Quantifying the safety performance of rural roadways using two models

Mehdi Hossein Pour
PhD Candidate, School of Civil Engineering, Universiti Sains Malaysia,
Penang, Malaysia, h_mehdi61@yahoo.com

Joewono Prasetijo
Senior Lecturer, School of Civil Engineering, Universiti Sains Malaysia,
Penang, Malaysia, cejoewono@eng.usm.my

Seyed Mohammad Reza Ghadiri
PhD Candidate , School of Civil Engineering, Universiti Sains Malaysia,
Penang, Malaysia, smrgh1@gmail.com

Submitted to the 3rd International Conference on Road Safety and Simulation,
September 14-16, 2011, Indianapolis, USA

ABSTRACT

There are a lot of studies to establish relationships between crash occurrences and road and roadside conditions on rural roads. Many researchers have developed statistical models such as Poisson or negative binomial regression models to analyze road accidents and their causative factors. Nevertheless, as far as the authors are aware, no extensive research work has been done to address the application of tree-based regression model, one of the most widely used non-parametric statistical techniques, for accident modeling and analysis. Thus, this study aims to develop a tree-based regression model and a negative binomial regression model separately to relate road accidents to road and roadside features. For this purpose, accident data, road geometry and road environmental characteristics were collected over a two-year period (2006-2007) along the Qazvin-Loshan intercity roadway in Iran. The candidate set of explanatory variables were: lane width (LW), shoulder width (SW), land use (LU), longitudinal grade (LG), mean horizontal curvature (CUR), minor access points (AP), horizontal curve density (HCD); all of them were obtained per 1-km length. Accident data was also obtained from police accident database for two years during the study period. The comparison of the prediction performance between the tree-based and negative binomial regression models shows that the negative binomial regression model has a better performance than the tree model.

Keywords: Accident prediction model, Negative binomial regression, Hierarchical Tree-based regression, Road geometry

INTRODUCTION

Road traffic accidents have been one of the top causes of death and injury around the world. Each year, more than 1.17 million people die in road accidents around the world and over 10 million are disabled or injured (WHO, 1999). According to the WHO, if no urgent action is taken during the next 10 years, especially in developing countries, it would result in more than 6 million to die and 60 million to be injured. In addition, traffic accidents often put serious costs on communities, such as welfare, damage to property, medical costs, and so on. The World Health Organization estimated that road accidents cost approximately 1 to 3 percent of a country's annual Gross National Product. Therefore, these issues mentioned above have motivated road safety authorities to attempt considerably to reduce traffic accidents (e.g., geometry improvement, traffic control, and enforcement). Transportation agencies are interested in identifying road hazardous locations and those factors (road and roadside characteristics, traffic, etc.) influencing accident occurrence to implement appropriate remedial improvements. It promotes the safety performance of roadway network and provides a safer condition for road users. Clearly, the success of road safety improvements strongly relies on the availability and reliability of methods that estimate the safety performance of a given roadway. Therefore, the need for efficient methods for identifying the contributing factors of accident occurrences is increasing. Most research has shown that the relationship between road accidents (as random and rare events) and road characteristics are often complicated. These findings have led most safety researchers to apply predictive models in order to quantify the safety effects of road elements where the accident frequency is considered as the dependent variable (Brijs et al., 2007; Montella et al., 2008). A number of studies have been carried out to quantify the safety effects of roadway geometric characteristics and traffic volume on accident occurrence. These range from conventional regression models to count-data models as well as non-parametric modeling approaches. Statistical models (e.g., linear regression models, count-data models, generalized linear models) have been the widely-used parametric techniques in traffic safety analysis for many years. However, these models have some restrictive assumptions and require their functional form to be specified in advance. If their assumptions are not met, the model could lead to incorrect estimation (Karlaftis and Golias, 2002; Karlaftis and Tarko, 1998). On the other hand, the parametric models can be easily affected by the problems such as multcollinearity among independent variables, existence of outliers and missing data. These issues may result in underestimating significancy of independent variables affecting accident likelihood. Non-parametric techniques are powerful tools for dealing with the above-mentioned problems and can be used as an alternative to parametric models whereas these techniques don’t require any functional form of the model to be specified in prior or any limitative assumption as well. However, the applications of non-parametric techniques to analyze traffic safety problems have been relatively few. Therefore, it is a justification for this study to employ this methodology for assessing the safety performance of a given roadway. Following a lot of research concerning identifying factors affecting roadway accidents, this paper attempts to establish the relationship between road characteristics and accident frequency by separately developing two techniques namely negative binomial regression model as a parametric model and hierarchical tree-based regression model (HTBR) as a non-parametric model. This study aims to reach two objectives: the first objective of this study, as mentioned above, is to separately develop a hierarchical tree-based regression model HTBR and a negative binomial regression model to assess the effects of various road characteristics on accident frequency. The second

objective is to compare the results of HTBR model with the analysis results of the constructed negative binomial regression model. The organization of this paper is as follows. The next section provides some literature background. Following this, the methodology framework and data that were used, details of the model estimation, and the comparison results are presented and discussed. The final section of the paper summarizes the findings and suggests some concluding remarks and recommendations.

LITERATURE REVIEW

The earlier models were regression-based models and initially developed by using ordinary or normal linear regression. These models assume a normal error structure for the response variable, a constant variance for the residuals, and the linear relationship between the response and explanatory variables (Ceder and Livneh, 1982; Mohamedshah et al., 1993). Many studies indicated whereas road accidents on a highway section are discrete, nonnegative, and rare events, multiple linear regressions are not suitable for such cases. To overcome these limitations, several researchers suggested Poisson regression models that is normally as first choice for modeling count data (e.g., Joshua and Garber, 1990; Poch and Mannering, 1996; McCarthy, 1999). For example, Blower et al. (1993) used a Poisson log-linear model to explain variations in accident rates. This Poisson regression model is especially suitable for handling data with large numbers of zero counts. Ivan and O'Mara (1997) applied Poisson regression for the prediction of traffic accidents using the Connecticut Department of Transportation's accident data. Results of the model suggested that the posted speed limit, AADT of the highway are critical accident prediction variables leading to the conclusion that the Poisson regression model is preferred than the linear regression model. However, using the Poisson regression model makes necessary that the mean and variance of the accident frequency (the response variable) be equal. On the other hand, in most accident data, the variance of the accident frequency exceeds the mean and caused the data to be overdispersed. Thus, in order to solve this problem, several authors such (e.g., Shankar et al., 1995; Maher and Summersgill, 1996; Abdel-Aty and Radwan, 2000) have used negative binomial regression models. Martin (2002), for instance, described the relationship between crash rate and traffic volume per hour (VH) and the influence of traffic on crash severity. A Negative Binomial distribution was used. Zhang and Ivan (2005) used Negative binomial generalized linear models to evaluate the effects of roadway geometric features on the incidence of head-on crashes on two-lane rural roads in Connecticut. The results suggested that to reduce the incidence of head-on crashes on two-lane roads, it is more effective to reduce the number and degree of horizontal and vertical curves than to widen the pavement. Ramírez et al. (2009) utilized negative binomial models to analyze the influence of traffic conditions, i.e. volume and composition on accidents on different types of interurban roads in Spain. More recently, zero-inflated Poisson and zero-inflated negative binomial models were also applied to solve the overdispersion problem which caused by the extra zero in traffic accident data. The findings have shown that zero-altered models can be appropriate choices for highway sections with extra zero in accident data (Shankar et al., 1995; Lee et al., 2002).

In addition to the parametric modeling techniques mentioned above, non-parametric modeling techniques such as artificial neural network ANN, fuzzy logic, and data mining have been widely used for road safety analysis (e.g., Sayed et al., 1995; Abdelwahab and Abdel-Aty, 2001; Chiou, 2006; Akgüngör and Doğan, 2009). Recently, among non-parametric modeling techniques, classification and regression tree CART has been of interest for transportation studies (Washington, 2000; Rakha et al., 2004; Juni et al., 2008). In the field of safety analysis, some

research applied tree-based models to analyze accident occurrence. For example, Kuhnert et al. (2000) used logistic regression, CART, and multivariate adaptive regression splines (MARS) to analyze motor-vehicle injury data. By comparing the analysis results, they demonstrated that CART and MARS are informative models for motor-vehicle accident analysis. They also suggested that CART and MARS can be used as a precursor to a more detailed logistic regression. Karlaftis and Golias (2002) applied hierarchical tree-based regression (HTBR) to analyze the effects of road geometric and traffic characteristics on accident rates for rural two-lane and multilane roads in Indiana from 1991 to 1995. Their study indicated that that HTBR as a nonparametric model has advantages over multiple linear and negative binomial regression models (parametric models) in analyzing highway accident rates. Park and Saccamonno (2005) identified the relationship between countermeasures and collision occurrence by using a sequential analytic strategy that combines the tree-based data stratification method with the generalized linear regression technique. They used tree-based regression model to isolate the mixed effects of the control factors like highway class, track type, and track number from the effects of countermeasures on collision occurrence at highway–railway grade crossings in Canada. Das et al. (2009) identified traffic, highway design, and driver information related with fatal/severe crashes on urban arterials for different crash types. They used an information discovery approach named Random Forests, which are ensembles of individual trees grown by CART (Classification and Regression Tree) algorithm. The results showed that the methodology is quite insightful in identifying the variables of interest like alcohol/drug use and higher posted speed limits contribute to severe crashes.

METHODOLOGY

Negative Binomial Regression

Since the last decades, statistical modeling techniques have been widely used in road safety modeling. Among these techniques, count-data modeling such as Poisson and negative binomial regression models have been commonly applied whereas accident frequencies on a specific highway section or intersection are discrete and non-negative integer. In applying Poisson regression model, the probability of having n_i accidents on highway section i is given by:

$$P(n_i) = \frac{\lambda_i^{n_i} \exp(-\lambda_i)}{n_i!} \quad (1)$$

where:

$P(n_i)$: the probability of n accidents occurring on highway section i over a period of time

λ_i : the expected accident frequency (i.e., $E(n_i)$) for highway section i.

When applying the Poisson regression model, the expected accident frequency is assumed to be a function of explanatory variables such that

$$\lambda_i = \exp(\beta \boldsymbol{X}_i) \quad (2)$$

where:

$\boldsymbol{X}_i$: vector of explanatory variables that include the geometric, traffic, and environment characteristics of highway section *i* that determine accident frequency

β: vector of estimable coefficients.

The coefficient vector β then can be estimated by the maximum likelihood method. With this form of λ_i, the coefficient vector β can be estimated by standard maximum likelihood methods with the likelihood function, L(β), being

$$L(\beta) = \prod_i \frac{\exp\left[-\exp\left(\beta \boldsymbol{X}_i\right)\right]\left[\exp\left(\beta \boldsymbol{X}_i\right)\right]^{n_i}}{n_i!} \quad (3)$$

The critical characteristic of Poisson probability distribution is that the mean and variance of a Poisson probability distribution are equal. However, past research has indicated that accident frequency data are probably to be overdispersed (i.e. having a variance that exceeds the mean, thus violating the underlying assumption made in the Poisson model). In such cases, the overdispersion problem may result in biased and inefficient coefficient estimates in Poisson regression models. To overcome this problem, negative binomial regression model has been commonly suggested by past research as an appropriate alternative. To do this, an error term is added to the expected accident frequency (λ_i) such that Eq. (4) becomes

$$\lambda_i = \exp\left(\beta \boldsymbol{X}_i + \varepsilon_i\right) \quad (4)$$

where:

exp(ε_i): a gamma-distributed error term with mean one and variance α. This gives a conditional probability:

$$P(n_i|\varepsilon) = \frac{exp[-\lambda_i \exp\left(\varepsilon_i\right)]\left[\lambda_i \exp\left(\varepsilon_i\right)\right]^{n_i}}{n_i!} \quad (5)$$

Integrating ε out of this expression produces the unconditional distribution of n_i. The formulation of this distribution (the negative binomial) is

$$P(n_i) = \frac{\Gamma(\theta+n_i)}{[\Gamma(\theta).n_i!)]} \cdot u_i^{\theta}(1-u_i)^{n_i} \quad (6)$$

Where $u_i = {}^{\theta}/_{(\theta+\lambda_i)}$ and $\theta = {}^{1}/_{\alpha}$, and $\Gamma(.)$ is a value of gamma distribution. The corresponding likelihood function is

$$L(\lambda_i) = \prod_i^N \frac{\Gamma(\theta+n_i)}{\Gamma(\theta).n_i!}\left[\frac{\theta}{\theta+\lambda_i}\right]^{\theta}\left[\frac{\lambda_i}{\theta+\lambda_i}\right]^{n_i} \quad (7)$$

The term of "*N*" is the total number of highway sections. This function is maximized to obtain coefficient estimates β and α. Compared with Poisson model, this model has an additional parameter α, such that it allows the mean to differ from the variance such that,

$$\text{var}[n_i] = E[n_i]\left[1 + \alpha E[n_i]\right] \quad (8)$$

The term of "α" is the variance of the gamma-distributed error term and used as a measure of dispersion. The choice between this negative binomial model and the Poisson model can largely be determined by the statistical significance of the estimated coefficient α. If α is not significantly different from zero, the negative binomial model simply reduces to a Poisson model with $\text{var}[n_i] = E[n_i]$. If α is significantly different from zero, the negative binomial model is the correct choice.

Hierarchical Tree-Based Regression Model (HTBR)

Hierarchical tree based regression model HTBR is a non-parametric technique that was first applied in the 1960s in the medical and the social sciences (Morgan and Sonquist, 1963). Later Breiman et al. (1984) carried out a comprehensive and extensive review of the methods as Classification and Regression Tree CART. Since the last decade, there has been increasing interest in applying tree based regression model in various fields. This method chooses the variables from a large number of those that are most important in determining the response variable (y) to be explained. This is done by building a tree structure, which partitions the data into mutually exclusive nodes as homogeneous as possible concerning their response variable. HTBR is essentially binary because parent nodes in the tree are always split into exactly two child nodes and is recursive whereas the process is repeated by treating each child node as a parent for new split. In HTBR, parent nodes are split according to impurity measure, and the splitting value is chosen such that the measure in each of the two child nodes is minimized.
Unlike conventional parametric models, HTBR does not require any assumptions or knowledge of the population's functional form in advance. It is also robust against multicollinearity between the predictive variables. The model is also capable of handling missing observations and identifying interactions, nonlinearities, and nonadditive behavior among variables prior to building the model. HTBR partitions the data into homogeneous nodes so that similarity within each terminal node is relatively high; it also takes the mean value of each node as its predicted value. Tree-based regression model is constructed by recursively partitioning data into relatively homogeneous terminal nodes with minimum impurities within the nodes. For this purpose, the values of all independent variables in the model, either discrete or continuous, are selected to maximize reduction in impurity measure in the terminal nodes. The method searches all the variables as well as their optimal split to reach the most reduction in impurity. It is worthy to note that if a response variable is categorical, a classification tree is constructed and the Gini Index is used as the impurity measure. If the response variable is continuous, a regression tree is developed and then the SSE is used as the impurity measure. If the response variable is count, a Poisson regression tree is developed and the log-likelihood ratio is used as the impurity measure.
In this study, since road accidents are count data in nature and are assumed to follow the Poisson distribution, we develop a Poisson regression tree in order to estimate safety effects of road elements on accident frequency. For Poisson regression tree, the log-likelihood ratio is used as impurity measures. RPART package in the R software is used for constructing a Poisson regression tree. The procedure is as follows (Park and Saccomanno, 2005):

1. PART starts with splitting each explanatory variable at all of its threshold values at the root node. For each split point, it split the parent node into two left and right child nodes for the given

explanatory variable, and then by applying a pre-defined impurity measure, determines the impurity measure for the parent node and both child nodes. For Poisson regression tree, the impurity measure (the likelihood ratio) is measured by within-node deviance, which is defined as:

$$D(t) = 2 \times \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i) \right] \quad (9)$$

where:
y_i: the observed accident frequency for section *i*
μ_i: the expected mean of accident frequency for section *i*
n : the total number of roadway sections.

2. By concerning reduction of impurity measure in child nodes, RPART then chooses the variable and its split point with the highest reduction in impurity measure and then partition the data set of the considered parent node into left and right subnodes. The decrease in the impurity (deviance) of parent node c and its children t_L and t_R can be estimated by using the following expression:

$$\Delta D(s,t) = D(t_C) - D(t_L) - D(t_R) \quad (10)$$

Where:
c: the current parent node
t_L and t_R: observations of the parent node c
$D(t_C)$, $D(t_L)$, and $D(t_R)$: the deviance at parent and children nodes, respectively

The property $D(t_C) \geq D(t_L) + D(t_R)$ indicates that the current deviance at the parent node is greater than or equal to the deviance of the child nodes (left and right subnodes) created by the current split. The best splitter is the one that maximizes $\Delta D(s,t)$.

3. RPART repeats recursively Steps 1 and 2 for each node as a new parent node, until the tree results in the largest maximum size such that no significant decrease become no longer possible at a certain point due to the lack of data for further splitting.

4. Prune the tree back to select a tree of right size from the pruned trees, by cutting off important nodes. The process starts with the maximal tree and prunes the tree in order to produce a sequence of sub-trees of the maximal tree. For this purpose, the *k*-fold cross-validation approach is applied to determine the optimal size tree structure. This approach depends on a complexity parameter which can be estimated through impurity measure of data and the size of the tree. In this approach, the data set is randomly divided into *k* (usually 10) subsets. One of the subsets is used as validation data set while the other *k*-1 subsets overlay used as learning data set. The method repeats tree growing and pruning procedure at *k* times, with a different subset as test set at each time. For each size of the tree, impurity measure (deviance) is calculated and averaged over all subsets. According to Breiman et al. (1984), the optimal sized tree is selected where its cost complexity measure is within one standard error of the cost complexity for the tree with the minimum cost complexity (cross-validation estimate error). This rule is known as the 1 S.E. rule, and usually allows selection of a smaller tree whose accuracy is still comparable to the maximal one (Questier et al., 2005).

STUDY AREA AND DATA COLLECTION

To achieve the objectives of this paper, a roadway with a wide variety of road characteristics must be selected. After reviewing several roadways, the intercity roadway located between Qazvin and Loshan was selected; this roadway connects Qazvin and Guilan Provinces in northern Iran (Fig. 1). This roadway passes through areas with industrial, manufacturing, and residential land uses, where the number of minor accesses is rather high. The terrain through the roadway varies from nearly flat at the beginning of the roadway to rolling in the middle of the roadway and then mountainous at the end of the roadway. All the factors mentioned above, along with other factors, cause this roadway to experience high accident frequencies. The study area is also long enough to construct an adequate number of sections to develop the model. We chose a 70-km stretch of roadway and divided it into 1-km fixed length sections. Accident data and road characteristics were collected over a two-year period between 2006−2007. Roadway information along the sections include mean horizontal curvature (MHC), shoulder width (SW), lane width (LW), land use (LU), access points (AP), longitudinal grade (LG), and horizontal curve density (HCD); all of these factors were obtained into 1-km lengths. Accident data were also obtained from the police accident database during the study period. Fortunately, the road had very limited changes during that period. Table 1 shows the candidate set of road and roadside features.

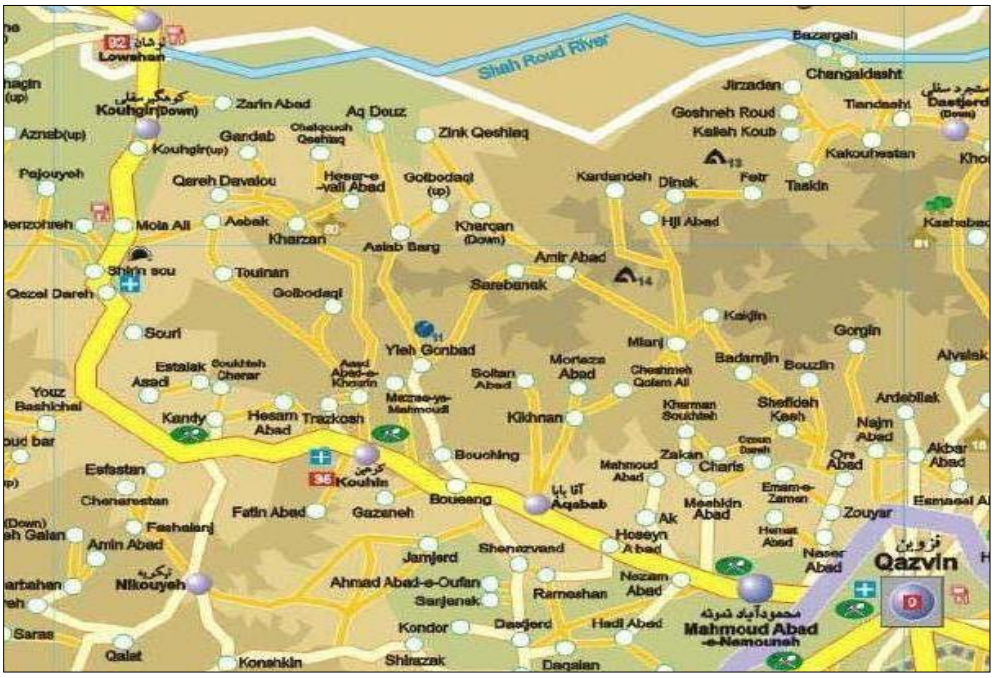

Figure 1 The map of Qazvin-Loshan intercity roadway

Table 1 The candidate set of road and roadside features

Variable	Symbol	Type	Description
Lane Widths	LW	Continuous	Widths of the two sides of the roadway (m)
Shoulder Widths	SW	Continuous	Sum of the left and right shoulders of the roadway (m)
Land Use	LU	Qualitative	Location of the roadway (the level of roadside development) (rural=1, semi-urban=2 , urban=3)[1]
Access Points	AP	Continuous	The number of driveways
Horizontal Curvature	MHC	Continuous	Weighted mean of horizontal curvature
Longitudinal Grade	LG	Continuous	Weighted mean of longitudinal grade (%)
Horizontal Curve Density	HCD	Continuous	The number of horizontal curves

RESULTS AND DISCUSSION

NB Model Result and Interpretation

The stepwise backward procedure was implemented to build model as well as select significant explanatory variables. The decision on whether to remove a non-significant variable from the model was based on pre-specified criterion. Since NB model is the selected parametric model, frequently used statistical test methods such as the likelihood ratio test, *F*-test, and *t*-test might not be appropriate to apply. As an alternative to these measures, the Akaike information criterion (AIC) was used for selecting the best of the models (Zhang and Ivan, 2005). AIC is defined as follows:

$$AIC = -2 * Loglik + 2 * P \qquad (11)$$

where:
Log-lik: the logarithm of maximum likelihood estimation for each model
P : number of the model parameters.

The smaller the AIC value, the better the model performs. Starting with maximal set of explanatory variables, the stepwise procedure selects the best model based on minimizing the AIC value. Table 2 summarizes the estimation results of the negative binomial regression model and the equation of the constructed NB model is

$$\lambda = \exp\begin{bmatrix} 2.604 - 0.198\,SW + 0.122\,AP + \\ 0.019\,LG + 0.059\,HCD \end{bmatrix} \qquad (12)$$

[1] This variable is qualitative and takes the value of 3 if the roadside has continuous development for at least 800 m on the two sides of the roadway, the value of 2 if the length is equal or greater than 300, and 1 if less than 300 m.

The term λ is the predicted accident frequency per km, SW the shoulder width, AP the access points, LG the longitudinal grade and HCD is the number of horizontal curve. As an example, for a section with AP=3, LG= 5.5, HCD= 3, and SW= 2, the estimated accident count is nearly λ= 17 accidents per two years

Table 2 Negative binomial estimation results

variable	Coefficient	*t*-statistics
Constant	2.604	12.632
Shoulder width (SW)	-0.198	-2.901
Access point (AP)	0.122	3.391
Longitudinal grade (LG)	0.019	1.792
Number of horizontal curve (HCD)	0.059	1.506
Over dispersion parameter (α)	0.102	3.207
Number of sections	70	
Restricted Log Likelihood (constant only)	-225.517	
Log Likelihood at converge	-212.559	

According to the table, out of seven candidate variables, four ones were found significant in determining accident occurrence: SW, AP, LG, and HCD. The results show that the sign for SW is negative, implying that an increase in the amount of SW will decrease the accident likelihood. The number of the horizontal curve per km, HCD, has a positive impact on the likelihood of accidents. It may be due to poor sight distance and loss of vehicle control for drivers on curves that altogether increase the potential for accident occurrence. A similar result was also found for the longitudinal grade LG. An increase in LG has a positive effect on the likelihood of accidents. Furthermore, the number of access points AP may be expected to affect positively accident likelihood. The fact that sections with the larger number of access points experience higher accident frequency as illustrated by the model may be due to the fact that turning into driveways would experience conflicts with approaching vehicles and potentially result in collisions. On the other hand, the increase in SW reduces the frequency of accidents because a large amount of shoulder width may give opportunity to a driver in the opposing lane to avoid the errant vehicle. It is also worthy the significance of the over-dispersion parameter (α) indicates that the Negative Binomial model is preferred to Poisson regression model.

HTBR Result and Interpretation

As mentioned before, the RPART package in the statistical software R was used to develop tree-based regression model for this study. Figure 2 shows the tree diagram produced by RPART. The tree can be used to determine the expected accident count for a specific section. For instance, to predict the expected number of accidents for a section with SW of 2, AP of 2, LG of 5.5, and CUR of 6.5, we begin from the root node (top of the tree), then branch left (SW > 1.25), left again (AP < 2.5), go to the branch LG < 8.95, right split (SW<2.5), and finally go to the branch CUR>= 2.39 to reach an average of 7.77 accidents for that section. The tree diagram can also help understand relative importance of variables where variables found to be significant were included in the tree, and insignificant variables were not kept in the tree. The interpretation of results is rather straightforward. The first optimal split in the root node is based on the SW, sending the sections with less than or equal to 1.25 m to the right forming a terminal node and

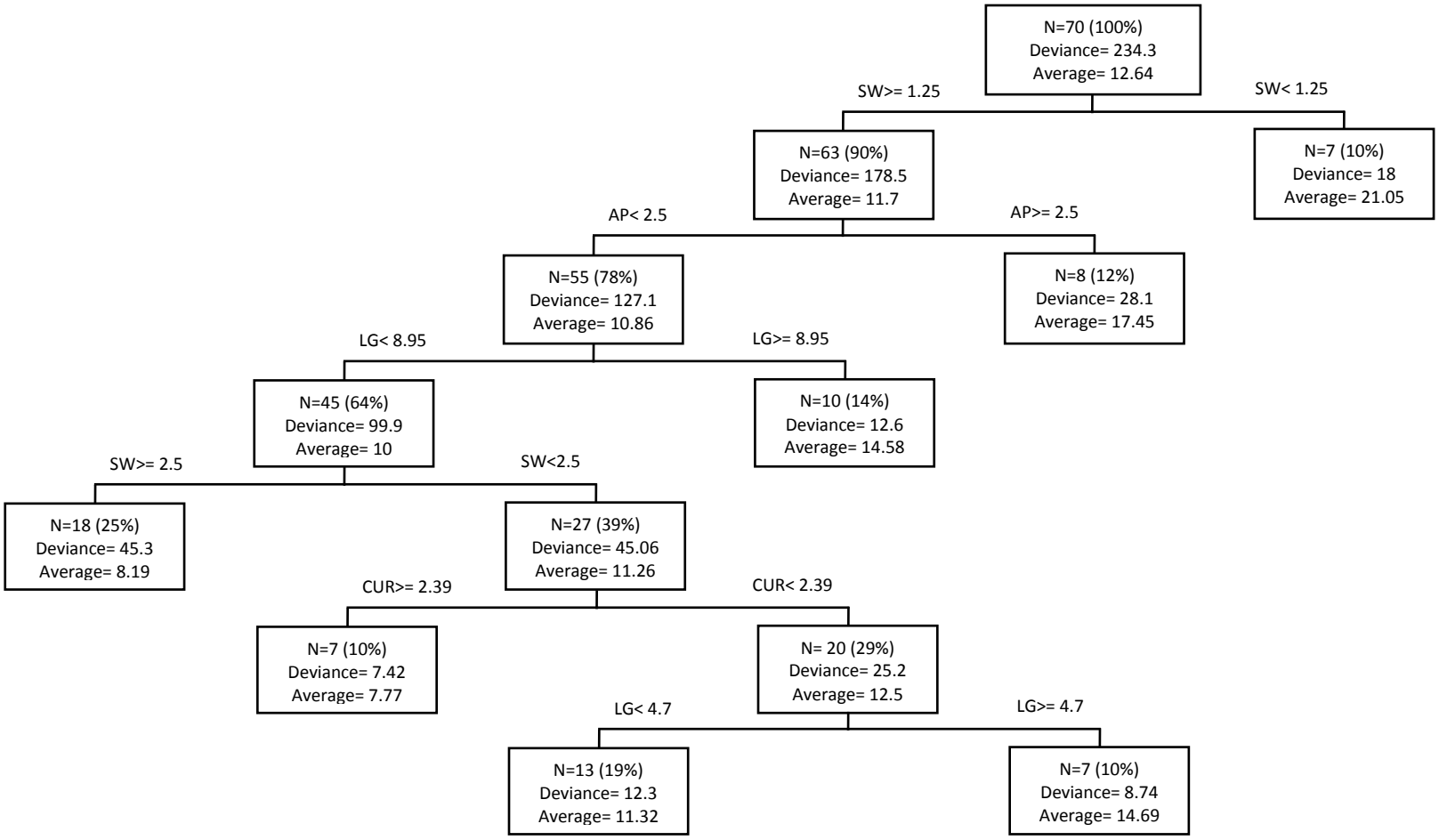

Figure 2 Tree model constructed by RPART for the expected number of accident over a 2-year period

others to the left. This implies that the single best variable to minimize the deviance most on the roadway sections is SW. For the sections with SW less than or equal to 1.25 m, the tree predicts an average of 21.05 accidents in a 2-year period based on 7 observations. Conditioned on the SW greater than 1.25 m, the second best variable to predict accident frequency is the number of access points AP. For the sections with AP greater than or equal to 2.5, the sections go to the right creating a terminal node with a mean of 17.45 accidents for 8 observations (sections). For the sections with AP less than 2.5, the RPART further split the road sections with LG greater than or equal to 8.95% to the right forming terminal node with an amount of 14.58 mean accident frequency based on 10 observations. For LG less than 8.95%, the remaining splits are made on SW, CUR, and LG by continuing down the splits of the tree in similar trend until a terminal node is reached. Since the variables RW, HCD, and LU did not appear in the tree diagram, thus considered as non-significant variables. On the other hand, it is worthy to note that variables SW and LG appeared twice in the tree diagram. It can be explained by the fact that these variables, in comparison to other ones, are relatively important variables for decreasing impurity measure within branches and thus come more than once, even if it never appears as a primary node splitter.

Comparison of HTBR and Negative Binomial Regression Models

In order to compare the performance of the tree and NB models in predicting the accident frequency, one needs to use pre-defined goodness-of-fit measures. For this purpose, several measures can be proposed. In ordinary least square models like linear regression models, the coefficient of determination R^2 based on Ordinary Least Squares estimation is often used. On the other hand, in the case of Poisson and NB regression models, since these models are based on maximum-likelihood estimation, the measure R^2 cannot be used. Instead, ML-based goodness-of-fit measures are used as alternative to R^2. This study employed two commonly-used ML-based measures as follows:

a) Akaike Information Criterion (AIC)

The AIC value is calculated as Equation (11). The smaller the value of AIC is, the better the model is. The first term of the AIC equation is the logarithm of likelihood function of the model that measures the badness of fit, when the maximum likelihood method is used for parameter estimation. The second term measures the complexity of the model by penalizing the model for using more parameters. The model with the best fit and the least complexity is selected as the best model (Montella et al., 2008).

b) Bayesian Information Criterion (BIC)

Another goodness-of-fit measure used was the Bayesian information criterion (BIC). This measure is estimated as follows:

$$BIC = -2 * Loglik + P * \ln(n) \quad (13)$$

Where n is the overall number of observations (n=70). The interpretation of BIC results is almost same as the AIC mentioned above. The smaller the value of BIC is, the better the model is. It is interesting to note that although these measures are used for model selection among a class of parametric models not for non-parametric models, but since our proposed tree model is based on maximum likelihood estimation, these measures can be used for the comparison purpose.

The prediction results are summarized in Tables 3.

Table 3 Prediction results of the negative binomial and tree models

Model	Deviance at zero	Deviance at convergence	Log-likelihood at zero	Log-likelihood at convergence	AIC	BIC
Tree model	234	132	-266	-215	444	460
NB model	103	71	-225	-212	437	450

Based on the comparison results in Table 3, the AIC and BIC values of the tree model are, respectively, 444 and 460, indicating a slightly worse performance for the model compared to the NB model with values of 437 and 450 for AIC and BIC, respectively. The results indicate that the negative binomial model performs slightly better than the tree model in accordance with the AIC and BIC values for the tree and NB models.

CONCLUSIONS AND RECOMMENDATIONS

A non-parametric tree-based (HTBR) model and a parametric (negative binomial) model were separately developed to establish the empirical relationship between roadway geometry characteristics together with road environmental factors and road accidents. The following conclusions and recommendations were obtained from the results of this paper:

- The results of the tree model indicated the significant variables affecting accidents were AP, SW, HCD, LG, and CUR whereas the variables RW and LU were found not to be statistically significant. For the NB regression model, the variables AP, SW, HCD, and LG had significant impact on accident occurrence on the roadway sections.

- By comparing the analysis and prediction results of both models, this study concluded that NB model is better than the tree model where the results of AIC and BIC for the NB model is less than those for the tree model, indicating a better prediction performance for the NB model.

- As mentioned earlier, statistical models such as Poisson or negative binomial models have been the commonly-applied techniques in road safety analysis. Future work by tree-based modeling techniques may be conducted for a better understanding of factors affecting road accident likelihood by more number of road sections and a larger pool of explanatory variables.

- It would also be interesting to employ other rarely-applied tree-based algorithms such as the Random Forests algorithm or GUIDE learning algorithm to develop accident prediction models and find the factors that influence accident frequency.

REFERENCES

Abdel-Aty, M. A. and A. E. Radwan (2000). "Modeling traffic accident occurrence and involvement", *Accident Analysis & Prevention* 32(5), 633-642.

Abdelwahab, H. T. and M. A. Abdel-Aty (2001). "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections", *Transportation Research Record*, 6-13.

Akgüngör, A. P. and E. Doğan (2009). "An artificial intelligent approach to traffic accident estimation: Model development and application", *Transport* 24(2), 135-142.

Blower, D., K. L. Campbell and P. E. Green (1993). "Accident rates for heavy truck-tractors in Michigan", *Accident Analysis and Prevention* 25(3), 307-321.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984). "Classification and regression trees. Wadsworth & Brooks", *Cole, Pacific Grove, California, USA*.

Brijs, T., D. Karlis, F. Van Den Bossche and G. Wets (2007). "A Bayesian model for ranking hazardous road sites", *Journal of the Royal Statistical Society. Series A: Statistics in Society* 170(4), 1001-1017.

Ceder, A. and M. Livneh (1982). "Relationships between road accidents and hourly traffic flow--I: Analyses and interpretation", *Accident Analysis & Prevention* 14(1), 19-34.

Chiou, Y. C. (2006). "An artificial neural network-based expert system for the appraisal of two-car crash accidents", *Accident Analysis and Prevention* 38(4), 777-785.

Das, A., M. Abdel-Aty and A. Pande (2009). "Using conditional inference forests to identify the factors affecting crash severity on arterial corridors", *Journal of Safety Research* 40(4), 317-327.

Ivan, J. N. and P. J. O'Mara (1997). " Prediction of traffic accident rates using Poisson regression". *Transportation Research Board Meeting*, Washington, DC, Transportation Research Board.

Joshua, S. C. and N. J. Garber (1990). "Estimating truck accident rate and involvements using linear and Poisson regression models", *Transportation Planning and Technology* 15(1), 41 - 58.

Juni, E., T. Adams and D. Sokolowski (2008). "Relating Cost to Condition in Routine Highway Maintenance", *Transportation Research Record: Journal of the Transportation Research Board* 2044(-1), 3-10.

Karlaftis, M. G. and I. Golias (2002). "Effects of road geometry and traffic volumes on rural roadway accident rates", *Accident Analysis & Prevention* 34(3), 357-365.

Karlaftis, M. G. and A. P. Tarko (1998). "Heterogeneity considerations in accident modeling", *Accident Analysis & Prevention* 30(4), 425-433.

Kuhnert, P. M., K.-A. Do and R. McClure (2000). "Combining non-parametric models with logistic regression: an application to motor vehicle injury data", *Computational Statistics & Data Analysis* 34(3), 371-386.

Lee, A. H., M. R. Stevenson, K. Wang and K. K. W. Yau (2002). "Modeling young driver motor vehicle crashes: data with extra zeros", *Accident Analysis & Prevention* 34(4), 515-521.

Maher, M. J. and I. Summersgill (1996). "A comprehensive methodology for the fitting of predictive accident models", *Accident Analysis and Prevention* 28(3), 281-296.

Martin, J. L. (2002). "Relationship between crash rate and hourly traffic flow on interurban motorways", *Accident Analysis and Prevention* 34(5), 619-629.

McCarthy, P. S. (1999). "Public policy and highway safety: a city-wide perspective", *Regional Science and Urban Economics* 29(2), 231-244.

Mohamedshah, Y. M., J. F. Paniati and A. G. Hoheika (1993). "Truck accident models for interstate and two-lane rural roads", *Transportation Research Board*(1407), 35-41.

Montella, A., L. Colantuoni and R. Lamberti (2008). "Crash prediction models for rural motorways", *Transportation Research Record*, 180-189.

Morgan, J. N. and J. A. Sonquist (1963). "Problems in the Analysis of Survey Data, and a Proposal", *Journal of the American Statistical Association* 58(302), 415-434.

Park, Y.-J. and F. Saccomanno (2005). "Collision Frequency Analysis Using Tree-Based Stratification", *Transportation Research Record: Journal of the Transportation Research Board* 1908(-1), 121-129.

Poch, M. and F. Mannering (1996). "Negative binomial analysis of intersection-accident frequencies", *Journal of Transportation Engineering* 122(2), 105-113.

Questier, F., R. Put, D. Coomans, B. Walczak and Y. V. Heyden (2005). "The use of CART and multivariate regression trees for supervised and unsupervised feature selection", *Chemometrics and Intelligent Laboratory Systems* 76(1), 45-54.

Rakha, H., K. Ahn and A. Trani (2004). "Development of VT-Micro model for estimating hot stabilized light duty vehicle and truck emissions", *Transportation Research Part D: Transport and Environment* 9(1), 49-74.

Ramírez, B. A., F. A. Izquierdo, C. G. Fernández and A. G. Méndez (2009). "The influence of heavy goods vehicle traffic on accidents on different types of Spanish interurban roads", *Accident Analysis and Prevention* 41(1), 15-24.

Sayed, T., W. Abdelwahab and F. Navin (1995). "Identifying accident-prone locations using fuzzy pattern recognition", *Journal of Transportation Engineering* 121(4), 352-358.

Shankar, V., F. Mannering and W. Barfield (1995). "Effect of roadway geometrics and environmental factors on rural freeway accident frequencies", *Accident Analysis & Prevention* 27(3), 371-389.

Washington, S. (2000). "Iteratively Specified Tree-Based Regression: Theory and Trip Generation Example", *Journal of Transportation Engineering* 126(6), 482-491.

World Health Organization (1999). "*The world health report 1999: Making a difference*". World Health Organization.

Zhang, C. and J. N. Ivan (2005). "Effects of geometric characteristics on head-on crash incidence on two-lane roads in connecticut", *Transportation Research Record*, 159-164.