Transportation Research Part C 50 (2015) 86–96

Contents lists available at ScienceDirect

Transportation Research Part C

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

Exploring the feasibility of classification trees versus ordinal discrete choice models for analyzing crash severity

Ghazan Khan [a,*], Andrea R. Bill [b,1], David A. Noyce [c,2]

[a] California State University Sacramento, 6000 J St., Sacramento, CA 95819-6029, United States
[b] Traffic Operations and Safety (TOPS) Laboratory, University of Wisconsin-Madison, B 243 Engineering Hall, 1415 Engineering Dr. Madison, WI 53706, United States
[c] Traffic Operations and Safety (TOPS) Laboratory, University of Wisconsin-Madison, 1204 Engineering Hall, 1415 Engineering Dr. Madison, WI 53706, United States

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form 26 June 2014
Accepted 8 October 2014
Available online 4 November 2014

Keywords:
Classification and Regression Tree
GUIDE
Cross-median crashes
Crash severity models
Multicollinearity

ABSTRACT

A cross-median crash (CMC) is one of the most severe types of crashes in which a vehicle crosses the median and sometimes collides with opposing traffic. A study of severity of CMCs in the state of Wisconsin was conducted by Lu et al. in 2010. Discrete choice models, namely ordinal logit and probit models were used to analyze factors related to the severity of CMCs. Separate models were developed for single and multi-vehicle CMCs. Although 25 different crash, roadway, and geometric variables were used, only 3 variables were found to be statistically significant which were alcohol usage, posted speed, and road conditions. The objective of this research was to explore the feasibility of GUIDE Classification Tree method to analyze the severity of CMCs to discover if any additional information could be revealed.

A dataset of CMCs in the state of Wisconsin between 2001 and 2007, used in the study by Lu et al. was used to develop three different GUIDE Classification Trees. Additionally, the effects of variable types (continuous or discrete), misclassification costs, and tree pruning characteristics on models results were also explored. The results were directly compared with discrete choice models developed in the study by Lu et al. showing that the GUIDE Classification Trees revealed new variables (median width and traffic volume) that affect CMC severity and provided useful insight on the data. The results of this research suggest that the use of Classification Tree analysis should at least be considered in conjunction with regression-based crash models to better understand factors affecting crashes. Classification Tree models were able to reveal additional information about the dependent variable and offer advantages with respect to multicollinearity and variable redundancy issues.

1. Introduction and Background

From 2001 to 2007, 298,131 people lost their lives on U.S. roads of which over 16,000 died in run-off-road (ROR) crashes when their vehicle departed from the travel lane and crashed (Data Summary and Traffic Safety Facts, 2005; Neuman et al., 2003; FARS, 2002). In recent years, approximately 55% of traffic fatalities were a result of ROR type crashes (FARS, 2002). A cross-median crash (CMC) is a type of roadway departure crash in which a vehicle crosses the median and sometimes collides with opposing traffic with potentially serious consequences for drivers. Wisconsin is no exception to the high

* Corresponding author.
E-mail addresses: ghazan.khan@csus.edu (G. Khan), bill@wisc.edu (A.R. Bill), noyce@engr.wisc.edu (D.A. Noyce).
[1] Tel.: +1 (608) 698 8519; fax: +1 (608) 262 5199.
[2] Tel.: +1 (608) 265 1882.

http://dx.doi.org/10.1016/j.trc.2014.10.003
0968-090X/

number of ROR and CMCs experienced nationally and a number of studies have addressed these crashes with the aim of understanding the nature and factors affecting their frequency and severity (Wisconsin Traffic Crash Facts, 2005; Noyce et al., 2005; Witte et al., 2007). CMCs present one of the most serious safety issues of all the ROR crashes.

Traffic safety researchers employ a variety of different statistical method to develop crash frequency and severity models in order to identify and understand factors that cause crashes. Traditionally, classical regression analysis techniques have been the mainstay of most crash modeling studies and same has been the case for CMCs. A study of severity of CMCs in the state of Wisconsin was conducted by Lu et al. in 2010 which explored the severity of crashes with respect to various geometric and roadway characteristics (Lu et al., 2010). Discrete choice models, namely ordinal logit and probit models were used to analyze factors related to the severity of CMCs. Separate models were developed for single and multi-vehicle CMCs. Although 25 different crash, roadway, and geometric variables were used, only 3 variables were found to be statistically significant which were alcohol usage, posted speed, and road conditions.

2. Objectives

The objective of this research was to explore the feasibility of Classification Tree methods to analyze the severity of CMCs and examine if the use of Classification Tree method can reveal information which can improve, replace, or compliment more traditional methods of crash modelling. The idea was to take advantage that non-parametric methods such as Classification Tree facilitate the use of a multitude of crash attribute information without concern of data distribution or multicollinearity issues; inherent in most classical regression methods.

3. Literature review

3.1. Crossover median crash analysis using discrete choice models

There have been a number of research studies using discrete choice models for the analysis of crash data in the past (Duncan et al., 1998; Renski et al., 1999; Abdel-Aty, 2003; Deng et al., 2006). Some studies have explored the issues of CMCs by modeling crash frequency and severity (Donnell and Mason, 2004, 2006; Lu et al., 2006; Shankar et al., 1998; Ulfarsson and Shankar, 2003; Miaou et al., 2005; Harkey, 2008). The focus of this research was on studies that used discrete choice models (ordinal logistic regression) to relate crash severity in CMCs, classified as fatal, injury, or property damage, to various geometric, traffic operation, and environmental conditions. Donnell and Mason used road inventory and crash record information collected on Pennsylvania Interstate highways for the five-year period between 1994 and 1998 to develop cross-median and median barrier crash logistic regression models (Donnell and Mason, 2004). The researchers modeled CMC severity as an ordinal response to reveal that the use of drugs and the presence of a curvilinear alignment increased the odds of a fatal CMC when compared to injury or property damage crashes.

In a 2006 study, ordinal logistic regression was also employed by Lu et al. to model the crash severity of CMCs in Wisconsin during the three year period between 2001 and 2003 (Lu et al., 2006). The researchers found that seasons had an effect on CMC severity, likely due to adverse weather and road conditions prevalent in Wisconsin during the winter months. Furthermore, driver age affected the severity of CMCs when the traffic volume was relatively high. However, road condition was the only significant variable identified under low traffic volumes. Given inadequate median width conditions, weather condition and emergency vehicle response time were found to be significant explanatory variables.

The results of research by Donnell and Mason and Lu et al. found few statistically significant variables amongst several used in the analyses (Lu et al., 2010; Donnell and Mason, 2004). Furthermore, while there have been examples of studies using classical regression-based methods to analyze CMCs, there were no examples related to the use of Classification Tree methods.

3.2. Comparison of discrete choice models and Classification Tree methods

A review of literature revealed some research on comparing the results of discrete choice models especially Logistic Regression models and Classification Tree methods (Breiman et al., 1984; Buntine, 1992; Cappelli et al., 2002, 1998; Siciliano and Mola, 2000). The results of the studies suggest varying conclusions where Logistic Regression methods outperformed Classification Tree methods in some cases and vice versa in other cases. Almost all of the examples found in the literature were from the field of medical sciences and no examples of crash data analysis were found.

In light of the literature review, the objective of this research was to compare the results of Discrete Choice and Classification Tree crash severity models using CMC dataset in Wisconsin. The results of the research by Lu et al. analyzing the severity of CMCs using logit and probit models were used as benchmark to compare with Classification Tree Results to identify factors affecting severity of these crashes.

4. Data collection

In this research, a CMC was defined as a crash where a vehicle crossed the median into opposing lanes. If the crossing vehicle collided with opposing traffic, it was considered a multi-vehicle CMC. In case of no collision after crossing into