Journal of Safety Research xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Journal of Safety Research

journal homepage: www.elsevier.com/locate/jsr

www.nsc.org

Commercial truck crash injury severity analysis using gradient boosting data mining model

Zijian Zheng, * Pan Lu, Brenda Lantz

Upper Great Plain Transportation Institute, North Dakota State University, NDSU Dept 2880 P. O. Box 6050, Fargo, ND 58108-6050, United States

ARTICLE INFO

Article history:
Received 20 September 2017
Received in revised form 8 February 2018
Accepted 6 March 2018
Available online xxxx

Keywords:
Truck crash injury severity
Commercial truck company
Data mining
Gradient boosting

ABSTRACT

This study seeks to identify the contributing factors affecting commercial truck crash severity using 2010 to 2016 North Dakota and Colorado crash data provided by the Federal Motor Carrier Safety Administration. To fulfill a gap of previous studies, broad considerations of company and driver characteristics, such as company size and driver's license class, along with vehicle types and crash characteristics are researched. Gradient boosting, a data mining technique, is applied to comprehensively analyze the relationship between crash severities and a set of heterogeneous risk factors. Twenty-five variables were tested and twenty-two of them are identified as significant variables contributing to injury severities, however, top 11 variables account for more than 80% of injury forecasting. The relative variable importance analysis is conducted and furthermore marginal effects of all contributing factors are also illustrated in this research. Several factors such as trucking company attributes (e.g., company size), safety inspection values, trucking company commerce status (e.g., interstate or intrastate), time of day, driver's age, first harmful events, and registration condition are found to be significantly associated with crash injury severity. Even though most of the identified contributing factors are significant for all four levels of crash severity, their relative importance and marginal effect are all different. Findings in this study can be helpful for transportation agencies to reduce injury severity, and develop efficient strategies to improve safety.

1. Introduction

Trucking is a well-known important element for freight movement and economic development. According to a 2012 commodity flow survey, trucks move 73.1% of commodities by value, 71.3% by tons, and 42.0% by ton-miles (USDOT/BTS, 2008). Truck crashes not only interrupt traffic flow, but also cause economic loss. Moreover, truck crashes contribute to a large number of injuries and fatalities due to additional risks, such as a larger vehicle size, heavier weight, and possible hazardous material release. In 2014, the total number of fatalities in truck crashes was 3903 (Federal Motor Carrier Safety Administration, 2014). Compared with the total number of fatalities in strictly passenger car crashes, 28,559, truck crashes do not seem as alarming. However, truck crashes are overall more likely to result in more severe outcomes such as a fatality. In 2014, there were 14 fatalities in large truck crashes per 100 million vehicle miles traveled by large trucks, while only 10.5 fatalities in passenger vehicle crashes per 100 million vehicle miles traveled by passenger vehicles. Additionally, there were 29.4 injury crashes involving large trucks per 100 million vehicle miles traveled by large trucks, compared with 58.5 for passenger vehicles (Federal Motor Carrier Safety Administration, 2014).

The need to improve commercial trucking company safety performance has been a major social concern in the United States for decades. Transportation agencies and other stakeholders must identify the complete picture of factors that contribute to the severity levels of commercial truck collision and provide directions for commercial truck operation policies that will reduce the severe crash rates of commercial trucks.

Previous studies on modeling truck crash severities provide great insights and findings (Lemp, Kockelman, & Unnikrishnan, 2011; Zhu & Srinivasan, 2011). However, some factors are overlooked and not considered in those studies. Intuitively thinking, characteristics of management, organization, culture, strategies, and financial situations in a trucking company should be closely associated with the company's safety performance. For example, safety culture shapes the attitude and behavior of their employees. Building a strong safety culture has a great effect on incident reduction (U.S. Department of Labor). Furthermore, a strong safety culture will result in better trained employees who will react better when they encounter a potential crash situation, and thus may result in a less severe crash outcome. Moreover, sufficient capital and profit promote truck maintenance and technology development, so that equipment is well-performing, which will minimize risk of equipment failure. In return, incident likelihood and crash severity level

* Corresponding author.
E-mail addresses: zijian.zheng@ndsu.edu (Z. Zheng), pan.lu@ndsu.edu (P. Lu), brenda.lantz@ndsu.edu (B. Lantz).

https://doi.org/10.1016/j.jsr.2018.03.002

would be reduced. Although several studies have been carried out to investigate contributing variables to truck crash severity outcomes, the literature review revealed that it is still not clear how some commercial trucking company and driver characteristics impact crash severity levels.

This paper seeks to investigate commercial truck crash severity and contributing factors, especially trucking company characteristics, through the application of a data mining model to commercial trucking crash data. The paper is organized with a literature review, data description, methodology, results analysis, and conclusions of the research.

2. Literature review

Vehicle crash studies have been completed by a substantial number of researchers focusing on crash frequencies and injury severity (Chen, Zhang, Yang, Milton, & Alcántara, 2016; Dong, Clarke, Richards, & Huang, 2014; Dong, Clarke, Yan, Khattak, & Huang, 2014; Gabauer & Li, 2015; Lu & Tolliver, 2016; Wood, Donnell, & Fariss, 2016; Wu, Zhang, Chen, et al., 2016; Wu, Zhang, Zhu, Liu, & Tarefder, 2016). The majority of them are focused on vehicle crashes in urban road tunnels. Meng and Qu (2012) examined rear-end vehicle crash frequency in urban road tunnels. Wu, Zhang, Chen, et al. (2016), Wu, Zhang, Zhu, et al. (2016) conducted a crash severity study examining the factors of weather condition, class of highway and drug use and their impact on single-vehicle crashes. An integrated study of crash frequency and severity was conducted by Chiou and Fu (2013). Freeway geometrics, traffic characteristics, neighborhood, and freeway facilities were found to significantly contribute to vehicle crash frequency and severity.

As a common understanding, vehicle types such as passenger cars or commercial trucks should have a different impact on crash severity outcomes. There are numerous studies focusing on truck crashes only. Most of them examined one specific influential factor of truck crashes, such as wind speed, driver turnover rate, presence of portable message sign, time of the day, truck configurations, and driver body mass (Anderson et al., 2012; Bai, Yang, & Li, 2015; Braver et al., 1997; Curnow, 2002; Pahukula, Hernandez, & Unnikrishnan, 2015; Staplin & Gish, 2005; Young & Liesman, 2007). The impact of the effective location of message signs in work zones on truck related crash is studied by Bai et al (2003). The impact of effects of time of day on truck related crashes is analyzed by Curnow (2002). Pahukula et al. (2015) conducted a study, and concluded that in clear weather nights, and nights with no illumination result in either no injury crashes or crashes with a severe result. Most of those studies focus on the effects on truck crash frequency but only a limited number of studies contribute to understanding truck crash severities (Campbell, 1991; Khattak, Schneider, & Targa, 2003; Naik, Tung, Zhao, & Khattak, 2016; Uddin & Huynh, 2017; Zou, Wang, & Zhang, 2017). Uddin and Huynh (2017) studied influential factors of crash severity involving hazardous materials trucks. They demonstrate that gender, day of week, rural highway, and illumination are associated with crash injury severity. Naik et al. (2016) investigated the impact of weather conditions on single-vehicle truck crash injury severity. Their results indicated that wind speed, rain humidity, and air temperature have a significant impact on single-vehicle truck crash injury severity. Zou et al. (2017) link truck crash severity with spatial location and time of day. Their results reveal that individual truck crashes are spatially dependent events for single and multi-vehicle crashes. Single-vehicle crashes in the afternoon and at night tend to be less severe, while multi-vehicle crashes at the same time are more severe.

An understanding of the influence of attributes of the trucking company and driver's license on crash injury severity is still unclear with the literature search. Several studies discussed that the little research on trucking company characteristics' impact on crash severity is due to the lack of available company data (Chen, 2008). This research focuses on risk factors for commercial truck crash severity, in particular how company related characteristics affect crash severity, with a more comprehensive truck crash dataset available through the Federal Motor Carrier Safety Administration (FMCSA). The detailed information regarding this database is described later in the data description section.

The literature search also reveals that most prior studies are based on logit, probit, and their extension statistical models (Charbotel, Martin, Gadegbeku, & Chiron, 2003; Lemp et al., 2011; Wu, Zhang, Chen, et al., 2016; Wu, Zhang, Zhu, et al., 2016; Zhu & Srinivasan, 2011). However, these statistical models are all based on certain assumptions. One of the common assumptions is that the effects of contributing factors are assumed identical across different severity levels. These assumptions are inappropriate and do not hold true in most circumstances. Once violated, numerous errors will be generated. In addition, truck crashes are affected by a set of heterogeneous variables (Kumar & Toshniwai, 2015). A good crash injury severity model is expected to be able to extract hidden, valuable information from large, complex datasets. Thus, instead of applying statistical models, the non-parametric gradient boosting (GB) model, a data mining technique, is selected in this study to overcome the shortcomings and achieve more convincing conclusions. The GB model does not have any pre-defined data assumptions like other statistical models do. Moreover, the GB model inherits most of the tree-based data mining models' advantages. It is also superior than most of the tree-based data mining models with its missing data handling techniques, robustness with data noise and resistance to over-fitting (Friedman & Meulman, 2003; Salford Systems). The GB model proves its success in crash prediction analysis (Chung, 2013; Saha, Alluri, & Gan, 2015); however, it has never been used in a truck crash injury severity explanatory study. Therefore, the authors decided to adopt a GB model to comprehensively analyze influential factors on truck crash injury severity.

3. Data description

In this study, truck crash data was obtained from the Federal Motor Carrier Safety Administration (FMCSA). Crash data file, census file, and inspection files from the Motor Carrier Management Information System (MCMIS) are selected for the research. The MCMIS datasets contain 1) records from state police crash reports including information on drivers, crash conditions, environment factors when the crash happened, and crash involved truck conditions; 2) motor carrier corporation variables and operational factors; and 3) motor carrier safety inspection records. This study examines truck crash related data for crashes that occurred in the states of North Dakota and Colorado in the past six years (from 2010 to 2016). The selection of the two states is due to the availability of data, research interest, and data size limitation; however, the research can be extended to national level or include additional states if it is of interest.

The authors exclude irrelevant, privacy variables and four redundancy variables from the raw data before performing mathematical analysis. Summarized in Table 1, 38 variables are removed from analysis.

The detailed information of the data analyzed in this research is shown in Table 2. In general the data variables can be grouped into the following five (5) categories:

1) Trucking company characteristics (e.g., total number of trucks, inspection value, registered date, and location);
2) Crash characteristics (e.g., first injury or damaging-producing event, day of week, time of day, and number of injuries);
3) Environment characteristics (e.g., road type, light condition, road surface condition, and weather condition);
4) Driver characteristics (e.g., age, driver license class, and driver license state); and
5) Truck characteristics (e.g., cargo type, configuration, and gross vehicle weight).

There are 24 variables selected to be investigated and tested. Twenty-one (21) of them are categorical variables (labeled with "$" in Table 2), and two (2) of them are numeric variables. In this study, the

Please cite this article as: Zheng, Z., et al., Commercial truck crash injury severity analysis using gradient boosting data mining model, *Journal of Safety Research* (2018), https://doi.org/10.1016/j.jsr.2018.03.002