Big Data and Transportation Safety: Connecting the Dots

Subasish Das, Ph.D.
(Corresponding Author)
Associate Transportation Researcher
Texas A&M Transportation Institute
3135 TAMU, College Station, TX 77843
Email: s-das@tti.tamu.edu
ORCID: 0000-0002-1671-2753

Greg P. Griffin, Ph.D.
Assistant Professor
Urban and Regional Planning
The University of Texas at San Antonio
501 W. César E. Chávez Blvd
San Antonio, TX 78207
Email: gregpgriffin@utexas.edu

TOTAL WORDS: 4,640 words

4,140 words = text (including abstract and references)
500 = 2 tables

ABSTRACT

Emerging big data resources and practices provide opportunities to improve transportation safety planning and outcomes. However, researchers and practitioners recognize that big data includes biases that need to be explored and examined before performing data-driven decision making. This study systematically reviews both the sources of bias and approaches to mitigate them through a review of published studies and interviews with experts. The study includes a quantified analysis of topic frequency and an evaluation of the reliability of concepts through two independent trained coders. To identify the trends in the unstructured textual contents, the research team developed a text mining pipeline to identify trends, patterns, and biases. The results show a need to maintain the central location of transportation experts and public to determine the proper goals and metrics to evaluate transportation safety, develop new methods that relate big data to the total population needed for transportation safety, solve difficult problems using big data, and work ahead of emerging trends and technologies.

Keywords: traffic volume, low-volume roads, interpretable machine learning.

INTRODUCTION

Obtaining high-quality data for transportation safety planning has been expensive and slow. Big data is generally "not about society, but about users and markets"—inherently including a range of biases *(1).* Research has found far-reaching bias problems in big data sources, but this study focuses on those with an impact on planning for transportation safety. Using interviews with expert practitioners and a synthetic literature review, results suggest implications for transportation safety research and practice to distinguish and diminish bias in big data. The project addresses two critical issues: the sources of the bias and the approaches to mitigate it.

One practical definition of big data in a planning context is that "it is too large and too complex to be stored, transferred, shared, curated, queried, and analyzed by traditional processing applications" *(2).* Some suggest "the volume of data continues to double every three years as information pours in from digital platforms, wireless sensors, and billions of mobile phones" (*3*). Despite the challenges, big data "can reveal new dynamics, can allow for the study of certain processes in real-time and can highlight relationships and correlations that may pass unnoticed using classical methods and data" *(1).* Leveraging new research on the use of big data in transportation and smart cities, this study supports both research and practice of transportation planning.

Leveraging results of semi-structured interviews with big data experts (*1*), this study performs three forms of textual analytics to respond to three emerging questions in big data for transportation safety:

- What are the key terms experts use when describing the role of big data in transportation safety?
- How are these terms related to the big data experts' language?
- How do clusters of bigrams (a string of two words) relate to each question asked by the researchers?

The next section briefly explains how the researchers performed the study. Following the methodology, a brief review of the literature is interweaved with qualitative interview results before describing the text mining results and conclusions for research and practice on the use of big data for transportation safety.

METHODOLOGY

Interview Design

Researchers conducted semi-structured interviews with ten experts to gain insights on how they used big data resources for transportation safety. An interview guide kept the discussion focused on research questions while allowing informants to emphasize their own experiences (*5*). Researchers identified experts on the topic by searching for "big data" in the conference materials for the 2017 and 2018 Transportation Research Board (TRB) Annual Meetings in addition to the 2017 American Planning Association (APA) National Planning Conference. Researchers further filtered these candidates by prioritizing conference speakers who: 1) presented research on big data or discussed the topic in round tables; 2) worked in a position suggesting substantial experience in big data, either directly or in a management role; and 3) had current contact information on an organizational website. Four interviewees worked in universities, three in transportation consulting, two from state departments of transportation, and one in a city transportation department, across the United States and one from Canada. Interviews were conducted online using synchronous text, except for one who preferred a video call. Interviews were anonymized before analysis or sharing on the Virginia Tech Dataverse (*4*).

Text Mining Methods

Following a qualitative review of interview results alongside current literature, researchers analyzed the corpus of interview responses through three text mining approaches. To summarize the words used in interview responses quantitatively, researchers applied Term Frequency-Inverse Document Frequency (TF-IDF), which normalizes the frequency of words considering repetition of words through documents (interviews in this study). Transportation researchers have used this approach to understand the use of Twitter for public engagement to weight responses considering some individuals post more frequently than others (*6*). Network analysis shows how key terms in the interview results relate to each other by frequency and has been used to understand communication in crisis events such as Hurricane Sandy (*7*). Multiple Correspondence Analysis (MCA) quantifies topical proximity, revealing clusters of similar words in the text corpus, and has been used in environmental and epidemiological studies (*8*). The two methods are focused on bigrams to associate object-subject pairings for meaning, further differentiating our approach from previous textual analytics research (*9*).

SUMMARY OF LITERATURE AND INTERVIEWS

Interviews extended findings of the literature review and confirmed challenges of bias in big data and different approaches on how to mitigate biases. Mobile phones are the most prevalent source of big data for transportation but do not characterize entire populations for transportation study (*10*). Passive collection of location data through cell tower proximity is obstructed when calls are made. Furthermore, the application of this data to transportation problems may not fit the given need as well as custom-designed data solutions (*11*). Therefore, phone data misrepresenting transportation system users can lead to problems for transportation planning (*10*).

Beyond these clear challenges, mobile phone data also represents a great deal of uncertainty (*11*). Transportation models using this data may then have even more 'unknown unknowns' regarding the likelihood of representing current and future travel correctly. Privacy restrictions and lack of mobile carrier compatibility introduce even further known margins at multiple scales. Data aggregators may help mitigate these challenges by combining datasets to balance individual challenges, but this may further obscure errors from original data, rather than fix them.

Representation biases, such as sampling and demographic biases, occur when the people buying the products tracked, like a car or phone, does not represent the total population. GPS-based travel surveys may be more accurate than traditional travel diaries regarding time and routing of trips but can introduce problems with correct identification of travel mode and trip purpose (*12*)—key inputs for travel modeling. However, big data can be particularly useful for tracking complex travel behaviors such as ride splitting (*13*). A consultant noted that "local planners have come to us once we have started data warehouses to get data for their needs," suggesting this area as an emerging field that depends making big data resources more refined and accessible.

In active data collection through purpose-built apps such as Strava, social desirability bias occurs when users only share information that shows accomplishment. Pedestrian travel observation through big data is in its infancy, other than simple counts using automatic detectors (*15*). New approaches that can track pedestrian trips using accelerometers or other sensors may support a broad representation of a pedestrian community.

Despite these biases, interviewees generally reported they support improvements to transportation planning. When asked how big data can help, one consultant stressed speed and

cost savings. Instead of spending the bulk of funds on finding data, planners and researchers can now use most of their funds to solve problems using existing data and data fusion platforms. Also, web-based data visualization and analytics made planning efforts significantly more efficient and accessible to a larger audience.

RESULTS

Knowledge Extraction by Text Mining

In recent years, text mining methods have been widely adopted by many transportation researchers. Keywords in terms of a sequence of one or more words can provide a condensed representation of a document's textual content. The keyword extraction methods are combined with supervised learning, machine-learning algorithms, or statistical methods. The research team developed a text corpus for interviewee responses for each of the questions shown in Table 1. Several text mining algorithms have been applied to determine the trends and clusters in the unstructured textual responses of the interview participants.

Table 1 Interview Questions

No.	Question
1	When did you start working with big data in transportation?
2	Why did your organization decide to use new sources of big data?
3	Has using big data helped improve transportation planning?
4	Are there ways that the data does not represent the entire population of interest in transportation planning?
5	How do you mitigate the impact of big data not representing the population?
6	Overall, has using big data has improved planning for transportation safety in your applications?
7	Is there anything else you would like to add?

Term Frequency Inverse Document Frequency (TF-IDF)

In information retrieval, TF-IDF has been widely used to differentiate between documents (one response by an individual can be considered as a document, and all responses can be considered as the document set or corpus) by estimating how relevant their contents are to a set of terms in a search string. It combines two different weighting parameters to determine the relevance of the keywords: TF and IDF. The basic concept of the TF-IDF is described here based on the book of Silge and Robinson *(16).*

In TF-IDF framework, the terms are viewed as having different levels of importance; some terms are weighted more while others are weighted less *(16).* The parameter TF can be denoted as $tf(t, d)$. It indicates the number of occurrences of the term (t) in the document (d). The other parameter IDF of a term (t) can be defined as (1):

$$idf(t, D) = log \frac{|D|}{df(t, D)} \quad (1)$$

where $|D|$ is the total number of documents in the corpus (set consists of all documents) *D*, and $df(t, D)$ is the number of documents that contain the term t:

$$df(t, D) = |\{d \in D : t \in d\}| \quad (2)$$

The concept behind the parameter IDF is to provide additional weight to terms that are found in only a few documents. It is important as it can be used to differentiate between

documents and to reduce the weightage of terms that frequently appear in all documents. The TF-IDF score for a term t is the product of both of these parameters, so for a string containing the set of terms q, the TF-IDF score of document d in corpus D can be expressed as following (3):

$$TF - IDF(q, D) = \sum_{t \in q} tf(t, d) \times idf(t, D) \quad (3)$$

Figure 1 illustrates the top 15 keywords by the question. The keywords are sorted based on the TF-IDF values. For the first question, the keyword *'years'* is more visible in this corpus due to the nature of the question. The top words associated with question 2 are value and interested. It indicates that industries are interested in being poised in big data due to the stakeholder interest and market values. Question 4 and question 5 are about the sample size and biases. The top three keywords with high TF-IDF in these two corpora are *gps, population, bias, customer, companies, and understand*. In answering the question regarding the effect of big data in improving transportation safety, the top words with high tf-idf values are *these, take,* and *range.* It indicates a long-term effect instead of immediate or short-term improvement. The open-end responses in Question 7 show that keywords such as *want* and *sector* are more used compared to other question.

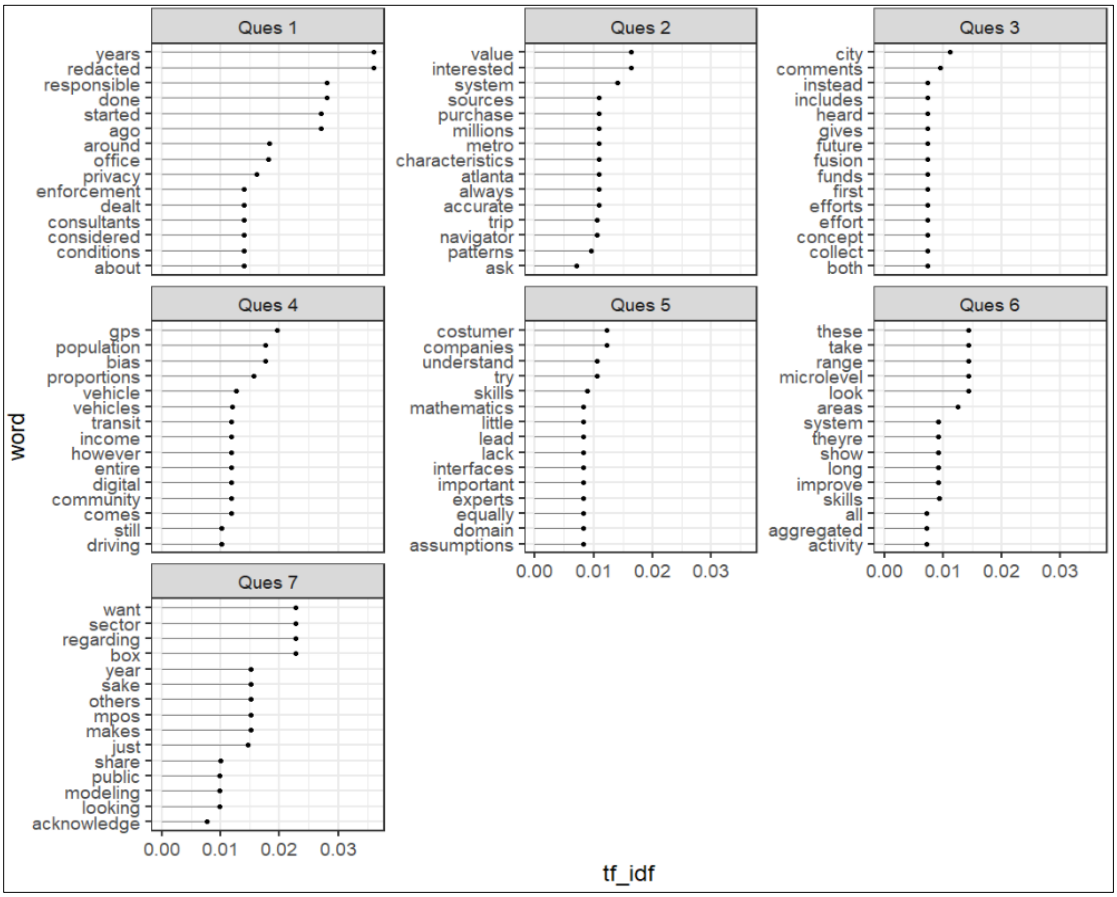

Figure 1 Term frequency-inverse document frequency (tf-idf) of the top 15 words in each question.

Network Analysis

One easy of way understanding the association between words is to construct a network plot. Within the context of text mining, network plots can show relationship strength or term cohesion, leading to a cluster of word groups with meaning or trend (see Figure 2).

When the beginning of the word is large in numbers, the plots can be very dense and hard to interpret. The lines connecting the circles in a network is known as the edge. The arrows indicate the directionalities of the terms. For example, *'purchase'* is directed towards *'inrix';* an indication of the word group as *'purchase inrix.'* The bottom of this network plot has a large number of nodes. The direction of the nodes shows two specific topics: *Waze doesn't represent the entire population*, and *exponentially growing domains.*

Multiple Correspondence Analysis (MCA)

MCA is an unsupervised learning algorithm that 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. The brief theoretical background of MCA has been based on Das et al. study *(17).*

The conceptual framework can be developed by considering *P* as the number of variables (i.e., columns) and *I* as the number of transactions (i.e., rows). This will generate a matrix of *'I multiplied by P.'* The total number of categories for all variables is denoted as $L = \sum_{p=1}^{p} L_p$, where: L_p is the number of categories for variable *p*. It will generate another matrix *'I multiplied by L,'* in which each of the variables will contain several columns to show all of their possible categorical 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 consideration of weightage n_j. For each of the variables, the sum of the weights of category points is considered as *n*. In this way, for the whole set *J* the sum is nP. The relative weight w_j for point C^j is $w_j = n_j/(nP) = f_j/P$. The sum of the relative weights of category points is 1/P, which makes the sum of the whole set as 1.

$w_j = \frac{n_j}{nP} = \frac{f_j}{P}$ with $\sum_{j \in J_q} w_j = \frac{1}{P}$ and $\sum_{j \in J} w_j = 1$

Here, n_{jj} is the number of individual records that have both categories (k and k'). The squared distance between two categories C^j and $C^{j'}$ is represented in (4):

$$(C^j C^{j'})^2 = \frac{n_j + n_{j'} - 2n_{jj'}}{n_j n_{j'}/n} \quad (4)$$

The numerator of (4) is the number of individual records associating with either j or j' (not both). For two different variables, p, and p', the denominator is the familiar *theoretical frequency* for the cell (j, j') of the $J_p \times J_{p'}$ two-way table.

To define different clusters in the attributes, MCA generates several parameters such as coordinates of the attributes. The two-dimensional coordinates indicate the clustering patterns of the attributes. It is important to note that all the points are not equally well displayed in the two dimensions. The quality of the representation is called the squared cosine (cos2), which measures the degree of association between variable categories and an axis. If a variable category is well represented by two dimensions, the sum of the cos2 is closed to one. For some of the row items, more than 2 dimensions are required to perfectly represent the data. The variable categories with the larger value, contribute the most to the definition of the dimensions.

We created n-grams (continuous sequence of n words from a document) to determine the trends of the textual content. A group of two words in sequence can be defined as a bigram.

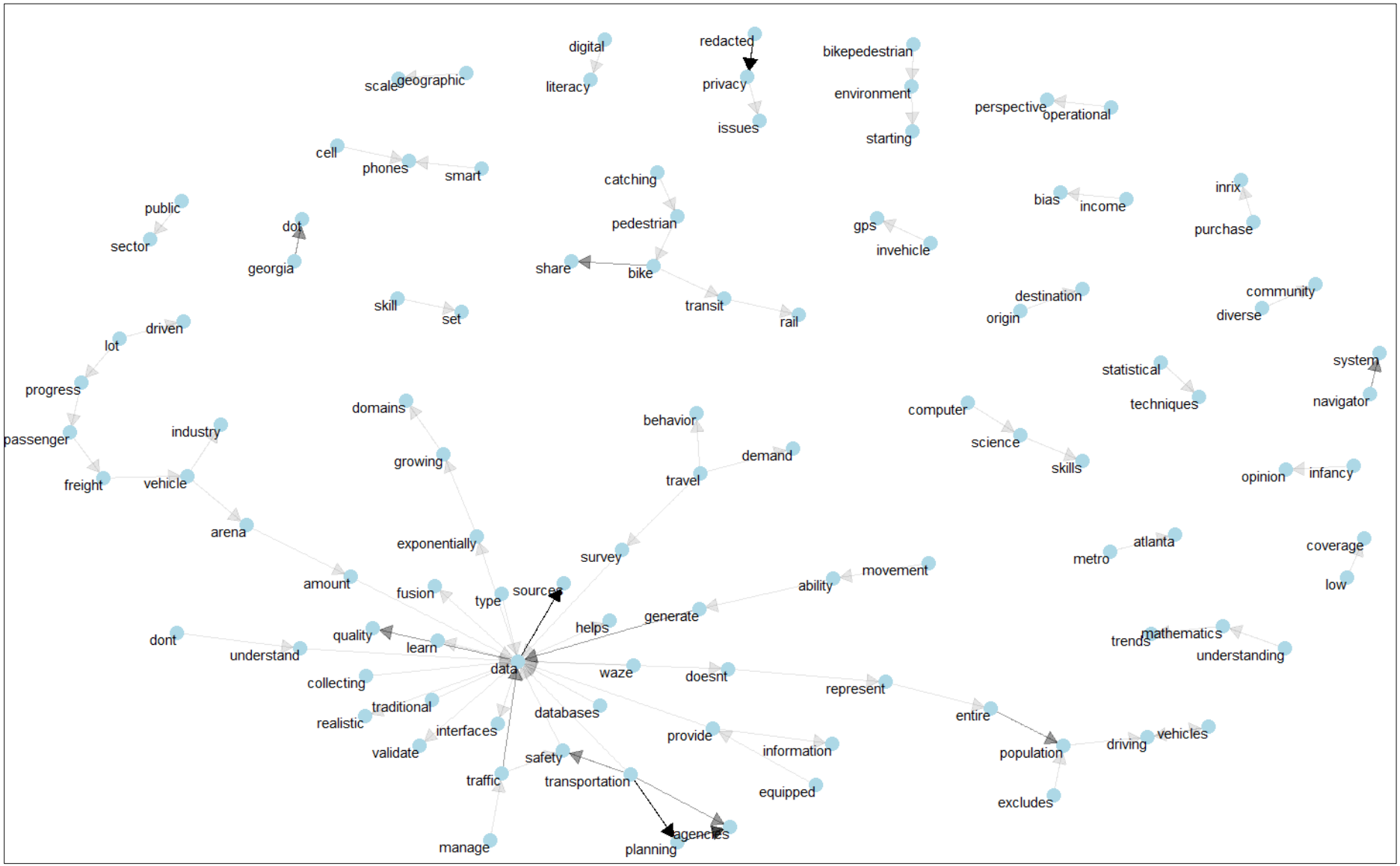

Figure 2 Network plot of the words generated from the complete corpus.

From a group of seven document groups, 51 bigrams have been identified. Table 2 lists the parameters of the MCA analysis on the generated bigrams.

Table 2 MCA Parameters for the Bigrams

Bigrams	Coordinate		cos2		Contribution		Quadrant
	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2	
data sources	0.023	0.343	0.001	0.125	0.001	0.342	1
looking data	0.622	0.381	0.229	0.086	0.957	0.426	1
better data	0.722	0.498	0.331	0.157	1.278	0.721	1
planning agencies	0.913	0.152	0.532	0.015	2.045	0.067	1
agencies except	0.923	0.732	0.074	0.047	2.110	1.576	1
geographic scale	1.184	0.971	0.398	0.268	3.472	2.768	1
traffic safety	1.184	0.971	0.398	0.268	3.472	2.768	1
accuracy data	1.445	1.209	0.451	0.316	5.173	4.295	1
acknowledge lot	1.445	1.209	0.451	0.316	5.173	4.295	1
activity-based modeling	1.445	1.209	0.451	0.316	5.173	4.295	1
advancing way	1.445	1.209	0.451	0.316	5.173	4.295	1
agencies really	1.445	1.209	0.451	0.316	5.173	4.295	1
ability generate	-1.428	0.891	0.364	0.142	5.049	2.331	2
actual number	-1.428	0.891	0.364	0.142	5.049	2.331	2
advanced methods	-1.428	0.891	0.364	0.142	5.049	2.331	2
cell phones	-1.300	0.579	0.723	0.143	4.187	0.984	2
generate data	-1.300	0.579	0.723	0.143	4.187	0.984	2
make sense	-1.300	0.579	0.723	0.143	4.187	0.984	2
operational perspective	-1.300	0.579	0.723	0.143	4.187	0.984	2
think data	-1.300	0.579	0.723	0.143	4.187	0.984	2
acquire data	-1.173	0.267	0.239	0.012	3.406	0.209	2
additionally need	-1.173	0.267	0.239	0.012	3.406	0.209	2
smart phones	-0.815	0.222	0.345	0.025	1.645	0.144	2
travel behavior	-0.815	0.222	0.345	0.025	1.645	0.144	2
collecting data	-0.687	-0.090	0.240	0.004	1.170	0.024	3
manage traffic	-0.687	-0.090	0.240	0.004	1.170	0.024	3
provide information	-0.687	-0.090	0.240	0.004	1.170	0.024	3
for travel	-0.528	-0.297	0.146	0.046	0.691	0.259	3
data quality	-0.401	-0.609	0.083	0.191	0.398	1.090	3
traditional data	-0.401	-0.609	0.083	0.191	0.398	1.090	3
add value	-0.202	-0.448	0.010	0.049	0.101	0.588	3
agencies as	-0.202	-0.448	0.010	0.049	0.101	0.588	3
agencies evaluations	-0.202	-0.448	0.010	0.049	0.101	0.588	3
agencies interested	-0.202	-0.448	0.010	0.049	0.101	0.588	3
big data	0.065	-0.028	0.015	0.003	0.010	0.002	4
origin destination	0.085	-0.966	0.005	0.610	0.018	2.742	4
two years	0.160	-0.905	0.005	0.147	0.063	2.403	4
safety data	0.180	-1.139	0.018	0.722	0.121	5.716	4
actually making	0.371	-1.485	0.034	0.549	0.341	6.476	4
again comments	0.371	-1.485	0.034	0.549	0.341	6.476	4
agencies behind	0.371	-1.485	0.034	0.549	0.341	6.476	4
agencies effort	0.371	-1.485	0.034	0.549	0.341	6.476	4
transportation planning	0.403	-0.640	0.094	0.237	0.403	1.205	4
bike share	0.446	-1.423	0.036	0.366	0.493	5.950	4
redacted privacy	0.446	-1.423	0.036	0.366	0.493	5.950	4
transportation agencies	0.538	-0.241	0.386	0.077	0.710	0.169	4

transportation safety	0.722	-0.315	0.071	0.013	1.292	0.291	4
data learn	0.908	-0.138	0.496	0.011	2.043	0.056	4
privacy issues	0.908	-0.138	0.496	0.011	2.043	0.056	4

Figure 3 illustrates the two-dimensional biplots based on the locations of the bigrams and columns as questions. The color of the bigrams is based on cos2 values. Four different clusters have been identified This analysis shows four clusters from the analytics in Table 2 and Figure 3.

- Cluster 1 (upper right) indicates data quality terms. Coordinates of Question 6 and Question 7 are also located in this cluster. This cluster overall represents data quality issues and usage in transportation safety planning.
- Cluster 2 (upper left) indicates data analysis related terms. This cluster contains the coordinates of Question 4 and Question 5. These two questions are the data analysis method related.
- Cluster 3 (lower left) indicates implementation and uses. This cluster contains the coordinates of Question 2, which is about the reasoning of big data usage.
- Cluster 4 (lower right) indicates big data and other innovative data sources. The coordinates of Question 1 and Question 3 are in this cluster.

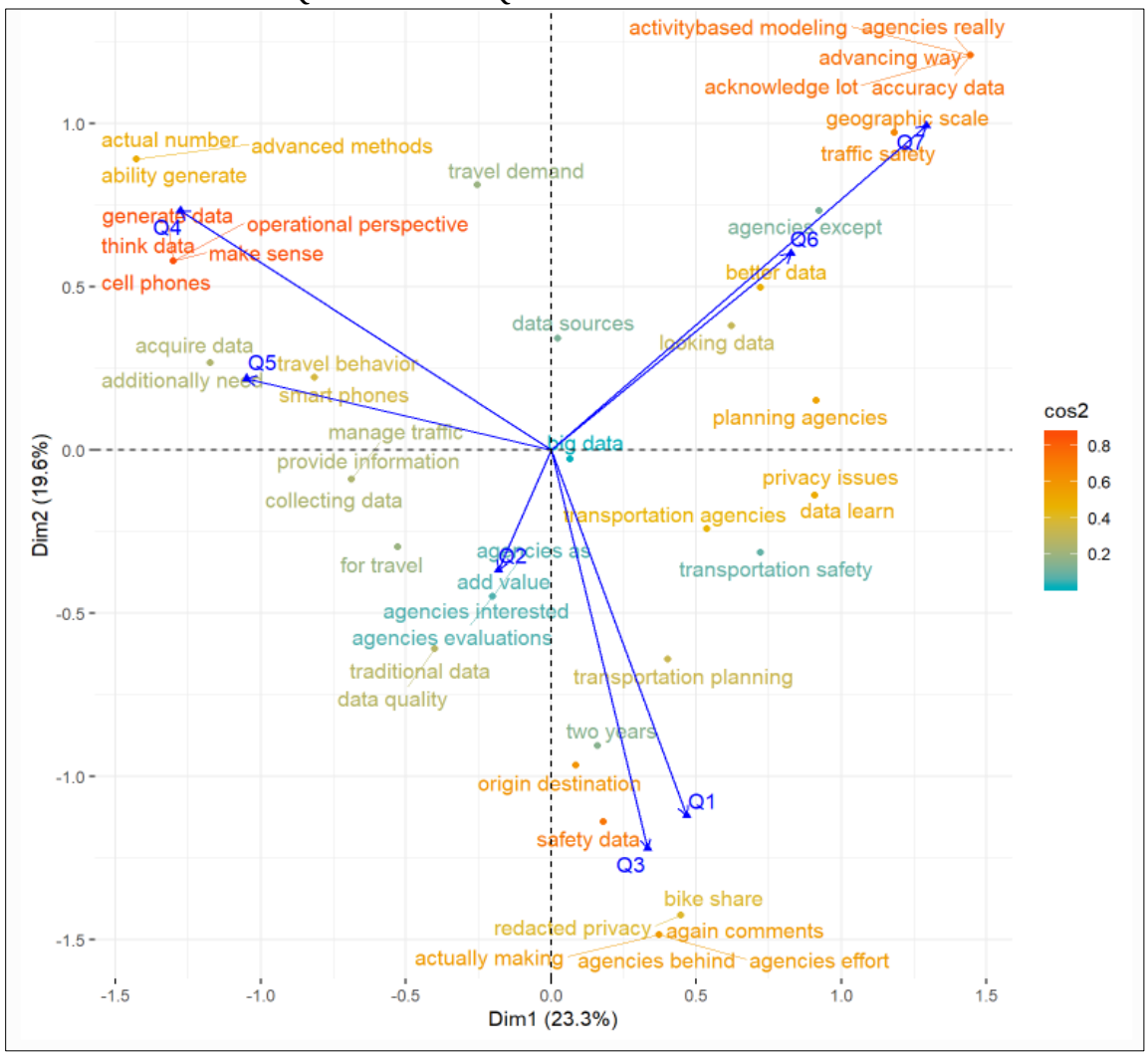

Figure 3 MCA plot showing clusters of bigrams related to interview questions.

DISCUSSION

Analysis of interview data from big data experts revealed four clusters of meaning related to challenges with big data. In the order of a typical transportation safety workflow, these include big data sources (Cluster 4), quality (Cluster 1), analysis (Cluster 2), and implementation (Cluster 3). This discussion focuses on the implications of biases in big data for transportation for research and practice.

Sources of Big Data

Previous research showed that big data could provide new information about all travel modes, but the data reflects electronic signatures of specific devices or electronic records, rather than the person making the trip. In the case of mobile phone data used for origin-destination trip analysis, this could mean that an individual phone shared by two people or used by one person but sometimes left at home. These effects get obscured when data is aggregated across populations, and therefore, the accuracy may not be comparable against more traditional data sources such as travel surveys. Our analysis of interviews with experts on big data showed that varying sources of big data likely have a different range of impacts on accuracy for a given purpose. Trip data from bike-share docking stations could reasonably be expected to differ from dockless trips, which have more flexible trip ends. Furthermore, changes in privacy rules and policies can impact the spatial or temporal accuracy of big data transportation sources and need to be considered in both transportation research and practical applications. Big data sources serve a primary role in how human trips are represented.

Quality of Big Data

Analysis of interview results related showed several ways which experts work to improve the quality of big data for transportation safety, including improving the accuracy of data products, geographic scale, and alignment of big data quality to transportation planning challenges. Expert practitioners challenge big data companies' methods and give them ideas to improve data collection approaches and processing algorithms, which may serve to improve the quality of big data resources available to others. Similarly, researchers evaluate and publish study results on quality that show quality issues that may be inherent to data source or may be mitigated through methods such as fusion with more accurate data from surveys or observational data. Since big data often reflects transportation actions as recorded through devices, it may help improve data quality by spanning larger geographic scales than traditional travel studies. Big data reflected in mobile data is not sensitive to language or political boundaries in the same ways as travel surveys, for instance, and could improve the quality of data for studies of border regions. Lastly, interview analysis showed an alignment of quality issues with planning agencies, which continually work to improve data for agency purposes. In this way, transportation agencies are chiefly responsible for the appropriate application of big data resources to a given need, which ultimately affects the quality of results.

Analysis of Big Data

Interview data show that the analysis of big data remains a challenge in practice and research. Advanced methods used by researchers are not available to all in transportation agencies, though corporate data products are serving as mediators to improve access through data dashboards and customized transportation safety analysis. Results serve as an additional resource to improve

travel demand and therefore, the common numerator of safety metrics (e.g. crashes per kilometers traveled).

Implementing Big Data

As suggested previously, agency implementation of big data is the key opportunity for improving transportation safety through improved data resources. The cluster analysis showed the that relationship of agency implementation could include a role in data collection in addition to how they apply big data in the context of traditional data sources. As agencies' interest in big data sources grow, they could compare analysis against travel surveys to test the potential impacts of changing data sources. Conversely, they could combine data sources, perhaps with the assistance of researchers or consultants, to balance the benefits of geographic and temporal scale of big data with the behavioral representation of traditional data sources.

CONCLUSIONS

A growing number of transportation researchers have been facing the influx of big data while solving the research problems. There is a need to raise awareness within the transportation profession to understand the biases and issues of big data for the improvement in the performance areas: improved safety, increased efficiency, and enhanced end-user experiences. This study aimed to mitigate the research gap by using an innovative data mining technique. In sum, the findings show that transportation organizations have four issues of crucial concern when thinking about biases in big data, and how to mitigate:

1. Keep transportation experts and public central in determining the right goals and metrics to evaluate transportation safety.
2. Develop new methods to relate big data to the total population needed for transportation safety.
3. Leverage big data to answer intractable questions.
4. Work ahead to transfer emerging knowledge to future problems.

The current study is not without limitation. The number of survey participants is limited in this study. There is a need for a comprehensive study with a large number of survey participants from a wide array of transportation safety professionals with a wider variety of bug data issues and concerns.

ACKNOWLEDGMENTS

Data collection for this study was supported by the Safety through Disruption University Transportation Center through the project Sources and Mitigation of Bias in Big Data for Transportation Safety.

AUTHOR CONTRIBUTION STATEMENT

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

REFERENCES

1. Shearmur, R. Dazzled by Data: Big Data, the Census and Urban Geography. Urban Geography, No. August 2015, 2015, pp. 1–4.
2. Desouza, K. C., and K. L. Smith. PAS Report 585 Big Data and Planning. American Planning Association, Chicago: IL, 2016.
3. Henke, N., J. Bughin, M. Chui, J. Manyika, T. Saleh, B. Wiseman, and G. Sethupathy. The Age of Analytics : Competing in a Data-Driven World, 2016.
4. Greg P. Griffin, Megan Mulhall, and Chris Simek. Sources and Mitigation of Bias in Big Data for Transportation Safety (02-026) - Safe-D UTC. https://dataverse.vtti.vt.edu/dataset.xhtml?persistentId=doi:10.15787/VTT1/KRTX66. Accessed Dec. 14, 2018.
5. Adams, W. C. Conducting Semi-Structured Interviews. In Handbook of Practical Program Evaluation, John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 492–505.
6. Kinawy, S. N., M. Nik Bakht, and T. E. El-Diraby. Mismatches in Stakeholder Communication: The Case of the Leslie and Ferrand Transit Stations, Toronto, Canada. Sustainable Cities and Society, Vol. 34, 2017, pp. 239–249.
7. Chatfield, A. T., and C. G. Reddick. All Hands on Deck to Tweet #sandy: Networked Governance of Citizen Coproduction in Turbulent Times. Government Information Quarterly, Vol. 35, No. 2, 2018, pp. 259–272.
8. Espinoza Espinoza, S. E., A. E. Vivaceta De la Fuente, and C. A. Machuca Contreras. Valparaiso's 2014 Fire: Evaluation of Environmental and Epidemiological Risk Factors During the Emergency Through a Crowdsourcing Tool. Disaster Medicine and Public Health Preparedness, Vol. 11, No. 2, 2017, pp. 239–243.
9. Sadri, A. M., S. Hasan, S. V. Ukkusuri, and M. Cebrian. Crisis Communication Patterns in Social Media during Hurricane Sandy. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2672, 2018, pp. 125–137.
10. Chen, C., J. Ma, Y. Susilo, Y. Liu, and M. Wang. The Promises of Big Data and Small Data for Travel Behavior (Aka Human Mobility) Analysis. Transportation Research Part C: Emerging Technologies, Vol. 68, 2016, pp. 285–299.
11. Toole, J. L., S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González. The Path Most Traveled: Travel Demand Estimation Using Big Data Resources. Transportation Research Part C: Emerging Technologies, Vol. 58, 2015, pp. 162–177.
12. Vij, A., and K. Shankari. When Is Big Data Big Enough? Implications of Using GPS-Based Surveys for Travel Demand Analysis. Transportation Research Part C, Vol. 56, 2015, pp. 446–462. https://doi.org/10.1016/j.trc.2015.04.025.
13. Chen, X., M. Zahiri, and S. Zhang. Understanding Ridesplitting Behavior of On-Demand Ride Services: An Ensemble Learning Approach. Transportation Research Part C, Vol. 76, 2017, pp. 51–70. https://doi.org/10.1016/j.trc.2016.12.018.
14. Mondschein, A. Five-Star Transportation: Using Online Activity Reviews to Examine Mode Choice to Non-Work Destinations. Transportation, Vol. 42, No. 4, 2015, pp. 707–722. https://doi.org/10.1007/s11116-015-9600-7.
15. Bergman, C., and J. Oksanen. Estimating the Biasing Effect of Behavioural Patterns on Mobile Fitness App Data by Density-Based Clustering. In Geospatial Data in a Changing World (T. Sarjakoski, M. Y. Santos, and L. T. Sarjakoski, eds.), Springer, Cham, pp. 199–218.

16. Silge, J., and D. Robinson. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. 2016. Journal of Open Source Software, 1(3), 2016.
17. Das, S., and X. Sun. Factor Association with Multiple Correspondence Analysis in Vehicle–Pedestrian Crashes. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2519, No. 1, 2015, pp. 95–103.
18. Griffin, G. P., K. Nordback, T. Götschi, E. Stolz, and S. Kothuri. Monitoring Bicyclist and Pedestrian Travel and Behavior. Transportation Research Board, Washington, D.C., 2014.
19. Erhardt, G. D., and A. Dennett. Understanding the Role and Relevance of the Census in a Changing Transportation Data Landscape, 2017.
20. Kwan, M.-P. The Uncertain Geographic Context Problem. Annals of the Association of American Geographers, Vol. 102, No. 5, 2012, pp. 958–968.