TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Article

Transportation Research Record
1–11

DOI: 10.1177/0361198118788464
journals.sagepub.com/home/trr
SAGE

Vehicle Consumer Complaint Reports Involving Severe Incidents: Mining Large Contingency Tables

Subasish Das[1], Abhisek Mudgal[1], Anandi Dutta[2], and Srinivas R. Geedipally[1]

Abstract

According to 2010–2014 Fatality Analysis Reporting System (FARS) data, nearly 6.35% of fatal crashes happened as a result of vehicles' pre-existing manufacturing defects. The National Highway Traffic Safety Administration's (NHTSA) vehicle complaint database incorporates more than 1.37 million complaint reports (as of June 1, 2017). These reports contain extended information on vehicle-related disruptions. Around 5% of these reports involve some level of injury or fatalities. This study had two principal objectives, namely (1) perform knowledge discovery to understand the latent trends in consumer complaints, and (2) identify clusters with high relative reporting ratios from a large contingency table of vehicle models and associated complaints. To accomplish these objectives, 67,201 detailed reports associated with injury or fatalities from the NHTSA vehicle complaint database were examined. Exploratory text mining and empirical Bayes (EB) data mining were performed. Additionally, this study analyzed five years (2010–2014) of FARS data to examine the research findings. Results show that major vehicular defects are associated with air bags, brake systems, seat belts, and speed controls. The EB metrics identified several key '*vehicle model with major defect*' groups that require more attention. This study demonstrates the applicability of consumer complaints in identifying major vehicular defects as well as key groups of '*vehicle model with major defect*.' The findings of this study will provide a significant contribution to the reduction of crashes from vehicle-related disruptions. The research presented in this paper is crucial given the ongoing advancement of connected and automated vehicle technologies.

Traffic safety is a major concern because of the economic and social costs of traffic crashes. According to National Safety Council (NSC), more than 40,000 people died in traffic crashes in 2016, which is the highest count since 2007. This count is a 6% rise from 2015 and is up 14% from the 2014 figure. The estimated annual death rate is 1.25 deaths per 100 million vehicle miles traveled, an increase of 3% from 2015. The estimated cost of motor-vehicle deaths, injuries, and property damage in 2016 was $432.5 billion, an increase of 12% from 2015 (*1*). The costs include wage and productivity losses, medical cost, administrative prices, employer costs, and property damage costs. These statistics make Vision Zero goals difficult to attain. Countless research efforts on conventional crash data have been conducted to understand better the factors that influence the frequency and severity of crashes and to provide more effective safety-related countermeasures. However, the number of crashes is still at an unacceptable level, which is evident by the sharp rise in fatalities in the last two years. This shows that, in addition to current efforts, research needs to be conducted with additional resources and in newer directions. Newer sources of data (for example, naturalistic driving data, U.S. census demographic data, American Community Survey, National Highway Traffic Safety Administration [NHTSA] consumer complaints data) can fill the current research gap.

According to 2010–2014 Fatality Analysis Reporting System (FARS) data, in nearly 6.35% of fatal crashes, vehicles' pre-existing defects were involved (*2*). A limited number of studies were conducted to identify risk factors from vehicle defects. An in-depth analysis of the vehicle defect-related issues would help in understanding the association between vehicle condition and automotive safety. The NHTSA vehicle complaint database (*3*)

[1]Texas A&M Transportation Institute, Texas A&M University System, College Station, TX
[2]Computer Science and Engineering, Texas A&M University System, College Station, TX

Corresponding Author:
Address correspondence to Subasish Das: s-das@tti.tamu.edu

incorporates more than 1.37 million complaint reports (as of June 1, 2017). Around 5% of these reports involve some level of injury or fatalities. The complaints file contains all safety-related defect complaints received by NHTSA since January 1, 1995. This research sought to better understand the implications of the correlations present across injury outcomes in the vehicular complaint database. The current study design has two major goals: (1) conduct knowledge discovery to identify the latent trends from consumer complaints, and (2) detect clusters with high relative reporting ratios from a large contingency table of vehicle models and associated complaints. To accomplish the research goals, this study performed exploratory text mining and empirical Bayes (EB) data mining.

Earlier Work and Research Context

A limited number of studies exist where the relationship between vehicle defects and associated safety implications are studied. Moodley and Allopi performed a random survey of parked vehicles to investigate if the vehicles were adequately safe for driving (*4*). They conducted the observational study without touching the vehicles, and found that 24% of them had tire defects and 11% of them had defective lights. In order to devise methods to evaluate vehicles for safety it may be important to find if vehicle defects are contributing factors to crashes and, if so, which defects are most prominent. In the past, defects in the brake system or tires have been cited as the most frequent cause of crashes (*5*). Wolf found in 1962 that 6% of crashes in trucks were as a result of mechanical failure, and the Road Research Laboratory in the United Kingdom (UK) found that vehicle defects caused at least 18% of crashes (*6*). Researchers have found through crash reconstruction that vehicle defects contributed to 9% of the crashes (*7*).

More recently, Cuerden et al. investigated the effects of vehicle defects in traffic crashes (*8*). This study estimated that vehicle defects are likely to be a contributing factor in around 3% of crashes in the UK. The findings also showed that reducing the frequency of testing of newer vehicles is likely to have adverse road safety consequences. Schoor and Niekerk attempted to establish the contribution of mechanical failures in traffic crashes (*9*). Data obtained from accident response units indicate that tires and brakes were the main contributors to mechanical failures resulting in crashes. Barry et al. found that the air bag is effective for the survival of vehicle occupants in case of a crash. This implies that in the event of a crash, occupants in vehicles with a defective air bag system would have lesser protection and a higher chance of trauma. Barry et al. also found that the effectiveness of the air bag becomes less clear when seat belts are employed (*10*).

The NHTSA complaint data have great potential for identifying how a particular vehicle defect is related to crash outcome, as data regarding many instances of vehicle defects are available. With the NHTSA vehicle consumer complaint data, the frequency of the combination of vehicle defects and crash severity (no injury, less severe injury, or fatality) can be studied. In recent years, several studies have incorporated text mining in transportation engineering research: consumer complaint analysis (*11*, *12*), understanding public sentiments on safety enhancement and bike sharing (*13*, *14*), identifying research trends from transportation engineering conference papers and journals (*15–19*), crash narrative investigation (*20–23*), text analytics using social media mining (*24–27*), and teen driver survey (*28*). Ghazizadeh et al. used text mining and latent Dirichlet analysis to examine the hidden trends in the NHTSA complaint database. The analysis was done using data from January 1995 to June 2012 over a small sample of 2000 comments (*11*).

This paper is unique in several aspects: (1) the complete data set of vehicle defect complaints involving some level of injury or fatalities is used instead of a sample, (2) complaint data are contrasted with FARS data to ensure complaint data is not biased because of under-reporting, and (3) the analysis goes beyond that of Ghazizadeh et al. (*11*) by analyzing a large contingency table comprising vehicle defects and occurrence of crashes.

Methodology

To accomplish the research goals, the current study is divided into two major tasks: (1) perform descriptive statistics and text mining to determine latest trends in complaint texts and associated data, and (2) conduct EB data mining to identify key groups of "vehicle model with complaint" from a large contingency table with 67,201 complaint reports.

Descriptive Statistics and Text Mining

The data used in this research are based on consumer provided complaints collected by NHTSA. As of June 1, 2017, this database incorporates more than 1.37 million complaint reports in structured form with 33 variables. Around 5% of these reports involve some level of injury or fatalities. The complaints file contains all safety-related defect complaints received by NHTSA since January 1, 1995. Table 1 lists the descriptions of the final variables. The number of reports associated with some level of injury or deaths is 67,201 (around 4.9% of total complaint entries). However, approximately 25% of those complaint reports with injury or death are not associated with traffic crashes.

Table 1. Description of the Variables

Code	Description	Categories
MAKETXT[a]	Vehicle/equipment make	–
MODELTXT[b]	Vehicle/equipment model	–
YEARTXT[c]	Model year	9999 if unknown
CRASH	Was vehicle involved in a crash	"Yes" or "No"
INJURED	Number of persons injured	–
DEATHS	Number of fatalities	–
COMPDESC	Specific component's description	–
CDESCR	Description of the complaint	–
CMPL_TYPE[d]	Source of complaint code	CAG = consumer action group CON = forwarded from a congressional office DP = defect petition EVOQ = hotline VOQ EWR = early warning reporting INS = insurance company IVOQ = NHTSA web site LETR = consumer letter MAVQ = NHTSA mobile app MIVQ = NHTSA mobile app MVOQ = optical marked VOQ RC = recall complaint RP = recall petition SVOQ = portable safety complaint form (pdf) VOQ = NHTSA vehicle owners questionnaire

[a,b,c]Combined together for Vehicle Model for EB data mining.

[d]In the final dataset, the complaint code is divided into three broader groups: H_VOQ = hotline vehicle owner's questionnaire (VOQ) [original code: EVOQ]; NHTSA = NHTSA web site [original code: IVOQ]; other = other complaint codes [original codes: CAG, CON, DP, EWR, INS, LETR, MAVQ, MIVQ, MVOQ, RC, RP, SVOQ, VOQ].

Figure 1 shows key risk factors from three data sources: (1) dataset 1 [NHTSA: FI]: 1995–2017 NHTSA complaint data with severities (both crash and no crash incidents), (2) dataset 2 [NHTSA: FI + Crash]: 1995–2017 NHTSA complaint data with severities (only traffic crash incidents; subset of dataset 1), and (3) dataset 3 [FARS]: 2010–2014 FARS data. NHTSA provides two columns on complaints. "COMPDESC" provides short and specific problem-related complaints, and "CDESCR" provides a detailed description of the complaints. Risk factors from NHTSA were performed by the sentence stemming (identifying the key theme) method. In 2010–2014 FARS data (dataset 3), 224,319 vehicles were involved in fatal crashes. Of these, 14, 222 vehicles (around 6.35%) had vehicular defects. FARS uses a variable named "MFACTOR" to classify vehicle defects. NHTSA database shows that "*air bag*" is the most dominant attributor to crash. FARS does not include "*air bag*" in "MFACTOR" classification. It has separate coding for "air bag" deployment. FARS shows that around 38% of fatal crashes involve an "*air bag*" deployment issue, which is similar to what the NHTSA complaint data show. "Brake system" is the dominant attribute in all three data sources. "Tire/Wheels" is the second most significant factor in FARS crashes, which is also among the top five factors in dataset 1 and dataset 2. The top five risk factors ("*air bags*," "*brake system*," "*vehicle speed limit*," "*seat belts*," *and* "*tires/wheels*") contribute to 62% of all reports in dataset 1. Dataset 2 shows that top five risk factors ("*air bags*," "*brake system*," "*seat belts*," "*seat*," and "*vehicle speed limit*") are included in 77% of all reports. This is corroborated by Schoor and Niekerk, who found that defects in tires and brake system were the two most dominant factors that resulted in mechanical failure leading to a crash (*9*).

Dataset 1 and dataset 2 require further exploration for a clear understanding of the percentage distribution of the key categories. These two databases (involve severities irrespective of crash or no crash involvement) have five major contributing factors: airbag, seat/seat belts, braking system, speed control, and tires/wheels (as listed in Table 2). For airbag-related complaints, two major groups (airbags, and airbags: frontal) contribute nearly 95% of the complaints. Similarly, the top three categories (tires, tires: tread/belt, and wheels) comprise 95% of the tire or wheel-related complaints. Vehicle speed control and accelerator pedal are the dominating categories in the speed control variable. Hydraulic brake failures are the key contributing category in braking system complaints. Seats and seat beats contribute nearly 60% of the

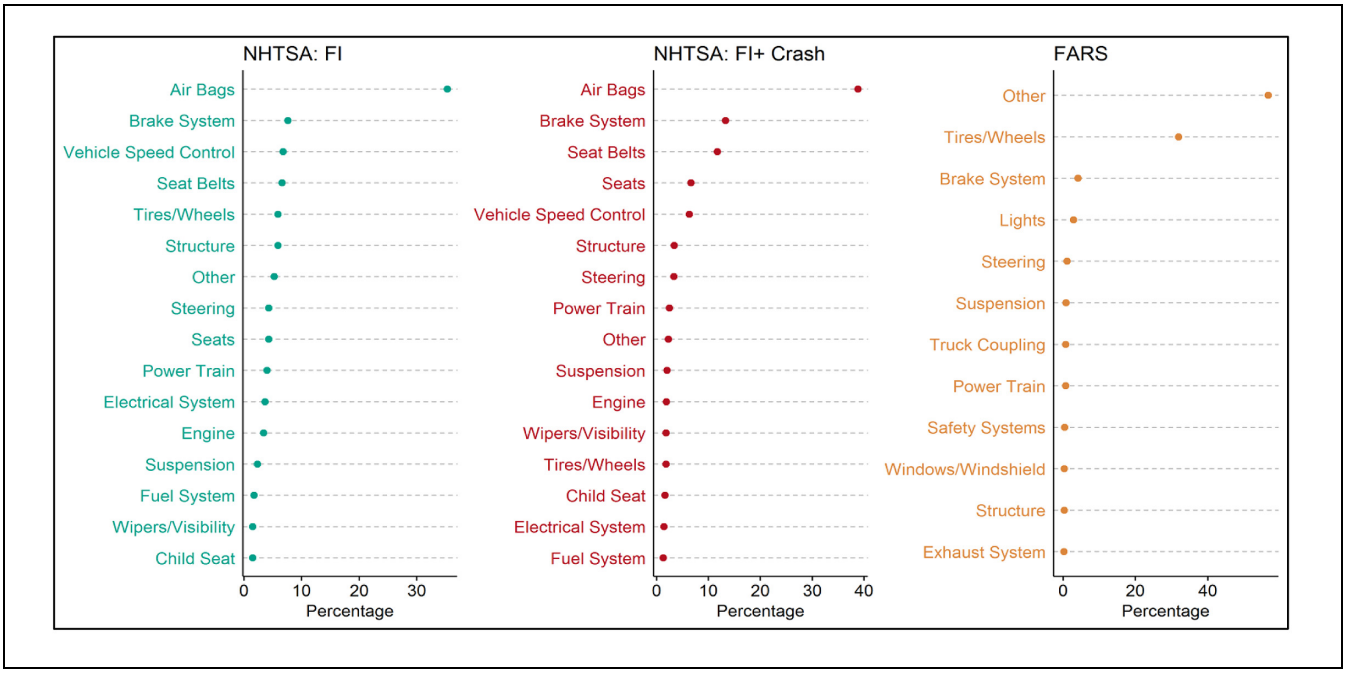

Figure 1. Comparison of major risk factors between NHTSA complaints and FARS.

Table 2. Top Categories under Major Complaints in NHTSA Compliant Databases

Complaint category	Percentage
Major complaint: airbag	
Air bags: frontal	47.2
Air bags	47.0
Air bags: side/window	4.7
Air bags: frontal: sensor/control module	0.7
Air bags: frontal: driver side inflator module	0.4
Major complaint: seat/seat belts	
Seat belts	31.9
Seats	28.2
Seats: front assembly: recliner	15.8
Seat belts: front: anchorage	13.2
Seat belts: front: retractor	10.8
Major complaint: tires/wheels	
Tires	52.3
Tires: tread/belt	30.4
Wheels	12.2
Tires: sidewall	3.5
Wheels: lugs/nuts/bolts	1.6
Major complaint: speed control	
Vehicle speed control	84.8
Vehicle speed control: accelerator pedal	8.8
Vehicle speed control: linkages	3.0
Vehicle speed control: cruise control	2.4
Vehicle speed control: cables	1.0
Major complaint: brakes	
Service brakes, hydraulic: antilock	41.5
Service brakes	25.9
Service brakes, hydraulic	18.9
Service brakes, hydraulic: foundation components	11.2
Parking brake	2.4

seat/seat belt complaint cases. Front seats and front seat belts comprise 40% of the seat/seat belt complaint cases. Future studies can focus on any of these major factors to identify more insights from the complaint database.

A comparison word cloud helps in studying the differences or similarities between two or more corpora by presenting the word cloud of each against the other (*29*). Consider that $p_{x,y}$ is the rate at which word x occurs in document y, and p_y is the average rate across documents $\left(\frac{\sum_y p_{x,y}}{n}\right)$, where n is the number of documents. In comparison clouds, the size of each word is mapped to its maximum deviation ($max_x(p_{x,y} - p_y)A = \pi r^2$) and its angular position is determined by the document in which that maximum occurs (*15*). For this analysis, the research team used complaint texts by broadly dividing them into two groups: crash, and no crash. Figure 2*a* shows comparison cloud for the corpora developed with vehicle models, and Figure 2*b* shows the comparison cloud for the corpora developed with vehicle complaints. The centrality of the word positioning indicates the presence of a centered word in both corpora. In Figure 2*a*, "*unknown*" and "*Toyota*" are dominant and central words in the "*No Crash*" group. This indicates that these two words are present in both groups, but the probability of presence in "*No Crash*" group is higher than in "*Crash*" group data (word frequencies were normalized before generating the word cloud).

In Figure 2*b*, dominant and central words are "*air*," "*bags*," "*speed*," "*service*," "*frontal*," and "*structure*." Five of these words fall in the "*No Crash*" group. Only

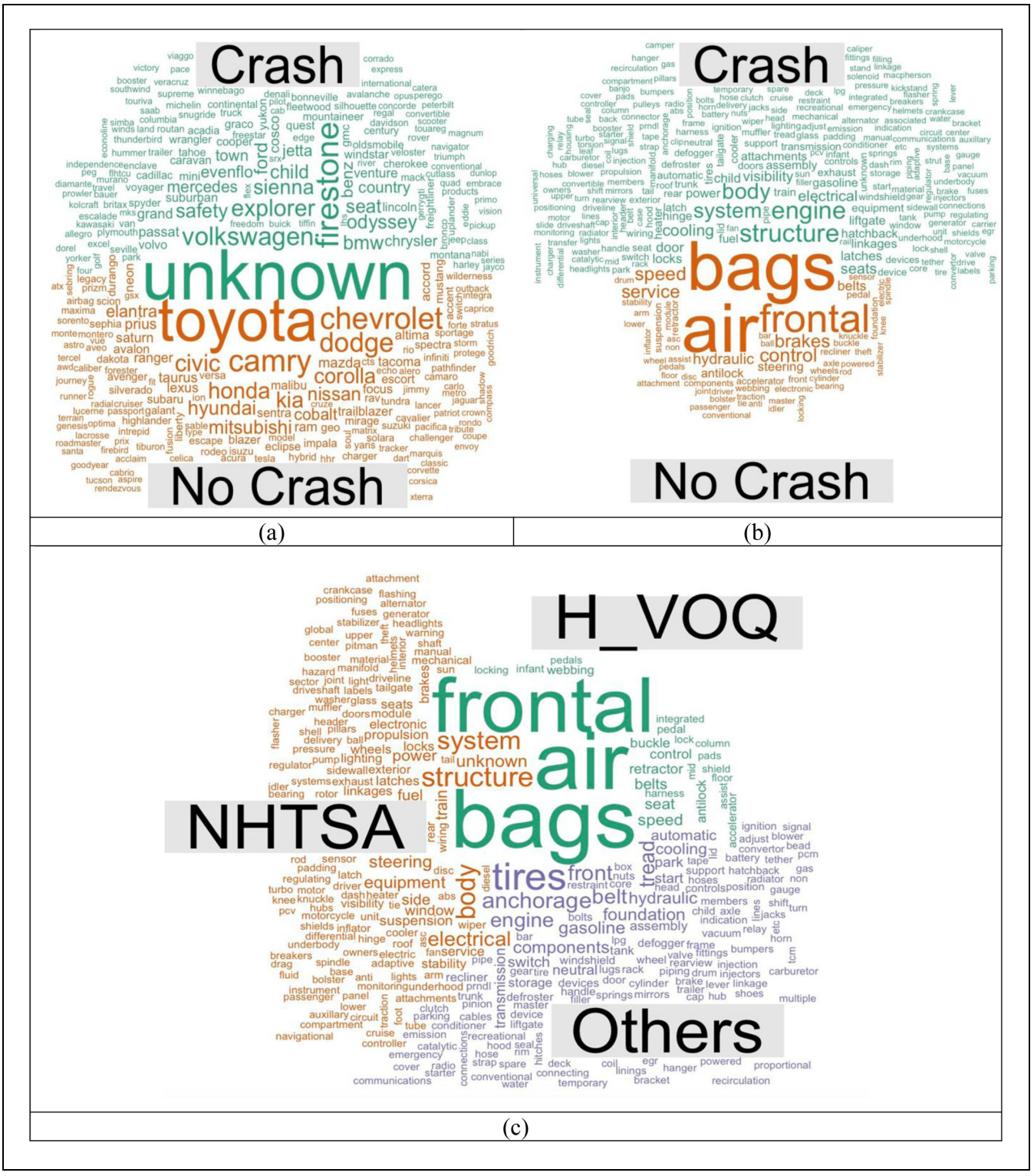

Figure 2. Comparison word clouds: (*a*) vehicle model; (*b*) customer complaints; (*c*) customer complaints clustered by complaint resources.

the word "structure" is inside the cloud boundary of "Crash"-related words. The findings indicate that the word group "*air bags*" is common in both corpora, but the probability of presence in "*No Crash*" is relatively higher. It also shows that body or structure, engine system, seat, and visibility are the dominant risk factors in

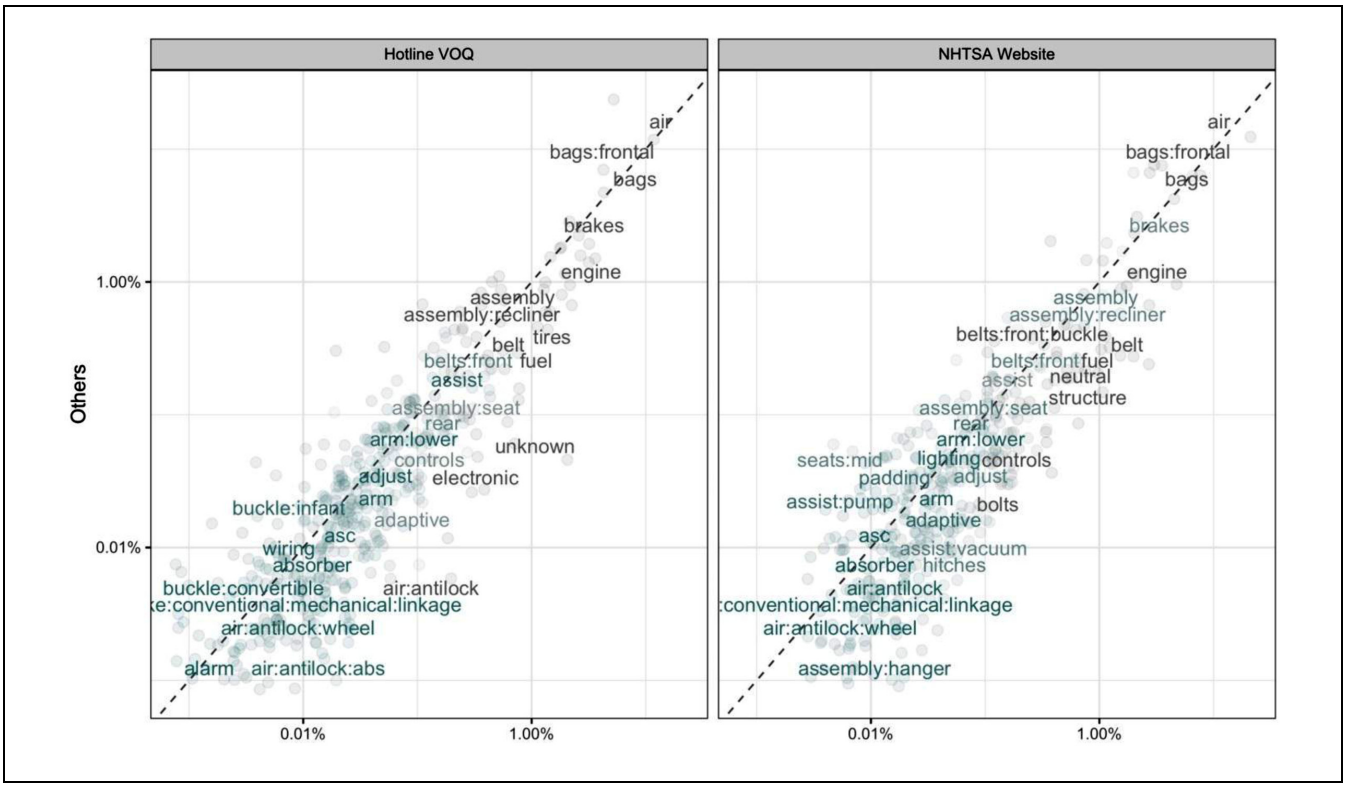

Figure 3. Comparing word frequencies of corpora based on complaint code.

the "Crash" group. Figure 2*c* shows comparison cloud for three groups based on the major complaint code. "*Air*," "*bag*," "*frontal*," "*tires*," "*structure*," "*belt*," "*body*," and "*electrical*" are the words with larger fonts. Most of these larger font words are in the center that represents their presence in all of the clusters.

Another approach is to examine the word frequencies of each corpus (as shown in Figure 3). Words that are close to the line in these plots have similar frequencies in both sets of texts, for example, in both "Hotline VOQ vs. Others" and "NHTSA website vs. Others" texts ("*air*," "*front bags*," "*bags*," "*brakes*" at the upper frequency end). Words that are far from the line are words that are found more in one set of texts than another is. For example, "*unknown*" word is a little far away from the center line in the "*Hotline VOQ vs. Others*" group (this word is not visible in "NHTSA website vs. Others" texts). The findings show that "Hotline VOQ" are associated with emergent defects like "*tire*," "*seat*," and "*engine*," and "*NHTSA website*" complaints are associated with the body or structural ("*structure*," "*bolts*," and "*lighting*") complaints.

Empirical Bayes Data Mining

The final NHTSA complaint database is a large contingency table. The association between complaints and vehicle model were analyzed using the EB geometric mean (EBGM) method, which is an association rules learning (ARM) method. ARM is a rule-based machine learning method that is used for discovering interesting connections between variables in a large frequency table. ARM was widely used in transportation safety research (*30–35*). Interested readers can consult for a short overview on the concepts of ARM.

The EBGM method helps to identify how significant is the frequency of an outcome given that it has a high occurrence. For example, it would help identify how significant is the occurrence of a certain defect in the vehicle given its model and model year. Complaint data may be affected by reporting bias, and therefore the association of complaints and vehicle model may not reflect the actual situation. However, the result of this analysis will create direction to further investigate the vehicle model and the occurrence of the given corresponding defects.

The analysis presented in this paper is performed using open source R package *openEBGM*, which estimates the significance of unusually large cell counts in large, sparse contingency table (*36*, *37*). This method can be used to find unusually high reporting rates of vehicle defects associated with the vehicle model.

Contingency tables can be analyzed using relative reporting ratio *RR* that is expressed by $^{N}/_{E}$, where N is the real count for the cell in the table and E is the expected frequency with an assumption of independency

of rows and columns. However, *RR* tends to overstate the significance of a count when the real count is small. The *EBGM* method can be thought as a Bayesian approach to *RR* since *EBGM* adds Bayesian shrinkage corrections to *RR* such that high values that result from low E and low N are no longer identified as significant. Small N could occur by chance and N/E becomes high while it is not significant. The *EB* approach shrinks large *RR* toward 1 (the null hypothesis value) when N is small. The shrinkage is smaller for large counts and becomes almost negligible for very large counts. Therefore, the *EB* approach provides more stable results as compared with the *RR* measurement. The *EB* approach also accounts for sampling variation.

Theory. Consider a contingency table with cell count N_{ij} that follows a Poisson distribution with unknown mean μ_{ij}. Here i and j represent row and column numbers respectively. To evaluate cell counts N_{ij}, baseline or null hypothesis frequency, E_{ij} for each cell should be defined along with a statistical measure to rank cells according to their significance. A stratification variable, k is also defined to utilize the within group correlation. Consider the following group of equations defining E_{ij}:

$$N_{ij} = N_{ij.} = \sum_k N_{ijk} \tag{1}$$

$$N_{i.j} = \sum_j N_{ijk} \tag{2}$$

$$N_{.jk} = \sum_i N_{ijk} \tag{3}$$

$$N_{..k} = \sum_i \sum_j N_{ijk} \tag{4}$$

$$E_{ij} = \sum_{jk} \frac{N_{i.k} N_{.jk}}{N_{..k}} \tag{5}$$

E_{ij} is the expected count assuming that the rows and columns variables are independent conditional on stratification variable. Consider $\lambda_{ij} = \frac{\mu_{ij}}{E_{ij}}$ as the decision statistics for evaluating unusually large frequencies for each cell in the contingency table. The statistic λ is drawn from a mixture of two gamma densities with (α_1, α_2) and (β_1, β_2) as the shape parameters and rate parameters and P as proportion constant of the two densities. Prior probability density for λ is given by $\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, P = P \times g(\lambda; \alpha_1, \beta_1) + (1 - P) \times g(\lambda; \alpha_2, \beta_2)$ where g is gamma distribution. The resultant density is a five parameter probability density function. Assuming that $E, \alpha_1, \beta_1, \alpha_2, \beta_2, and\ P$ are known, the marginal distribution of N follows a mixture of two negative binomial distributions. Consider $EBlog2_{ij}$, the posterior expectation of $log2(\lambda_{ij})$ as the Bayesian version of *RR*. It can be derived in the following form:

$$E[\log(\lambda)|N = n] = Q_n[\psi(\alpha_1 + n) - \log(\beta_1 + E)] + (1 - Q_n)[\psi(\alpha_2 + n) - \log(\beta_2 + E)] \tag{6}$$

where ψ is digamma function, the derivative of $\log[\Gamma(.)]$, and Q_n is the posterior probability that λ came from the first component of the mixture, given $N = n$.

For large values of E and N/E, ψ(n) approaches $\log(n)$, then $EBlog2_{ij}$ or $EB[\log(\lambda)|N = n]$ approaches $\log(\alpha + N) - \log(\beta + N)$, *or* $\log\left(N/E\right)$or $\log(RR)$. However, when E and N/E are not large, $EBlog2_{ij}$ would shrink toward a lower value. Finally, to make $EBlog2_{ij}$ of the same scale as *RR* and obtain a value that is easily comparable and interpretable, DuMuchel computes EBGM ($EBGM_{ij}$), the geometric mean of $EBlog2_{ij}$. It is given by the equation $EBGM_{ij} = 2^{EBlog2_{ij}}$. In EB methods, choice of prior parameters is obtained from the data themselves. In this methodology, $\alpha_1, \beta_1, \alpha_2, \beta_2$, and P are obtained by maximizing the likelihood of these parameters taken together. This likelihood is the marginal distribution of N_{ij}.

Model and Estimates. A significantly large proportion of vehicle owners' complaints were identified using an R package *openEBGM*, which uses an EB data mining approach developed by DuMouchel. The EBGM method provides information about the significance of frequency of a given combination of effect and response in the contingency table. It works well even with very large but sparse tables. The final database contains 67,201 complaint records. There are 37,312 unique combinations of "Variable 1-Variable 2". Table 3 lists the variable used for the final analysis.

The variables are strata used for the modeling:

- Variable 1: Vehicle model (code developed by combining MAKETXT, MODELTXT, YEARTXT)
- Variable 2: Vehicle defect (COMPDESC)
- Strata 1: Crash involvement (CRASH)
- Strata 2: Complaint code (CMPL_TYPE)

The first step is to conduct actual counts of each vehicle model and defect combination, expected counts (E) under the row/column independence assumption, *RR*, and proportional reporting ratio (*PRR*). Stratification is beneficial in controlling confounding variables. For instance, complaints are affected by different complaint submission methods or whether the complaints were associated with crashes or not. Stratification will affect E and *RR*, but not *PRR*. The Es are calculated by summing the expected counts from every stratum. Ideally, each stratum should contain several unique reports to ensure

Table 3. Top 10 Combination Groups with High PRRs

Vehicle model/component model	Complaint	E	RR	PRR	EBGM
Accubuilt-Limousine (2001)	Wheels: rim	0.0018	1110.1	1080.94	1.16
Aai Motorsports-Combination Lamps (unknown)	Equipment: motorcycle: helmets	0.0020	1009.18	1015.44	1.16
Acura-Acura (1999)	Seats: mid/rear assembly	0.0005	2018.36	957.44	1.06
A-1 Alternative Fuel Syst-Cng Fuel System (unknown)	Equipment: electrical	0.0043	231.46	744.68	1.06
Ace Electric-Ace (unknown)	Equipment: electrical	0.0043	231.46	744.68	1.06
Acura-Acura (1996)	Suspension: front: control arm: lower ball joint	0.0015	648.17	609.28	1.06
Ac Delco-18M615 (unknown)	Electrical system: battery	0.0029	347.19	478.71	1.06
Accessory-Helmet (unknown)	Equipment: other: labels	0.0046	217.95	462.19	1.06
Accubuilt-Limousine (2001)	Parking brake	0.0067	300.03	331.77	1.16
Ac Delco-45A3073 (unknown)	Equipment: mechanical	0.0073	137.05	272.43	1.06

good estimates of E. The actual counts (N) and expected counts (E) are used to estimate the hyperparameters of the prior distribution. A large contingency table will have many cells, resulting in computational difficulties for the optimization routines needed for estimation. Data squashing transforms a set of 2-dimensional points to a smaller number of 3-dimensional points. The idea is to reduce a large number of points (N, E) to a smaller number of points (N_k, E_k, W_k), where $k = 1, \ldots, M$ and W_k is the weight of the kth "superpoint". To minimize information loss, only points close to each other should be squashed. To determine the hyper parameters efficiently, around 10% of the reports were used for squashed data. The hyper parameters are estimated by minimizing the negative log-likelihood function. The optimized hyper parameters are $(\alpha_1, \beta_1, \alpha_2, \beta_2, P) = (2.32 \times 10^{-08}, 0.1052, 10.0441, 9.9493, 0.3362)$. EBGM is a measure of central tendency of the posterior distributions $\lambda_{ij}|N = n$. Scores much larger than 1 indicate vehicle model and complaint event pairs that are reported at an unusually high rate. Table 3 lists the top 10 combination groups with high PRRs.

EBGM is the antilog of the mean of the *ln*-transformed posterior distribution. It can be used as a measure of central tendency of the posterior distribution. The EBGM scores indicate adjusted estimate for the relative reporting ratio. For example, "*Ford-Explorer (2002) - Structure: Body: Hatchback/Liftgate: Hinge and attachments*" pair has an EB score of 55.14. The interpretation is that this pair occurs in the data 55.14 times more frequently than expected under the assumption of no association between the vehicle model and complaint. The 5% and 95% quantiles of the posterior distributions can be used to create two-sided 90% credibility intervals for λ_{ij}, given N_{ij}. As a result of the Bayesian shrinkage property, the EB scores are much more stable than *RR* for small counts. Table 4 lists the top 20 combination groups with high EBGM and quantiles (the list is sorted in descending order based on quantile 5% values). Figure 4 illustrates the values of Table 4 for easy interpretation.

A closer look at these vehicle model–complaint combinations can be helpful in identifying combination groups that may need further investigation. It is important to note that the EB scores find statistical associations, not causal relationships of any kind. Future investigations are required before a solid quantification of the scale of adverse effects of these combination groups can be performed. It is interesting that except one recent vehicle, all of the vehicles are older vehicles. As the new vehicles have been available to people for fewer years, these vehicle models would usually not show up in the top 20 groups. The authors envision using this database for future exploration to investigate whether there are other plausible reasons.

Conclusion

All vehicles depreciate in performance over time, which can have an adverse effect upon safety and roadway environment. Crashes are random incidents, and crashes associated with vehicle defects are more infrequent. Conventional crash reports do not include vehicle defects in detail. The NHTSA complaint database can be helpful in filling the current knowledge gap in identifying vehicle defects associated with crashes. The paper presented an analysis of a high-level evaluation of risk factors associated with vehicle defects. This analysis is based on the part of NHTSA consumer complaint database that includes records involving either injury or deaths. Some of the findings were compared with the FARS database.

The study has concluded that:

- The vehicle defects are likely to be a contributing factor for severe crashes for nearly 5% of the complaint database. For FARS, the rate is slightly higher. Vehicle defects contribute to 6.35% of all fatalities in FARS.

Table 4. Problem Groups with High EBGM Scores and Quantiles

Vehicle model/component model	Complaint	EBGM	Quantile 5%	Quantile 95%
Ford-Explorer (2002)	Structure: body: hatchback/liftgate: hinge and attachments	55.14	41.69	71.82
Jeep-Grand Cherokee (1998)	Power train: automatic transmission: park/neutral start switch	44.38	36.54	53.48
Jeep-Grand Cherokee (1999)	Power train: automatic transmission: park/neutral start switch	40.13	27.8	56.43
Jeep-Grand Cherokee (1996)	Power train: automatic transmission: park/neutral start switch	34.01	26.26	43.47
Jeep-Grand Cherokee (2001)	Power train: automatic transmission: park/neutral start switch	39.94	26.22	58.81
Jeep-Liberty (2002)	Suspension: front: control arm: lower ball joint	40.41	26.16	60.21
Hyundai-Veloster (2012)	Visibility: sun roof assembly	44.53	25.4	73.86
Jeep-Grand Cherokee (1997)	Power train: automatic transmission: park/neutral start switch	32.34	24.14	42.59
Volkswagen-Jetta (2002)	Seats: front assembly: seat heater/cooler	29.16	21.77	38.41
Pontiac-Trans Sport (1996)	Engine and engine cooling	38.14	21.69	63.94
Volkswagen-Passat (2003)	Seats: front assembly: seat heater/cooler	30.24	21.45	41.66
Volkswagen-Jetta (2003)	Seats: front assembly: seat heater/cooler	27.09	20.03	35.98
Chevrolet-S10 (2000)	Structure: body: tailgate: hinge and attachments	35.2	19.9	59.75
Toyota-Sienna (2004)	Structure: body: hatchback/liftgate: support device	25.69	18.35	35.18
Firestone-ATX (unknown)	Tires: tread/belt	21.7	16.64	27.89
Firestone-Wilderness (unknown)	Tires: tread/belt	20.91	16.29	26.52
Chevrolet-Suburban (1994)	Service brakes, hydraulic: antilock	22.08	15.87	30.05
GMC-Jimmy (1994)	Service brakes, hydraulic: antilock	15.78	23.46	33.85
Toyota-Sienna (2004)	Latches/locks/linkages: hatchback/liftgate: lock	15.7	25.04	38.39
Jeep-Grand Cherokee (1995)	Power train: automatic transmission: park/neutral start switch	15.53	22.43	31.54

- Both NHTSA and FARS show that air bags were deployed significantly (around 35%) in severe crashes.
- Tire/wheel was found as a key contributing factor in FARS data, but not in NHTSA complaints. The findings of FARS are in line with Schoor and Niekerk's study (*9*).
- The comparison cloud for vehicle models shows that some vehicle models were overrepresented in the "crash" group, although this study implies an association, not causation.
- According to NHTSA complaints, the key factors of vehicular defects include "*air bags*," "*brake system*," "*seat belts*," and "*speed control*." "*Seat belt*" and "*brake system*" were not found as key contributing factors in FARS. "*Brake system*" was found to be a key contributing factor in Schoor and Niekerk's study (*9*). These findings require further investigations.
- The NHTSA complaint database forms a large contingency table with 37,312 unique combination group records associated with vehicle model and complaint type. Simple count analysis will not consider the effect of other strata and additional local and global effects. EBGM is an appropriate tool for analyzing such database. By using EBGM, this study identifies the top 20 combination groups that were more likely to be associated with higher risks.

The findings from the NHTSA complaint data are based on reporting rates and not occurrence rates of a certain combination of vehicle and defects. This is because, unlike FARS data, NHTSA complaint data can suffer from under-reporting. Nevertheless, since the summary parameters (such as key defects associated with crashes and percentage of severe crashes involving vehicle defects) from both NHTSA complaint data and FARS data are comparable, the NHTSA data can be assumed to be a representative sample of the population of vehicle defects and crash data. Consequently, the results obtained from EB data mining provide a significant contribution to the area of safety in reducing crashes from vehicular manufacturing defects. Future studies with different subset data from NHTSA complaints (for example, tire or fuel system complaints; vehicle service life; restraint) could add value to the related research areas. With the advancement of connected and autonomous vehicle technologies, research such as that presented in this paper might be critical in maximizing the benefits of these technologies while minimizing system malfunctioning and associated injuries and fatalities.

This study aimed to play the role of a starting point in conducting research by performing both text mining and EBGM in analyzing complex datasets like the NHTSA vehicle consumer complaint database. This study can open doors for future studies on this database, which has

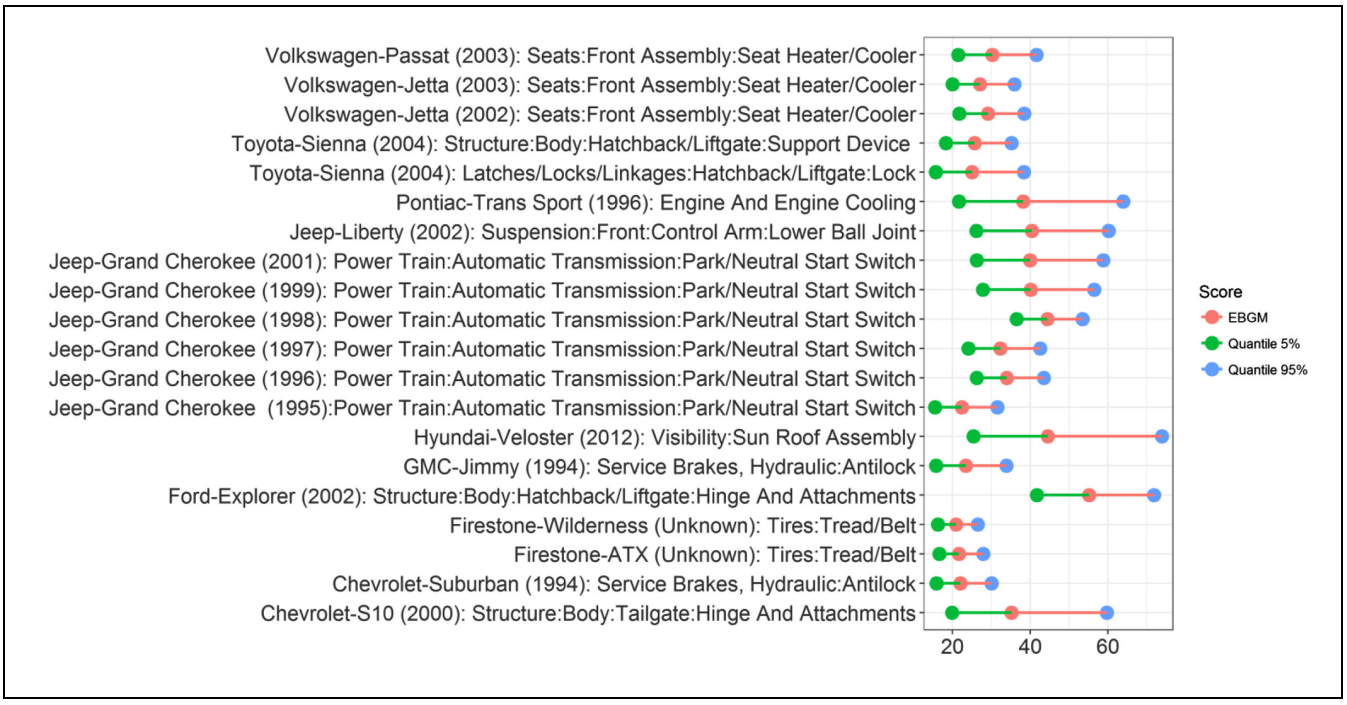

Figure 4. Vehicle/component model and complaint groups with higher EBGM.

not been explored in depth in earlier studies. The current study has several limitations. This study has not used the complete database of NHTSA vehicle consumer complaints; rather, it used complaint cases that involve at least one injury. An extended analysis on the complete database would identify the nature and distribution of various complaint categories to identify key patterns for possible solutions. Another limitation is that the database was compared with FARS that involves only fatal crashes. It would be more intuitive to compare this database with General Estimates System (GES) data of the National Automotive Sampling System, as GES represents national estimates for all levels of severities. Moreover, extended analysis on vehicle/component model and complaint groups would identify more vehicle models and associated major complaints. Current limitations in this study can be addressed in future studies.

Acknowledgments

The authors appreciate the assistance provided by Mr. Shashank Mehrotra.

Author Contributions

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

References

1. *NSC Motor Vehicle Fatality Estimates- 2016*, 2015.
2. National Highway Traffic Safety Administration. *Fatality Analysis and Reporting System*. ftp://ftp.nhtsa.dot.gov/fars/. Accessed July 31, 2017.
3. National Highway Traffic Safety Administration. *Office of Defects Investigation*. http://www-odi.nhtsa.dot.gov/downloads/. Accessed July 31, 2017.
4. Moodley, S., and D. Allopi. An Analytical Study of Vehicle Defects and Their Contribution to Road Accidents. *Proc., 27th Southern Africa Transport Conference*, Pretoria, South Africa, 2008.
5. U.S. Department of Transportation. *National Traffic Safety Newsletter*, NHTSA Washington D.C., 1975.
6. Wolf, R. *Truck Accidents and Traffic Safety – An Overview*. SAE Meeting, Detroit, 1968.
7. Kinsey, G. Contribution of Unroadworthy Vehicles to Accidents. *Proc., 2nd Conference*, National Institute of Transport and Road Research, Pieterburg, South Africa, 1976.
8. Cuerden, R., M. Edwards, and M. Pittman. Effect of Vehicle Defects in Road Accidents. *Transport Research Laboratory Report No. PPR565*, TRL, Berkshire, UK, 2011.
9. Schoor, O. V., and J. L. Niekerk. Mechanical Failures as a Contributing Cause to Motor Vehicle Accidents-South Africa. *Accident Analysis & Prevention*, Vol. 33, No. 6, 2001, pp. 713–721.
10. Barry, S., S. Ginpil, and T. J. O'Neill, The Effectiveness of Air Bags. *Accident Analysis & Prevention*, Vol. 31, No. 6, 1999, pp. 781–787.

11. Ghazizadeh, M., A. McDonald, and J. Lee. Text Mining to Decipher Free-Response Consumer Complaints: Insights from the NHTSA Vehicle Owner's Complaint Database. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 56, No. 6, 2014, pp. 1189–1203.
12. Mehrotra, S., and S. Roberts. Identification and Validation of Themes from Vehicle Owner Complaints and Fatality Reports using Text Analysis. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C., 2018.
13. Das, S., X. Sun, and A. Dutta. Investigating User Ridership Sentiments for Bike Sharing Programs. *Journal of Transportation Technologies*, Vol. 5, 2015, pp. 69–75.
14. Chen, F., and R. Krishnan. *Transportation Sentiment Analysis for Safety Enhancement*. Report: DTRT12GUTG11. 2013.
15. Das, S., X. Sun, and A. Dutta. Text Mining and Topic Modeling of Compendiums of Papers from Transportation Research Board Annual Meetings. *Transportation Research Record: Journal of the Transportation Research Board*, 2016. 2552: 48–56.
16. Das, S., K. Dixon, X. Sun, A. Dutta, and M. Zupancich. Trends in Transportation Research: Exploring Content Analysis in Topics. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2552: 27–38.
17. Boyer, R., W. Scherer, and M. Smith. Trends Over Two Decades of Transportation Research: A Machine Learning Approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2552: 27–38.
18. Sun, L., and Y. Yin. Discovering Themes and Trends in Transportation Research Using Topic Modeling. *Transportation Research Part C: Emerging Technologies*, Vol. 77, 2017, pp. 49–66.
19. Das, S., and A. Dutta. Knowledge Extraction from Transportation Research Thesaurus. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C., 2018.
20. Das, S., A. Dutta, and M. Zupancich. Text mining on 100 years of Air Crash Narratives: Key Findings. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.
21. Das, S., B. Brimley, T. Lindheimer, and M. Zupancich. Association of Reduced Visibility with Crash Outcomes. *IATSS Research*, 2017. https://doi.org/10.1016/j.iatssr.2017.10.003
22. Brown, D. Text Mining the Contributors to Rail Accidents. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, 2016, pp. 346–355.
23. Bazargan, M., M. Johnson, and A. Vijayanarayanan. *An Evaluation of AIRES and STATISTICA Text Mining Tools as Applied to General Aviation Accidents*. Report: DOT/FAA/TC-TN13/7, FHWA, Washington, D.C., 2013.
24. Das, S., X. Sun, M. Zupancich, and A. Dutta. Twitter in Circulating Transportation Information: A Case Study on Two Cities. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.
25. Gu, Y., Z. Qian, and F. Chen. From Twitter to Detector: Real-time Traffic Incident Detection Using Social Media Data. *Transportation Research Part C: Emerging Technologies*, Vol. 67, 2016, pp. 321–342.
26. Das, S., L. Minjares-Kyle, K. Dixon, A. Palanisamy, and A. Dutta. #TRBAM: Understanding Communication Patterns and Research Trends by Twitter Mining. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C., 2018.
27. Das, S., G. Medina, L. Minjares-Kyle, and Z. Elgart. Social Media Hashtags Associated with Bike Commuting: Applying Natural Language Processing Tools. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C., 2018.
28. Minjares-Kyle, L., S. Das, G. Medina, and R. Henk. Knowledge about Crash Risk Factors and Self-Reported Driving Behavior: Exploratory Analysis on Multi-State Teen Driver Survey. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C., 2018.
29. Gohil, A. *R Data Visualization Cookbook*. Packt Publishing, 2015.
30. Geurts, K., G. Wets, T. Brijs, and K. Vanhoof Profiling of High-Frequency Accident Locations by Use of Association Rules. *Transportation Research Record: Journal of the Transportation Research Board*, 2003. 1840: 123–130.
31. Pande, A., and M. Abdel-Aty. Discovering Indirect Associations in Crash Data Through Probe Attributes. *Transportation Research Record: Journal of the Transportation Research Board*, 2008. 2083: 170–179.
32. Das, S., and X. Sun. Investigating the Pattern of Traffic Crashes under Rainy Weather by Association Rules in Data Mining. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
33. Das, S., L. Minjares-Kyle, R. Avelar, K. Dixon, and B. Bommanayakanahalli. Improper Passing-Related Crashes on Rural Roadways: Using Association Rules Negative Binomial Miner. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.
34. Das, S., A. Dutta, R. Avelar, K. Dixon, X. Sun, and M. Jalayer. Supervised Association Rules Mining on Pedestrian Crashes in Urban Areas: Identifying Patterns for Appropriate Countermeasures. *International Journal of Urban Sciences*, 2018. https://doi.org/10.1080/12265934.2018.1431146
35. Das, S., A. Dutta, M. Jalayer, A. Bibeka, and L. Wu. Factors influencing the Patterns of Wrong Way Driving Crashes on Freeway Exit Ramps and Median Crossovers: Exploration using 'Eclat' Association Rules to Promote Safety. *International Journal of Transportation Science and Technology*, Vol. 7, No. 2, 2018, pp. 114–123. https://doi.org/10.1016/j.ijtst.2018.02.001.
36. Ihrie, J, I. Ahmed, and A. Poncet. *EBGM Scores for Mining Large Contingency Tables. Version 0.1.0*, 2017.
37. The R Foundation for Statistical Computing. *R: A Language and Environment for Statistical Computing, Version 3.4.1*, 2017.

The Standing Committee on Transportation Safety Management (*ANB10*) *peer-reviewed this paper* (*18-03567*).