# Discovering themes and trends in transportation research using topic modeling

**Working Paper** · June 2016



**2 authors:**


Lijun Sun
McGill University



Yafeng Yin
University of Michigan

# Discovering themes and trends in transportation research using topic modeling

Lijun Sun[a,*], Yafeng Yin[b]

[a]*The Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142, USA*
[b]*Department of Civil and Coastal Engineering, University of Florida, Gainesville, FL 32611, USA*

**Abstract**

Transportation research is a key area in both science and engineering. In this paper, we present an empirical analysis of 17,163 articles published in 22 leading transportation journals from 1990 to 2015. We apply a latent Dirichlet Allocation (LDA) model on article abstracts to infer 50 key topics. We show that those characterized topics are both representative and meaningful, mostly corresponding to established sub-fields in transportation research. These identified fields reveal a research landscape for transportation. Based on the results of LDA, we quantify the similarity of journals and countries/regions in terms of their aggregated topic distributions. By measuring the variation of topic distributions over time, we find some general research trends, such as topics on sustainability, travel behavior and non-motorized mobility are becoming increasingly popular over time. We also carry out this temporal analysis for each journal, observing a high degree of consistency for most journals. However, some interesting anomaly, such as special issues on particular topics, are detected from temporal dynamics as well. By quantifying the temporal trends at the country/region level, we found that countries/regions display clearly distinguishable patterns, suggesting that research communities in different regions tend to focus on different sub-fields. Our results could benefit different parties in the academic community—including researchers, journal editors and funding agencies—in terms of identifying promising research topics/projects, seeking for candidate journals for a submission, and realigning focus for journal development.

*Keywords:* transportation research, topic modeling, publication data, research policy

## 1. Introduction

With the rapid urbanization globally, transportation has become an increasingly important ingredient in the quality of life, making a major impact on human well-being. Aiming to provide better transportation systems and services, transportation research has long been a key topic in both science and engineering. This has been reflected in both the rising application of emerging technologies, the growth in interdisciplinary collaborations, and the increasing number of conferences organized, journals created and research articles published (Banister, 2014; Button, 2015).

Scientific publication is often considered a key proxy to reflect the trend of research development in both theory and practice. In terms of transportation research, the problems and challenges we encountered have been constantly changing over time, and the scope of transportation research has also become more diverse, with a widening and inter-disciplinary coverage of topics, ranging from those long lasting questions such as traffic congestion and signal control, to emerging technologies such as autonomous vehicles, connected vehicles, big data analytics, and artificial intelligence, to societal problems such as sustainability and environmental justice. The field is evolving given the specific questions raised and the advances in solutions/technologies developed. As a result, transportation research has witnessed an explosion of research publications in last decades.

There exists a great body of literature studying publication data with quantitative methods, which are often referred to as scientometrics (e.g., see Heilig and Voß (2015) for a study on public transportation). Although scientometric analysis offers a good tool to quantify the importance of articles and authors from citation data, it fails to provide topic related information for us to better understand different research context in detail. In fact, the content of scientific publications is often of more importance to study a field, in the sense that it could help us to obtain solutions to targeted problems, understand the development of particular technology, and learn the

*Corresponding author. Address: 75 Amherst Street, E14-574A, Cambridge, MA 02142, USA. Tel.: +1-6173243782.
*Email addresses:* sunlijun@mit.edu (Lijun Sun), yafeng@ufl.edu (Yafeng Yin)

motivation and creation of new ideas. The abstract of an article is the first but concise piece of content-related information we can get, since it essentially reveals the whole picture of an article from a reader's point of view. In other words, an abstract can be considered a condensed representation of an article, and it has been successfully used to identify and interpret scientific themes. For example, Griffiths and Steyvers (2004) investigated abstract data from articles published in the *Proceedings of the National Academy of Science (PNAS)* from 1991 to 2001 and compared research topics/areas obtained from topic modeling with existing categories. Blei and Lafferty (2006) applied dynamic topic models on historical literature from the journal *Science* during 1880-2000 to investigate how individual topics change over time. Gatti et al. (2015) applied topic modeling on article metadata from 20 journals in the field of operations research and management science, and quantified the generality and specificity of different journals. To the best of our knowledge, there is little work done in the field of transportation with an exception that Das et al. (2016) applied topic modeling on a sample of abstracts from papers presented at the *Transportation Research Board (TRB) Annual Meeting* and investigated topics changes from 2008 to 2014.

In this paper, we investigate research topics and their trends to gauge the evolution of transportation research from 1990 to 2015 using publication metadata obtained from 22 scientific journals. We follow a similar framework as what Gatti et al. (2015) has applied in the field of operations research and management science. The purpose of this work is to better identify, quantify and understand themes and trends in transportation research over the last 25 years, and to provide a valuable tool to researchers, journal editors, publishers and funding agencies to make more informed decisions. We also hope this work could stimulate more discussion on the state of publishing in transportation research (e.g., see a recent discussion in Button (2015)).

The remainder of this paper is organized as follows. Section 2 summarizes the notations used throughout this paper. In Section 3, we introduce the concept of topic modeling and latent Dirichlet allocation (LDA). We also present various measures to quantify topic distribution by aggregating the result at the levels of journal, country/region, and time. Section 4 introduces the article abstract data extracted from Web of Science and the software package we used for topic inference. In Section 5, we conduct extensive analysis on the extracted topic and word distributions using those defined measures. Finally, Section 6 summarizes our study and suggests some future research directions.

## 2. Notations

We use the notations listed in Table 1 throughout this paper.

## 3. Methodology

In this section, we first introduce the concept of latent Dirichlet allocation and its application in topic modeling. We follow a similar analytical framework and use similar measures as the work of Gatti et al. (2015), which focuses on the field of operations research and management science, to quantify the variation of topics across journals, countries/regions and time. In doing so we introduce various measures based on the posterior document-topic distribution $\theta_d$: (1) the topic composition of each journal, (2) the topic composition of each country/region, and (3) topic evolution over time for each journal or country/region. These measures are used in the analyses presented in Section 4.

### 3.1. Latent Dirichlet allocation (LDA)

LDA is a generative probabilistic model introduced by Blei et al. (2003) for the purpose of topic modeling. It is built on the classical probabilistic latent semantic analysis (pLSA) model (Hofmann, 1999) and focuses on discovering main themes from multinomial document-word observations. However, LDA itself is a general statistical model and can be applied in various domains and settings, such as finding patterns in genetic data, images, music, and social networks (see Blei (2012) for a short review). For example, in travel behavior and activity research, LDA has been used to analyze human location and activity data to discover structural daily routines (Huynh et al., 2008; Farrahi and Gatica-Perez, 2011; Hasan and Ukkusuri, 2014). As an unsupervised model, LDA does not require any prior annotations or labeling of the documents. All the topics emerge naturally from the statistical structure of document-word data itself.

Fig. 1 shows the graphical representation of LDA in plate notation. The LDA model first defines $K$ topics, with each topic $k$ associated with a distribution $\psi_k$ over words in the vocabulary. In particular, $\psi_k$ is picked from a Dirichlet distribution $\text{Dirichlet}_V(\beta)$. Based on these created topics, a document $d$ (namely a collection

Table 1: Notations of variables and parameters

| Notation | Description |
|---|---|
| indices | |
| $d$ | Index of documents |
| $k$ | Index of topics |
| $i$ | Index of words |
| $j$ | Index of journals |
| $t$ | Index of years |
| in LDA | |
| $\alpha$ | Dirichlet prior on the per-document topic distributions (hyperparameter) |
| $\beta$ | Dirichlet prior on the per-topic word distributions (hyperparameter) |
| $\theta_d$ | Topic distribution of document $d$ |
| $\theta_{dk}$ | Proportion of topic $k$ in document $d$ |
| $\psi_k$ | Word distribution of topic $k$ |
| $\psi_{kw}$ | Probability of word $w$ occurring in topic $k$ |
| $w_d$ | Word collection of document $d$ |
| $w_{di}$ | Word $i$ in $w_d$ |
| $z_{di}$ | Topic assignment for word $w_{di}$ from document $d$ |
| $K$ | Number of topics |
| $V$ | Number of words in the vocabulary |
| $M$ | Number of documents |
| $N_d$ | Number of words in document $d$ |
| $N$ | $N = \sum_{d=1}^{M} N_d$ total number of words in all documents |
| derived | |
| $t_d$ | Publication year of document $d$ |
| $j_d$ | Journal of document $d$ |
| $c_d$ | Country/region of document $d$ |
| $\theta_k^{[t]}$ | Proportion of topic $k$ at year $t$ |
| $\theta_k^{j}$ | Proportion of topic $k$ in journal $j$ |
| $\theta_k^{j[t]}$ | Proportion of topic $k$ in journal $j$ at year $t$ |
| $\theta_k^{(c)}$ | Proportion of topic $k$ from country/region $c$ |
| $\theta_k^{(c)[t]}$ | Proportion of topic $k$ from country/region $c$ at year t |

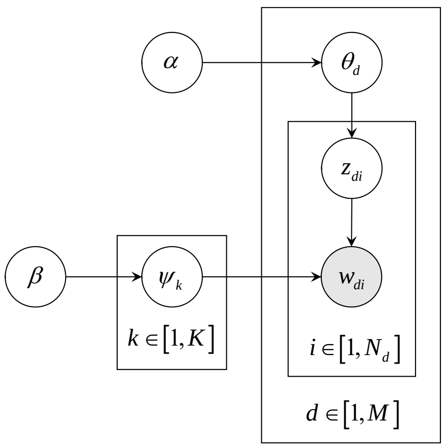



Figure 1: Graphical model representation of LDA

of words $w_d$) is generated by first sampling a distribution $\theta_d$ over $K$ topics from another Dirichlet distribution $\text{Dirichlet}_K(\alpha)$, which determines topic assignment for each word in $w_d$, and then choosing each word $w_{di}$ based on $\theta_d$. In generating each word $w_{di}$, LDA first samples a particular topic $z_{di} \in [1, K]$ from multinomial distribution $\text{Multinomial}_K(\theta_d)$, and then the word $w_{di}$ is selected from multinomial distribution $\text{Multinomial}_V(\psi_{z_{di}})$. This process can be summarized into three steps:

**Step 1:** Word distribution of each topic $k$ is determined by $\psi_k \sim \text{Dirichlet}_V(\beta)$

**Step 2:** Topic distribution for each document $d$ is determined by $\theta_d \sim \text{Dirichlet}_K(\alpha)$

**Step 3:** For each document $d$, for each word $w_{di}$ in $d$

1. Choose a topic $z_{di} \sim \text{Multinomial}_K(\theta_d)$;

2. Choose a word $w_{di} \sim \text{Multinomial}_V(\psi_{z_{di}})$.

The inference of LDA models can be done by applying the variational expectation-maximization (VEM) algorithm (Blei et al., 2003) or through Gibbs sampling (Griffiths and Steyvers, 2004). Both methods can infer the posterior of document-topic distribution $\theta$ and topic-word distribution $\psi$ efficiently. The results from the inference allow us to discover the latent thematic structure from a large collection of documents. In the meanwhile, using a trained model we can also infer topic compositions of new/unseen documents with folding-in.

### *3.2. Topic variation with journal, country/region and time*

Using the posterior document-topic distribution $\theta_d$ and article information (i.e., journal name, the location of the corresponding author's affiliation and publishing year) of each document $d$, we can analyze how each inferred topic differs across journals, country/region and varies with time. To measure these quantitatively, we define some derived terms and clustering distance measures as follows.

*Topic distribution over time*

We denote $\theta^{[t]}$ as the topic distribution at time $t$ for all articles and $\theta_k^{[t]}$ as the proportion of topic $k$ within $\theta^{[t]}$:

$$\theta_k^{[t]} = \frac{\sum_{d=1}^{M} \theta_{dk} \times \mathbb{I}(t_d = t)}{\sum_{d=1}^{M} \mathbb{I}(t_d = t)}. \tag{1}$$

*Journal topic distribution*

We denote $\theta^j$ as the topic distribution in journal $j$ and $\theta_k^j$ as the proportion of topic $k$ within $\theta^j$:

$$\theta_k^j = \frac{\sum_{d=1}^{M} \theta_{dk} \times \mathbb{I}(j_d = j)}{\sum_{d=1}^{M} \mathbb{I}(j_d = j)}, \tag{2}$$

where $\mathbb{I}(e) = 1$ if $e$ is true and 0 otherwise.

As can be seen, $\theta^j$ is the averaged topic distribution across all articles in journal $j$. The overall topic distribution $\theta^j$ can be considered a signature of journal $j$. This distribution also allows us to quantify the similarity and difference between journals by performing hierarchical clustering. We use Jensen-Shannon divergence (JSD) as a measure to quantify the difference between the signatures ($\theta^u$ and $\theta^v$) of two journals ($u$ and $v$):

$$JSD(\theta^u, \theta^v) = \frac{1}{2} KLD\left(\theta^u, \bar{\theta}\right) + \frac{1}{2} KLD\left(\theta^v, \bar{\theta}\right), \tag{3}$$

where $\bar{\theta} = \frac{1}{2}(\theta^u + \theta^v)$ and $KLD(\theta, \theta') = \sum_{k=1}^{K} \theta_k \log \frac{\theta_k}{\theta'_k}$ is the Kullback-Leibler divergence between two topic distributions $\theta$ and $\theta'$.

In measuring the distance between two journals, we use Jensen-Shannon distance, which is the square root of the Jensen-Shannon divergence as a metric (Endres and Schindelin, 2003):

$$d_{u,v}^j = \sqrt{JSD(\theta^u, \theta^v)}. \tag{4}$$

With the measured distance, we can perform hierarchical clustering by using a particular linkage method to compute distances between paired clusters.

*Journal topic distribution over time*

In order to analyze temporal topic variation within each journal, we define $\theta^{j[t]}$ as the topic distribution in journal $j$ at time $t$, and each element:

$$\theta_k^{j[t]} = \frac{\sum_{d=1}^{M} \theta_{dk} \times \mathbb{I}(t_d = t, j_d = j)}{\sum_{d=1}^{M} \mathbb{I}(t_d = t, j_d = j)}. \tag{5}$$

*Country/region topic distribution*

Similar to previous definition for the journal level, we define $\theta^{(c)}$ as topic distribution of country/region $c$, and $\theta_k^{(c)}$ as the proportion of topic $k$ in country/region $c$:

$$\theta_k^{(c)} = \frac{\sum_{d=1}^{M} \theta_{dk} \times \mathbb{I}(c_d = c)}{\sum_{d=1}^{M} \mathbb{I}(c_d = c)}. \tag{6}$$

With the definition of $\theta^{(c)}$ (signature of country/region $c$), we can also quantify topic similarity between paired countries/regions. In doing so, we define the distance between country/resion $u$ and $v$ as

$$d_{u,v}^{c} = \sqrt{JSD\left(\theta^{(u)}, \theta^{(v)}\right)}, \tag{7}$$

where $JSD$ is also computed as Eq. 3.

*Country/region topic distribution over time*

In the same way as we quantify the journal topic over time, we define $\theta_k^{(c)[t]}$ as the proportion of topic $k$ in country/region $c$ at time $t$:

$$\theta_k^{(c)[t]} = \frac{\sum_{d=1}^{M} \theta_{dk} \times \mathbb{I}(t_d = t, c_d = c)}{\sum_{d=1}^{M} \mathbb{I}(t_d = t, c_d = c)}, \tag{8}$$

## 4. Topic modeling in transportation research

In this section, we applied LDA model on an article-abstract data set extracted from 22 scientific journals in the field of transportation research from 1990 to 2015. In doing so, we considered article abstracts the "documents" in LDA. Therefore, the two terms "abstract" and "document" are interchangeable hereinafter.

### 4.1. Data

We selected 22 journals listed in Table 2 in the field of transportation research. These journals are chosen as top tiers from Science Citation Index (SCI) under category "Transportation Science & Technology" and from Social Science Citation Index (SSCI) under category "Transportation". Some journals, such as *Computer-aided Civil and Infrastructure Engineering* and *Accident Analysis and Prevention*, are not chosen since they are substantially shared with other fields. We also excluded articles published in *Transportation Research Record*, although it is an important journal in transportation research. Firstly, a considerable portion of articles in this journal are in other fields such as structure engineering, geotechnology, and hydraulics. Secondly, the high volume of publications in this journal dominates the analysis and thus distorts the results for other journals. Lastly, topic analysis on the proceedings of the Transportation Research Board has been investigated in Das et al. (2016). The abstract data of each selected journal was extracted from Web of Science (`https://apps.webofknowledge.com/`) using scraping scripts. Web of Science did not register article abstract data before year 1990, and we thus only took those articles published after 1990 into account. Before the analysis, we first cleaned the data set by removing those non-content articles, articles with no abstract information, and articles with short abstracts being less than 10 words. In total, we obtained a collection of $M = 17,163$ articles, which span 26 years from 1990 to 2015. The total number of articles from each journal is provided in Table 2. As can be seen, the number of articles published has increased dramatically since year 2000. This is due to two reasons: (1) the number of articles published in each journal is generally increasing over time, and (2) the introduction of new journals.

Note that *Journal of Transportation Engineering* was formerly named *Journal of Transportation Engineering ASCE* before year 2013, and the journal *Transportmetirca* was split into two sister journals *Transportmetrica A - Transport Science* and *Transportmetrica B - Transport Dynamics* in year 2013. To correct these journal names, we combined records from *Journal of Transportation Engineering* and *Journal of Transportation Engineering*

Table 2: Journal article data in this study

| Journal | Abbreviation | Articles | Year |
|---|---|---|---|
| IEEE Transactions on Intelligent Transportation Systems | IEEE Trans Intell Transp Syst | 1480 | 2000-2015 |
| International Journal of Sustainable Transportation | Int J Sus Transp | 199 | 2007-2015 |
| International Journal of Transport Economics | Int J Transp Econ | 174 | 2005-2015 |
| Journal of Advanced Transportation | J Adv Transp | 508 | 1994-2015 |
| Journal of Intelligent Transportation Systems | J Intell Transp Syst | 210 | 2006-2015 |
| Journal of Transport Economics and Policy | J Transp Econ Policy | 477 | 1992-2015 |
| Journal of Transport Geography | J Transp Geogr | 898 | 2006-2015 |
| Journal of Transportation Engineering | J Transp Eng | 2115 | 1991-2015 |
| Network & Spatial Economics | Netw Spat Econ | 312 | 2003-2015 |
| Transport Policy | Transp Policy | 852 | 2005-2015 |
| Transport Reviews | Transp Rev | 615 | 1991-2015 |
| Transportation | Transportation | 801 | 1990-2015 |
| Transportation Letters | Transp Lett | 144 | 2009-2015 |
| Transportation Research Part A: Policy and Practice | Transp Res Part A | 1607 | 1991-2015 |
| Transportation Research Part B: Methodological | Transp Res Part B | 1525 | 1990-2015 |
| Transportation Research Part C: Emerging Technologies | Transp Res Part C | 1314 | 1995-2015 |
| Transportation Research Part D: Transport and Environment | Transp Res Part D | 1082 | 1996-2015 |
| Transportation Research Part E: Logistics and Transportation Review | Transp Res Part E | 1066 | 1997-2015 |
| Transportation Research Part F: Traffic Psychology and Behavior | Transp Res Part F | 711 | 2011-2015 |
| Transportation Science | Trans Sci | 784 | 1991-2015 |
| Transportmetrica A - Transport Science | Transportmetrica A | 253 | 2005-2015 |
| Transportmetrica B - Transport Dynamics | Transportmetrica B | 36 | 2013-2015 |

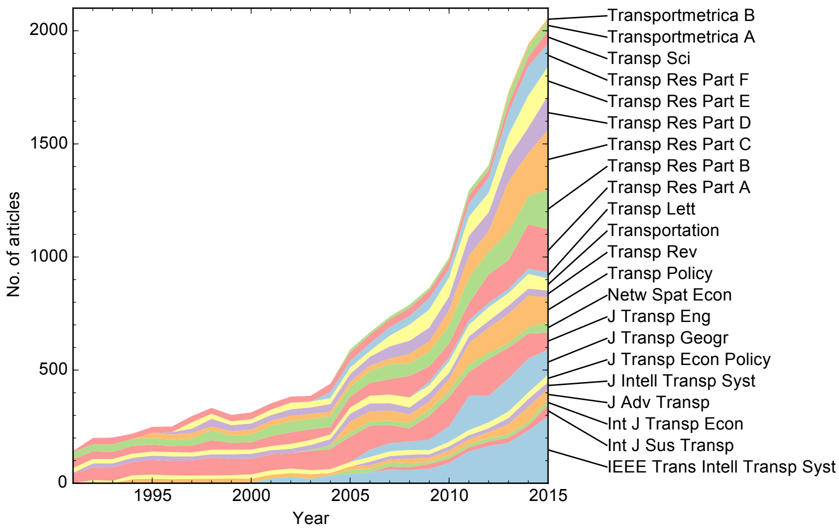



Figure 2: Number of articles of each journal from 1990 to 2015 in the processed data set

*ASCE*. In terms of *Transportmetrica* (2005-2013), we aggregated it with the recent *Transportmetrica A - Transport Science* and considered *Transportmetrica B - Transport Dynamics* a new journal. Fig. 2 shows the final number of articles per journal from 1990 to 2015.

As mentioned, we used an article abstract as a proxy to a document, since the abstract is a compact representation of the whole article and it normally contains enough key words about research themes (Griffiths and Steyvers, 2004). To extract word data from those filtered articles, we split a full abstract into words using any delimiting character, such as space and hyphen. We also removed those words that appeared in less than 5 abstracts or belonged to a standard "stop" list in natural language processing (in this study, we used the stop list provided by Natural Language Toolkit (NLTK), see `http://www.nltk.org/`). We also removed those common words appearing more than 6,000 times, including "model, traffic, paper, time, data, travel, results, using, transport, study, system, models, analysis, problem, based, used, use, transportation, approach, different, proposed, two, new, systems". After this process, we obtained a vocabulary of $V = 13,499$ words that occurred $N = 1,635,206$ times in total in the collection.

### 4.2. *Model inference*

As mentioned in Section 3, the inference of LDA can be done by either applying VEM or via Gibbs sampling. There are various sophisticated software packages implementing these algorithms. In this study, we used the MALLET package (McCallum, 2002), which provides an efficient collapsed Gibbs sampler, to conduct LDA inference on the processed abstract-word data set.

The LDA algorithm requires some basic input parameters, such as number of topics $K$ and the Dirichlet topic distribution prior $\alpha$. In this study, we chose the number of topics $K = 50$. This number is selected given our subjective analysis of the results. The hyperparameter $\alpha$ controls the mean shape and sparsity of $\theta_d$ from the underlying Dirichlet distribution. A larger $\alpha$ prefers distributions that are more uniform over topics, while a smaller $\alpha$ favors sparser distributions. Griffiths and Steyvers (2004) suggested to use $\alpha = 50/K$ for general analysis. In this study, we implemented a smaller value $\alpha = 5/K = 0.1$ to prefer sparse topic distributions, since research themes for general transportation articles are quite focused and concentrated. This value is also the default suggestion of MALLET. The parameter $\beta$ (hyperparameter on topic word distribution $\psi_k$) is set to 0.01. We started 10 runs with different random seeds and initialization and chose the one with maximum posterior probability. The experiment was run on a PC with an Intel Xeon E5 processor (with 8 cores, 16 threads). We ran the sampling for 2000 iterations. The process took about 3 min with 16 threads, using about 650MB of RAM.

## 5. Results and analysis

We show and interpret out main result in this section. In so doing, we focus on analyzing the posterior document-topic distribution $\theta$ and posterior topic word distribution $\psi$. By aggregating $\theta$ at the levels of journal, country/region and publishing year, we estimated the topic distributions of each journal and country/region, and their variation/evolution over time.


Figure 3: Wordcloud of Topic #1– Topic #25


Figure 4: Wordcloud of Topic #26– Topic #50

After running the LDA model, we obtained two types of posterior distributions, i.e., $\theta_d$ – posterior topic distribution of each document $d$, and $\psi_k$ – posterior word distribution of each topic $k$. We show $\psi_k$ as wordcloud in Fig. 3 and Fig. 4. For each topic, we only present those top words with highest posterior probability $\psi_{kw}$. The size of each word is in proportion to its probability.

The topics we inferred here are purely resulted from the statistical structure of the data. Based on the word distribution $\psi$, we can link topics with some specific research areas intuitively. The result could be used as a classification scheme for area/field in transportation research and its literature, as such latent topics normally correspond to research area classification schemes very well (Griffiths and Steyvers, 2004). For example, Topic #1: "prediction, incident, forecasting, neural, flow, performance, shortterm, accuracy, detection, realtime, ..." is mostly related to traffic operations and incident management and Topic #2: "equilibrium, dynamic, network, assignment, solution, user, link, algorithm, flows, networks, ..." centers on network modeling.

Apart from those established research areas, we also found some general topics that are frequently used in academic writing, such as: Topic #26: "performance, method, evaluation, measures, methodology, methods, assessment, criteria, fuzzy, measure, ...", Topic #42: "research, literature, studies, review, issues, recent, future, methods, discussed, approaches, ..." and Topic #43: "may, however, one, many, often, even, possible, whether, would, much, ...". These topics cannot be mapped to particular research fields, but they do cover a substantial proportion in writing an abstract.

### *5.2. Topics distribution over time*

After labeling the inferring topics, we analyzed the temporal trend of each topic using Eq. 1. In this sense, we focus on the overall temporal dynamics of each topic without discriminating different journals or different countries/regions. We show the result in Fig. 5.

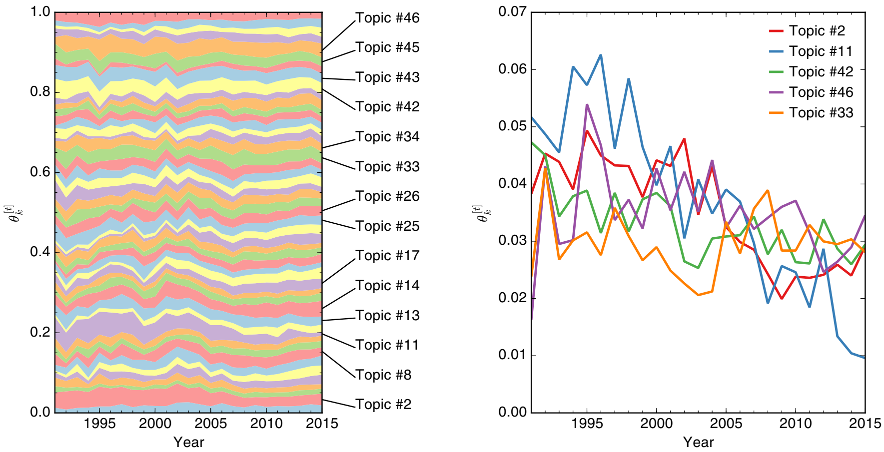



Figure 5: Topic distribution over time

The left panel of Fig. 5 displays the proportion of all the 50 topics from 1991 to 2015. The topics are shown in order (i.e., Topic #1 to #50) from the bottom to the top. The most popular five topics are: #46: routing algorithm "routing, algorithm, problems, solution, heuristic, scheduling, vehicle, computational, solutions, ...", #14: optimization "optimization, optimal, design, solution, algorithm, network, programming, cost, method, ...", #42: common words in academic writing "research, literature, studies, review, issues, recent, future, methods, discussed, ...", #2: network modeling "equilibrium, dynamic, network, assignment, solution, user, link, algorithm, flows, ...", and #33: policy and planning for sustainability "policy, planning, policies, sustainable, environmental, urban, development, public, strategies, ...". The figure on the right presents a closer look at the temporal trends of the most popular five topics. From this figure we can clearly tell that some topics have been declining over time, e.g., Topic #11 "pavement, concrete, pavements, asphalt, design, test, performance, surface, temperature, ..." on pavement, concrete and asphalt.

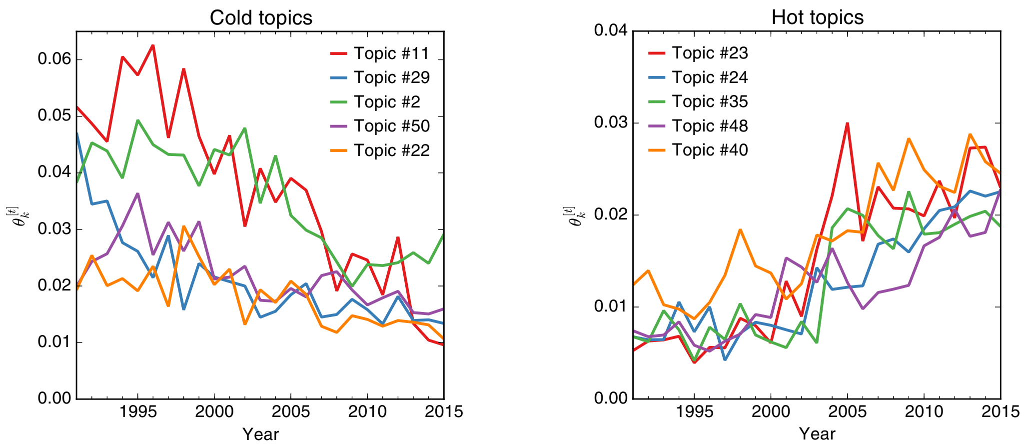
To further investigate hot/cold topics, we computed

$$r_k = \frac{\sum_{t=1991}^{1995} \theta_k^{[t]}}{\sum_{t=2011}^{2015} \theta_k^{[t]}}, \tag{9}$$

as an increase index between two time windows for each topic $k$.

Table 3: Increase index $r_k$ for all topics

| topic | $r_k$ | topic | $r_k$ | topic | $r_k$ | topic | $r_k$ |
|---|---|---|---|---|---|---|---|
| Topic #11 | 0.36 | Topic #32 | 0.84 | Topic #7 | 1.04 | Topic #44 | 1.58 |
| Topic #29 | 0.44 | Topic #46 | 0.86 | Topic #30 | 1.05 | Topic #20 | 1.67 |
| Topic #2 | 0.56 | Topic #31 | 0.88 | Topic #9 | 1.09 | Topic #49 | 1.87 |
| Topic #50 | 0.61 | Topic #38 | 0.89 | Topic #47 | 1.12 | Topic #6 | 1.90 |
| Topic #22 | 0.64 | Topic #34 | 0.89 | Topic #3 | 1.14 | Topic #18 | 2.05 |
| Topic #4 | 0.65 | Topic #27 | 0.90 | Topic #15 | 1.26 | Topic #12 | 2.26 |
| Topic #25 | 0.66 | Topic #26 | 0.90 | Topic #21 | 1.29 | Topic #40 | 2.27 |
| Topic #36 | 0.69 | Topic #13 | 0.91 | Topic #37 | 1.32 | Topic #48 | 2.56 |
| Topic #42 | 0.69 | Topic #28 | 0.93 | Topic #14 | 1.38 | Topic #35 | 2.78 |
| Topic #43 | 0.74 | Topic #19 | 0.95 | Topic #8 | 1.39 | Topic #24 | 2.79 |
| Topic #45 | 0.74 | Topic #39 | 0.95 | Topic #1 | 1.39 | Topic #23 | 4.11 |
| Topic #16 | 0.76 | Topic #33 | 0.97 | Topic #17 | 1.46 | | |
| Topic #10 | 0.77 | Topic #41 | 1.00 | Topic #5 | 1.55 | | |




Topic #11: pavement, concrete, pavements, asphalt, design, ...
Topic #29: maintenance, highway, cost, projects, pavement, ...
Topic #2: equilibrium, dynamic, network, assignment, solution, ..
Topic #50: public, private, services, regulation, market, ...
Topic #22: speed, design, highway, road, distance, ...

Topic #23: driving, drivers, driver, behavior, task, ...
Topic #24: bicycle, cycling, mobility, school, walking, ...
Topic #35: drivers, driving, driver, behaviour, age, ...
Topic #48: information, communication, mobile, technologies, technology, ..
Topic #40: factors, behaviour, attitudes, perceived, survey, ...

Figure 6: Five coldest/hottest topics identified from increase ratio

Therefore, $r_k > 1$ indicates that topic $k$ has increased from 1991-1995 to 2011-2015, while $r_k < 1$ suggests a decreasing trend. Table. 3 provides the estimated $r_k$ for all topics in an increasing order. We identified 5 topics with lowest and highest $r_k$ values in Fig. 6, as coldest and hottest topics respectively. The coldest topics are Topics #11, #29, #2, #50 and #22, which correspond to pavement engineering, highway maintenance, network modeling, infrastructure project financing, and highway design. The hottest topics are Topics #23, #24, #35, #48 and #40, corresponding to driving psychology and behavior (cognitive and simulation), non-motorized mobility, driving behavior of people with different socioeconomic characteristics, the implication of information and communication technologies, and travel survey and analysis. In general, the trends reveal that the topic proportion of traditional engineering problems are decreasing and replaced gradually by human-centered behavioral research questions, sustainable transportation, and emerging transportation technologies and mobility services.

The definition of hot/cold is purely based on the $r_k$ measure and we used all articles to estimate the values. It should be noted that the revealed trend could be resulted from two factors: (1) the natural topic evolution of different journals, and (2) the inflation brought by new journals (new specific research areas). We cannot tell it using only the aggregated topic distribution over time. To distinguish natural trend from the inflation, we further quantified the indicator on both the journal dimension and the temporal dimension.

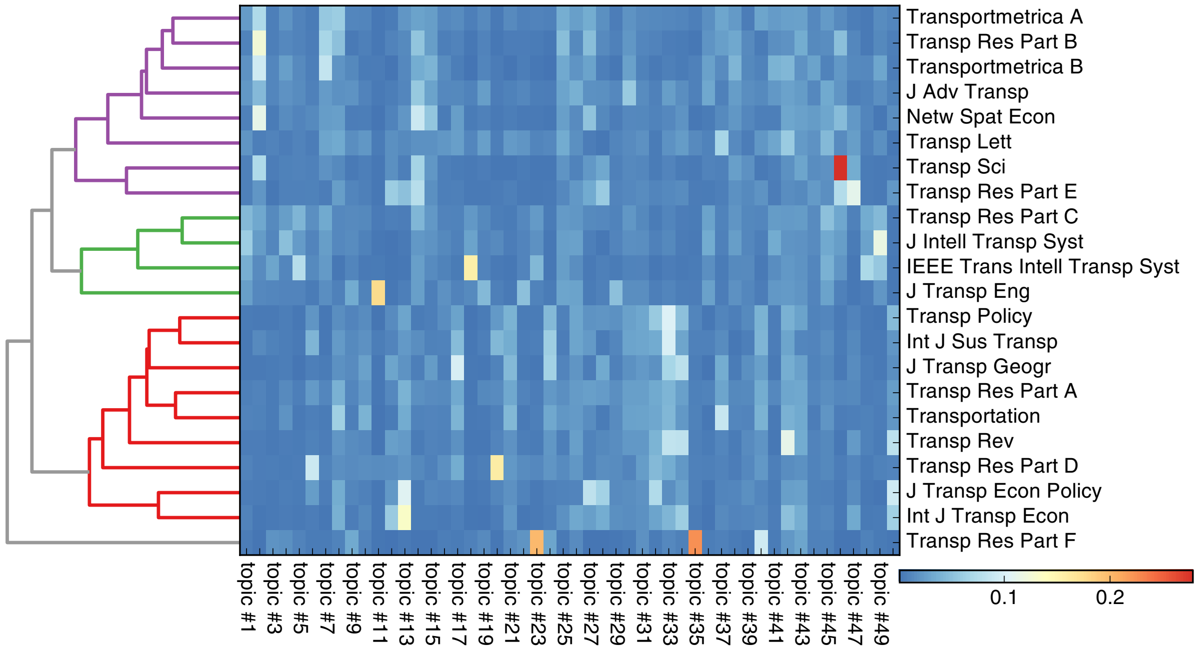



Figure 7: Journal topic distribution and journal similarity

### 5.3. *Journal topic distribution*

At the journal level, we first labeled all articles using journal names and investigated the aggregated topic distribution for each journal. As defined previously in Eq. 2, a journal topic distribution $\theta^j$ is the mean topic distribution of all those articles published in journal $j$. Fig. 7 shows $\theta_k^j$ as a matrix, with each row representing the topic distribution of a particular journal. We found that for most journals topics are widely distributed, while a few journals demonstrate sparse patterns, focusing on a specific set of research topics (e.g., *Transportation Science* and *Transportation Research Part F: Traffic Psychology and Behavior*). To show these differences, we list journal-topic combinations with $\theta_k^j \geq 0.12$ in Table 4.

Table 4: Journals focusing on particular topics

| journal | topic | prob | words |
|---|---|---|---|
| IEEE Trans Intell Transp Syst | 18 | 0.16 | detection, vehicle, road, method, tracking, image, features, algorithm, images, video |
| Int J Transp Econ | 13 | 0.14 | demand, efficiency, price, cost, effects, productivity, effect, period, scale, changes |
| J Transp Eng | 11 | 0.18 | pavement, concrete, pavements, asphalt, design, test, performance, surface, temperature, load |
| Transp Res Part B | 2 | 0.12 | equilibrium, dynamic, network, assignment, solution, user, link, algorithm, flows, networks |
| Transp Res Part D | 20 | 0.16 | emissions, emission, air, environmental, carbon, pollution, vehicle, quality, reduction, greenhouse |
| Transp Res Part F | 23 | 0.20 | driving, drivers, driver, behavior, task, performance, participants, simulator, vehicle, visual |
| Transp Res Part F | 35 | 0.22 | drivers, driving, driver, behaviour, age, young, group, older, road, risk |
| Transp Sci | 46 | 0.28 | routing, algorithm, problems, solution, heuristic, scheduling, vehicle, computational, solutions, search |

By comparing the topic word distribution and journal topic distribution, we found an interesting pair of topics: #23 and #35, discussing about driver and driving extensively. We compare the top 20 words of these two topics in Table. 5. As can be seen, these two topics are quite similar in terms of top words “driver”, “driving” and “behavio(u)r”. However, when looking at the journal topic distribution in Fig. 7, we found that Topic #23 is substantially covered in both *Transportation Research Part F: Traffic Psychology and Behavior* and *IEEE Transactions on Intelligent Transportation Systems*, while Topic #35 barely appears in other journals except *Transportation Research Part F: Traffic Psychology and Behavior*. In fact, by taking a closer look at the two distributions over other words, we can tell that Topic #23 is more about driving simulator studies, focusing on drivers’ reaction and cognition, while #35 mainly discusses the impacts of socioeconomic characteristics of different groups of people on their driving behaviors. In other words, we can tell that Topic #23 is both a psychological and technological problem, while Topic #35 is more on the behavioral side. This is a good example to show that LDA has successfully distinguished the minor differences between these two topics.

Table 5: Two topics about driver and driving

| Topic #23 | | Topic #35 | |
|---|---|---|---|
| driving | 0.0562 | drivers | 0.0599 |
| drivers | 0.0404 | driving | 0.0574 |
| driver | 0.0356 | driver | 0.0168 |
| behavior | 0.0149 | behaviour | 0.0111 |
| task | 0.0115 | age | 0.0092 |
| performance | 0.0099 | young | 0.0091 |
| participants | 0.0097 | group | 0.0090 |
| simulator | 0.0096 | older | 0.0081 |
| vehicle | 0.0090 | road | 0.0068 |
| visual | 0.0084 | risk | 0.0067 |
| road | 0.0072 | reported | 0.0066 |
| effects | 0.0072 | speeding | 0.0065 |
| safety | 0.0069 | groups | 0.0063 |
| speed | 0.0067 | training | 0.0061 |
| warning | 0.0066 | participants | 0.0055 |
| experiment | 0.0056 | among | 0.0054 |
| cognitive | 0.0056 | differences | 0.0053 |
| situations | 0.0052 | safety | 0.0051 |
| tasks | 0.0052 | motorcycle | 0.0051 |
| invehicle | 0.0051 | questionnaire | 0.0050 |

After computing $d^j_{u,v}$ using Eq. 4 for every pair of journals, we measured distance between paired clusters using the complete linkage method. The result of hierarchical clustering is shown as the dendrogram on the left panel of Fig. 7. From the dendrogram we can see that the most unique journal among the set of 22 is *Transportation Research Part F: Traffic Psychology and Behavior*, which has the maximum distance to other clusters. The rest journals can be essentially categorized into three clusters colored in purple, green and red, respectively. A smaller distance suggests a higher degree of similarity. This corresponds to the authors understanding as the journals in each cluster do have similar contents and interests. For example, we can clearly identify four pairs of journals with small distances: the pair of *Transportmetrica A - Transport Science* and *Transportation Research Part B: Methodological* in the purple cluster, the pair of *Transportation Research Part C: Emerging Technologies* and *Journal of Intelligent Transportation Systems* in the green cluster, the pair of *Transport Policy* and *International Journal of Sustainable Transportation* in the red cluster, and the pair of *Transportation Research Part A: Policy and Practice* and *Transportation* in the red cluster (see Fig. 7).

### 5.4. *Journal topic distribution over time*

As mentioned, although Fig. 5 shows the importance of each topic over time, it is still unknown to us whether the trend is due to intrinsic variation or caused by the inflation of new journals. In order to investigate this, we applied the same procedure for each journal using Eq. 5. We plot the final aggregated temporal topic for each of the 22 journals in Fig. 8. Similar to Fig. 5, topics are shown in order from the bottom to the top.

Essentially, we found that topic distribution in most journals are consistent over time. The major difference between panels is the topic proportion, indicating that different journals possess different scopes, which is also

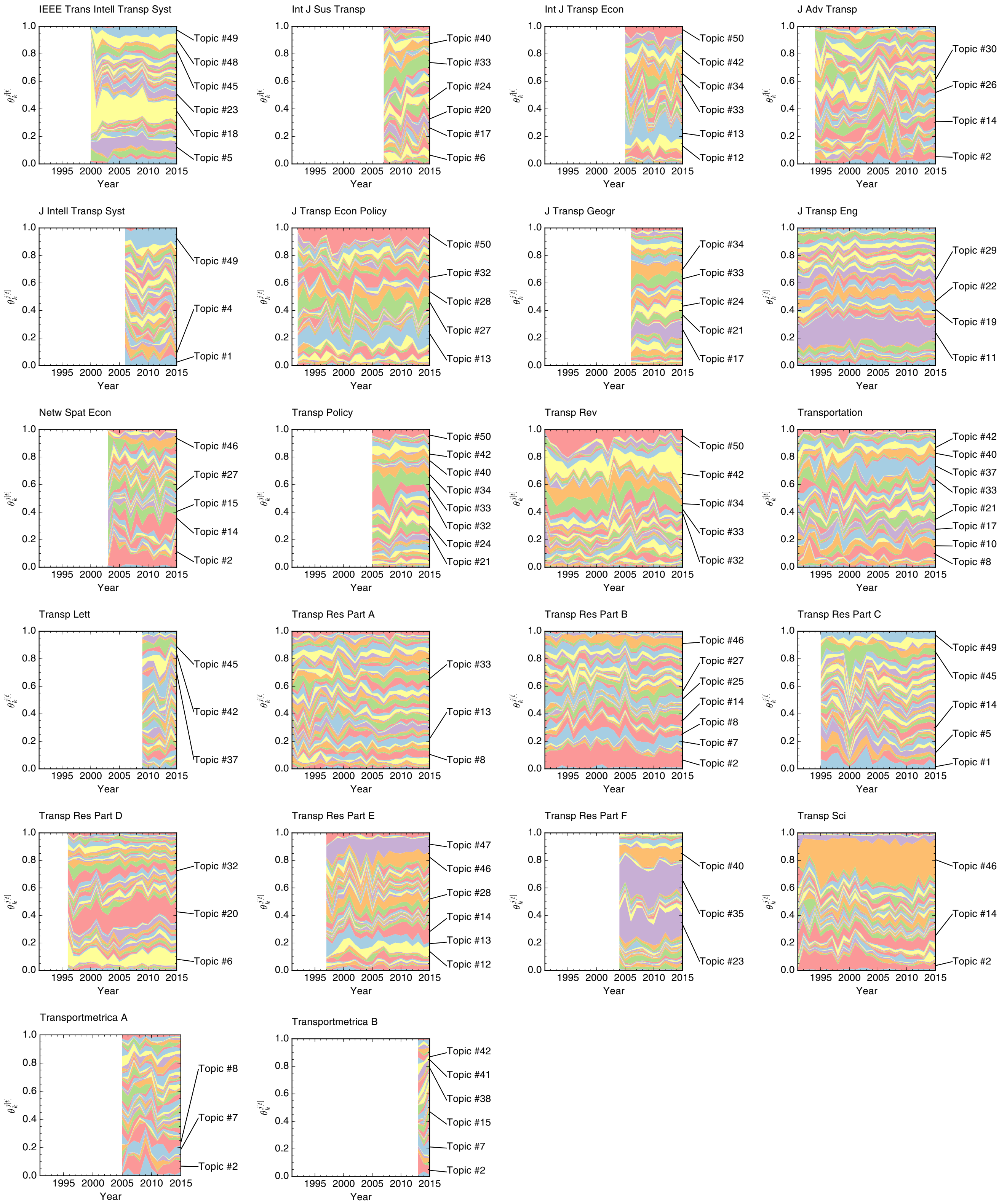



Figure 8: Topic distribution over time for each individual journal

shown in Fig. 7. For example, *Transportation Research Part F: Traffic Psychology and Behavior* mainly covers Topics #23, #35 and #40, and *Transportation Science* mainly covers Topic #46. The temporal trend can further help us to identify the intrinsic evolution over time for each journal. For example, we can see Topic #11 is the most important topic in *Journal of Transportation Engineering*; however, its importance has been decreasing over the last 20 years. This suggests that, despite the inflation from new journals, the decline of Topic #11 in Fig. 6 is also because of this natural evolution in research. In the meanwhile, some topics have grown considerably in some particular journals. For instance, Topic #40: "factors, behaviour, attitudes, perceived, survey, influence, satisfaction, behavior, variables, ...” about travel surveys in *Transportation* and Topic #12: “port, container, shipping, ports, freight, terminal, terminals, intermodal, maritime, ...” about maritime transportation and ports have become two central topics *Transportation Research Part E: Logistics and Transport Reviews*.

Moreover, this temporal trend could also help to detect some anomaly in the history of a journal. Taking *Transportation Research Part C: Emerging Technologies* as an example, we observed a sharp transition in year 2000, when Topic #45 suddenly became prominent. We traced back this observation to the publication metadata and found that it was indeed a special year for the journal. In fact, *Transportation Research Part C: Emerging Technologies* published 22 articles in 2000 as a special volume with the same theme, which is about “transportation and geographic information systems (GIS)”. As a result, words such as “framework, information, tool, management, application, software, ...” were heavily used. And thus, Topic #45, which includes “simulation, framework, modeling, design, integrated, developed, process, development, planning, management, tool, application, presents, decision, presented, support, dynamic, software, information, tools, complex, microscopic, describes, agents, dynamics, various, integration, generation, gis, microsimulation, agentbased, case, methodology, concept, scenarios, level, applied, evaluation, ...” became substantial. Interestingly, we found the word “geographic” is not well presented in Topic #45. To investigate this, we computed the conditional distribution over topics for word “geographic” and found three major topics $P(\text{Topic \#15}|\text{“geographic”}) = 0.164$, $P(\text{Topic \#17}|\text{“geographic”}) = 0.476$ and $P(\text{Topic \#45}|\text{“geographic”}) = 0.210$. From this conditional distribution, the word “geographic” does have a great chance coming from Topic #45, while it is also highly presented in Topic #15 “network, networks, road, links, path, link, urban, paths, hong, kong, shortest, nodes, structure, connectivity, spatial, large, routes, flow, flows, ...” and Topic #17 “urban, spatial, accessibility, areas, land, area, city, location, residential, population, metropolitan, environment, density, built, patterns, region, access, development, characteristics, ...”. Similarly, the conditional distribution of word “GIS” is almost fully covered by two topics $P(\text{Topic \#17}|\text{“GIS”}) = 0.362$ and $P(\text{Topic \#45}|\text{“GIS”}) = 0.638$. This suggests that, although “geographic” and “GIS” are not top words in Topic #45, we still have great confidence to state they are mainly represented in the topic. And the reason of “geographic” and “GIS” not being top words is simply because their overall occurrence is low compared with other words. Using this special case, we further confirmed the extracted topics are both representative and meaningful.

### 5.5. Country/region topic distribution

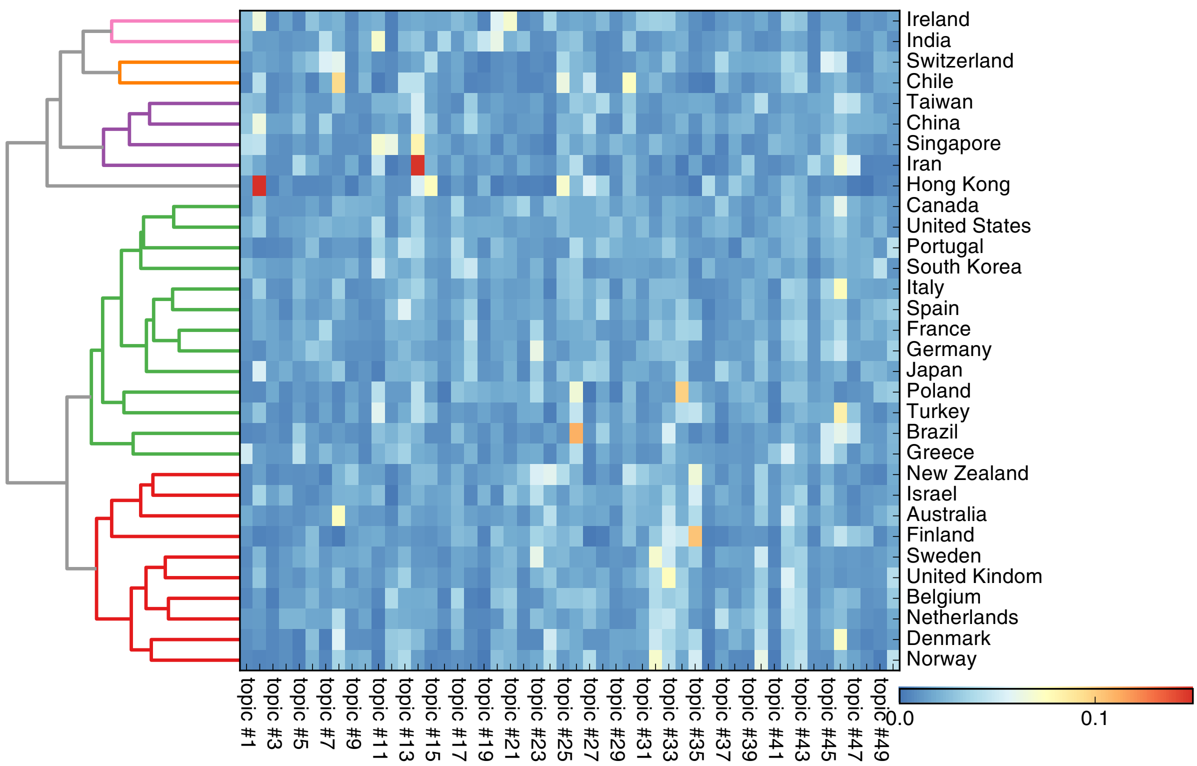



Figure 9: Country/region topic distribution and country/region similarity

Similar to the exercise we have performed at the journal level, we aggregated topics distribution using the correspondence address using Eq. 6. In the full data set, there exist 185 articles without correspondence address, so we used the rest 16,978 articles to perform the following analysis at the country/region level. We present $\theta_k^{(c)}$ of countries/regions with more than 50 articles and the hierarchical clustering result in Fig. 9. The topic distribution at country/region level shows great diversity. We hardly observed any pairs of countries/regions sharing a similar distribution; instead, different countries/regions appear to focus on different topics. We listed the top 10 country/region and topic pairs in Table 6.

Table 6: Countries/regions focusing on particular topics

| country | topic | prob | words |
|---|---:|---:|---|
| Hong Kong | 2 | 0.17 | equilibrium, dynamic, network, assignment, solution, user, link, algorithm, flows, networks |
| Iran | 14 | 0.15 | optimization, optimal, design, solution, algorithm, network, programming, cost, method, location |
| Brazil | 26 | 0.11 | performance, method, evaluation, measures, methodology, methods, assessment, criteria, fuzzy, measure |
| Finland | 35 | 0.10 | drivers, driving, driver, behaviour, age, young, group, older, road, risk |
| Poland | 34 | 0.10 | countries, development, economic, growth, cities, infrastructure, regional, european, china, regions |
| Chile | 8 | 0.09 | choice, logit, preference, stated, utility, mode, preferences, attributes, discrete, value |
| Turkey | 46 | 0.08 | routing, algorithm, problems, solution, heuristic, scheduling, vehicle, computational, solutions, search |
| Singapore | 14 | 0.08 | optimization, optimal, design, solution, algorithm, network, programming, cost, method, location |
| Italy | 46 | 0.08 | routing, algorithm, problems, solution, heuristic, scheduling, vehicle, computational, solutions, search |
| Australia | 8 | 0.08 | choice, logit, preference, stated, utility, mode, preferences, attributes, discrete, value |

This diversity is also reflected on the clustering results, in which only a few pairs of countries displaying strong similarity: (US and Canada), (Italy and Spain), (Germany and France), (UK and Sweden), and (Belgium and Netherlands). Although the similarity is weak, we still noticed that the hierarchical clustering of countries/regions basically corresponds to their geographical locations and development stages. We identified several clusters there, with the largest one colored in green including US, Canada, Japan, South Korea, Brazil and most European countries such as Germany and France; the cluster in red consisting of UK, Australia, Belgium, Israel, and some other Northern European countries (e.g., Netherlands, Denmark and Sweden). There is also an Asian cluster of China, Taiwan, Singapore and Iran shown in purple. The rest countries/regions are not well captured these clusters given their unique signatures.

### *5.6. Country/region topic distribution over time*

We next performed the same analysis to investigate the temporal topic variation for each country/region. In doing so, we estimated $\theta_k^{(c)[t]}$ using Eq. 8. We present the temporal trend of the top 8 countries/regions in the data set. This plot is very interesting in the sense that it reveals temporal variation of research focus/strategy of different countries/regions. For example, we found that US has a wide distribution with some particular focus on Topic #2: "equilibrium, dynamic, network, assignment, solution, user, link, algorithm, flows, networks, ...", Topic #11: "pavement, concrete, pavements, asphalt, design, test, performance, surface, temperature, load, ...", Topic #14: "optimization, optimal, design, solution, algorithm, network, programming, cost, method, location, ..." and Topic #46: "routing, algorithm, problems, solution, heuristic, scheduling, vehicle, computational, solutions, search, ...". For China, we found that Topic #2 on network modeling and Topic #27: "pricing, toll, congestion, optimal, road, welfare, tolls, demand, cost, price, ..." on congestion pricing cover a much larger proportion than other countries/regions. The focus of UK has been Topic #33: "policy, planning, policies, sustainable, environmental, urban, development, public, strategies, process, ...” and Topic #40: “factors, behaviour, attitudes, perceived, survey, influence, satisfaction, behavior, variables, perceptions, ...”. The research Canada has been centered on is Topic #17: “urban, spatial, accessibility, areas, land, area, city, location, residential, population, ...”, Topic #37: “activity, activities, behavior, household, social, patterns, individuals, individual, participation, daily, ...” and Topic #46: “routing, algorithm, problems, solution, heuristic, scheduling, vehicle, computational, solutions, search, ...”. These examples show clearly that, within the field of transportation research, research focuses of different countries/regions demonstrate clearly distinguishable patterns. Some topics stand out at some regions, probably because there are a group of active researchers on these topics or the topics are of particular relevance or more importance to the regions, as transportation is an applied discipline.

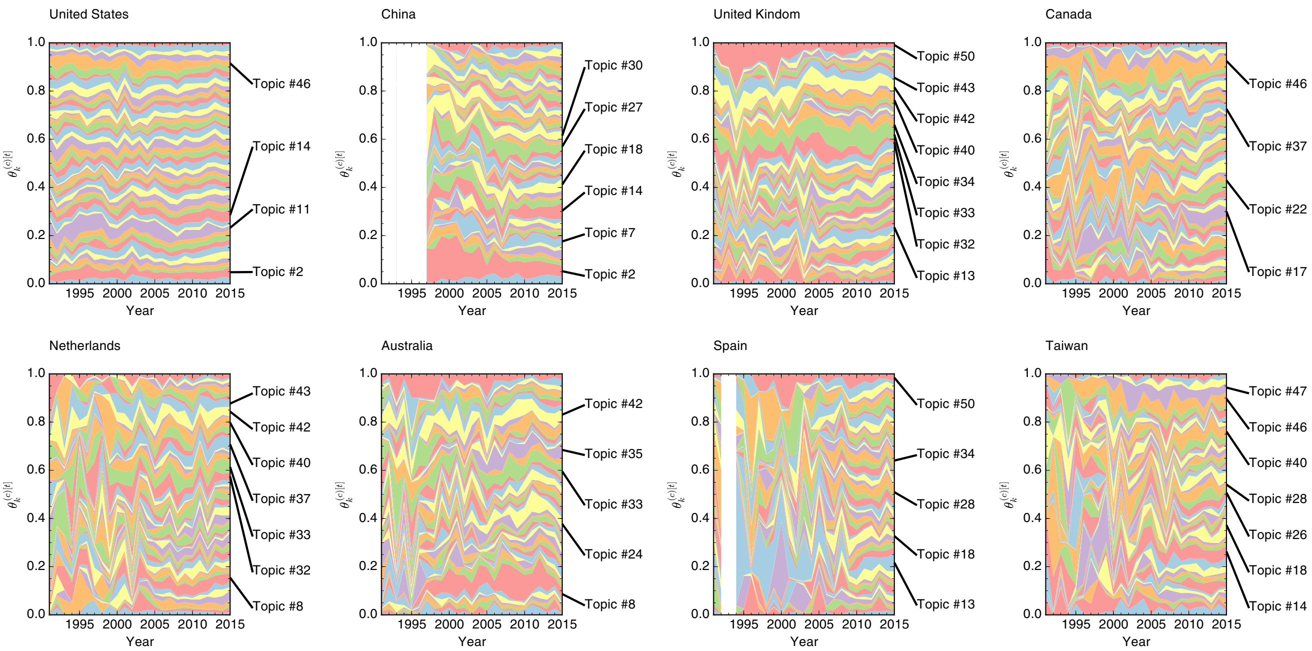



Figure 10: Topic distribution over time for the top 8 countries/regions

## 5.7. *Network of word co-presence*

The analysis of topic word distribution reveals that some words may have strong interconnections to each other. This is also true in terms of the co-occurrence/co-presence of words in different topics. In this part, we present a network visualization of co-presence structure of words across all topics. The network is defined as follows. Firstly, we define a binary matrix $B = \left[\psi_k^v \geq 0.075\right]$ of size $V \times K$, with each element $b_{vk} = 1$ if $\psi_k^v \geq 0.075$ and 0 otherwise. Thus, $b_{vk}$ characterizes that whether word $v$ is a substantial component of topic $k$. Next, based on this binary matrix, we compute an adjacency matrix of words (with a size of $V \times V$) as $A = BB^\top$. And thus, $a_{uv}$ in this matrix represents the number of topics with both $\psi_k^u \geq 0.075$ and $\psi_k^v \geq 0.075$ (in other words, the two words $u$ and $v$ tend to be co-present in a topic). In Fig. 11 we visualize the structure of word co-presence network defined by matrix $A$. For better visualization, we only show the largest connected component of the network, which consists of 512 vertices and 5,113 edges.

As edges in this network are obtain from each distribution $\psi_k$, the network captures the topic structure to a certain degree. For example, we see clearly clusters of words which are highly connected. In the meanwhile, there are some words serving as bridges in this network. For example, the word “operations” appears heavily in three topics “rail operation”, “airline operation” and “port operation” (see the three clusters in green on the right of Fig. 11). This network provides an interesting tool to show how far a pair of words are from each other. It is also useful in detecting words with with diverse representations in different areas in transportation research. For example, the word “capacity” is a general term in transportation, but it may refer to different meanings in research areas of traffic flow theory and public transport. In general, the network presented here unveils how topics are connected by their word distribution and how different topics are allocated in this word landscape. This could be used as a tool to measure conception distance between topics.

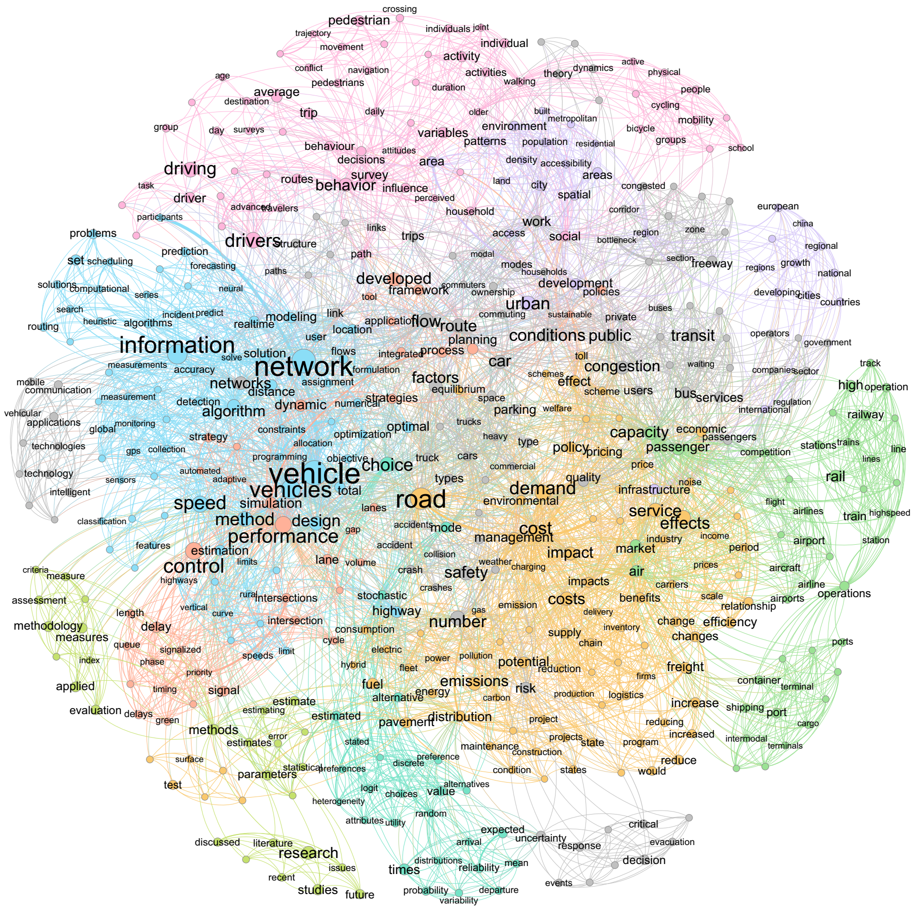



Figure 11: Co-presence structure of words across topics

## 6. Conclusion and discussion

In this paper, we have presented an empirical work about discovering research topics and their trends in the field of transportation research. In detail, we applied topic modeling (LDA) on article abstract data from 1990 to 2015. We focused on 22 scientific journals included in Science Citation Index and Social Science Citation Index under the category "Transportation science & technology" and "Transportation", respectively.

As an unsupervised learning algorithm, LDA does not require any prior knowledge and the inferred topics are purely resulted from the statistical structure of the abstract-word data. Using the posterior document-topic and topic-word distribution, we investigated the context of each topic and the evolution of topics across time and journals. By aggregating the result at the journal level, we found that most journals essentially cover various topics, while some show specific and well-defined scopes.

Our results could benefit different parties in the academic community, such as researchers, journal editors, publishers, conference organizers and funding agencies. One direct application of our study is to provide a classification scheme for transportation research. And this could be used as a better structure for conference organizers and journal editors to define the scope of a conference or a journal (see http://www.wctrs-conference.com/submit-abstract.asp for an example of session and tracks from the 14th World Conference on Transport Research). Such a structure could make it easier for conference organizers to distribute talks with similar themes to the same session. And journal editors could use it to categorize a new submission and find the right set of experts as reviewers.

The temporal evolution reveals topic importance dynamics (e.g., hot/code topics) of transportation research. This information could help researchers understand the research trends and decide their research focuses. From the analysis, we found that sustainable transportation, non-motorized mobility and travel/driving behavior seem to attract more and more attention overtime. In the meanwhile, the journal topic distribution can help researchers to identify targeted journals when preparing a manuscript/submission. For example, a researcher may infer posterior topic distribution of a new paper, and then decide which journal is the best venue for the submission based on scopes of different journals. This is easy for special topics such as transport psychology, since there is only one journal having this scope. However, for topics such as network modeling or optimization, one needs to evaluate journals carefully before submission (e.g., journal impact, review duration), since there exist multiple journals covering them. On the other hand, journal editors and publishers could use this information to consider the need of adjustment of focus and scope and form strategies for future development of their journals (e.g., more specific v.s. more general; and if specific, which topics to focus on). For instance, we observed that the journal *Transportation Science* has been following a consistent strategy over the last decades, with a particular focus on routing algorithms.

By aggregating topic distribution using correspondence address, we found that different countries/regions do pay special attention to different sub-fields. In a sense this trend is a proxy to reflect the actual demand and what the country/region seeks in research, since transportation is an applied discipline. Funding agencies could use it to evaluate potential of different topics and prioritize their funding supports given the research need of the country/region.

Despite that transportation research is continuously growing in terms of number of publications and number of journals, our results show that the scope of transportation research has become broader and more interdisciplinary as well. With the natural evolution of topics and the introduction of new journals, we found that human-centered research and sustainable development are becoming research hotspot nowadays. These research topics requires not only knowledge from traditional engineering, but also advance in social science and the integration with behavioral, environmental and economic research. Although we have analyzed publications quantitatively, we still cannot address the question raised by Button (2015) on how the quality of research has changed with increasing number of publications. Nevertheless, we hope our work can stimulate more discussions about publishing in transportation research and the future of the field.

In summary, this study provides a tool for us to have a better understanding about transportation research in general and different subareas and scopes in detail. Possible future directions include to integrate other data sources to further quantify the growth and evolution of research content and measure the impact and potential of new emerging topics. Apart from this, it is also interesting to adopt other variants of topic models, such as author topics models (Rosen-Zvi et al., 2004), dynamic topic model (Blei and Lafferty, 2006) and relational topic model (Chang and Blei, 2009) to find more insights.

## Acknowledgment

The authors greatly appreciate the discussions and input from Dr. Kay W. Axhausen from Institute for Transport Planning and Systems (IVT), ETH Zurich and Dr. Hai Wang from School of Information Systems, Singapore Management University. The work described in this paper was partially supported by research grants from the National Science Foundation (CMMI-1362631; CMMI-1562420)

## References

Banister, D., 2014. Where to start? Transport Reviews 34 (1), 1–3.

Blei, D. M., 2012. Probabilistic topic models. Communications of the ACM 55 (4), 77–84.

Blei, D. M., Lafferty, J. D., 2006. Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp. 113–120.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Button, K., 2015. Publishing transport research: Are we learning much of use? Transport Reviews 35 (5), 555–558.

Chang, J., Blei, D. M., 2009. Relational topic models for document networks. In: International conference on artificial intelligence and statistics. pp. 81–88.

Das, S., Sun, X., Dutta, A., 2016. Text mining and topic modeling on compendium papers from transportation research board annual meetings. In: Transportation Research Board 95th Annual Meeting. No. 16-3009. Transportation Research Board.

Endres, D. M., Schindelin, J. E., 2003. A new metric for probability distributions. IEEE Transactions on Information theory.

Farrahi, K., Gatica-Perez, D., 2011. Discovering routines from large-scale human locations using probabilistic topic models. ACM Transactions on Intelligent Systems and Technology (TIST) 2 (1), 3.

Gatti, C. J., Brooks, J. D., Nurre, S. G., 2015. A historical analysis of the field of or/ms using topic models. arXiv preprint arXiv:1510.05154.

Griffiths, T. L., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101 (suppl 1), 5228–5235.

Hasan, S., Ukkusuri, S. V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transportation Research Part C: Emerging Technologies 44, 363–381.

Heilig, L., Voß, S., 2015. A scientometric analysis of public transport research. Journal of Public Transportation 18 (2), 8.

Hofmann, T., 1999. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 50–57.

Huynh, T., Fritz, M., Schiele, B., 2008. Discovery of activity patterns using topic models. In: Proceedings of the 10th international conference on Ubiquitous computing. ACM, pp. 10–19.

McCallum, A. K., 2002. Mallet: A machine learning for language toolkit.
URL http://mallet.cs.umass.edu/

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P., 2004. The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, pp. 487–494.