Accident Analysis and Prevention 132 (2019) 105274

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Application of a model-based recursive partitioning algorithm to predict crash frequency

Houjun Tang*, Eric T. Donnell

Department of Civil and Environmental Engineering, The Pennsylvania State University, 212 Sackett Building, University Park, PA 16802, United States

ARTICLE INFO

Keywords:
Roadway safety
Crash frequency
Heterogeneity
Data mining
Model-based recursive partitioning

ABSTRACT

Count regression models have been applied widely in traffic safety research to estimate expected crash frequencies on road segments. Data mining algorithms, such as classification and regression trees, have recently been introduced into the field to overcome some of the assumptions associated with statistical models. However, these data-driven algorithms usually provide non-parametric output, making it difficult to draw statistical inference or to evaluate how independent variables are associated with expected crash frequencies. In this paper, the model-based recursive partitioning (MOB) algorithm is applied in a crash frequency application. The algorithm incorporates the concept of recursive partitioning data in tree models and develops user-defined statistical models as outputs. The objective of this paper is to explore the potential of the MOB algorithm as a methodological alternative to parametric modeling methods in crash frequency analysis. To accomplish the objective, a standard negative binomial (NB) regression model, a NB model developed using the MOB algorithm, adjusted NB models which incorporate variables identified by the MOB algorithm, and a random parameters NB model are compared using 8 years of data collected from two-lane rural highways in Pennsylvania. The models are compared in terms of data fitness, sign and magnitude of statistical association between the independent and dependent variables, and predictive power. The results show that the MOB-NB model yields better data fitness than other NB models, and provides similar performance to the RPNB model, suggesting that the MOB-NB model may be capturing unobserved heterogeneity by dividing the data into subgroups. The presence of a passing zone and posted speed limit are two covariates identified by the MOB algorithm that differentiate variable effects among subgroups. In addition, the MOB-NB model provides the highest prediction accuracy based on the training and test data sets, although the difference among models is small. The comparison results reveal that the MOB algorithm is a promising alternative to identify covariates, evaluate variable associations and instability, and make predictions in a crash frequency context.

1. Introduction

Roadway safety is among the top goals of the United States Department of Transportation. According to data published by the National Highway Traffic Safety Administration (NHTSA), 37,461 people lost their lives due to 34,439 fatal crashes in 2016. Both figures have increased in consecutive years since 2014. Additionally, more than 2 million people were injured in traffic crashes in recent years (NHTSA, 2017). The economic and societal impact of traffic crashes is high. A research report revealed that the total economic cost of motor vehicle crashes was $242 billion in 2010 and, when considering societal impacts, the cost was $836 billion (Blincoe et al., 2015). In order to manage safety on the road and street network, predictive tools are needed to estimate the expected performance of roadways on the network.

Crash frequency models relate the number of crashes to site-specific data, such as traffic volume, horizontal and vertical alignment information, as well as cross-section dimensions. Count regression, such as Poisson or negative binomial (NB) models (Jovanis and Chang, 1986; Poch and Mannering, 1996; Shankar et al., 1995; Milton and Mannering, 1998; Hadi et al., 1995), zero-inflated Poisson or NB models (Miaou, 1994; Shankar et al., 1997; Lee and Mannering, 2002), have been widely applied in the crash frequency context. These traditional safety models offer an opportunity to make inferences on the relationships between site-specific roadway characteristics and expected crash frequencies but may have limited predictive power due to omitted variable bias and other limitations associated with parametric modeling assumptions. In order to improve statistical inference and

* Corresponding Author.
E-mail addresses: hut26@psu.edu (H. Tang), edonnell@engr.psu.edu (E.T. Donnell).

https://doi.org/10.1016/j.aap.2019.105274
Received 13 June 2019; Received in revised form 12 August 2019; Accepted 16 August 2019
Available online 22 August 2019

predictions from traditional safety models, recent empirical research has considered heterogeneity models (Lord and Mannering, 2010; Mannering et al., 2016; Mannering, 2018), such as random parameters models (Anastasopoulos and Mannering, 2009; El-Basyouny and Sayed, 2009). These models permit each observation in a sample to have its own parameter in crash frequency modeling context, which likely improves statistical inference and the predictive capability of such models relative to traditional safety models (Mannering, 2018).

In addition to heterogeneity and traditional safety models, researchers have introduced data mining algorithms into the crash frequency modeling framework with the aim of making accurate predictions. Popular examples of these data-driven methods include decision trees (DT) (Karlaftis and Golias, 2002; Abdel-Aty and Keller, 2005; Park and Saccomanno, 2005; Chang and Chen, 2005) and random forests (RF) (Siddiqui et al., 2012). These algorithms do not require assumptions on the variables or the data (Karlaftis and Golias, 2002; Abdel-Aty and Keller, 2005). While the DT and RF algorithms are powerful predictive tools, they are non-parametric methods, and are limited with regards to statistical inference.

To overcome the challenge of making statistical inference using data mining algorithms, a model-based recursive partitioning (MOB) algorithm was recently developed (Zeileis et al., 2008). The approach combines the advantages of non-parametric tree models and parametric statistical models, and shows promising results in social science (Kopf et al., 2010) and medical (Thomas et al., 2018; Seibold et al., 2016; Pirkle et al., 2018) applications. MOB has not been applied in the traffic safety field. Therefore, the objective of this paper is to apply MOB in a crash frequency context and assess its potential as a predictive model, as well as a tool to consider statistical inference. This is accomplished by comparing outcomes between the MOB model, standard NB models (traditional safety models), and random parameters NB models (heterogeneity models).

2. Literature review

Crash frequency is defined as the number of crashes reported at a specific location (e.g., roadway segment, intersection, interchange) over a defined time period. The responses (i.e., crashes) are non-negative integers, which are often modeled using count regression methods. Lord and Mannering (Lord and Mannering, 2010) reviewed the development of statistical and econometric methodologies used to estimate models of crash frequency. Among the methods included in the review were the following: Poisson and NB regression models (Jovanis and Chang, 1986; Poch and Mannering, 1996; Shankar et al., 1995; Milton and Mannering, 1998; Hadi et al., 1995), zero-inflated Poisson and NB models (Miaou, 1994; Shankar et al., 1997; Lee and Mannering, 2002), panel data models (Johansson, 1996; Shankar et al., 1998), random parameters models (Anastasopoulos and Mannering, 2009; El-Basyouny and Sayed, 2009), Poisson-lognormal models, generalized estimating equations, and multivariate models (Lord and Mannering, 2010). Some data mining algorithms, such as neural networks and support vector machines, were identified as predictive tools previously used in a crash frequency context. In addition to relaxing model assumptions, the authors' noted that data mining algorithms are difficult to estimate, and the results do not provide interpretable parameters (Lord and Mannering, 2010).

From a methodological point of view, count regression models of crash frequency were initially applied in the 1980s by Jovanis and Chang (1986), using the Poisson distribution. The authors' used the method to examine the relationship between crash frequency and vehicle miles of travel. The researchers concluded that the Poisson model was superior to conventional linear regression as the normal distribution assumption underlying linear regression provided biased confidence limits. Subsequently, in an effort to handle over-dispersion commonly found in crash data, many researchers estimated expected crash frequencies on various roadway types using the NB regression model. For example, Poch and Mannering (1996) investigated the relationship between crash frequency and traffic and geometric factors based on seven-years of crash data from 63 intersections in Bellevue, Washington. Shankar et al. (1995) evaluated the effects of roadway and environmental characteristics on crash frequency using data from a rural Interstate in Washington State. Milton and Mannering (Milton and Mannering, 1998) estimated the relationship between crash frequency and roadway geometrics and traffic-related variables using data from principal arterials in Washington State. Hadi et al. (1995) assessed the statistical association between crash frequency and highway cross-section configurations using data from Florida. Collectively, these studies established the appropriateness of the NB model to estimate expected crash frequencies as a function of traffic volumes and other roadway and roadside features.

More recently, traffic safety researchers have identified several limitations of NB regression modeling, especially in handling issues such as the existence of excess zeroes in the data, temporal and spatial correlations, and unobserved heterogeneity. To overcome these limitations, other modeling methods have been introduced in the traffic safety literature. Zero-inflated models have been used to account for excess sampling zeros in the crash data. These models consist of two parts: a binary model to determine the possibility of observing a zero-state for each observation and a count model to predict crash frequency. For instance, Miaou (1994) employed the method when studying the relationship between truck accidents and roadway geometric design variables. Shankar et al. (1997) used a zero-inflated model to identify "near-safe" roadway segments from the segments that happened to have zero crashes during the observation period. Lee and Mannering (2002) employed zero-inflated count models to analyze run-off-road crash frequencies in urban and rural areas. The findings all revealed that zero-inflated models were more flexible in identifying variables affecting crash frequency with zero and non-zero counts. However, Lord and Mannering (2010) recently concluded that zero-inflated count models have limitations, such as potential theoretical inconsistencies and an adverse influence resulting from low sample mean values from the zero state.

Random effects NB models have been applied to address spatial and temporal correlation in traffic safety data, which is caused when data are collected on the same roadway segment or intersection over multiple years (repeated observations) or on adjacent roadway segments or intersections. By estimating the effect of speed limits on crash frequency using Swedish roadway data, Johansson (Johansson, 1996) confirmed that the random effects NB model outperformed the fixed effects NB model when predicting crash frequency. Shankar et al. (1998) compared the random effects NB and fixed effects NB models in a study of median crossover frequency. The researchers concluded that random effects NB models performed better than fixed effects NB models only when spatial and temporal indicators were not included in the model specification.

Recent empirical research using heterogeneity models has focused on temporal instabilities in models of expected crash frequencies (Lord and Mannering, 2010; Mannering, 2018). Examples of these models include random parameters models, which allow estimated parameters to vary among observations. Anastasopoulos and Mannering (2009) explored the appropriateness of the random parameters NB model and compared the results to a fixed parameters NB model using data from rural Interstate highways in Indiana. El-Basyouny and Sayed (2009) compared the standard Poisson lognormal (PLN) model to the random parameters PLN model. Both studies reported improvements in the model goodness-of-fit when random parameters were introduced into model estimation. The researchers concluded that the random parameters model could provide unique insights into expected crash frequencies and recommended the method as an alternative to the standard NB approach.

Data mining approaches, although employed in many scientific fields, have not been widely used in traffic safety research. Among the popular data mining algorithms, decision trees (DT) and random forests (RF) have been applied in some safety contexts. There are two common tree model types: (i) classification trees if the dependent variable is categorical and (ii) regression trees if the dependent variable is continuous. In either case, the tree model can be viewed as a series of "if-then" criteria. The full dataset represents the root node in a tree model – data from the root node are then split based on whether they satisfy the condition as each splitting point. The splitting process concludes at the terminal node. For a classification tree, the dominant class of observations in each terminal node is selected as the predicted value for that node, while for a regression tree, the mean value in each terminal node is considered the predicted response. More details about tree models can be found in Breiman et al. (1984).

In recent years, DT and RF algorithms have received increasing attention in the traffic safety literature. For example, Karlaftis and Golias (2002) developed a tree model using data from rural roads in Indiana, with a focus on quantifying the effects of traffic volume and geometric features on crashes. Abdel-Aty and Keller (2005) analyzed crash severity outcomes at signalized intersections using tree models. Park and Saccomanno (2005) evaluated the relationship between crash rates at highway-rail grade crossings and countermeasure effectiveness using a stratified decision tree model, where data were separated into multiple subgroups based on control factors such as highway class (i.e., arterial / collector, local road and others), track type (i.e., mainline and others), and track number (i.e., multiple and single track). Chang and Chen (2005) compared freeway crash frequencies using a classification and regression tree (CART) model and a NB model based on roadway geometry, traffic volume, and environmental factors. Siddiqui et al. (2012) explored critical variables associated with total and severe crashes within traffic analysis zones (TAZ) using both CART and RF algorithms. These data mining applications identified the potential of the selected algorithms as an alternative to provide reliable crash predictions. An advantage of using data mining algorithms in traffic safety research is that they do not require assumptions on the variables included in the analysis, while the outcome is straight-forward to interpret and visualize.

A disadvantage of the DT and RF algorithms results from the trade-off between predictive power and statistical inference (Mannering, 2018). These data-driven models do not enable researchers to draw statistical inferences. To overcome this challenge, a model-based recursive partitioning (MOB) algorithm has been developed, which can produce a tree-based statistical model to visualize, analyze, and interpret data. MOB has been applied in several research fields, including the social sciences (Kopf et al., 2010) and medicine (Thomas et al., 2018; Seibold et al., 2016; Pirkle et al., 2018), and shows promise as an effective tool to identify covariates and subgroup patterns in data. This paper considers an exploratory application of MOB in traffic safety and compares the outcome to the results from a traditional safety model (NB regression) and a heterogeneity model (random parameters model).

3. Methodology

The objective of this research is to explore the appropriateness and potential of the MOB algorithm to predict crash frequency on roadway segments by comparing it to the fixed effects NB model (traditional safety model) and random parameters NB model (heterogeneity model). Five models are estimated in the present study. These include the standard (fixed effects) NB model, NB model incorporating the MOB algorithm, two adjusted NB models, and a random parameters NB model. The adjusted NB models use the splitting variables identified by the MOB-NB model as indicator variables in the model specification, while the other adjusted NB model considers interaction terms of the splitting variables. The methodology described in this section begins with a brief introduction to the standard NB and random parameters NB models, followed by an overview of the MOB algorithm.

3.1. Negative binomial regression model

The NB model effectively handles non-negative count data, and also accounts for overdispersion commonly found in crash data. The generalized functional form of a NB model, as well as the corresponding mean-variance relationship, is shown in Eqs. (1) and (2), respectively.

$$\ln \lambda_i = X_i \beta + \varepsilon_i, \tag{1}$$

where λ_i is the expected number of crashes for observation i; X is the vector of independent variables; β is the vector of estimated coefficients; $exp(\varepsilon_i)$ is error term following a gamma distribution.

$$Var\,(y_i) = E\,(y_i) + \alpha E\,(y_i)^2, \tag{2}$$

where y_i is the number of crashes for observation i; α is the over-dispersion parameter.

Under such assumptions, the probability distribution and likelihood function of NB models can be written as shown in Eqs. (3) and (4).

$$P\,(y_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(\theta) y_i!}\left(\frac{\theta}{\theta + \lambda_i}\right)^{\theta}\left(\frac{\lambda_i}{\theta + \lambda_i}\right)^{y_i} \tag{3}$$

where θ is the inverse of the over-dispersion parameter; Γ is the gamma function.

$$L\,(\lambda_i) = \prod_{i=1}^{N} \frac{\mathit{\Gamma}\,(\theta + y_i)}{\mathit{\Gamma}\,(\theta) y_i!}\left(\frac{\theta}{\theta + \lambda_i}\right)^{\theta}\left(\frac{\lambda_i}{\theta + \lambda_i}\right)^{y_i} \tag{4}$$

where N is the total number of observations in the sample.

A maximum likelihood estimation procedure is used to derive the coefficients and overdispersion parameter in the NB regression model.

3.2. Random parameters negative binomial model

The random parameters NB model assumes that the independent variable effects vary over observations in the data, while following the same functional form of the standard NB model in Eq. (1). The random parameters can be described using both a fixed and random component, as shown in Eq. (5) (Anastasopoulos and Mannering, 2009).

$$\beta_i = \beta + \varphi_i, \tag{5}$$

where β_i is the coefficient for observation i; and φ_i is the random term following some known distribution (e.g., normal distribution).

The expected crash frequency λ_i is then conditioned on the randomly distributed error term and can be written as $\lambda_i | \varphi_i = \exp(\beta X_i + \varepsilon_i)$. The log-likelihood of the RPNB model is shown in Eq. (6) (Anastasopoulos and Mannering, 2009).

$$LL = \sum_{\forall i} ln \int_{\varphi_i} g\,(\varphi_i) P\,(n_i | \varphi_i) d\varphi_i, \tag{6}$$

where $g\,(\bullet)$ is the probability density function of φ_i and $P\,(n_i | \varphi_i)$ is the Poisson probability of observation i having n_i crashes conditioned on φ_i.

Considering that the random terms in the coefficients do not have a closed form expression, the random parameters NB model is usually estimated using a simulation-based maximum likelihood approach, such as 200 Halton draws, which was found to be appropriate by previous researchers (Bhat, 2003; Milton et al., 2008; Anastasopoulos and Mannering, 2009).

3.3. Model-based recursive partitioning (MOB)

As noted in the literature review section, in standard decision tree algorithms such as CART, the predicted crash frequency is a constant value for each terminal node. Although past research indicates that CART models have utility in traffic safety research, it is difficult to draw statistical inference or factor effects from tree models. In this study, a model-based recurve partitioning (MOB) algorithm is applied to overcome the disadvantage of decision tree algorithms, where the final

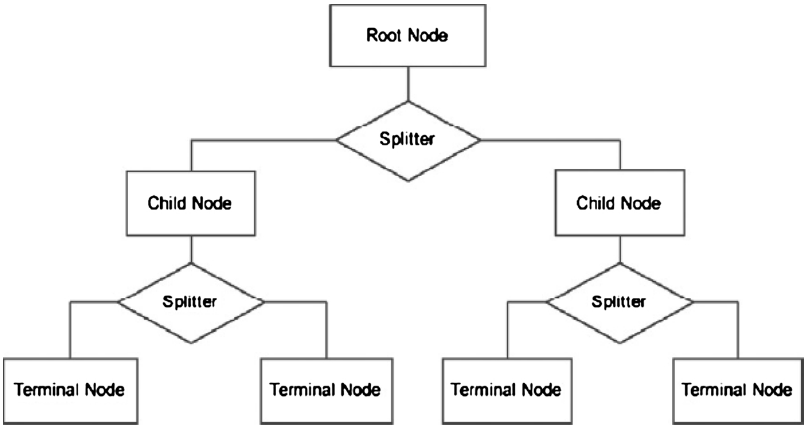

Fig. 1. Generalized structure of tree model (Kashani and Mohaymany, 2011).

model has the tree structure and each terminal node in the tree is associated with a parametric model estimated from observations of that node. The data-driven method intends to offer a high-level of prediction accuracy, while capturing unobserved heterogeneity when dividing the data sample into multiple subgroups (i.e., terminal nodes in the tree). An example illustration of a generalized tree model is shown in Fig. 1. The root node is the full dataset used to estimate the model. The data are then split using decision rules, to form child nodes. Splitting continues until reaching the terminal node of the tree.

The development of MOB models incorporates the concept of recursive partitioning data in the decision tree with statistical models. The process consists of three key steps. The first step is fitting a parametric model to all observations in the data, which occurs at the root node in Fig. 1. The parametric model can be estimated using ordinary least squares, maximum likelihood, or other methods. The generalized optimization equation is shown in Eq. (7) (Seibold et al., 2016).

$$\hat{\theta} = arg\ min \sum_{i=1}^{N} \Psi((y, x)_i, \theta) \tag{7}$$

where θ is parameter vector; $\Psi()$ is the objective function depending on estimation method.

The coefficients are estimated using a partial score function, which is shown in Eq. (8) (Seibold et al., 2016).

$$\sum_{i=1}^{N} \frac{\partial\Psi((y, x)_i, \theta)}{\partial\beta} = \sum_{i=1}^{N} \psi_\beta((y, x)_i, \theta) = 0 \tag{8}$$

where $\psi()$ is the score function; β is the estimated parameter.

In the second step of the MOB process, coefficient stability is tested over selected splitting variables using a generalized M-fluctuation test. The null hypothesis states that partial score functions from Eq. (8) are independent of partitioning variables, suggesting that a global estimation of an independent variable is appropriate. The functional form of the hypothesis is shown in Eq. (9) (Seibold et al., 2016). Details about the hypothesis test can be found in references by Zeileis et al. (Zeileis et al., 2008) and Zeileis and Hornik (Zeileis and Hornik, 2007).

$$H_0^{\beta,j}: \psi_\beta((Y, X), \hat{\theta}) \perp Z_j,\ j = 1,\ ...,J \tag{9}$$

where Z is splitting variable; J is the number of splitting variables.

In total, $J \times P$ null hypotheses are tested at each data split, where P is the number of estimated parameters. To adjust for the risk of incorrectly rejecting one hypothesis when multiple hypotheses are tested, a Bonferroni correction is used. If at least one of the null hypotheses is rejected on the pre-defined significance level, it is likely that parameter instability exists. The algorithm then selects the splitting variable which has the highest correlation with partial score functions to separate data, and the variable is shown as the splitter in Fig. 1. The optimal cut point of the splitting variable is determined by evaluating segmented objective functions, as shown in Eq. (10) (Zeileis et al., 2008). The one with the minimum value is selected.

$$SCORE = \sum_{b=1}^{B} \sum_{i \in I_b} \Psi(Y_i, \theta_b) \tag{10}$$

where I_b is the set of observations that belong to segment b under splitting rule B.

Although an exhaustive search over all possible cut points provides optimal splitting, Zeileis et al. (Zeileis et al., 2008) recommended employing a binary split search throughout the entire model construction process for efficiency purposes.

Once the optimal cut point is determined, previous procedures are repeated recursively on each identified child node, as shown in Fig. 1, using subgroup data only, until no instability is detected. The final node with no coefficient instability is treated as the terminal node. Similar to a decision tree model, each observation can find a path in the final model from the root node down to one of the terminal nodes. The associated statistical model of that terminal node is applied to predict the response of that observation. More details about the algorithm can be found in Zeileis et al. (Zeileis et al., 2008).

4. Data

Roadway inventory and crash data from a previous research project (Donnell et al., 2014) were used in the present study. The roadway data were collected from 21,340 two-lane rural roadway segments from the Pennsylvania Department of Transportation (PennDOT) Roadway Management System (RMS) database, PennDOT's online vehicle photolog system, and Google Earth. The RMS database provided information such as traffic volume, segment length, cross-section width, and posted speed limit, while PennDOT's online video photolog system and Google Earth were used to collect supplemental data for each segment. Example elements include roadside hazard rating, presence of passing zone, presence of low-cost safety improvements (e.g., shoulder and centerline rumble strips), and access density (including driveways and intersections). Crash data were collected over eight years (2005 through 2012) on state-owned, two-lane rural highways in Pennsylvania, and were merged with the roadway inventory data. Additional details concerning the crash frequency data used in the present study can be found in Donnell et al. (Donnell et al., 2014).

In total, the dataset consisted of 169,500 observations, after excluding cases with unknown or missing values and short segment lengths (less than 0.1 mile). The list of dependent and independent variables considered for the models, along with their definitions, are shown in Tables 1 and 2. The summary statistics of all variables are shown in Tables 3 and 4.

The full dataset was randomly split into two parts, a training set and

Table 1
Continuous Variables Descriptions.

Variable	Description
Total	Total number of crashes per year
AADT	Average annual daily traffic (veh/day)
Length_mi	Segment length in miles
Speed	Posted speed limit (mph)
Ls_pave	Left paved shoulder width in feet
Rs_pave	Right paved shoulder width in feet
Curve_den	Horizontal curve density (curves per mile)
Curdeg_seg	Degree of curve per mile
Curlen_seg	Length of curve in feet per mile
Acc_den	Access density (access points and intersections per mile)

a test set. The training set consisted of 118,650 observations (70 percent of full dataset) and was used to construct the models. The test dataset comprised of 50,850 observations (30 percent of full dataset) and was used to examine the predictive power of the models.

5. Results

Five models were estimated, including a standard NB model, NB model based on the MOB algorithm, two adjusted NB models (one directly utilizing splitting variables identified by the MOB-NB model and the other evaluating interactions between splitting variables), and a random parameters NB model following the same functional form as the first adjusted NB model. When searching for the optimal final model, all independent variables listed in Tables 1 and 2 were examined in the MOB-NB model, while only categorical variables and the posted speed limit were tested as splitting variables to separate data into subgroups. The reason to exclude continuous variables from the candidate splitting variable set is that the searching time for the optimal cut point over continuous variables is significantly longer than categorical variables. In addition, the MOB-NB model utilized a global overdispersion parameter estimated from the standard NB model to guarantee stabilized output from the software. The maximum depth of the tree structure was set to three in order to restrict the tree size and ensure adequate observations in each terminal node. An overview of the MOB-NB model structure is shown in Fig. 2.

The upper part of the figure shows the tree structure of the final model. The interpretation proceeds in a manner consistent with other decision tree algorithms. The first splitting rule is associated with the passing zone indicator. If no passing zone exists in the segment, the observation is classified into node 2. Otherwise, the observation is classified into node 5. Conditioned on no passing zone in the segment, the data can be further separated according to the posted speed limit. If the speed limit is less than or equal to 50 miles per hour (mph), the observation goes to node 3, and if the speed limit is greater than 50 mph, the observation goes to node 4. Similarly, conditioned on having a passing zone in the segment, the data can be divided into node 6 and node 7, based on whether the posted speed limit is greater than 45 mph. In addition, the p-value in each inner node indicates the significance level of the associated splitting variable when conducting the coefficient stability test.

The lower part of Fig. 2 shows a partial scatter plot of each independent variable relative to the total crash frequency within each terminal node. In this case, each row represents an independent variable in the final model, including the natural logarithm of AADT, access density, curve density, degree of curve per mile, presence of shoulder rumble strips, and the natural logarithm of segment length, respectively. The natural logarithm of the segment length was specified as an offset variable in this analysis so that the segment length is directly proportional to the expected crash frequency. Each column in Fig. 2 represents one terminal node in the final model, node 3, node 4, node 6 and node 7, respectively, with the number of observations in the node shown in parentheses. The partial scatter plot provides a direct relationship between each predictor and total crash frequency. The MOB algorithm is a helpful tool to identify potential outliers in the dataset. For example, in the third row of the second column in Fig. 2, which represents the distribution of curve density in node 4, an outlier is

Table 3
Statistics Summary of Continuous Variables.

Variable	Mean	Standard Deviation	Minimum	Maximum
Total	0.668	1.145	0	23
AADT	3276	2928	74	28,674
Length_mi	0.477	0.125	0.100	1.476
Speed	47.461	7.615	15	55
Ls_pave	3.003	2.299	0	22
Rs_pave	3.049	2.298	0	19
Curve_den	2.293	2.462	0	42.581
Curdeg_seg	18.910	43.102	0	1263.478
Curlen_seg	1006.163	1231.607	0	29256.370
Acc_den	16.290	14.006	0	179.660

Table 4
Statistics Summary of Categorical Variables.

Variable	Category	Proportion in Sample (in percent)
RHR	1	0.1
	2	0.5
	3	5.1
	4	21.6
	5	53.1
	6	19.4
	7	0.2
Pass_zone	0 (No)	71.5
	1 (Yes)	28.5
Cl_rs	0 (No)	78.9
	1 (Yes)	21.1
Sh_rs	0 (No)	91.8
	1 (Yes)	8.2
Curve_warn	0 (No)	98.6
	1 (Yes)	1.4
Int_warn	0 (No)	99.5
	1 (Yes)	0.5
Agg_dots	0 (No)	99.9
	1 (Yes)	0.1

Table 2
Categorical Variables Descriptions.

Variable	Description	Categories
RHR	Roadside hazard rating	Estimated on the 1 to 7 scale, and 1 represents least hazardous
Pass_zone	Passing zone indicator	1 = yes and 0 = no
Cl_rs	Centerline rumble strip indicator	1 = yes and 0 = no
Sh_rs	Shoulder rumble strip indicator	1 = yes and 0 = no
Curve_warn	Curve warning pavement marking indicator	1 = yes and 0 = no
Int_warn	Intersection warning pavement marking indicator	1 = yes and 0 = no
Agg_dots	Aggressive driving dots indicator	1 = yes and 0 = no

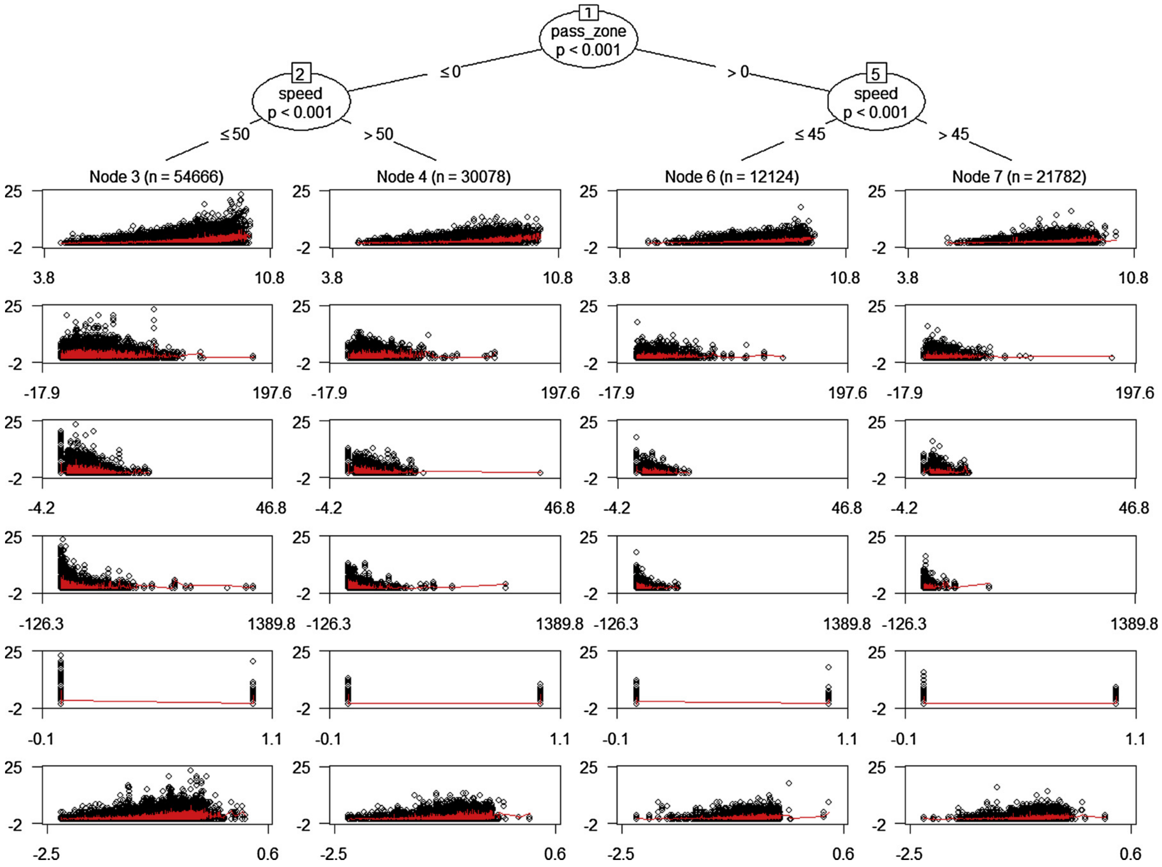

Fig. 2. MOB-NB model structure.

Table 5
Model Estimation Results (T-Statistics in Parentheses).

Variable	MOB-NB				Standard NB	Adjusted NB 1	Adjusted NB 2	RPNB
	Node3	Node 4	Node 6	Node 7				
Intercept	−5.861	−6.032	−5.941	−5.781	−5.994	−5.561	−5.942	−5.707
	(−89.423)	(−63.131)	(−37.619)	(−42.631)	(−130.330)	(−102.620)	(−128.182)	(−118.439)
Lnaadt	0.770	0.762	0.744	0.703	0.759	0.759	0.758	0.760
	(96.252)	(65.836)	(38.128)	(42.081)	(135.040)	(136.000)	(136.063)	(156.226)
Acc_den	0.006	0.009	0.007	0.011	0.009	0.007	0.007	0.007
	(14.186)	(11.247)	(6.167)	(9.946)	(30.290)	(20.480)	(22.009)	(23.147)
Curve_den	0.023	0.037	0.039	0.032	0.044	0.031	0.031	0.026
	(7.191)	(7.821)	(3.573)	(3.632)	(19.090)	(13.300)	(13.011)	(11.895)
Curdeg_seg	0.002	0.002	0.003	0.006	0.002	0.002	0.002	0.002
	(10.505)	(8.703)	(3.768)	(6.495)	(14.810)	(13.440)	(14.032)	(13.474)
Sh_rs	−0.168	−0.169	−0.230	−0.119	−0.217	−0.175	−0.167	−0.176
	(−5.187)	(−6.147)	(−4.586)	(−3.829)	(−13.840)	(−11.080)	(−10.616)	(−12.197)
Speed						−0.006		−0.007
						(-10.410)		(−11.568)
Pass_zone						−0.218	−0.198	−0.217
						(−20.850)	(−14,139)	(−22.924)
Pass_speed1							0.132	
							(12.007)	
Pass_speed3							0.110	
							(6.263)	
Standard Deviation of Random Parameters								
Intercept								0.534
								(138.562)
Lnaadt								0.012
								(24.917)
Curve_den								0.041
								(34.238)
Log-likelihood	−121,307.5				−121,712.8 (df = 7)	−121,396.3 (df = 9)	−121,360.2 (df = 10)	−121,260.1
	(df = 27)							(df = 12)
AIC	242,669.0				243,439.6	242,810.5	242,740.3	242,544.2

identified to the far-right of the subplot.

Table 5 summarizes the estimation results of the five models. Specifically, variable pass_speed1 takes a value of 1 if there is no passing zone and the posted speed limit is less than or equal to 50 mph, and 0 otherwise. Variable pass_speed3 takes a value of 1 if there is a passing zone and the posted speed limit is less than or equal to 45 mph, and 0 otherwise. Both indicators represent the interactions between splitting variables (i.e., the presence of passing zone and posted speed limit), and correspond to the identification of observations in terminal node 3 and node 6 of the MOB-NB model, respectively. Traffic volume and horizontal curve density were specified as random parameters in the random parameters model.

The interpretation of the coefficients in the MOB-NB model is the same as the traditional NB model, except a prior condition is added to each node, as shown in the tree structure in Fig. 2. The results show that the coefficients from different models have the same sign and similar magnitudes, suggesting that the MOB-NB model produces results similar to the results from a parametric regression model such as the standard NB model. The log-likelihood in the NB model is -121,712.8—it improves to a value of -121,307.5 in the MOB-NB model, while the AIC is improved from 243,439.6 to 242,669.0 when comparing the NB to the MOB-NB models. Meanwhile, both adjusted NB models offer improvements relative to the standard NB model, although they do not perform as good as the MOB-NB model based on the log-likelihood and AIC values. The log-likelihood of the adjusted NB models are -121,396.3 and -121,360.2, respectively. The AIC of adjusted NB models are 242,810.5 and 242,740.3, respectively. The significant improvement of the MOB-NB model in log-likelihood and AIC leads to better model performance and better data fitness when comparing the MOB-NB to traditional safety models. This is achieved by capturing unobserved heterogeneity and estimating variable coefficients of each terminal node accordingly, which is identified by different covariate conditions. These results indicate that the MOB algorithm is effective in recognizing subgroup patterns and describing the full dataset. Moreover, by introducing interaction effects between two identified covariates (in this case, the passing zone indicator and posted speed limit), the adjusted NB model 2 provides better data fitness than the adjusted NB model 1, which directly evaluates the association between the posted speed limit and the predicted total crash frequency. This suggests that posted speed limit may have a non-linear influence on crash frequency, and direct estimation is not able to fully capture the statistical association.

When comparing the MOB-NB model to the random parameters (heterogeneity) NB model, the results are similar in the present study. The log-likelihood in the RPNB model is -121,260.1, while the AIC is 242,544.2, both of which are nominally better than the same metrics produced from the MOB-NB model. This is as expected since the RPNB model allows for the highest coefficient variation (i.e., distinct coefficient of random parameter for each observation) in the data, while the MOB-NB model only allows the coefficients to vary among different subgroups. The similarity in model performance suggests that the MOB-NB model is capturing some unobserved heterogeneity via data separation.

The 95 percent confidence intervals and elasticities of each variable in the MOB-NB and standard NB models are shown in Tables 6 and 7, respectively. For each cell in Table 6, the first number represents the lower bound of 95 percent confidence interval of estimated variable coefficient, and the second number represents the upper bound of the confidence interval. Note that in this crash frequency analysis, the original data were partitioned into four terminal nodes, meaning that coefficient instability existed between node 2 and node 5, node 3 and node 4, and node 6 and node 7, respectively. The RPNB model is not included in this comparison because this model estimates different coefficients for each observation in the training dataset, while the confidence intervals and elasticities from the other models are calculated using fixed coefficients and mean values of the independent variables.

Confidence intervals are used to identify which variables have statistically different coefficients. For the coefficients in the MOB-NB model, the confidence intervals of traffic volume, access density and degree of curve per mile between node 2 and node 5 do not overlap with each other. The comparison means that, conditioned on the presence of a passing zone, the effects of the three variables on crash frequency are statistically different. Similarly, the confidence intervals of access density between node 3 and node 4, and between node 6 and node 7, do not overlap with each other. This reveals the different influences of access density on segments with low speed limits and segments with high speed limits, even if holding the presence of a passing zone constant.

When comparing the confidence intervals from two inner nodes (nodes 2 and 5) of the MOB-NB model with those from the standard NB model, it is worth noting that, the estimated effects of access density and curve density in the standard NB model tends to be higher than the MOB-NB model estimation, when there is no passing zone present in the segment (node 2 vs. standard NB). The effect of traffic volume in the standard NB model is higher than the MOB-NB model estimation, and the effect of curve degree per mile is lower than the MOB-NB model estimation, when a passing zone is present in the segment (node 5 vs. standard NB).

A comparison between the terminal nodes and the standard NB model showed different influential patterns with respect to the posted speed limit. Conditioned on no passing zone presence, the effects of access density and curve density in the standard NB model are larger than the MOB-NB model estimations on segments with low posted speed limits (node 3 vs. standard NB). The effects are similar on segments with high posted speed limits, where the confidence intervals from the MOB-NB and standard NB models overlap with each other (node 4 vs. standard NB). Conditioned on the presence of a passing zone, the global coefficients from the standard NB model tend to yield larger access density effects on segments with low posted speed limits (node 6 vs. standard NB). The global estimation also produces higher effects of traffic volume and shoulder rumble strips, and lower effects of curve degree per mile on segments with high posted speed limits than the MOB-NB model (node 7 vs. standard NB). Note that each model (i.e., each column in Table 6) in the MOB-NB structure (inner node or terminal node) is independent. They can be treated as SPFs using specific data subgroups, except that the MOB algorithm automatically identifies and determines sub-datasets. The comparison results suggest that potential bias may exist in global estimation if omitting the differences between subgroups, and the MOB-NB model can efficiently detect and capture these different patterns.

Statistics from Table 7 can also be applied to understand variable effects under different conditions. For continuous variables, the elasticity represents the percent increase in the expected crash frequency based on a one percent increase in that variable. For indicator variables, the elasticity represents the percent increase in expected crash frequency if the variable value is changed from zero to one. Elasticities in Table 5 are calculated using coefficient and mean values in the corresponding data subgroups.

The elasticity provides a direct estimation of the variable sensitivity within each subgroup and reveals relative variable effect differences between subgroups. For example, a one percent increase in curve density leads to a 0.101 percent increase in total crash frequency in the standard NB model, while the increment reduces to 0.065 percent, 0.092 percent, 0.056 percent and 0.037 percent, for terminal node 3, node 4, node 6 and node 7 in the MOB-NB model, respectively. The magnitude of effects differs slightly, while the relative difference is high, ranging from a 8.91 percent difference (node 4 vs. standard NB) to a 63.37 percent difference (node7 vs. standard NB). The presence of shoulder rumble strips is expected to decrease the total crash frequency by 19.51 percent in the standard NB model, while the influence is 15.46 percent, 15.55 percent, 20.55 percent and 11.22 percent, for node 3,

Table 6
Summary of 95 Percent Confidence Intervals for MOB-NB and Standard NB Models.

Variable	MOB-NB						Standard NB
	Node 2	Node5	Node 3	Node 4	Node 6	Node 7	
Lnaadt	0.7520, 0.7778	0.6939, 0.7435	0.7540, 0.7854	0.7390, 0.7844	0.7061, 0.7826	0.6700, 0.7355	0.7476, 0.7697
Acc_den	0.0066, 0.0080	0.0084, 0.0113	0.0048, 0.0064	0.0074, 0.0106	0.0045, 0.0088	0.0090, 0.0134	0.0083, 0.0094
Curve_den	0.0246, 0.0350	0.0236, 0.0504	0.0168, 0.0294	0.0277, 0.0462	0.0174, 0.0597	0.0148, 0.0495	0.0391, 0.0481
Curdeg_seg	0.0014, 0.0019	0.0032, 0.0055	0.0012, 0.0018	0.0017, 0.0026	0.0015, 0.0048	0.0039, 0.0072	0.0015, 0.0019
Sh_rs	−0.2446, −0.1628	−0.2087, −0.1054	−0.2317, −0.1046	−0.2223, −0.1148	−0.3282, −0.1317	−0.1801, −0.0581	−0.2480, −0.1865

node 4, node 6 and node 7 in MOB-NB model, respectively. The relative difference among coefficients ranges from 5.33 percent (node 6 vs. standard NB) to 42.49 percent (node 7 vs. standard NB). This result also reveals some potential parameter bias in the global estimation and indicated that the MOB algorithm is able to identify various subgroups within the sample.

To further explore the potential of the MOB algorithm, the predictive power of the five models were also examined using both the training and test datasets, followed by a simple case study example to illustrate how the MOB-NB model may be applied in practice. Table 8 summarizes the prediction accuracy on both datasets using the mean square error. Note that mean values of the random parameters in the RPNB model are directly applied to derive the prediction accuracy on the training and test datasets.

Table 8 shows similar levels of predictive accuracy across each model in the training and test datasets. In the training dataset, the MOB-NB model has the lowest MSE value, followed by the adjusted NB 1 and NB 2 models, suggesting the highest level of prediction accuracy. The adjusted NB 2 model offers improved prediction accuracy over the NB 1 model as a result of including interaction effects of the splitting variables in the model. The standard NB and RPNB models offer the lowest level of prediction accuracy in the training dataset. The NB model has fewer independent variables than the adjusted NB models, while the mean value of the random parameters were used to estimate the prediction accuracy of the RPNB model (due to software restrictions associated with recording individual observations), which eliminate the value of having a distribution associated with the parameter estimate. Similar results were found when comparing the prediction accuracy of the models in the test dataset.

With regards to practical application, consider an example two-lane rural roadway segment with the following features: 5000 vehicles per day, one mile long with 40 mph posted speed limit, having two access points, one 10-degree horizontal curve, and no passing zone or shoulder rumble strip in the segment. In order to derive the expected crash frequency using the MOB model, the first step is to determine which terminal node the observation falls into. According to Fig. 2, the segment is classified into node 3, since it has no passing zone and the posted speed limit is less than 50 mph. Then the predicted crash frequency is calculated using a safety performance function derived from traditional NB model, with coefficients corresponding to node 3 only, as shown in Table 5. In this case, the prediction using the MOB model is 2.12 crashes per year, while the predictions using the other three NB and RPNB models shown in Table 5 are 1.74, 2.07, 2.04, and 1.74, respectively.

Table 8
Model Predictive Power Summary.

	Standard NB	MOB-NB	Adjusted NB 1	Adjusted NB 2	RPNB
Trainset	1.062	1.049	1.053	1.051	1.067
Testset	1.048	1.036	1.039	1.037	1.048

In summary, Table 8 shows that the overall differences in prediction accuracy among the five models is relatively small, suggesting no preference of one approach over the other if the primary research objective is making predictions. However, if the research objective is more than predicting crash frequency, the MOB algorithm is recommended as a promising alternative in crash frequency studies with the following advantages: (1) the model can recognize covariates and subgroups, capture unobserved heterogeneity, and further identify interactions if more than one splitting rule is uncovered; (2) the model can significantly improve performance when compared to the standard NB models; (3) the model is easy to understand and interpret because the functional form is the same as the standard NB models, except a prior condition is introduced for each terminal node; (4) the model can provide statistical inference and factor effects compared to other data mining algorithms; (5) the model can also provide a slightly better prediction accuracy than the NB and RPNB models on new observations.

6. Conclusions

In this paper, a novel data mining algorithm, model-based recursive partitioning, was employed in a crash frequency case study. The objective of this study was to explore the appropriateness and potential of this methodological alternative in frequency analysis. To accomplish the objective, a standard NB regression model, NB model using the

Table 7
Summary of Elasticities for MOB-NB and Standard NB Models.

Variable	MOB-NB						Standard NB
	Node 2	Node5	Node 3	Node 4	Node 6	Node 7	
Lnaadt	0.765	0.719	0.770	0.762	0.744	0.703	0.759
Acc_den	0.121	0.136	0.119	0.116	0.120	0.129	0.147
Curve_den	0.081	0.047	0.065	0.092	0.056	0.037	0.101
Curdeg_seg	0.048	0.027	0.053	0.037	0.030	0.030	0.038
Sh_rs	−18.45	−14.53	−15.46	−15.55	−20.55	−11.22	−19.51

MOB algorithm, two adjusted NB models which incorporated splitting variables identified by the MOB, and the random parameters NB model were developed using 8 years of data collected from two-lane rural highways in Pennsylvania. The models were compared in terms of data fitness, variable effects, and predictive power.

The results showed that the MOB-NB model provided some unique insights into factor effects on crash frequency under different covariate conditions. The presence of passing zones and the posted speed limit were identified as two covariates that offered different effects for different subgroup classes, with improved the overall fitness of the MOB-NB model relative to the NB models estimated in this study. For example, the effects of traffic volume, access density and degree of curve per mile were different between segments with a passing zone and segments without a passing zone. The MOB-NB model provided similar results to the RPNB model based on two fitness metrics (log-likelihood and AIC), suggesting that the MOB-NB model is able to capture unobserved heterogeneity by dividing data into subgroups. In addition, the MOB-NB model provided the highest prediction accuracy on the training and test sets, although the differences among models were quite small. Given the model results, if the primary objective is predicting crash frequency, the MOB algorithm is suitable as an alternative. If considering the statistical association between the dependent variable (crash frequency) and independent variables, the MOB algorithm is recommended as a promising alternative to count regression models as a means to identify covariate subgroup patterns.

Future research should further consider the MOB algorithm in the context of crash frequency modeling by applying the method to other datasets. In addition, it would be valuable to consider applications of the MOB-NB model in the context of temporal or spatial instabilities in order to assess how it handles these issues relative to heterogeneity models. Future modeling efforts using the MOB algorithm should also consider continuous variables as covariates, the evaluation of variables as both predictors and covariates, and the application of a local overdispersion parameter for each identified subgroup.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the Pennsylvania Department of Transportation for providing the data used in this analysis." and the following Disclaimer: "The contents of this paper reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Federal Highway Administration or the Commonwealth of Pennsylvania at the time of publication. This paper does not constitute a standard, specification or regulation.

References

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accid. Anal. Prev. 37, 417–425.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41, 153–159.

Blincoe, L.J., Miller, T.R., Zaloshnja, E., Lawrence, B.A., 2015. The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised). Report No. DOT HS 812 013. . https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013.

Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences. Transp. Res. Part B Methodol. 37, 837–855.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

Chang, L., Chen, W., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Safety Res. 36, 365–375.

Donnell, E., Gayah, V., Jovanis, P., 2014. Safety Performance Function. Report No. FHWA-PA-2014-007-PSU WO 1.

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accid. Anal. Prev. 41, 1118–1123.

Hadi, M.A., Aruldhas, J., Chow, L., Wattleworth, J.A., 1995. Estimating safety effects of cross-section design for various highway types using negative binomial regression. Transp. Res. Rec. J. Transp. Res. Board 1500, 169–177.

Johansson, P., 1996. Speed llimitation and motorway casualities: a time series count data regression approach. Accid. Anal. Prev. 28, 73–87.

Jovanis, P.P., Chang, H., 1986. Modeling the relationship of accidents to miles traveled. Transp. Res. Rec. J. Transp. Res. Board 1068, 42–51.

Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accid. Anal. Prev. 34, 357–365.

Kashani, A.T., Mohaymany, A.S., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Saf. Sci. 49, 1314–1320.

Kopf, J., Augustin, T., Strobl, C., 2010. The Potential of Model-based Recursive Partitioning in the Social Sciences – Revisiting Ockham's Razor. Technical Report No. 88. Department of Statistics, Ludwig-Maximilians University, Munich. https://doi.org/10.5282/ubm/epub.11933.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accid. Anal. Prev. 34, 149–161.

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, 291–305.

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

Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13.

Miaou, S., 1994. The relationship between truck accidents and geometric design of road sections: poisson versus negative binomial regressions. Accid. Anal. Prev. 26, 471–482.

Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25, 395–413.

Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40, 260–266.

National Highway Traffic Safety Administration, 2017. National Center for Statistics and Analysis Motor Vehicle Traffic Crash Data Resource Page. Quick Fact 2016. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812451.

Park, Y., Saccomanno, F.F., 2005. Collision frequency analysis using tree-based stratification. Transp. Res. Rec. J. Transp. Res. Board 1908, 121–129.

Pirkle, C.M., Wu, Y.Y., Zunzunegui, M., Gómez, J.F., 2018. Model-based recursive partitioning to identify risk clusters for metabolic syndrome and its components: findings from the international mobility in aging study. BMJ Open 8. https://doi.org/10.1136/bmjopen-2017-018680.

Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-accident frequencies. J. Transp. Eng. 122, 105–113.

Seibold, H., Zeileis, A., Hothorn, T., 2016. Model-based recursive partitioning for subgroup analyses. Int. J. Biostat. 12, 45–63.

Shankar, V., Albin, R., Milton, J., Mannering, F., 1998. Evaluating median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. Transp. Res. Rec. J. Transp. Res. Board 1635, 44–48.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accid. Anal. Prev. 27, 371–389.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accid. Anal. Prev. 29, 829–837.

Siddiqui, C., Abdel-Aty, M., Huang, H., 2012. Aggregate nonparametric safety analysis of traffic zones. Accid. Anal. Prev. 45, 317–325.

Thomas, M., Bornkamp, B., Seibold, H., 2018. Subgroup identification in dose-finding trials via model-based recursive partitioning. Stat. Med. 37, 1608–1624.

Zeileis, A., Hothorn, T., Hornik, K., 2008. Model-based recursive partitioning. J. Comput. Graph. Stat. 17, 492–514.

Zeileis, A., Hornik, K., 2007. Generalized M-Fluctuation tests for parameter instability. Stat. Neerl. 61, 488–508.