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
Available online 12 October 2018
0001-4575/

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).

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

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 log-normal 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

connected vehicle traffic would be congested. The research team extracted vehicle trajectory data from the vehicle communications data. Park et al. (2018) estimated a Bayesian network model to predict the traffic conditions, whether congested or uncongested. The authors configured the model formulation in such a way to provide interpretable results. Wang et al. (2018) proposed a deep learning method to model Uber trip frequencies. Kan et al. (2018) employed the gradient boosting and RF methods to impute missing ramp flow records in the Shanghai expressway data.

2.6. Transferring safety performance functions

As it is essential to note the multitude of modeling frameworks used for crash frequency predictions, it is equally important to examine past studies in which SPFs, developed for particular conditions, were applied elsewhere. In the majority of past research efforts, the HSM's default SPFs were calibrated to specific jurisdictions using the HSM calibration method. The HSM's SPFs were calibrated to rural divided segments of Missouri (Sun et al., 2014), North Carolina (Srinivasan and Carter, 2011), Oregon (Xie et al., 2011) and Alabama (Mehta and Lou, 2013). For Alabama's conditions, the research team undertook an additional step by comparing the performances of the HSM's calibrated SPFs with those of ones developed for local conditions. Researchers have also attempted to calibrate the HSM's SPFs of other roadway facilities to data of jurisdictions such as Regina, Saskatchewan, Canada (Young and Park, 2012) and Utah (Brimley et al., 2012). The HSM's SPFs were also applied abroad North America including in Fortaleza City, Brazil (Cunto et al., 2015), the Messina-Catania region, Italy (Cafiso et al., 2012) and Riyadh, Saudi Arabia (Al Kaaf and Abdel-Aty, 2015).

Studies involving SPFs that were developed and transferred elsewhere, are scarce. Persaud et al. (2002) estimated stop-controlled and signalized intersection SPFs for Toronto's conditions. The SPFs were then applied to California's and Vancouver's data. Recently, Ambros and Sedoník (2016) assessed the transferability of SPFs of undivided and divided roads of the South Moravian and Zlín regions in the Czech Republic.

2.7. Contribution to the literature

In general, there is a growing interest in investigating the transferability of SPFs. As previously highlighted, the aim of this paper is to ascertain the modeling structures that perform the best when transferred. From the findings interpreted, suggestions are made regarding the SPF types recommended for transferring. That is to aid practitioners willing to cut down on resources and borrow SPFs instead of develop them. This paper is focused on building upon the research effort of a previous one (Farid et al., 2018) in which the transferability of NB SPFs of the seven previously mentioned states is evaluated. In this paper, the analysis is expanded extensively and re-conducted multiple times by implementing not only NB models to build the SPFs but also ZINB, Tobit, regression tree, RF, boosting and PLN models. Inferences are drawn regarding the predictive performances of the transferred SPFs of the variety of the different modeling structures.

3. Data preparation

The data comprise of crash, AADT and geometric design characteristics' records of the seven aforementioned states. Contiguous rural divided multilane roadway segments with homogeneous features are conjoined as a single segment. Note that intersections delimit the segments. Also, as per the HSM, segments shorter than 0.1 mi are not considered for analysis. The Florida crash data are collected from the Crash Analysis Reporting System (CARS) while the geometric characteristics and traffic data are collected from the Roadway Characteristics Inventory (RCI). Both the CARS and the RCI belong to the Florida Department of Transportation (FDOT). All other states' data are collected from the Highway Safety Information System (HSIS), a national database (State Data, 2017). Exploratory descriptive statistics of the states' data are summarized in Table 1 and the crash rates are presented in Fig. 1. As shown in the table, the crash categories, of which data are summarized, include total (KABCO) crashes, injury (KABC) crashes, injury crashes excluding possible injury (KAB) crashes, fatal-and-incapacitating injury (KA) crashes and single-vehicle (SV) crashes. The SPFs are built for such crash categories. The North Carolina KA crash count sampled is one. Therefore, no analysis is conducted for North Carolina's KA crashes. Also, no lane width data in North Carolina are available in the HSIS database. However, the lane width parameter is not found to influence crash occurrence as per results of all SPFs estimated for the other states. Likewise, it is discovered that the shoulder pavement variable has no impact on crash frequency in any SPF developed.

4. Research methodology

For each state's conditions, the SPFs are developed for KABCO, KABC, KAB, KA and SV crashes. The SPFs are transferred to the other states' data. The estimated models are of different types namely NB, ZINB, tree, RF, boosting, Tobit and PLN regression. It should be noted that variable transformations are attempted as well. The NB and PLN structures are appropriate for modeling overdispersed crash data which is the case. The ZINB and Tobit models are well suited for excess zero crash counts especially those of severe crashes observed in the data. The tree, RF and boosting models are selected since they are non-parametric. The Poisson model is not employed since it does not accommodate overdispersion. Other data mining techniques, discussed in the literature review section including NN, Bayesian NN, KNN and MARS models are not employed since their results cannot be directly interpreted. Also, Lord and Mannering (2010) maintain that NN, Bayesian NN and SVM models are difficult to transfer to data of jurisdictions elsewhere. Details of the procedure for transforming variables prior to the modeling process, modeling frameworks and goodness of fit (GOF) measures are discussed.

4.1. Variable transformations

The independent variables incorporated are the AADT, shoulder width and median width. The lane width and shoulder pavement variables are not found to contribute to crashes as previously mentioned. The Box-Cox (Box and Cox, 1964) transformation is employed to reduce the absolute values of the skewnesses of the independent variables' distributions. The Box-Cox family of transformations is expressed as follows.

$$X_i^{(\lambda)} = \begin{cases} \frac{X_i^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(X_i), & \lambda = 0 \end{cases} \quad (2)$$

The exponent, λ, is the transformation tuning parameter. The attempted transformations are the reciprocal ($\lambda = -1$), reciprocal square root ($\lambda = -0.5$), natural log ($\lambda = 0$), square root ($\lambda = 0.5$), square ($\lambda = 2$) and cubic ($\lambda = 3$) transformations. An additional transformation specifically intended for the AADT variable is proposed by normalizing the variable's values by the thousands of vehicles per day. For any variable, in any SPF, the chosen transformation is the one that best corrects for skewness.

4.2. Negative binomial model

In the NB model formulation, the probability of observing y_i crashes at a segment is given as the following (Lord and Miranda-Moreno, 2008).

Table 1
Descriptive Statistics of the Data.

Florida (436 segments, 350.641 mi, crash years: 2009–2011)

Variable	Frequency	Mean	Standard Deviation	Minimum	Maximum
Segment Length	–	0.804	1.459	0.100	18.078
AADT (vpd)	–	12,681.930	8,709.680	2,500	49,500
Lane Width (ft)	–	12.008	0.205	10	13
Presence of Paved Shoulder	427	–	–	–	–
Absence of Paved Shoulder	9	–	–	–	–
Shoulder Width (ft)	–	4.752	1.669	1.5	12
Median Width (ft)	–	40.429	23.504	8	140
KABCO	1,114	2.555	5.335	0	44
KABC	648	1.486	3.346	0	29
KAB	403	0.924	2.189	0	21
KA	198	0.454	1.149	0	11
SV	571	1.310	3.041	0	30

Ohio (1248 segments, 661.716 mi, crash years: 2009–2011)

Variable	Frequency	Mean	Standard Deviation	Minimum	Maximum
Segment Length	–	0.530	0.584	0.100	9.422
AADT (vpd)	–	9941.160	5,591.140	233	38,710
Lane Width (ft)	–	11.757	0.422	10	12
Presence of Paved Shoulder	1248	–	–	–	–
Absence of Paved Shoulder	0	–	–	–	–
Shoulder Width (ft)	–	6.494	2.480	0	8
Median Width (ft)	–	43.293	21.529	10	100
KABCO	2,527	2.025	4.043	0	59
KABC	795	0.637	1.512	0	22
KAB	577	0.462	1.072	0	15
KA	144	0.115	0.383	0	5
SV	1,356	1.087	2.103	0	26

Illinois (780 segments, 189.61 mi, crash years: 2009–2010)

Variable	Frequency	Mean	Standard Deviation	Minimum	Maximum
Segment Length	–	0.243	0.170	0.100	1.850
AADT (vpd)	–	8356.510	5,008.920	550	32,650
Lane Width (ft)	–	11.977	0.237	11	13
Presence of Paved Shoulder	736	–	–	–	–
Absence of Paved Shoulder	44	–	–	–	–
Shoulder Width (ft)	–	10.485	2.131	0	14
Median Width (ft)	–	43.233	16.676	4	88
KABCO	236	0.303	0.686	0	6
KABC	95	0.122	0.382	0	3
KAB	82	0.105	0.350	0	3
KA	31	0.040	0.195	0	1
SV	161	0.206	0.526	0	4

Minnesota (946 segments, 356.767 mi, crash years: 2009–2011)

Variable	Frequency	Mean	Standard Deviation	Minimum	Maximum
Segment Length	–	0.377	0.353	0.100	3.594
AADT (vpd)	–	11,871.68	6,272.046	1,655	30,392.67
Lane Width (ft)	–	12.032	0.355	10	13
Presence of Paved Shoulder	938	–	–	–	–
Absence of Paved Shoulder	8	–	–	–	–
Shoulder Width (ft)	–	9.445	1.398	0	12
Median Width (ft)	–	62.008	26.382	4	99
KABCO	1,282	1.355	1.770	0	16
KABC	454	0.480	0.832	0	6
KAB	155	0.164	0.453	0	5
KA	15	0.016	0.125	0	1
SV	957	1.012	1.422	0	12

California (1149 segments, 595.217 mi, crash years: 2009–2010)

Variable	Frequency	Mean	Standard Deviation	Minimum	Maximum
Segment Length	–	0.517	0.572	0.1	5.329
AADT (vpd)	–	21,258.250	14,665.696	2,325	79,500
Lane Width (ft)	–	11.995	0.204	11	13
Presence of Paved Shoulder	1,128	–	–	–	–
Absence of Paved Shoulder	21	–	–	–	–
Shoulder Width (ft)	–	8.563	2.150	0	13
Median Width (ft)	–	42.722	30.288	5	99
KABCO	3,883	3.379	6.192	0	71
KABC	1,443	1.256	2.385	0	26
KAB	704	0.613	1.209	0	9
KA	190	0.165	0.460	0	4
SV	1,784	1.553	3.183	0	41

Washington (292 segments, 114.004 mi, crash years: 2009–2011)

Variable	Frequency	Mean	Standard Deviation	Minimum	Maximum
Segment Length	–	0.390	0.402	0.103	2.373
AADT (vpd)	–	14,914.530	7,578.160	3,947	42,310
Lane Width (ft)	–	12.017	0.130	12	13
Presence of Paved Shoulder	292	–	–	–	–
Absence of Paved Shoulder	0	–	–	–	–
Shoulder Width (ft)	–	9.700	0.820	4.500	10.500
Median Width (ft)	–	50.116	27.647	4	180
KABCO	893	3.058	3.794	0	24
KABC	239	0.818	1.237	0	7
KAB	154	0.527	0.972	0	6
KA	32	0.110	0.391	0	3
SV	671	2.298	3.074	0	22

North Carolina (168 segments, 58.947 mi, crash years: 2009–2011)

Variable	Frequency	Mean	Standard Deviation	Minimum	Maximum
Segment Length	–	0.351	0.292	0.100	1.675
AADT (vpd)	–	12,150.397	3,180.136	3,600	23,000
Lane Width (ft)	–	Not Available	Not Available	Not Available	Not Available
Presence of Paved Shoulder	84	–	–	–	–
Absence of Paved Shoulder	84	–	–	–	–
Shoulder Width (ft)	–	4.31	4.176	0	12
Median Width (ft)	–	12.899	3.686	4	30
KABCO	84	0.500	1.089	0	7
KABC	34	0.202	0.575	0	4
KAB	15	0.089	0.286	0	1
KA	1	0.006	0.077	0	1
SV	22	0.131	0.388	0	2

$$p_i(y_i) = \frac{\Gamma\left(y_i + \frac{1}{k_i}\right)}{y! \times \Gamma\left(\frac{1}{k_i}\right)} \times \left(\frac{\frac{1}{k_i}}{\frac{1}{k_i} + \mu_i}\right)^{\frac{1}{k_i}} \times \left(\frac{\mu_i}{\frac{1}{k_i} + \mu_i}\right)^{y_i} \quad (3)$$

The log-likelihood function is computed as the sum of the natural logarithm of each segment's probability function. As previously noted, the mean function, N_{SPFi}, is $\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon_i)$ and $\exp(\varepsilon_i) \sim \Gamma[1, k_i]$. Two NB model formulations are proposed. One is expressed as follows.

$$\hat{N}_{SPFi} = \exp\left[\hat{a} + \hat{b} \times \ln(AADT_i) + \hat{d} \times (S_{wi}) + \hat{e} \times (M_{wi}) + \ln(L_i \times yr_i)\right] \quad (4)$$

The overdispersion is segment specific and is computed as follows as per the HSM.

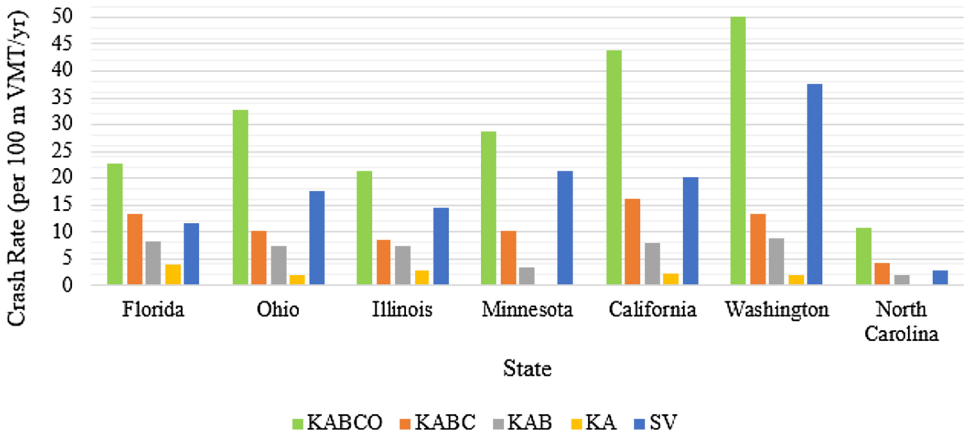

Fig. 1. State's crash rates.

$$\hat{k} = \frac{1}{\exp[\hat{c} + \ln(L_i \times yr_i)]} \tag{5}$$

The predictors, $S_{w\ i}$, $M_{w\ i}$, L_i and yr_i are the shoulder width (ft), median width (ft), segment length (mi) and number of crash years at segment i respectively. The parameter coefficients, denoted by their hats, are estimated via the MLE method. The logarithm of the product of the segment length and the number of years is an offset term that accounts for the period during which crashes occurred. The second model formulation is the same except that the most appropriate Box-Cox transformations are applied to the independent variables. The models are run using the PROC NLMIXED procedure (2015) in base SAS. Once the models are estimated, they are applied to conditions of all other states and the crash count predictions are normalized by the product of the segment length and the number of years.

4.3. Zero-inflated negative binomial model

The ZINB model, is used to compute the probabilities of zero and non-zero crashes, known as the dual state. The model structure is expressed as follows (Cai et al., 2016).

$$p_i(y_i) = \begin{cases} \pi_i + (1-\pi_i) \times \left(\frac{1}{1+k_i\mu_i}\right)^{\frac{1}{k_i}}, & y_i = 0 \\ (1-\pi_i) \times \frac{\Gamma\left(y_i + \frac{1}{k_i}\right) \times (k_i\mu_i)^{y_i}}{\Gamma(y_i+1) \times \Gamma\left(\frac{1}{k_i}\right) \times (1+k_i\mu_i)^{\left(y_i+\frac{1}{k_i}\right)}}, & y_i > 0 \end{cases} \tag{6}$$

The term, π_i, represents a logistic regression model used to compute the probability of observing zero crashes. It is the following.

$$\pi_i = \frac{\exp[\hat{f}]}{1 + \exp[\hat{f}]} \tag{7}$$

The predicted crash counts are calculated as follows where the exponential expression represents μ_i in Eq. (6).

$$N_{SPFi} = (1-\pi_i) \times \exp[\hat{a} + \hat{b} \times \ln(AADT_i) + \hat{d} \times (S_{wi}) + \hat{e} \times (M_{wi}) + \ln(L_i \times y_i)] \tag{8}$$

Another ZINB model is estimated with the parameters transformed. The model runs are conducted using the base SAS PROC GENMOD procedure (2008). After the models are developed, they are transferred to each state. The predicted crash frequencies are normalized by the segment length and the number of crash years as is the case of the NB models.

4.4. Regression tree model

The regression tree model splits the data into subsets depending on whether segments satisfy a range of values of a particular independent variable. For instance, segments with an AADT of less than 5000 vpd are extracted into a subset while the remaining segments remain in the main set. Both subsets are referred to as terminal nodes. The independent variable and its cut-point value are chosen to minimize the residual sum of squares of the differences between observed crash counts and average observed crash counts of segments of which records are in the subset. Similarly, the data splitting process is repeated for each node until either only one record remains in the final node or a specified number of nodes is reached. For more information about regression trees, the reader is referred to James et al. (2013). The dependent variable modeled is the crash count normalized by both the crash years and the segment length. Since an advantage of the regression tree is that no specific relationship is assumed between the dependent variable and the independent variables, no transformations of the independent variables are designated. The "tree" function of the R freeware package "tree" (Ripley, 2016) is used for model runs.

4.5. Random forest model

Random forest regression is an extension of regression trees in that multiple bootstrap samples are collected with replacement from the data. Each sample has a size of the total number of observations in the data. For each bootstrap sample, a regression tree is fit. Yet, in the tree building process, only a subgroup of the independent variables is considered when splitting the data into nodes. That is to reduce the correlation among the nodes in the tree. For a road segment, the average of the crash count predictions of all tree models fit to the bootstrap data is the output predicted crash frequency (James et al., 2013). Similar to the case of the regression tree models, the observed crashes per mile per year are modeled as a function of the AADT, shoulder width and median width. All variables included are not transformed. The analysis is undertaken using the "randomForest" function belonging to the package of the same name in R software (Brieman et al., 2015). The chosen number of bootstrap samples is 500 and the number of independent variables used for the splitting procedure is set to 1.

4.6. Boosting model

Boosting regression is also a tree based method. Yet, unlike RF regression, the boosting model operates by first fitting a tree to the data and then fitting another one onto the residuals. The process of regressing the residuals is iterated until a preset iteration limit is reached (James et al., 2013). The same parameters of the RF model are used with no transformation for the boosting model. The dependent variable is the crash count per mile per year as well. The models are built in R software using the "gbm" function of the package having the same name (Ridgeway, 2017) and the maximum number of boosted trees is set to 5000.

4.7. Tobit model

Tobit regression is similar to OLS regression except that it permits the censorship of specific values of the outcome that are beyond a chosen threshold (Zeng et al., 2017a). Typically, the Tobit framework is employed for modeling crash rates, which are crash counts normalized by the vehicle mileages of travel (Zeng et al., 2017b; Anastasopoulos et al., 2012b, 2008). However, in this paper, the outcome, $N^*_{SPF\ norm\ i}$, is the ratio of the crash frequency to the product of the segment length and the number of crash years. That is to compare the Tobit model's performance with those of the other models. Since the Tobit model's outcome is non-negative, a lower boundary is set to 0 for the regression. The model's configuration is expressed as the following.

$$N_{SPF\ norm\ i} = \begin{cases} N^*_{SPF\ norm\ i}, & N^*_{SPF\ norm\ i} > 0 \\ 0, & N^*_{SPF\ norm\ i} \le 0 \end{cases} \quad (9)$$

Where

$$N^*_{SPF\ norm\ i} = \hat{a} + \hat{b} \times \ln(AADT_i) + \hat{c} \times (S_{wi}) + \hat{d} \times (M_{wi}) \quad (10)$$

Another Tobit formulation is implemented by transforming the parameters of Equation (10) using the most appropriate Box-Cox transformation for each parameter. The Tobit models are estimated in base SAS using the PROC QLIM procedure (2014).

4.8. Poisson lognormal model

The PLN model is the same as the NB model except that the error term follows a lognormal distribution instead of a gamma distribution (Lord and Mannering, 2010). The model structure is the following.

$$N_{SPFi} = \exp[\hat{a} + \hat{b} \times \ln(AADT_i) + \hat{d} \times (S_{wi}) + \hat{d} \times (M_{wi}) + \varepsilon_i] \quad (11)$$

In Eq. (11), $\exp(\varepsilon_i) \sim N[0, \tau]$. The parameter coefficients cannot be estimated by direct computation. Bayesian inference with the use of Markov Chain Monte Carlo (MCMC) simulations is employed to obtain the posterior distributions of the coefficients. A script code in R software is prescribed to run the simulations in WinBUGS open source software with the aid of the package "R2WinBUGS" (Gelman, 2015). The independent variables' coefficients are assumed to follow normal distributions with means and inverse variances of 0 and 10^{-6} as priors respectively. On the other hand, τ, is assumed to be gamma distributed with a mean prior and an inverse variance prior of 0.001. The medians of the estimated independent variables' coefficients are used for crash count prediction. Furthermore, separate PLN models are developed with variables transformed via the most suited Box-Cox transformations as well. All models are transferred to each state and the resulting crash predictions are normalized by the product of the number of years and the segment length as well.

4.9. Hybrid model

The normalized predicted crash frequency results of the Tobit model with the transformed variables, $N_{SPF\ norm\ Tobit\ Tri}$, and of the NB model, $N_{SPF\ norm\ NBi}$, having the configuration described in Eq. (4), are manipulated to achieve predictions of a hybrid model. The computation is the following.

$$N_{SPFi} = \begin{cases} N_{SPF\ norm\ NB\ i}, & N_{SPF\ norm\ Tobit\ i} > 0 \\ 0, & N_{SPF\ norm\ Tobit\ i} = 0 \end{cases} \quad (12)$$

4.10. Goodness of fit measures

All transferred models' predictive performances are compared in terms of mean absolute deviation (MAD), mean squared prediction error (MSPE) and mean absolute percentage deviation (MAPD) defined as shown. The predicted and observed number of crashes (per mile per year) are denoted by $N_{SPF\ norm\ i}$ and $N_{obs\ norm\ i}$ respectively.

$$MAD = \frac{\sum_{i=1}^{n} |N_{SPF\ norm\ i} - N_{obs\ norm\ i}|}{n} \quad (13)$$

$$MSPE = \frac{\sum_{i=1}^{n} (N_{SPF\ norm\ i} - N_{obs\ norm\ i})^2}{n} \quad (14)$$

$$MAPD = \frac{\sum_{i=1}^{n} |N_{SPF\ norm\ i} - N_{obs\ norm\ i}|}{\sum_{i=1}^{n} N_{obs\ norm\ i}} \quad (15)$$

5. Empirical analysis

The SPFs, of the model configurations described previously, are developed successfully for each state and all crash categories except for North Carolina's KA crashes. Only one KA crash is sampled from North Carolina. Also, the KA crash ZINB SPFs of Minnesota failed to converge. Sample SPF results are presented in Tables A1 and A2 in Appendix A. The complete modeling results are available in (Farid, 2018). The SPFs are built following the backward elimination procedure to remove variables that are not statistically significant at a confidence level of 95%. From the model results, it is found that the relationship between the AADT and crash occurrence is not necessarily linear. The shoulder width and median width parameters' results indicate that widening shoulders and medians enhances road safety.

All models, estimated, are transferred to each state's conditions and

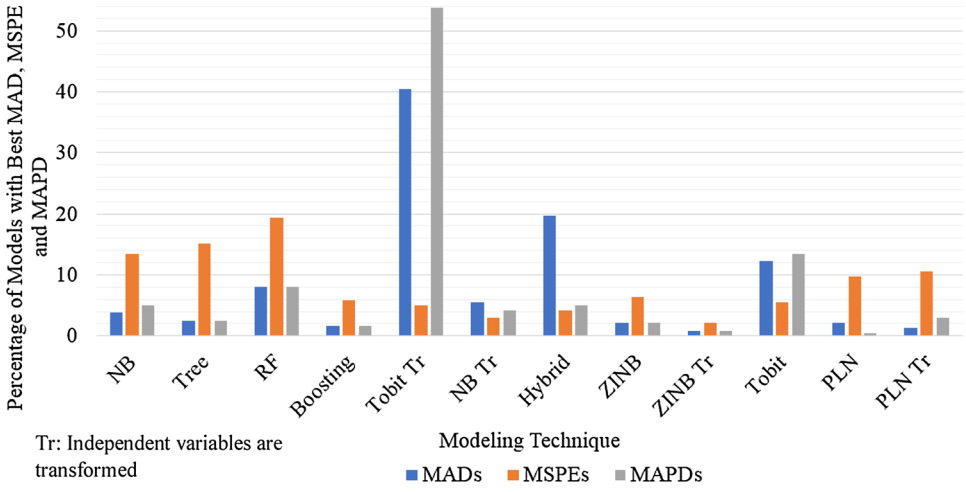

Fig. 2. Comparison of models' performances.

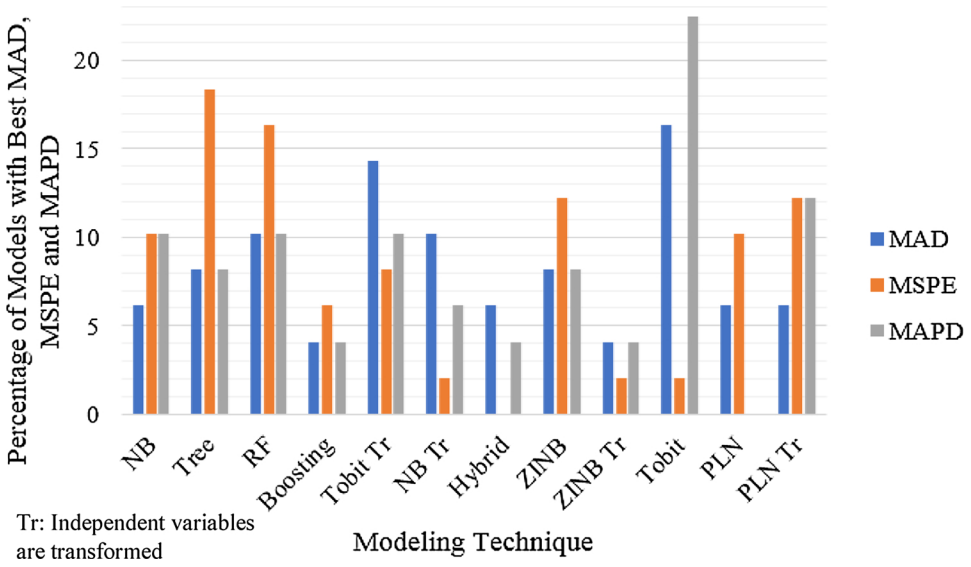

Fig. 3. Comparison of KABCO models' performances.

the models' performances are compared in terms of the MADs, MSPEs and MAPDs. For each model structure, thirty-five SPFs are developed (seven states each having five crash categories) which are transferred to the seven states' conditions. That is a total of 238 SPF transfers since no analysis is conducted on North Carolina's KA crashes. Fig. 2 depicts the comparison of model performances.

From Fig. 2, it can be interpreted that not a single model type is the superior one. Yet, the Tobit SPFs with transformed variables exhibit the best MAPD results for 53.8% (128/238) of the model transfers and the best MAD results for 40.3% (96/238) of the model transfers. The second largest proportion of best MADs is attributed to the hybrid model. The largest proportion of transferred models with best MSPEs is that of RF SPFs followed by those of the tree and NB SPFs. The hybrid model underperforms the Tobit model with transformed parameters even though it combines the results of both the Tobit model, having transformed variables, and the NB model. A further investigation is made by comparing the model performances by crash category as illustrated in Figs. 3–7. Also, detailed GOF results of the Tobit models, with transformed variables, RF models, hybrid models and NB models of KABCO crashes are presented in Tables B1–B3 in Appendix B.

As shown in Figs. 3–7, the Tobit SPFs with transformed variables consistently perform the best in the majority of model transfers especially for the KAB and KA crash categories. That is in terms of MAD and MAPD. The hybrid model is the next top SPF in terms of MAD results specifically for KABC, KAB and KA crashes. A plausible explanation for the remarkable performance of the Tobit model is that it captures the excess zero observed crash counts when transferred particularly when negative predictions are assigned as zero. That is typical of the records of the severe crash categories observed in the data analyzed. Even though the ZINB model is intended to capture excess zero crash counts it does not perform as well as the Tobit model when transferred to conditions elsewhere. As previously noted, the ZINB model structure is subject to criticism because it fails to properly represent crash patterns assuming excess zero observed crash records (Lord, 2006). The RF, tree and NB models exhibit large proportions of best MSPEs for most crash categories. The RF model's performance is better than those of the regression trees and the boosting models possibly because an advantage of the RF model is that it controls for correlations among tree nodes. The NB model demonstrates satisfactory performance. Yet, it is crucial to note that the HSM's SPFs are traditional NB models which may not be

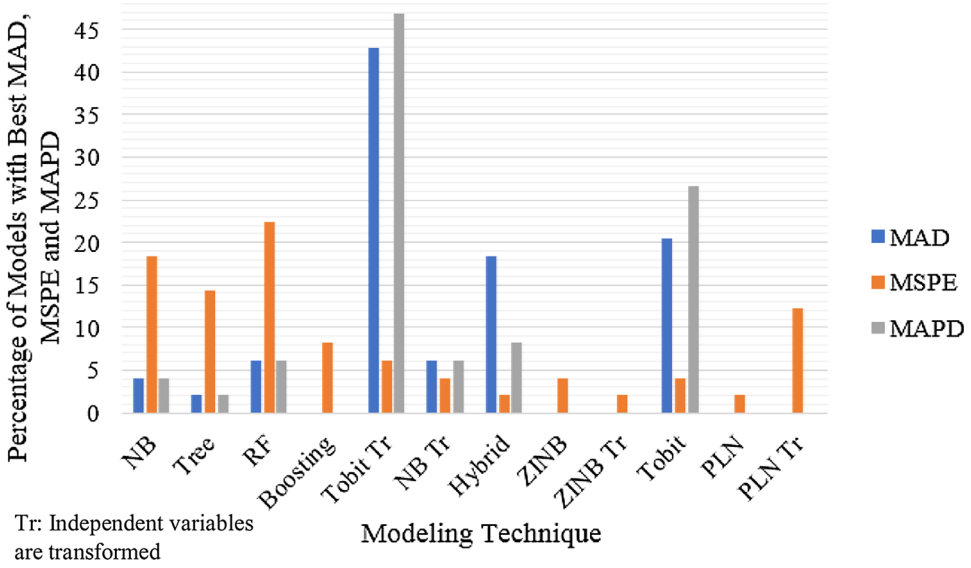

Fig. 4. Comparison of KABC models' performances.

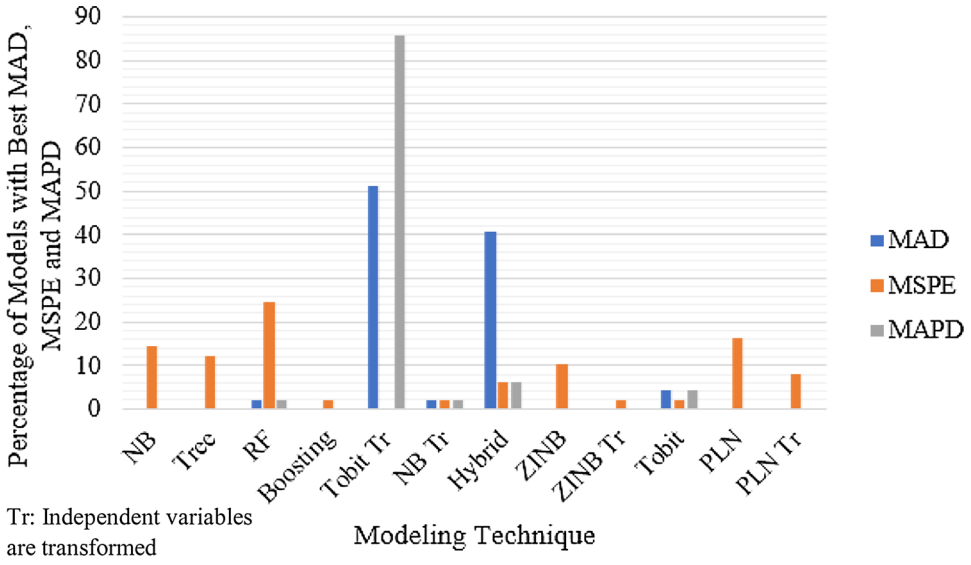

Fig. 5. Comparison of KAB models' performances.

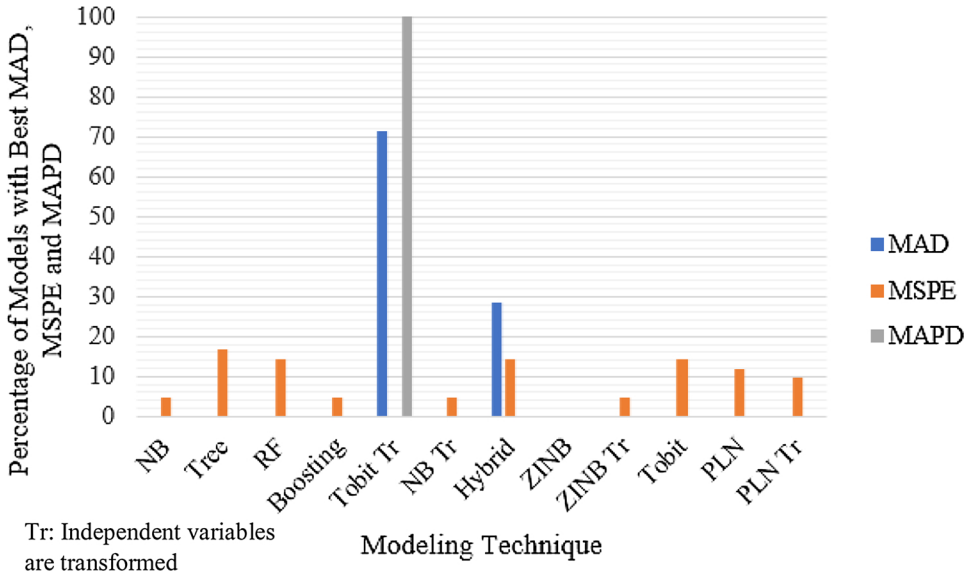

Fig. 6. Comparison of KA models' performances.

transferable relative to Tobit models with independent variables being transformed. Boosting and PLN models demonstrate mediocre results. Cumulative residual (CURE) plots (Hauer and Bamfo, 1997) are attempted to gauge the magnitudes of the residuals along the range of prevalent AADTs. However, the results are not shown because of space limitations. According to the plot results, for AADTs less than 5000 vpd, the 95th percentile confidence band contains 0 in numerous cases. The 99th percentile confidence limits contain 0 for AADTs up to 10,000 vpd.

6. Conclusions

In the road safety literature, diverse modeling methods are available for predicting crash counts, the common one of which is the NB model. The HSM's SPFs are NB models. Transferring SPFs is beneficial for jurisdictional agencies preferring to adopt SPFs instead of developing them thereby reducing costs. Hardly have researchers delved into investigating the transferability of SPFs let alone the transferability of SPFs of various modeling techniques. In this paper, the transferability of NB, ZINB, PLN, tree, RF, boosting and Tobit SPFs, developed for rural divided multilane highway segments for seven states, is investigated. That is to ascertain which model is the most versatile. The states are Florida, Ohio, Illinois, Minnesota, California, Washington and North Carolina. The SPFs are of KABCO, KABC, KAB, KA and SV crashes. Independent variables considered are traffic flow and geometric design elements. Box-Cox transformations are also applied to the independent variables to reduce the skewnesses of the variables' distributions.

The chief finding is that there is no single model type that is demonstrated to outdo all others, in terms of fit, when transferred from one jurisdiction to another. Yet, as per the findings, a large proportion of Tobit SPFs, with the independent variables transformed using the Box-Cox transformations, exhibit better MAD and MAPD results than those of the other models when transferred elsewhere. All other types of

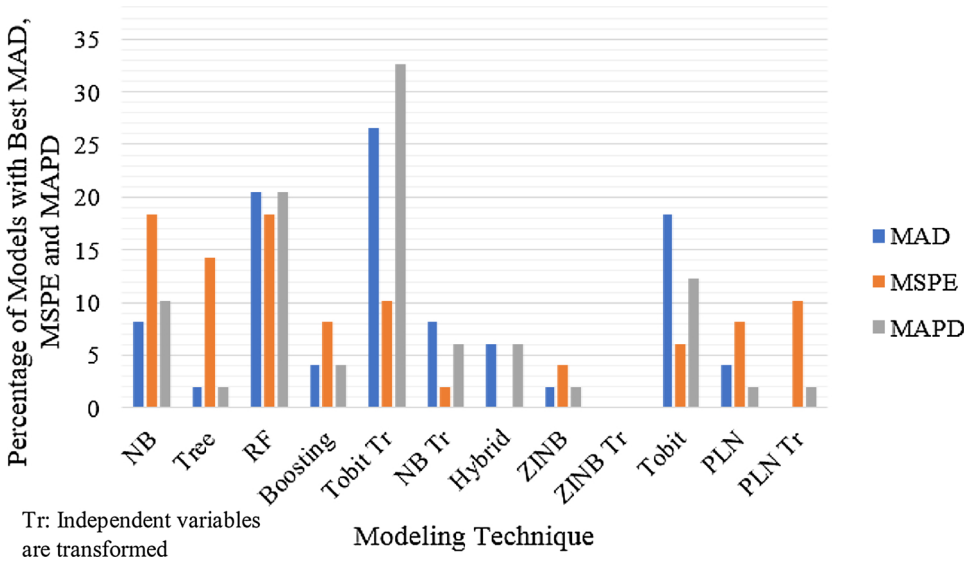

Fig. 7. Comparison of single-vehicle models' performances.

SPFs exhibit lower proportions of MADs and MAPDs. That is possibly because the Tobit's negative crash frequency predictions are set to zero and therefore accommodate the excess zero crash counts in the data of the jurisdictions to which the SPFs are transferred. That is particularly for severe crash count records observed in the data. On the other hand, the largest proportion of SPFs with the best MSPEs is that of RF models, tree and NB models. A hybrid model that manipulates the predictions of both the NB model and the Tobit model with transformed independent variables is proposed. It demonstrates the second best largest proportion of best MAD results after the Tobit model with transformed independent variables. The regression tree, boosting, ZINB and PLN models do not consistently demonstrate high performances for each crash category. As a recommendation, Tobit models, with the application of Box-Cox transformations, RF and hybrid modeling frameworks may be considered for developing and transferring SPFs alongside NB models. The Tobit structure with transformed variables is suggested specifically for developing and transferring SPFs of crash data having excess zero counts. In this context, excess zero counts of severe (KAB and KA) crashes are observed. Data belonging to specific rural multilane roads, not used for this paper, or data of other roadway facility types may not have excess zero counts of severe crashes. The Tobit, RF, tree, NB and hybrid models demonstrate better predictive performances than those of the other methods in a considerably large proportion of the transferred SPFs. The RF, NB and tree frameworks are suggested for developing and transferring SPFs of crash data characterized by overdispersion. In the data, used for the analyses, overdispersed crash records are those of KABCO, KABC and SV crashes. For data of other settings, the crash distributions can be different. For instance, specific crash records can be underdispersed. For data having low sample means or sizes, none of the models employed are recommended.

Undoubtedly, this paper's research is not without limitations. All SPFs, developed and transferred, suffer from omitted variable bias (Lord and Mannering, 2010) which results in errors associated with the estimates of the independent variables' coefficients. Collecting data about more variables common to all the seven states is a challenge but worth the endeavor. Random parameters are also not taken into consideration. Mannering et al. (2016) assert that omitting random parameters inhibits the capturing of unobserved heterogeneity effects. However, incorporating random parameters renders the SPFs to be not transferable to other settings (Mannering et al., 2016). Incorporating finite mixture effects and random effects into the SPFs may also deter SPF transferability (Lord and Mannering, 2010). Furthermore, the generalized additive model and the hierarchical model structures are not attempted because they are also difficult to transfer to data of jurisdictions elsewhere (Lord and Mannering, 2010). Other than the regression techniques that are difficult to transfer, it should be noted that there are several viable techniques that could've been attempted. The transferability of Conway-Maxwell Poisson, gamma, negative multinomial, multivariate and generalized estimating equation SPFs (Lord and Mannering, 2010) is worth investigating. Another shortcoming is that the Tobit, RF, tree and hybrid models, recommended, are not applicable to before-and-after countermeasure deployment analysis using the empirical Bayes (EB) method prescribed by the HSM. The EB method depends on weights which are a function of the overdispersion of the NB model. Yet, Tobit, RF, tree and hybrid structures are applicable for evaluating the safety of alternative road designs.

Acknowledgments

The authors appreciate the help of NCHRP 17-62 team, including our team, Drs. John Ivan, Raghavan Srinivasan and Bhagwant Persaud. The authors also appreciate the efforts of the FDOT and the HSIS research center for sharing their data with us as well. All opinions, stated in this paper, are those of the authors only.

Appendix A

Table A1
Negative Binomial SPF Results.

Florida (436 segments, 350.641 mi, crash years: 2009–2011)

Crash Category	Parameter Estimate with P-Value (in Parentheses)					
	Constant	Ln(AADT)	Shoulder Width	Median Width	$\hat{c}$	−2LL
KABCO	−6.226 (< 0.001)	0.720 (< 0.001)	−0.098 (0.006)	–	−0.440 (0.001)	1485.2
KABC	−5.109 (< 0.001)	0.552 (< 0.001)	−0.112 (0.006)	–	−0.373 (0.038)	1139.7
KAB	−4.325 (< 0.001)	0.361 (< 0.001)	–	–	−0.542 (0.022)	883.9
KA	−4.316 (< 0.001)	0.285 (0.023)	–	–	−0.449 (0.299)	595.1
SV	−0.611	–	–	–	−0.515	1,070.5

Ohio (1248 segments, 661.716 mi, crash years: 2009–2011)

Crash Category	Parameter Estimate with P-Value (in Parentheses)					
	Constant	Ln(AADT)	Shoulder Width	Median Width	$\hat{c}$	−2LL
KABCO	−9.709 (< 0.001)	1.125	−0.074 (< 0.001)	–	0.199 (0.036)	3909.5
KABC	−9.642 (< 0.001)	1.008 (< 0.001)	−0.092 (< 0.001)	–	−0.186 (0.264)	2263.8
KAB	−9.107 (< 0.001)	0.906 (< 0.001)	−0.076 (0.001)	–	−0.265 (0.196)	1923.9
KA	−8.891 (< 0.001)	0.679 (< 0.001)	–	–	0.162 (0.847)	795.7
SV	−7.182 (< 0.001)	0.736 (< 0.001)	–	–	0.203 (0.144)	2986.7

Illinois (780 segments, 189.61 mi, crash years: 2009–2010)

Crash Category	Parameter Estimate with P-Value (in Parentheses)					
	Constant	Ln(AADT)	Shoulder Width	Median Width	$\hat{c}$	−2LL
KABCO	−9.221 (< 0.001)	0.974 (< 0.001)	–	–	0.877 (0.010)	1000.1
KABC	−9.333 (< 0.001)	0.886 (< 0.001)	–	–	1.075 (0.204)	552.2
KAB	−9.552 (< 0.001)	0.894 (< 0.001)	–	–	2.185 (0.420)	492.5
KA	−11.359 (< 0.001)	0.985 (0.003)	–	–	12.306 (0.950)	240.1
SV	−8.494 (< 0.001)	0.852 (< 0.001)	–	–	0.871 (0.060)	801.3

Minnesota (946 segments, 356.767 mi, crash years: 2009–2011)

Crash Category	Parameter Estimate with P-Value (in Parentheses)					
	Constant	Ln(AADT)	Shoulder Width	Median Width	$\hat{c}$	−2LL
KABCO	−9.800	1.067	–	–	0.524	3,064.0
KABC	−9.754	0.953	–	–	0.501	1,795.5
KAB	−7.748	0.626	–	–	0.054	905.8
KA	−4.268	–	–	–	10.279	149.5
SV	−9.291	0.983	–	–	0.432	2,671.7

California (1149 segments, 595.217 mi, crash years: 2009–2010)

Crash Category	Parameter Estimate with P-Value (in Parentheses)					
	Constant	Ln(AADT)	Shoulder Width	Median Width	$\hat{c}$	−2LL
KABCO	−9.253	1.111	−0.053	−0.006	0.698	4,530.0
KABC	−9.885 (< 0.001)	1.088 (< 0.001)	−0.0661 (0.001)	−0.006 (< 0.001)	0.722 (< 0.001)	2906.0
KAB	−9.486 (< 0.001)	0.973 (< 0.001)	−0.085 (< 0.001)	–	0.786 (< 0.001)	2014.0
KA	−8.357 (< 0.001)	0.690 (< 0.001)	–	−0.007 (0.016)	1.172 (0.193)	954.0
SV	−7.966 (< 0.001)	0.906 (< 0.001)	−0.075 (< 0.001)	–	0.501 (< 0.001)	3250.6

Washington (292 segments, 114.004 mi, crash years: 2009–2011)

Crash Category	Parameter Estimate with P-Value (in Parentheses)					
	Constant	Ln(AADT)	Shoulder Width	Median Width	$\hat{c}$	−2LL
KABCO	−3.417	0.480	–	−0.004	1.154	1,132.2
KABC	−0.358	–	–	–	0.858	631.9
KAB	−0.798	–	–	–	0.330	485.4
KA	−1.293 (0.001)	–	–	−0.024 (0.005)	−0.897 (0.245)	173.2
SV	−3.577	0.447	–	–	0.978	1,016.9

North Carolina (168 segments, 58.947 mi, crash years: 2009–2011)

Crash Category	Parameter Estimate with P-Value (in Parentheses)				
	Constant	Ln(AADT)	Shoulder Width	Median Width	$\hat{c}$	−2LL
KABCO	−14.761	1.486	–	–	−0.544	323.6
KABC	−1.220 (< 0.001)	–	−0.122 (0.032)	–	−0.803 (0.184)	175.0
KAB	−2.467	–	–	–	11.409	103.6
KA	Only 1 observed KA crash - no SPF estimated					
SV	−19.857	1.884	–	–	11.045	133.4

-2LL: $-2 \times$ SPF log-likelihood.
-: statistically insignificant variable at 95th percentile confidence level removed.

Table A2
Tobit SPFs with Transformed Variables.

Florida (436 segments, 350.641 mi, crash years: 2009–2011)							Ohio (1248 segments, 661.716 mi, crash years: 2009–2011)					
Crash Category	Parameter Estimate with P-Value (in Parentheses)						Parameter Estimate with P-Value (in Parentheses)					
	Constant	AADT	Shoulder Width	Median Width	σ	−2LL	Constant	AADT	Shoulder Width	Median Width	σ	−2LL
Variable Transformation	NA	Natural Log	Natural Log	Natural Log			NA	Natural Log	Cubic	None		
KABCO	−12.434	1.537	−1.286	–	4.138	1,648.7	−21.883	2.517	−0.005	–	2.617	4,142.0
KABC	−5.925	0.736	−1.015	–	2.83	1,264.7	−11.692	1.246	−0.003	–	1.717	2,386.0
KAB	−1.040	–	–	–	2.303	980.8	−10.270	1.046	−0.002	–	1.659	2,048.0
KA	−1.296	–	–	–	1.747	680.8	−8.308	0.689	–	–	1.545	941.2
SV	−0.643	–	–	–	2.411	1,145.3	−10.410	1.121	–	–	1.877	3,158.0
Illinois (780 segments, 189.61 mi, crash years: 2009–2010)							Minnesota (946 segments, 356.767 mi, crash years: 2009–2011)					
Crash Category	Parameter Estimate with P-Value (in Parentheses)						Parameter Estimate with P-Value (in Parentheses)					
	Constant	AADT	Shoulder Width	Median Width	σ	−2LL	Constant	AADT	Shoulder Width	Median Width	σ	−2LL
Variable Transformation	NA	Natural Log	Square	None			NA	Natural Log	Cubic	None		
KABCO	−25.543	2.457	–	–	4.548	1,413.7	−25.047	2.772	–	–	4.004	3,732.0
KABC	−22.012	1.839	–	–	4.446	796.1	−15.509	1.542	–	–	2.816	2,192.0
KAB	−18.272	1.612	−0.030	–	4.143	708.0	−10.891	0.831	–	–	2.943	1,126.3
KA	−23.559	1.774	–	–	4.362	353.0	−6.910	–	–	–	3.218	188.2
SV	−23.225	2.102	–	–	4.533	1,141.4	−21.050	2.272	–	–	3.487	3,230.0
California (1149 segments, 595.217 mi, crash years: 2009–2010)							Washington (292 segments, 114.004 mi, crash years: 2009–2011)					
Crash Category	Parameter Estimate with P-Value (in Parentheses)						Parameter Estimate with P-Value (in Parentheses)					
	Constant	AADT	Shoulder Width	Median Width	σ	−2LL	Constant	AADT	Shoulder Width	Median Width	σ	−2LL
Variable Transformation	NA	Natural Log	Square	Square Root			NA	Natural Log	Cubic	Square Root		
KABCO	−62.104	6.807	−0.065	–	9.127	6,197.0	−20.206	2.368	–	–	4.043	1,364.8
KABC	−33.137	3.460	−0.039	–	4.717	3,879.7	−0.427	–	–	–	2.44	796.1
KAB	−19.836	1.942	−0.019	–	3.130	2,599.0	−1.109	–	–	–	2.246	609.4
KA	−12.658	1.007	–	–	2.493	1,262.3	−2.890	–	–	–	2.11	218.4
SV	−32.763	3.299	–	–	5.412	4,387.5	−13.683	1.587	–	–	3.247	1,178.2

North Carolina (168 segments, 58.947 mi, crash years: 2009–2011)

Crash Category	Parameter Estimate with P-Value (in Parentheses)					
	Constant	AADT	Shoulder Width	Median Width	σ	−2LL
Variable Transformation	NA	Thousandths	Square Root	Inverse Square Root		
KABCO	−4.824	0.234	–	–	3.41	364.5
KABC	−5.906	0.233	–	–	2.819	200.4
KAB	−3.100	–	–	–	2.322	133.8
KA	Only 1 observed KA crash - no SPF estimated					
SV	−3.042	–	–	–	2.541	165.2

σ: square root of the variance of the residuals.
−2LL: −2 × SPF log-likelihood.
-: statistically insignificant variable at 95th percentile confidence level removed.

Appendix B

Table B1
MADs of Transferred KABCO Crash SPFs.

Tobit Model with Variable Transformation							
SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	1.298	2.108	1.055	2.066	3.962	2.862	10.546
Ohio	1.444	1.023	0.653	1.645	3.238	2.393	1.212
Illinois	1.355	1.240	0.621	1.840	4.033	2.888	0.613
Minnesota	1.362	1.034	0.779	1.588	3.201	2.195	1.042
California	1.908	1.139	0.680	1.703	3.244	2.459	1.247
Washington	1.591	1.170	1.116	1.695	3.292	2.106	1.697
North Carolina	1.412	1.201	0.628	1.791	3.324	2.750	0.609

Random Forest Model

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	1.076	1.530	1.074	1.669	3.593	2.264	1.234
Ohio	1.699	0.934	1.152	1.725	3.325	2.133	1.679
Illinois	1.498	1.130	0.807	1.645	3.535	2.490	1.273
Minnesota	1.504	1.420	1.385	1.486	3.544	2.063	2.179
California	2.042	1.650	1.825	2.015	2.475	2.339	2.396
Washington	1.711	1.733	1.952	2.005	3.638	1.716	1.869
North Carolina	1.446	1.370	1.839	2.074	3.795	2.372	0.719

Hybrid Model

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	1.277	1.171	0.636	1.786	3.666	2.788	0.999
Ohio	1.567	1.043	0.725	1.621	3.086	2.234	1.553
Illinois	1.398	1.230	0.621	1.840	3.713	2.828	0.613
Minnesota	1.377	1.058	0.952	1.565	3.143	2.194	0.924
California	1.746	1.143	0.684	1.648	3.140	2.277	1.833
Washington	1.956	1.509	1.472	1.886	3.392	2.122	2.389
North Carolina	1.346	1.190	0.628	1.769	3.501	2.726	0.606

Negative Binomial Model

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	1.307	1.085	0.857	1.670	3.588	2.485	1.168
Ohio	1.602	1.083	0.985	1.619	3.100	2.180	1.591
Illinois	1.291	1.046	0.941	1.610	3.348	2.288	1.010
Minnesota	1.400	1.052	1.048	1.611	3.150	2.144	1.188
California	1.871	1.180	1.139	1.652	3.159	2.100	2.148
Washington	2.022	1.641	1.856	1.925	3.411	2.094	2.430
North Carolina	1.282	1.094	0.750	1.662	3.413	2.496	0.795

Table B2
MSPEs of Transferred KABCO Crash SPFs.

Tobit Model with Variable Transformation

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	8.863	27.019	13.051	17.932	76.125	19.462	256.091
Ohio	8.422	3.450	2.194	9.712	60.746	14.903	2.579
Illinois	10.079	5.434	2.403	12.076	77.385	19.842	2.307
Minnesota	8.410	3.596	2.171	8.727	59.362	12.759	2.177
California	12.007	3.964	2.604	9.976	50.735	14.878	3.301
Washington	8.728	3.385	2.650	7.949	57.647	11.078	3.939
North Carolina	9.614	4.932	2.359	11.565	58.357	18.332	2.291

Random Forest Model

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	3.960	5.831	2.594	9.476	68.040	13.864	2.977
Ohio	9.387	2.114	2.258	8.355	58.298	12.019	4.131
Illinois	8.541	3.727	1.523	9.157	67.423	15.486	2.910
Minnesota	8.398	4.040	2.883	5.721	59.803	10.553	7.448
California	11.569	4.572	5.028	9.279	23.570	11.494	7.293
Washington	8.598	4.788	4.517	8.402	60.258	7.039	4.745
North Carolina	8.317	4.273	4.202	9.765	71.852	14.241	1.397

Hybrid Model

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	8.426	4.538	2.361	11.454	69.793	18.709	2.386
Ohio	8.908	3.250	2.209	9.031	53.693	13.213	3.564
Illinois	9.832	5.315	2.403	12.076	68.881	19.299	2.308
Minnesota	8.738	3.819	2.154	8.570	56.949	12.792	2.103
California	10.344	3.776	2.449	9.339	48.905	13.204	5.406
Washington	9.441	4.170	3.674	8.149	57.399	10.835	6.730
North Carolina	9.348	4.933	2.356	11.409	65.557	18.000	2.278

Negative Binomial Model

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	7.930	3.717	1.965	9.746	68.451	15.747	2.296
Ohio	8.743	3.113	2.018	8.451	53.404	12.685	3.520
Illinois	8.113	3.593	1.944	8.932	63.363	13.899	1.966
Minnesota	8.130	3.264	2.068	8.283	56.794	12.367	2.301
California	10.063	3.169	2.271	8.095	48.257	11.480	5.811
Washington	9.330	4.308	4.064	8.114	57.365	10.667	6.798
North Carolina	8.633	4.193	2.055	10.017	64.160	15.829	1.882

Table B3
MAPDs of Transferred KABCO Crash SPFs.

Tobit Model with Variable Transformation

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	0.956	1.696	1.698	1.123	0.955	0.986	17.211
Ohio	1.063	0.823	1.051	0.894	0.780	0.825	1.978
Illinois	0.998	0.998	1.000	1.000	0.972	0.995	1.000
Minnesota	1.003	0.832	1.253	0.863	0.771	0.757	1.700
California	1.405	0.916	1.094	0.926	0.782	0.848	2.035
Washington	1.172	0.942	1.796	0.921	0.793	0.726	2.769
North Carolina	1.040	0.967	1.010	0.974	0.801	0.948	0.995

Random Forest Model

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	0.793	1.232	1.729	0.907	0.866	0.780	2.014
Ohio	1.251	0.752	1.854	0.937	0.801	0.735	2.741
Illinois	1.104	0.909	1.298	0.894	0.852	0.858	2.078
Minnesota	1.108	1.143	2.229	0.808	0.854	0.711	3.557
California	1.504	1.328	2.937	1.095	0.597	0.806	3.911
Washington	1.260	1.395	3.141	1.090	0.877	0.591	3.051
North Carolina	1.065	1.102	2.959	1.127	0.915	0.818	1.174

Hybrid Model

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	0.940	0.942	1.024	0.971	0.884	0.961	1.631
Ohio	1.155	0.839	1.167	0.881	0.744	0.770	2.535
Illinois	1.030	0.990	1.000	1.000	0.895	0.975	1.000
Minnesota	1.014	0.852	1.532	0.850	0.757	0.756	1.507
California	1.286	0.920	1.100	0.896	0.757	0.785	2.991
Washington	1.441	1.215	2.369	1.025	0.818	0.731	3.899
North Carolina	0.992	0.958	1.011	0.962	0.844	0.940	0.989

Negative Binomial Model

SPF	Application Data						
	Florida	Ohio	Illinois	Minnesota	California	Washington	North Carolina
Florida	0.962	0.873	1.379	0.908	0.865	0.856	1.906
Ohio	1.180	0.871	1.584	0.880	0.747	0.752	2.597
Illinois	0.951	0.842	1.513	0.875	0.807	0.789	1.649
Minnesota	1.031	0.847	1.686	0.876	0.759	0.739	1.938
California	1.378	0.950	1.833	0.898	0.761	0.724	3.506
Washington	1.489	1.320	2.986	1.046	0.822	0.722	3.965
North Carolina	0.944	0.881	1.206	0.903	0.823	0.861	1.297

References

Al Kaaf, K., Abdel-Aty, M., 2015. Transferability and calibration of highway safety manual performance functions and development of new models for urban four-lane divided Roads in Riyadh. Presented at 94th Annual Meeting of the Transportation Research Board.

Ambros, J., Sedoník, J., 2016. A feasibility study for developing a transferable accident prediction model for czech regions. Transp. Res. Procedia 14, 2054–2063.

American Association of State Highway and Transportation Officials, 2010. Highway Safety Manual. Washington, D.C.. .

Anastasopoulos, P., Tarko, A., Mannering, F., 2008. Tobit analysis of vehicle accident rates on interstate highways. Accid. Anal. Prev. 40 (2), 768–775.

Anastasopoulos, P., Mannering, F., Shankar, V., Haddock, J., 2012a. A study of factors affecting highway accident rates using the random-parameters tobit model. Accid. Anal. Prev. 45 (1), 628–633.

Anastasopoulos, P., Shankar, V., Haddock, J., Mannering, F., 2012b. A multivariate tobit analysis of highway accident-injury-severity rates. Accid. Anal. Prev. 45, 110–119.

Box, G., Cox, D., 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B 26, 211–252.

Brieman, L., Cutler, A., Liaw, A., Wiener, M., 2015. Package 'randomForest'. Package Version 4.6-12.Package 'randomForest'. Package Version 4.6-12.

Brimley, B., Saito, M., Schultz, G., 2012. calibration of highway safety manual safety performance function: development of new models for rural two-lane two-way highways. Transp. Res. Rec. J. Transp. Res. Board 2279, 82–89.

Cafiso, S., Di Silvestro, G., Di Guardo, G., 2012. Application of highway safety manual to Italian divided multilane highways. Procedia-Soc. Behav. Sci. 53, 910–919.

Cai, Q., Lee, J., Eluru, N., Abdel-Aty, M., 2016. Macro-level pedestrian and bicycle crash analysis: incorporating spatial spillover effects in dual state count models. Accid. Anal. Prev. 93, 14–22.

Cunto, F., Sobreira, L., Ferreira, S., 2015. Assessing the transferability of highway safety manual predictive method for urban roads in Fortaleza City, Brazil. Am. Soc. Civil Eng. J. Transp. Eng. 141 (1) Content ID 04014072.

Elfar, A., Talebpour, A., Mahmassani, H., 2018. Machine learning approach to short-term traffic congestion prediction in a connected environment. Presented at 97th Annual Meeting of the Transportation Research Board.

Farid, A., 2018. Investigating and Facilitating the Transferability of Safety Performance Functions. Transportation Engineering Doctoral Dissertation. University of Central Florida, Orlando, FL, pp. 135–183.

Farid, A., Abdel-Aty, M., Lee, J., 2018. Transferring and calibrating safety performance functions among multiple states. Accid. Anal. Prev. 117, 276–287.

Gelman, A., 2015. Package 'R2WinBUGS'. Package Version 2.1-21.

SAS/STAT(R) 9.2 User's Guide, 2008. The GENMOD Procedure, second edition. Statistical Analysis Software Institute, Cary, North Carolina Accessed December 28, 2017. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_genmod_sect009.htm.

Hauer, E., Bamfo, J., 1997. Two tools for finding what function Links the dependent variable to the explanatory variables. Proceedings of the International Co-Operation on Theories and Concepts in Traffic Safety Conference.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. Springer, New York City, NY.

Kan, Y., Wang, Y., Wang, D., Sun, J., Shao, C., 2018. A novel approach to missing ramp flow imputation using machine learning. Presented at 97th Annual Meeting of the Transportation Research Board.

Lord, D., 2006. Modeling motor vehicle crashes using poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accid. Anal. Prev. 38 (4), 751–766.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44 (5), 291–305.

Lord, D., Miranda-Moreno, L., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of poisson-gamma models for modeling motor vehicle crashes: a bayesian perspective. Saf. Sci. 46 (5), 751–770.

Lord, D., Guikema, S., Geedipally, S., 2008. Application of the conway-maxwell poisson generalized linear model for analyzing motor vehicle crashes. Accid. Anal. Prev. 40 (3), 1123–1134.

Mannering, F., Shankar, V., Bhat, C., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

Martz, P., Bill, A., Khan, G., Noyce, D., 2017. Safety performance function for undivided rural Two-lane roadways using regression tree analysis. Transportation Research Board 96th Annual Meeting Compendium of Papers 17-05652.

Mehta, G., Lou, Y., 2013. Calibration and development of safety performance functions for Alabama: two-lane, two-way rural roads and four-lane divided highways. Transp. Res. Rec. J. Transp. Res. Board 2398, 75–82.

Miaou, S., Song, J., Mallick, B., 2003. Roadway traffic crash mapping: a space-time modeling approach. J. Transp. Stat. 6 (1), 33–57.

SAS/STAT(R) 9.2 User's Guide, 2015. The NLMIXED Procedure, second edition. Statistical Analysis Software Institute, Cary, North Carolina Accessed December 28, 2017. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#nlmixed_toc.htm.

Park, H.-C., Kim, D.-K., Kho, S.-Y., 2018. Bayesian network for the traffic State prediction. Presented at 97th Annual Meeting of the Transportation Research Board.

Persaud, B., Lord, D., Palmisano, J., 2002. Calibration and transferability of accident prediction models for urban intersections. Transp. Res. Rec. J. Transp. Res. Board 1784, 57–64.

SAS/ETS(R) 9.3 User's Guide, 2014. The QLIM Procedure. Statistical Analysis Software Institute, Cary, North Carolina. http://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_qlim_sect001.htm Accessed December 28, 2017.

Ridgeway, G., 2017. Package 'gbm'. Package Version 2.1.3.

Ripley, B., 2016. Package 'Tree'. Package Version 1.0-37.

Srinivasan, R., Carter, D., 2011. Development of Safety Performance Functions for North Carolina. Publication FHWA/NC/2010-09. Research and Analysis Group, North Carolina Department of Transportation, North Carolina.

Srinivasan, R., Carter, D., Bauer, K., 2013. Safety Performance Function Decision Guide: SPF Calibration versus SPF Development. Publication FHWA-SA-14-004. Federal Highway Administration, U.S. Department of Transportation, Washington D.C.

State Data, 2017. Highway Safety Information System. Highway Safety Research Center, University of North Carolina, Chapel Hill, North Carolina Accessed May 23, 2017. http://www.hsisinfo.org/.

Sun, C., Brown, H., Edara, P., Claros, B., Nam, K., 2014. Calibration of the HSM's SPFs for Missouri. Publication CMR14-007. Missouri Department of Transportation, MO.

Sun, B., Cheng, W., Goswami, P., Bai, G., 2018. Short-term traffic forecasting using self-adjusting K-Nearest neighbors. Inst. Eng. Technol. Intell. Transp. Syst. 12 (1), 41–48.

Wang, C., Hao, P., Wu, G., Qi, X., Barth, M., 2018. Predicting the number of uber pickups by deep learning. Presented at 97th Annual Meeting of the Transportation Research Board.

Washington, S., Karlaftis, M., Mannering, F., 2003. Statistical and Econometric.

Xie, F., Gladhill, K., Dixon, K., Monsere, C., 2011. Calibration of the highway safety manual predictive models for oregon state highways. Transp. Res. Rec. J. Transp. Res. Board 2241, 19–28.

Young, J., Park, P., 2012. Comparing the highway safety manual's safety performance functions with jurisdiction specific functions for intersections in Regina. Presented at 2012 Annual Meeting of the Transportation Association of Canada Federation.

Zeng, Q., Wen, H., Huang, H., Abdel-Aty, M., 2017a. A bayesian spatial random parameters tobit model for analyzing crash rates on roadway segments. Accid. Anal. Prev. 100, 37–43.

Zeng, Q., Wen, H., Huang, H., Pei, X., Wong, S., 2017b. A multivariate random-parameters tobit model for analyzing highway crash rates by injury severity. Accid. Anal. Prev. 99 (Part A), 184–191.