Accident Analysis and Prevention 79 (2015) 133–144

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees

Dibakar Saha *, Priyanka Alluri [1], Albert Gan [2]

Department of Civil and Environmental Engineering, Florida International University, 10555 West Flagler Street, EC 3680, Miami, FL 33174, United States

ARTICLE INFO

Article history:
Received 15 September 2014
Received in revised form 14 February 2015
Accepted 10 March 2015
Available online xxx

Keywords:
Highway Safety Manual
Data mining
Boosted regression trees
Variable importance
Crash predictions
Calibration factor

ABSTRACT

The Highway Safety Manual (HSM) recommends using the empirical Bayes (EB) method with locally derived calibration factors to predict an agency's safety performance. However, the data needs for deriving these local calibration factors are significant, requiring very detailed roadway characteristics information. Many of the data variables identified in the HSM are currently unavailable in the states' databases. Moreover, the process of collecting and maintaining all the HSM data variables is cost-prohibitive. Prioritization of the variables based on their impact on crash predictions would, therefore, help to identify influential variables for which data could be collected and maintained for continued updates. This study aims to determine the impact of each independent variable identified in the HSM on crash predictions. A relatively recent data mining approach called boosted regression trees (BRT) is used to investigate the association between the variables and crash predictions. The BRT method can effectively handle different types of predictor variables, identify very complex and non-linear association among variables, and compute variable importance. Five years of crash data from 2008 to 2012 on two urban and suburban facility types, two-lane undivided arterials and four-lane divided arterials, were analyzed for estimating the influence of variables on crash predictions. Variables were found to exhibit non-linear and sometimes complex relationship to predicted crash counts. In addition, only a few variables were found to explain most of the variation in the crash data.

Published by Elsevier Ltd.

1. Introduction

The Highway Safety Manual (HSM), published by the American Association of State Highway and Transportation Officials (AASHTO) in 2010, is designed to "assist agencies in their effort to integrate safety into their decision-making processes" (AASHTO, 2010). Part C of the HSM presents predictive models to estimate predicted average crash frequency at individual sites on different roadway facilities including rural two-lane two-way roads, rural multilane highways, and urban and suburban arterials. The general form of the predictive models in the HSM can be expressed as follows:

$$N_{predicted,i} = N_{spf,i} \times (CMF_{1,i} \times CMF_{2,i} \times \ldots \times CMF_{n,i}) \times C_i \quad (1)$$

where $N_{predicted,i}$ is the predicted average crash frequency for a specific year for site type i; $N_{spf,i}$ is the predicted average crash frequency for a specific year for site type i for base conditions; $CMF_{1,i} \ldots CMF_{n,i}$ are crash modification factors for n geometric conditions or traffic control features for site type i; and C_i is the calibration factor to adjust SPF for local conditions for site type i.

As shown in Eq. (1), there are three components of the predictive models: base safety performance functions (SPFs), crash modification factors (CMFs), and calibration factors. Base SPFs are statistical models that are used to estimate predicted average crash frequency for a facility type with specified base conditions. CMFs are used to account for the effects of non-base conditions on predicted crashes. Calibration factors are required "to account for differences between the jurisdiction and time period for which the predictive models were developed and the jurisdiction and time period to which they are applied by HSM users" (AASHTO, 2010). Calibration factor is estimated as the ratio of the total number of observed crashes to the total number of predicted crashes calculated using the SPFs and CMFs provided in the HSM. The predictive models are most effective when calibrated to local conditions (Findley et al., 2012; Lu, 2013; Sun et al., 2006; Young and Park, 2013).

Very detailed roadway geometry, traffic, and crash characteristics data are needed to derive local calibration factors. Several of

* Corresponding author. Tel.: +1 786 488 7744.
E-mail addresses: dsaha003@fiu.edu (D. Saha), palluri@fiu.edu (P. Alluri), gana@fiu.edu (A. Gan).
[1] Tel.: +1 305 348 1896.
[2] Tel.: +1 305 348 3116.

http://dx.doi.org/10.1016/j.aap.2015.03.011
0001-4575/Published by Elsevier Ltd.

the variables are often unavailable in the states' databases. Collecting and maintaining all the data variables on the entire road network for the purpose of implementing the HSM is not cost-feasible. Therefore, a process to streamline the data requirements that minimizes the potential impacts to the quality of analysis is desirable. The objective of this study is to investigate the impact of the variables identified in the HSM on crash predictions. The study used five years of crash data from 2008 to 2012 on urban and suburban two-lane undivided arterials and urban and suburban four-lane divided arterials in Florida. Boosted regression tree (BRT), a data mining approach, is applied to evaluate variables' importance and analyze their marginal effects on crash predictions.

2. Literature review

Traditionally, statistical regression models are developed in highway safety studies to associate crash frequency with the most significant variables (for example, Hadi et al., 1995; Abdel-Aty and Radwan, 2000; Sawalha and Sayed, 2001; Hauer et al., 2004; Caliendo et al., 2007; Cafiso et al., 2010 etc.). The models, however, were limited in their scope to evaluate the influence of predictor variables on crash outcome. Few studies identified and ranked the influence of predictor variables on crash predictions using sensitivity analysis (Alluri and Ogle, 2012; Findley et al., 2012; Jalayer and Zhou, 2013). The typical approach used in sensitivity analysis is to alter the value of one predictor variable at its maximum, minimum, and/or average, and estimate the change in output relative to the output generated from using the actual values of the variable. The variables that produced substantial changes in predicted crash frequencies were identified as influential variables. The main limitation of this approach is that only a single variable is evaluated at one time and, therefore, the possible association between the variables is ignored while measuring the effect of each variable on crash predictions.

Data mining procedures are increasingly being applied in transportation safety studies to capture the complex and non-linear relation between data variables and crash characteristics. Tree-based method, typically known as classification and regression tree (CART) (Breiman et al., 1984) or decision tree, is a suitable data mining approach in this regard (Williams, 2011). The CART method provides several benefits over traditional regression models. It does not require a pre-specified function and variable transformation for developing models. It can intrinsically identify non-linear association among predictor variables. Also, the method provides interpretable results by demonstrating the relative influence of each variable on model prediction.

Karlaftis and Golias (2002) developed CART model, termed as hierarchical tree-based regression (HTBR) model, to estimate the relative contribution of variables on crash frequency for rural two-lane and multilane roads in Indiana. Yan and Radwan (2006) analyzed crashes involving two vehicles at signalized intersections by developing two classification tree models; one was built to obtain the causal features associated with rear-end crashes compared to non-rear-end crashes, and the other was built to identify the factors attributed to at-fault drivers/vehicles against not-at-fault drivers/vehicles for rear-end crashes. Chang and Wang (2006) and Kashani and Mohaymany (2011) applied the CART method to quantify the effects of vehicle, driver, and crash attributes on injury severity. Elmitiny et al. (2010) fitted a CART model to analyze the behavioral patterns of driver's stop-and-go situations and red-light running violations at an intersection.

Although CART models have several advantages, they can sometimes be unstable and produce output with high variance (Zhang et al., 2005). The ensemble approach that combines outputs from a collection of trees reduces prediction error and is usually more stable (Das et al., 2009; De'ath, 2007; Elith et al., 2008). There are two ensemble approaches based on decision trees: random forests and boosted regression trees (BRT). In random forests, predictions are computed by fitting a number of trees (typically 50–1000) using a bootstrap sample of data and a subset of predictors. Harb et al. (2009) used random forests technique to estimate the relative importance of variables on the binary outcome of drivers' crash avoidance maneuvers. Abdel-Aty and Haleem (2011) applied random forests method to determine the importance of the explanatory variables in predicting angle crash frequency at unsignalized intersections. Alluri et al. (2014) prioritized the variables in the HSM for several segment and intersection subtypes using random forests algorithm.

Unlike random forests, the BRT method produces the assembly of trees with a slow learning rate and in a sequential manner to extract more variability in data. One major advantage of the BRT approach over other tree-based models is that it can also rigorously deal with different types of response variable such as binomial,

Table 1
Variables identified in the HSM for urban and suburban arterials.

Variable	Type
Average annual daily traffic (AADT)	Continuous
Segment length	Continuous
Median width[a]	Categorical (10 levels: 10 ft, 20 ft, 30 ft, 40 ft, 50 ft, 60 ft, 70 ft, 80 ft, 90 ft, 100 ft)
Number of major commercial driveways	Continuous
Number of major residential driveways	Continuous
Number of major industrial driveways	Continuous
Number of minor commercial driveways	Continuous
Number of minor residential driveways	Continuous
Number of minor industrial driveways	Continuous
Number of other driveways	Continuous
Number of roadside objects	Continuous
Speed limit	Categorical (Two levels: $\leq$30 mph, >30 mph)
Presence of on-street parking	Categorical (Two levels: absent, present)
Presence of lighting	Categorical (Two levels: absent, present)
Presence of automated speed enforcement	Categorical (Two levels: absent, present)
Type of on-street parking	Categorical (Two levels: angle parking, parallel parking)
Parking by land use type	Categorical (Two levels: residential/other, commercial or industrial/institutional)
Curb length with on-street parking	Continuous
Offset to roadside objects	Categorical (Seven levels: 2 ft, 5 ft, 10 ft, 15 ft, 20 ft, 25 ft, $\geq$ 30 ft)

[a] Median width is applicable only for four-lane divided arterials.