Safety Science 63 (2014) 50–56

Contents lists available at ScienceDirect

Safety Science

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

Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data

Rongjie Yu [b,*], Mohamed Abdel-Aty [a]

[a] Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816-2450, United States
[b] School of Transportation Engineering, Tongji University, 4800 Cao'an Road, 201804 Shanghai, China

ARTICLE INFO

Article history:
Received 4 March 2013
Received in revised form 7 October 2013
Accepted 12 October 2013
Available online 23 November 2013

Keywords:
Crash injury severity
Random parameter logit model
Support vector machine
Traffic safety
Microscopic traffic data

ABSTRACT

This study focuses on developing crash injury severity analysis models for a mountainous freeway section. In addition to the data obtained from crash reports, real-time traffic and weather data were utilized. The introduction of real-time data would benefit model applications on crash injury severity prediction. Crash injury severity was classified as a binary outcome (severe and non-severe crashes) and random forest model was firstly estimated to select the most important explanatory variables associated with severe crash occurrence. Four most critical variables (snow season indicator, steep grade indicator, speed standard deviation, and temperature) were chosen by the random forest model as inputs for the modeling analyses. For the purpose of identifying actual relationships between severe crash occurrence and the chosen explanatory variables and enhancing model goodness-of-fit, a total of three models were developed to analyze crash injury severity: (1) fixed parameter logit model; (2) support vector machine (SVM) model with radial-basis kernel function to detect non-linearity; and (3) random parameter logit model with unrestricted variance–covariance matrix to account for individual heterogeneity and also to investigate potential correlations between the explanatory variables. The three models were compared based on the areas under the ROC curve (AUC) values and it was demonstrated that SVM model and random parameter model provide superior model fits than the fixed parameter logit model. Findings of this study demonstrate that real-time traffic and weather variables have substantial influences on crash injury severity, which could be utilized to predict crash injury severity. Moreover, it is important to consider possible non-linearity and individual heterogeneity when analyzing crash injury severity. In addition, potential applications of the modeling results, limitations of this study have been discussed.

1. Introduction

Traffic safety is a major concern in the transportation industry since crash occurrence causes immense losses from the human, economic, and social sides, especially the injury and fatal crashes. Tremendous efforts have been dedicated by researchers and practitioners to improve traffic safety. Among the various aspects of traffic safety studies, crash injury severity analysis is a key part as it unveils the relationships between crash injury severity and various explanatory variables (including driver behavior, environmental conditions, traffic flow, and geometry characteristics). This kind of study is vital since professionals in roadway design, freeway management, public health, enforcement, emergency and trauma, policy, and education and awareness could benefit from the results to reduce the occurrence of injury and fatality crashes from different aspects. Savolainen et al. (2012) documented a review study of the numerous methodological alternatives used in the highway crash injury severity studies. From the review, it can be seen that both advanced parametric models (e.g. random parameter logit model, markov switching multinomial logit model) and the emerging data mining techniques (e.g. neural network models and support vector machine models) have been introduced to address the crash injury severity issues.

In this study, we focus on traffic crash injury severity analysis for a mountainous freeway section on I-70 in Colorado. In addition to the traditional data obtained from crash reports and geometric roadway characteristics inventory (RCI) databases, this study also utilizes variables from real-time traffic and weather databases. With the benefits of real-time data, critical microscopic contributing factors affect severe crash occurrence can be identified and the results can be utilized to predict crash injury severity. Crashes were classified as severe crashes (injury and fatality crashes) and non-severe (property damage only) crashes in this study. Due to data limitations, we were not able to further classify the severe crashes by different levels of severity In order to accurately and efficiently identify the crucial variables affecting crash severity, multiple promising analytical techniques were applied. We

* Corresponding author. Tel.: +86 137 6157 2327.
E-mail addresses: rongjie.yu@gmail.com, yurongjie@tongji.edu.cn (R. Yu).

0925-7535/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ssci.2013.10.012

compared traditional crash injury analysis approach (fixed parameter logit model) with innovative support vector machine (SVM) model and sophisticated random parameter logit model. A random forest model was firstly developed to rank the decisive explanatory variables that affect crash injury severity. Then, three different models were developed to analyze crash injury severity: (1) fixed parameter logit model; (2) support vector machine model to account for possible non-linearity relationships; and (3) random parameter logit model to account for individual heterogeneity. Model comparisons have been made based on the areas under the ROC curves (AUC). Finally, according to the modeling results, potential applications of the modeling results and limitations of this study have been discussed.

2. Background

Traffic crash injury severity analyses such as severe vs. non-severe crashes or fatal vs. non-fatal crashes have natural discrete outcomes. Binary logit models (fixed parameter) have been widely employed to analyze crash injury severity (Farmer et al., 1997; Bédard et al., 2002; Lee and Abdel-Aty, 2008). However, although modeling procedures and result interpretations of fixed parameter logit models are straightforward, it is not sufficient to describe relationships between explanatory variables and the crash injury severity outcomes. Extensions of the binary models (e.g. hierarchical logit model (Huang et al., 2008) and hierarchical probit model (Yu and Abdel-Aty, 2014)) and other nonparametric models (e.g. neural network models) were introduced to account for unobserved heterogeneity and non-linearity.

2.1. Random parameter logit model

More recently, random parameter logit models (also called mixed logit model) have been widely used in injury severity analyses. Compared to the fixed parameter models, random parameter models allow parameters to vary across the population to account for the unobserved heterogeneity (Hensher and Greene, 2003). Milton et al. (2008) utilized a random parameter model to investigate the crash severities along with the frequency model. The model allows some variables to vary across different roadway segments and in this way the methodology could account for the unobserved effects (roadway characteristics, environmental factors and driver behavior). Gkritza and Mannering (2008) employed a mixed logit model (model with both fixed and random parameters) to achieve better understandings of the effects of safety belts usages in single- and multi-occupant vehicles. The mixed logit models were used to account for vehicle-specific variations of the independent variables' effects on safety-belt use probabilities. The authors claimed that this approach has its flexibility to capture individual-specific heterogeneity. Anastasopoulos and Mannering (2011) compared fixed and random parameter logit models to analyze crash injury severity. A random parameter multinomial logit model was estimated with simulated maximum likelihood approach and the random parameters were applied at the individual observation level. Results indicated that the random parameter models provided a better model fit relative to the traditional fixed parameter models. Kim et al. (2012) also utilized a random parameter model to analyze single-vehicle crash injury severity data in California. The purpose of using random parameter model was to account for the heterogeneous effects of age and gender.

However, abovementioned studies utilized random parameter logit models in a relatively restricted approach: distributions of the random parameters were assumed to be mutually independent. Recently, Xiong and Mannering (2013) utilized a more general approach to develop the random parameter model. The random parameter vector was set to follow a multivariate normal distribution with an unrestricted variance–covariance matrix. Correlation effects of the guardian indicator on other explanatory variables were able to be captured. In this study, similar unrestricted random parameter logit model would be used.

2.2. Support vector machine

Support vector machine (SVM), a newly introduced pattern classifier based on statistical learning theory (Vladimir and Vapnik, 1995), has already been employed in various aspects of transportation research studies. Yuan and Cheu (2003) introduced SVM in incident detection and they compared results from SVM models with the multi-layer feed forward neural network (MLFNN) and probabilistic neural network models. It was concluded that SVM models provided a lower misclassification rate, higher correct detection rate, and lower false alarm rate. Later on, Chen et al. (2009) similarly constructed SVM models to detect traffic incidents. Instead of building one SVM model, different SVM models have been estimated and combined using various ensemble methods (bagging, boosting, and cross-validated). The authors utilized ensemble methods to overcome the variability of a single SVM and improve the model's accuracy. Beside the incident detection, SVM has been utilized in crash frequency studies. Li et al. (2008) estimated safety performance functions for motor vehicle crashes with support vector machine models. SVM models have been estimated and compared to the traditional negative binomial models, and the results demonstrated that SVM models provide better goodness-of-fit than the negative binomial models. Yu and Abdel-Aty (2013) employed SVM technique to develop real-time crash risk evaluation models. SVM model with radial-basis kernel function was proved to be capable of capturing the non-linearity relationships between crash risk and explanatory variables and providing better prediction accuracy. As for the crash injury severity analysis, Li et al. (2011) investigated the feasibility of SVM models to analyze crash injury severity. SVM models were compared with the frequently used ordered probit model; it was concluded that SVM models outperformed the ordered probit models in model fit and more reasonable relationships between severity outcomes and the explanatory variables were provided. In this study, SVM models with radial-basis kernel functions (RBF) would be utilized to analyze crash injury severity. Moreover, sensitivity analyses would be conducted to investigate the effects of the explanatory variables on injury severity.

3. Data preparation

Four datasets were used to form the database used in this study, (1) I-70 crash data provided by Colorado Department of Transportation (CDOT), (based on the traffic data availability, crash data from Jul, 2007 to Jul, 2009 and Aug, 2010 to Apr, 2011 were used), (2) roadway geometric characteristic data obtained from roadway characteristics inventory (RCI), (3) real-time weather data recorded by 6 weather stations along the roadway segment under study, and (4) real-time traffic data detected by 20 Automatic Vehicle Identification (AVI) detectors located on the east and west bounds of I-70.

For each specific crash, the corresponding weather and traffic data were identified and matched. AVI system on I-70 provides average segment speed at 2-min interval; in order to decrease data noise, raw data were aggregated into 6-min intervals. The mean, standard deviation and coefficient of variation for the AVI speed during the 6-minute intervals were calculated at the data aggregation step and used as explanatory variables in the crash injury severity analysis to represent real-time traffic flow conditions prior