and M. I. Jordan. Variational inference for Dirichlet process mixtures. Bayesian Analysis, 1 (1), 2006, 121-144.
41. Girolami, M. and A. Kabán. On an equivalence between PLSI and LDA. SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003, 433-434.
42. Masada, T., S. Kiyasu, and S. Miyahara. Comparing LDA with PLSI as a dimensionality reduction method in document clustering. LKR'08 Proceedings of the 3rd international conference on Large-scale knowledge resources: construction and application, 2008, 13-26.
43. Kim, Y., and K. Shim. TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. Information Systems. Vol. 42, 2014, pp. 59-77.
44. Roberts M., B. Stewart and D. Tingley. stm: R Package for Structural Topic Models. 2016. http://www.structuraltopicmodel.com, Accessed: July, 2016.
45. Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian. Structural Topic Models for open-ended survey responses. American Journal of Political Science, 58(4), 2014, 1064–1082.
46. Bauer, P. C., P. Barberá, K. Ackermann, and A. Venetz. Is the left-right scale a valid measure of ideology? Political Behavior, 39(3), 2017, 553–583.
47. Farrell, J. Corporate funding and ideological polarization about climate change. Proceedings of the National Academy of Sciences, 113(1), 2016, 92–97.
48. Tingley, D. Rising power on the mind. International Organization, 71(S1), 2017, S165–S188.

49. Tvinnereim, E., and K. Fløttum. Explaining topic prevalence in answers to open ended survey questions about climate change. Nature Climate Change, 5(8), 2015, 744.
50. Blei, D. M., and J. D. Lafferty. A correlated topic model of science. Annals of Applied Statistics, 1(1), 2007, 17–35.
51. Das, S. Heatmap of the Most Prolific TRR Authors. https://rpubs.com/subasish/507543 Accessed: July 2019.
52. Das, S., and R. Wang. http://subasish.github.io/pages/gephi_html/TRR_C/network/ Accessed: July 2019.
53. Das, S., R. Avelar, K. Dixon, and X. Sun. Investigation on the Wrong Way Driving Crash Patterns Using Multiple Correspondence Analysis. Accident Analysis & Prevention, 2018. 111:43–55.
54. Das, S., and X. Sun. Exploring Clusters of Contributing Factors for Single-Vehicle Fatal Crashes Through Multiple Correspondence Analysis. Proceedings in the 93rd Transportation Research Board Annual Meeting, Washington DC, 2014.
55. Das, S., and X. Sun. Factor Association with Multiple Correspondence Analysis in Vehicle–Pedestrian Crashes. Transportation Research Record: Journal of the Transportation Research Board, 2015. 2519: 95–103.
56. Feinerer, I., K. Hornik, and D. Meyer. Text Mining Infrastructure in R. Journal of Statistical Software. 25(5), 1008, 1-54.
57. Sievert, C. and K. Shirley. LDAvis: Interactive Visualization of Topic Models. R package version 0.3.2. https://CRAN.R-project.org/package=LDAvis Accessed: July 2019.
58. Das, S. Interactive LDAvis tool for TRR abstracts. http://subasish.github.io/pages/trr_abstract/ Accessed: July 2019.
59. Das, S. Interactive LDAvis tool for TRR titles. http://subasish.github.io/pages/trr_title/ Accessed: July 2019.