Accident Analysis and Prevention 122 (2019) 85–98

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Comparative analysis of multiple techniques for developing and transferring safety performance functions

Ahmed Farid*, Mohamed Abdel-Aty, Jaeyoung Lee

Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL, 32816-2450, United States

ARTICLE INFO

Keywords:
Safety performance functions
Transferability
Negative binomial model
Tobit model
Data mining methods
Highway safety manual

ABSTRACT

Safety performance functions (SPFs) are crash count prediction models that are used for identifying high crash risk locations, evaluating road safety before and after countermeasure deployment and comparing the safety of alternative site designs. The traditional method of modeling crash counts is negative binomial (NB) regression. Furthermore, the Highway Safety Manual (HSM) provides analytical tools, including NB SPFs, to assess and improve road safety. Even though the HSM's SPFs are restricted to NB models, the road safety literature is rich with a variety of different modeling techniques. Researchers have calibrated the HSM's SPFs to local conditions using a calibration method prescribed by the HSM. However, studies in which SPFs are developed and transferred to other localities are uncommon. In this paper, we develop and transfer rural divided multilane highway segment SPFs of Florida, Ohio, Illinois, Minnesota, California, Washington and North Carolina to each state. For every state, NB, zero-inflated NB, Poisson lognormal (PLN), regression tree, random forest (RF), boosting and Tobit models are developed. A hybrid model that coalesces the predictions of both the Tobit and the NB model is proposed and developed as well. All SPFs are transferred to each state and their predictive performances are evaluated to discern which model type is the most transferable. According to the transferability results, there is no single superior model type. However, the Tobit, RF, tree, NB and hybrid models demonstrate better predictive performances than those of the other methods in a considerably large proportion of transferred SPFs.

1. Introduction

Safety performance functions (SPFs), are used for predicting crash counts by severity or type at any roadway facility class. The SPFs are used for detecting high crash risk locations, assessing efficacies of deployed countermeasures in before-and-after analyses and comparing the safety of alternative road designs. The traditional method, implemented for crash frequency prediction, is negative binomial (NB) regression since the NB model is not only a count model but also handles overdispersion. Overdispersion is the condition at which the variance of the crash counts is greater than the corresponding mean. Such state is typically observed in crash data (Lord and Mannering, 2010). The national Highway Safety Manual (HSM) published by the American Association of State Highway and Transportation Officials (2010) provides default NB SPFs for both public agencies and private firms to apply to local conditions. Prior to the widespread use of the NB model, researchers employed Poisson regression. Ordinary least squares (OLS) regression is inappropriate because crash counts are non-negative and discrete (Lord and Mannering, 2010). The Poisson model is more suited than both OLS and generalized linear regression models since it is a count model. Yet, the Poisson model suffers from the fact that it cannot accommodate overdispersed crash data because one of the model's main assumptions is that the mean and the variance of the crash frequencies are equal. In addition, a wide variety of modeling frameworks aimed at predicting crash counts exists in the road safety literature.

As it is critical to introduce SPFs so too is it crucial to discuss the knowledge gap in which this paper is aimed to fill. The goal is to investigate the transferability of SPFs of different modeling structures to aid roadway agencies and consulting firms, unwilling to invest in developing local SPFs, in adopting SPFs from elsewhere. Borrowing SPFs curtails expenditures of capital and labor resources considerably relative to developing the models. The cost of data collection and hiring of expert data analysts to process the data are slashed (Srinivasan et al., 2013). Researchers applied a technique provided by the HSM to calibrate its SPFs to local conditions. However, few developed and transferred SPFs from one locality to another's conditions. In this paper, we develop and transfer rural divided multilane highway segment SPFs of different model types among seven states. We also compare the SPFs' predictive performances. Multilane divided highways are four-lane bi-

* Corresponding author.
E-mail addresses: ahmedtf91@knights.ucf.edu (A. Farid), m.aty@ucf.edu (M. Abdel-Aty), jaeyoung.lee@ucf.edu (J. Lee).

https://doi.org/10.1016/j.aap.2018.09.024
Received 21 January 2018; Received in revised form 21 September 2018; Accepted 23 September 2018
0001-4575/ © 2018 Elsevier Ltd. All rights reserved.

directional roads with a median or a two-way left-turn lane separator. The states, of which conditions are analyzed, are Florida, Ohio, Illinois, Minnesota, California, Washington and North Carolina. The literature review, data, analysis methodology, analysis results and conclusions are all discussed in the following sections.

2. Literature review

Current crash frequency prediction methods are discussed followed by an overview of the research studies that were aimed at calibrating the HSM's SPFs to specific locations. Furthermore, a discussion about the limited number of studies, in which SPFs are developed and transferred from one jurisdiction to another's conditions, is provided. Subsequently, the shortcomings of all past studies are highlighted and this paper's contribution to the literature is described.

2.1. Poisson model

The Poisson regression model is considered the basic crash count model because linear regression is not well suited for accommodating non-negative crash count data. Under the Poisson framework, the probability of observing y_i crashes at road segment i is provided as follows (Lord and Mannering, 2010).

$$p(y_i) = \frac{exp(-\mu_i)\mu_i^{y_i}}{y_i!} \tag{1}$$

The mean of the crash counts at the segment is μ_i, which is the predicted number of crashes, N_{SPFi}. It is a function of crash contributing factors, X's, including traffic, geometric design and other characteristics associated with their respective coefficients, β's. The coefficients are typically obtained by the maximum likelihood estimation (MLE) method. The crash frequency prediction equation is expressed as $N_{SPFi} = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi})$. The limitations of the Poisson model are that it yields inaccurate results for not only overdispersed crash data but also data having low sample mean crash counts and underdispersed data (Lord and Mannering, 2010).

2.2. Negative binomial model

The NB model is a modification of the Poisson model in that the mean function is configured as $N_{SPFi} = \exp(\beta_0 + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \ldots + \beta_p \times X_{pi} + \varepsilon_i)$ such that $\exp(\varepsilon_i) \sim \Gamma[1, k_i]$. The variance of the crash frequencies is $\mathrm{Var}[y_i] = \mathrm{E}[y_i] \times (1 + k_i \times \mathrm{E}[y_i])$. The term, k_i, is referred to as the overdispersion parameter which allows for the NB model to accommodate overdispersed crash data. As $k_i \to 0$, the NB model reduces to the Poisson model. Even though the NB model is the conventional one used for crash frequency prediction and the HSM's SPFs are NB models, the modeling framework has its disadvantages. First, it is ineffective when it comes to accounting for underdispersion. Second, erroneous overdispersion parameters are produced when modeling is conducted using data with low sample sizes and counts (Lord, 2006).

2.3. Zero-inflated negative binomial model

The zero-inflated negative binomial (ZINB) model is an extension of the NB model in that it is configured to incorporate the preponderance of records of segments with zero observed crashes (Lord and Mannering, 2010). The ZINB structure is set such that each segment has two separate models. One represents the probability of observing zero crashes and the other represents the probability of observing one or more crashes. A logistic model is incorporated in the zero-inflated framework for determining the probability of whether the segments experience crashes or not (Washington et al., 2003). The probabilities of one or more crashes are modeled using the NB model. Even though the ZINB model is advantageous because it accommodates excess zero crash counts it has been subject to criticism (Lord, 2006). Its faulty underlying assumption of observing no crashes translates to a difficulty in correctly capturing crash occurrence trends. Variations of the ZINB model are the zero-inflated Poisson (ZIP) model (Lord and Mannering, 2010) and the hurdle model (Cai et al., 2016).

2.4. Poisson-lognormal model

The Poisson-lognormal (PLN) model is implemented as a substitute to the NB model (Lord and Miranda-Moreno, 2008). The PLN structure is the same as that of the NB model except that $\exp(\varepsilon_i)$ follows a lognormal distribution instead of a gamma distribution (Lord and Mannering, 2010) which renders the model estimation process to be more complex. The Poisson-lognormal model can account for oversdispered crash data better than the NB model. Yet PLN regression is not appropriate for underdispersed data, data with low sample sizes and data with low sample means (Miaou et al., 2003).

2.5. Miscellaneous models

Miscellaneous regression modeling techniques, aimed at predicting crash counts, are briefly discussed. The Conway-Maxwell Poisson model is derived from the Poisson model. It can accommodate both overdispersion and underdispersion. Yet, it is not appropriate for datasets with low sample means and sizes (Lord and Mannering, 2010; Lord et al., 2008). Tobit regression is another applicable technique. It is similar to OLS regression except that it is censored at either a lower or an upper limit. For instance, a lower boundary of zero is designated for crash count predictions. The Tobit model is not restricted to modeling crash frequencies. It may also be implemented for predicting crash counts normalized by the segment length and the number of years during which the crashes occurred (Zeng et al., 2017a; Anastasopoulos et al., 2012a).

Data mining techniques including neural networks (NN), support vector machines (SVM), K nearest neighbors (KNN), multivariate adaptive regression splines (MARS), regression trees and the techniques' variants are applicable for regressing crash frequencies as well. Such methods are non-parametric since no assumption is made about the relationship between crash occurrence and crash contributing factors. Neural network, Bayesian NN and SVM models typically exhibit more appropriate fits than parametric regression models. However, the model development processes are complex and the results are not interpretable. Also, the KNN model is intended to be used for predicting outcome crash frequencies instead of providing insights into the crash contributing factors. Similarly, interpreting the results of MARS models is challenging (Lord and Mannering, 2010). Regression tree analysis (James et al., 2013; Martz et al., 2017) fragments the data into subsets according to fragmentation rules that are influenced by independent variable values. For instance, segments having an average annual daily traffic (AADT) under a specific threshold are subset from the original data. Each subset is also split into other subsets. The objective of the fragmentation rule is to minimize the sum of the squares of the differences between the observed crash count per segment and the average of the crash counts of the segments in the subset. Tree model results are interpretable. However, the performances of tree models are mediocre. Hence, the random forest (RF) and boosting models, which are variants of the tree model, are introduced to address the shortcoming of the tree model. The RF model entails the application of regression trees in conjunction with bootstrapping while the boosting model involves an iterative process of fitting trees to crash data and to the resulting residuals (James et al., 2013).

Data mining methods are not restricted to traffic safety applications. For instance, Sun et al. (2018) implemented a modified KNN method for predicting traffic patterns in the short run. Elfar et al. (2018) employed the RF, logistic regression and NN methods to predict whether

Download English Version:

https://daneshyari.com/en/article/11012370

Download Persian Version:

https://daneshyari.com/article/11012370

Daneshyari.com