Accident Analysis and Prevention 96 (2016) 108–117

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

M5 model tree based predictive modeling of road accidents on non-urban sections of highways in India

Gyanendra Singh (Assistant Professor)[a,*], S.N. Sachdeva (Professor)[b], Mahesh Pal (Professor)[b]

[a] Civil Engineering Department, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonepat, Haryana, India
[b] Civil Engineering Department, National Institute of Technology, Kurukshetra, Haryana, India

ARTICLE INFO

Article history:
Received 22 March 2016
Received in revised form 1 August 2016
Accepted 4 August 2016
Available online 10 August 2016

Keywords:
Predictive modeling
Road safety
M5 model tree
Random effect negative binomial model

ABSTRACT

This work examines the application of M5 model tree and conventionally used fixed/random effect negative binomial (FENB/RENB) regression models for accident prediction on non-urban sections of highway in Haryana (India). Road accident data for a period of 2–6 years on different sections of 8 National and State Highways in Haryana was collected from police records. Data related to road geometry, traffic and road environment related variables was collected through field studies. Total two hundred and twenty two data points were gathered by dividing highways into sections with certain uniform geometric characteristics. For prediction of accident frequencies using fifteen input parameters, two modeling approaches: FENB/RENB regression and M5 model tree were used. Results suggest that both models perform comparably well in terms of correlation coefficient and root mean square error values. M5 model tree provides simple linear equations that are easy to interpret and provide better insight, indicating that this approach can effectively be used as an alternative to RENB approach if the sole purpose is to predict motor vehicle crashes. Sensitivity analysis using M5 model tree also suggests that its results reflect the physical conditions. Both models clearly indicate that to improve safety on Indian highways minor accesses to the highways need to be properly designed and controlled, the service roads to be made functional and dispersion of speeds is to be brought down.

1. Introduction

Indian road network of 4.96 million km is second largest in the world (NHAI, 2014). It has been experiencing a very fast and unprecedented growth. Registered motor vehicle growth rate in the last five decades has been above 10% (MORTH, 2013). But this growth has not resulted in safe travel. The widening and upgradation of roads have not been able to contain the fatal road accidents. Road accidents alone accounted for 1, 41,526 accidental deaths in 2014 (National Informatics Centre (NIC, 2014). The proportion of fatal to total accidents has consistently increased from 18.1% in 2003 to 25.2% in 2013 (MORTH, 2013). Studies suggest that conversion of roads from two to four lanes has resulted in an increase of fatality rate from 41% to 51% on the high-crash-rate sections (Shaheem and Das Gupta, 2005; Shaheem et al., 2006). On 4-lane divided roads, head-on collisions comprise 19% of the crashes due to wrong side traffic alone (MOST, 2000).

Accident prediction models are being used worldwide by engineers and planners as a useful tool to understand the causal factors of road accidents and to suggest measures for improvement. Accident prediction models try to understand the factors associated with accident occurrence by developing statistical relationships correlating various risk factors with the number of accidents occurring on a road section over a period of time. Accident occurrence being a complex phenomenon presents serious challenges to the modeller. The possibility of making causal inferences based on accident prediction models depends strongly on how well the assumptions reflect the reality, what functional relationship was chosen and which method was adopted to overcome disturbing factors. These data and methodological issues have been thoroughly discussed in the literature by various researchers (Miaou, 1996; Persaud, 2001; Hauer, 2010; Lord and Mannering, 2010; Elvik, 2011; Savolainen et al., 2011; Mannering and Bhat, 2014).

Indian scenario has added complexities in accident prediction due to heterogeneity of traffic (Landge et al., 2006; Mohan et al., 2009), under-reporting of accidents (Varghese and Mohan, 1991;

* Corresponding author.
E-mail addresses: singhgyan27@yahoo.in, singhgyan27@gmail.com (G. Singh), snsachdeva@yahoo.co.in (S.N. Sachdeva), mpce_pal@yahoo.co.uk (M. Pal).

http://dx.doi.org/10.1016/j.aap.2016.08.004
0001-4575/

Mohan, 2002; MORTH, 2007) and poor quality of available accident data (Sharma and Landge 2012, 2013; Sharma et al., 2013; Jacob and Anjaneyulu, 2013). Studies have also suggested that separate models must be developed according to the conditions of the individual countries as the traffic flow and its composition, road conditions, driver and road user behaviour are very different in different countries (Jacobs et al., 2000; Fletcher et al., 2006).

The present study was carried out in Haryana, a state in Northern India, lying on the border of the national capital region, New Delhi. In the last two decades (1991–2011) the road accidents in Haryana have increased six times, while the fatalities in road accidents have gone up by 11 times. With 3.5% and 2.2% share in road fatalities and road accidents respectively, Haryana is among top thirteen unsafe states on road safety indicators in India although its geographical area is around 1.34% and population is 2.09% of India (MORTH, 2013). Fig. 1 indicates that all road safety indicators in Haryana are worse than the national average.

Keeping above issues related to road safety in mind, this study focuses on development of predictive model for traffic accidents on non-urban sections of highways in Haryana which are home to about two-third of road accidents and three-fourth road fatalities in the state (MORTH, 2013), and examines the effectiveness of M5 model tree based regression model and conventionally used fixed/random effect negative binomial (FENB/RENB) regression models for prediction of road accidents and identification of the key risk factors.

2. Literature review

2.1. Predictive modeling of road accidents

The occurrence of accidents is a rare and random event and number of accidents is a non-negative discrete variable. Therefore the probability of occurrence of accidents could be better represented by Poisson distribution. As Poisson regression assumes that mean is equal to variance, which is not true for most of the highway safety problems, various variants of Poisson model have been in use for accident prediction according to the data specifications. Negative Binomial regression (NB) models (Lord et al., 2005; Fletcher et al., 2006; Robert and Veeraragavan, 2007; Cafiso et al., 2010; Jacob and Anjaneyulu, 2013; Divakaran and Sreelatha, 2013; Sharma et al., 2014) are best suitable to model over-dispersed data. When accident data has a large number of zero accident sites, zero-inflated Poisson or/and NB models (Sharma et al., 2013; Sharma and Landge, 2013; Jacob and Anjaneyulu, 2013) are employed. Poisson-Weibull models (Maher and Mountain, 2009; Chikkakrishna et al., 2013) provide flexibility in choosing the distribution of error terms. Conway-Maxwell-Poisson models (Lord et al., 2008) also provide flexibility in choosing the distribution of error terms and can handle over- as well as under-dispersion.

Accident data are generally produced by repeated measurements in time over road sections. Thus the data may have spatial and temporal correlations due to some unobserved effects like regional correlation in the data, variation in traffic and driver related effects which are particularly significant in mixed traffic conditions. Poisson and NB distributions cannot handle these unobserved heterogeneities arising from spatial and temporal effects, as the accident distributions for the sites with similar observed characteristics are considered the same in these models. Furthermore, accident counts for a specific location at different time periods are also assumed to be independent of each other. Without appropriately accounting for the location-specific effects and potential temporal correlations, the estimates of the standard error in the regression coefficients may be underestimated. To tackle this problem RENB model was proposed by Shankar et al. (1998). Generalised estimation equations were used by Lord and Persaud (2000) to model crash data with serially correlated repeated measurements. Two-state Markov switching NB models (Malyshkina et al., 2009) which assume switching of roadway segment between two unobserved states of roadway safety to account for unobserved effects, also resulted in a better fit as compared to regular NB models. Anastasopoulos and Mannering (2009) and Dinu and Veeraragavan (2011) applied Random Parameter model by employing a normally distributed error term in the coefficients to allow them to vary across observations. Finite mixture NB regression models with fixed and varying weight parameters were also employed to address the unobserved heterogeneity problem in accident data (Zou et al., 2014). Although these models provide a statistical fit that is significantly better than traditional NB model but they are very complex, may not necessarily improve predictive capability, and model results may not be transferable to other data sets because the results are observation specific (Shugan, 2006; Washington et al., 2010).

The major advantage of applying these statistical models is their ability to identify a broad range of risk factors that can contribute significantly to accidents. However, most of the statistical models have their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, the model could lead to erroneous estimation of accident likelihood (Chang, 2005; Savolainen et al., 2011). NB models, though most widely used in predictive modeling of accidents, are unable to handle under-dispersed data and the low sample mean problem (Lord, 2006; Lord et al., 2008). These models can easily and significantly be influenced by outliers, cannot handle discrete independent variables with more than two levels, and can be adversely affected by multi-colinearity among independent variables (Karlaftis and Golias, 2002).

Another class of predictive algorithms, which does not require any pre-defined underlying relationship between dependent and independent variables, has also been reported in the literature. The main algorithms belonging to this category are Hierarchical Tree based regression (Karlaftis and Golias, 2002; Fletcher et al., 2006), Artificial Neural Network (ANN) (Chang, 2005; Riviere et al., 2006; Xie et al., 2007; Sikka, 2014) and Support Vector Machine (SVM) (Li et al., 2008). Applications of these algorithms are new to the field of highway safety but found to be performing well in comparison to most widely used NB regression approach.

Few applications of tree based regression are reported in highway safety analysis literature. Hierarchical Tree based regression (Karlaftis and Golias, 2002) was tested in Indian conditions by Fletcher et al. (2006) but due to the availability of limited amount of data, the results were found inferior to generalised linear models. In a comparative study (Chang and Chen, 2005), Classification and Regression Tree (CART) was proposed as a good alternative method to NB regression to analyse freeway accident frequencies. Emerson et al. (2011) used M5 model tree based regression (Quinlan 1992; Wang and Witten, 1997) and other data mining approaches to predict crash counts based upon skid resistance values and suggested that M5 regression tree produced high classification rates of instances with a low rule count. M5 model tree can tackle tasks with very high dimensionality (up to hundreds of attributes) and can learn efficiently from large datasets with a small computational cost. The advantage of M5 over CART (Breiman et al., 1994) is that model trees are generally much smaller than regression trees and have proven to be more accurate, due to their ability to exploit local linearity in the data. M5 model tree can have multivariate linear models at its terminal nodes; and thus analogous to piecewise linear functions. Regression trees never give a predicted value lying outside the range observed in the training cases, whereas model trees are found to extrapolate well (Quinlan, 1992). Keeping in view the encouraging performance of M5 model tree, this study

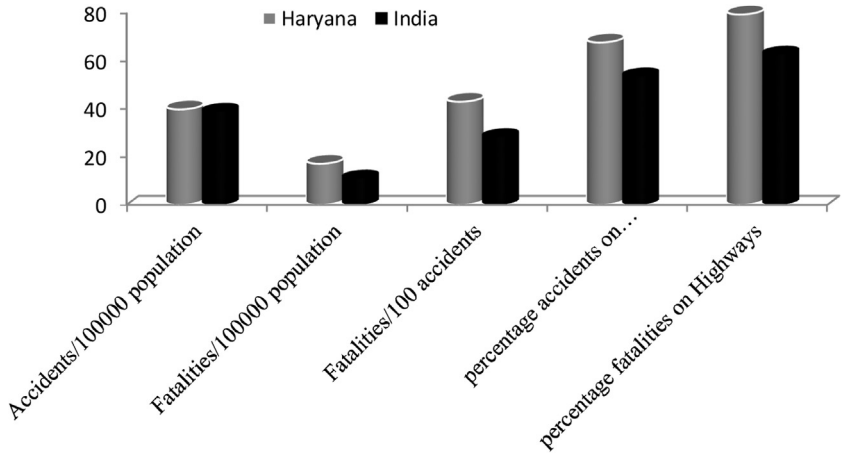

Fig. 1. Road safety indicators in Haryana, India.

examines the effectiveness of M5 model tree approach in the analysis of causal factors of accidents, along with the most widely used FENB/RENB regression approach for accident data from India.

2.2. Choice of explanatory variables

The usual basis for choosing explanatory variables in the accident prediction models has been the availability of the data, the variables that have been found significant in the previous studies and the variables that can be measured in a reliable way (Eenink et al., 2008).

Traffic volume and section length are considered important explanatory variables by most of the researchers in India and found positively associated with accidents (Dinu and Veeraragavan, 2011; Chikkakrishna et al., 2013). The proportion of trucks in traffic was found positively associated with fatal and major accidents (Srinivas et al., 2007). By developing separate models for day and night time, Dinu and Veeraragavan (2011) concluded that increase in the proportion of trucks caused an increase in *day time* accidents but a decrease in night-time accidents. The proportion of cars in traffic was found negatively associated with accidents. Speed and its variance were also considered significantly correlated with accidents involving two-wheeler riders and cars (Sharma and Landge, 2013; Sharma et al., 2014). Contrary to this, Rokade et al. (2010) found speed negatively correlated with accidents. In addition to above variables Jacob and Anjaneyulu (2013) found carriageway width and number of curves also significant and concluded that increased carriageway width beyond certain limit, results in higher speeds and higher unsafe overtaking manoeuvres resulting accidents. They also concluded that each additional curve causes 23% increase in accident frequency but increase beyond certain number reduces accidents. Number/density of driveways, minor accesses and median openings are also found having significant positive association with accidents (Sharma and Landge, 2012, 2013; Chikkakrishna et al., 2013; Sharma et al., 2014).

Table 2
Performance indicators of various Models.

Model	CC	RMSE value	MAE value
M5 model tree	0.916	7.343	4.889
RENB2 model	0.891	8.862	5.529

3. Methodology

3.1. M5 model tree

M5 model tree is a conventional decision tree with the linear regression function at the terminal nodes (Quinlan, 1992). It is constructed by a divide-and-conquer method and used for predicting continuous numeric variables, unlike conventional decision tree classifier which is used to predict the discrete classes. A model tree generation involves two stages. Stage one involves in using a splitting criterion to create a decision tree whereas stage two involves in using a pruning method to prune back an overgrown tree. M5 model tree algorithm uses the standard deviation of the class val-

Table 1
Description of Model Variables.

Sr. No.	Variable with measurement units	Designation in the model	Min.	Max.	Mean	Std. Deviation
1	Accident Frequency (Dependent variable)	A	0	96	12.07	15.780
2	Annual Average Daily Traffic (1000 PCU/day)	AADT	6.397	93.752	32.668	26.103
3	Section length (km)	L	0.400	13.200	4.891	3.314
4	Carriageway width (m)	RW	5.5	28.0	10.977	6.190
5	Paved shoulder width (m)	PSW	0.0	6.0	1.989	1.313
6	Median width (m)	MW	0.0	10.50	1.197	1.847
7	Number of Minor access	M_Acc	0	30	9.08	7.554
8	Number of horizontal curves	HC	0	19	3.84	4.493
9	Number of Median openings	MO	0	15	2.34	4.023
10	Length of service road (km)	SR	0.000	18.100	1.070	3.203
11	Percentage of trucks in AADT	trucks_perc	11.34	56.19	34.47	12.61
12	Percentage of cars in AADT	Cars_perc	9.32	72.5	27.94	14.05
13	Driveways and commercial units along the road (numbers)	DW	0	54	9.41	9.932
14	Narrow Bridge and culverts	BC	0	9	2.00	2.195
15	98th percentile speed (KMPH)	98th speed	66	123	80.52	12.149
16	Standard Deviation of speed (KMPH)	STD	13.00	25.71	17.82	3.60

Table 3
Comparative statistics and Parameter estimates of FENB/RENB models.

Sr. No.	Model characteristics	FENB model	RENB Model with	
			Time specific random effects	Location-time specific random effects
A	Comparative statistics		RENB1	RENB2
1	Bayesian Information Criterion (BIC)	323.459	343.739	324.501
2	Akaike Information Criterion - Finite Sample Corrected (AICC)	326.371	321.305	299.411
3	-2 Log pseudo likelihood	321.430	304.206	280.026
B	Parameter estimates			
1	(Intercept)	−0.621 (0.472)	−0.700 (0.356)	−0.643 (0.393)
2	ln AADT	0.264 (0.091)	0.238 (0.084)	0.240 (0.093)
3	ln L	0.154 (0.072)	0.128 (0.066)	0.143 (0.078)
4	M_Acc	0.035 (0.008)	0.036 (0.007)	0.038 (0.006)
5	PSW	−0.123 (0.040)	−0.133 (0.035)	−0.141 (0.036)
6	SR	0.097 (0.011)	0.097 (0.011)	0.105 (0.011)
7	STD	0.084 (0.018)	0.094 (0.015)	0.085 (0.016)
	Over-dispersion parameter	0.174	0.174	0.174
C	Covariance parameters			
a.	Residual effect	1.325 (0.158)		
1	2007		0.524 (0.513)	0.109 (0.280)
2	2008		3.375 (1.453)	3.038 (1.403)
3	2009		0.347 (0.152)	0.415 (0.251)
4	2010		0.912 (0.247)	0.324 (0.144)
5	2011		2.239 (0.512)	0.941 (0.279)
6	2012		0.861 (0.257)	0.688 (0.255)
7	2013		1.089 (0.336)	0.587 (0.276)
8	2014		0.504 (0.306)	0.291 (0.383)
b.	Random effects			0.133 (0.049)

ues reaching a node as a measure of the error at that node. It further calculates the expected reduction in the error as a result of testing each variable at that node. The standard deviation reduction *(SD*R) formula used in the design of M5 model tree is represented by:

$$SDR = sd(K) - \sum \frac{|K_i|}{|K|} sd(K_i) \quad (1)$$

where *K* represents number of examples reaching the node; K_i represents the number of examples having i^{th} outcome of the potential set; and *sd* represents the standard deviation. This splitting process forces child node to have smaller value of standard deviation as compared to parent node thus making them more pure (Quinlan, 1992). The design of M5 model tree chooses the split that maximizes the expected error reduction after examining all the possible splits. This data division during M5 model creation produces a large tree which may be the cause of over fitting with testing data. To remove the problem of over fitting, Quinlan (1992) suggested using some pruning method to prune back the over grown tree. In general, this pruning is achieved by replacing a *sub tree* with a linear regression function. For further details, the readers are referred to Quinlan (1992) and Witten and Frank (2005).

3.2. Random effect negative binomial (RENB) model

Regular negative binomial model or FENB model (Lord, 2006; Lord et al. 2008) does not allow for location-specific effects or serial correlation over time for clustered accident counts as each observation of year *t* within the i^{th} location group is treated as an independent observation resulting in a total of NxT independent observations. Shankar et al. (1998) suggested that when location specific effects are random and have temporal effects within the location group, the same set of observations must be modelled as a penal with N location groups and T periods.

According to RENB model specification used in this study (Agresti et al., 2000; Diggle et al., 2002), the form of RENB model for the target λ (expected accident frequencies) with the random effects $\boldsymbol{\gamma}$ is given by Eq. (2).

$$\eta = \ln E(\lambda|\gamma) = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon, \lambda|\gamma \sim NB, \quad (2)$$

where η is the linear predictor; *ln* is link function; $\mathbf{X}$ is an $(N \times p)$ design matrix for the *p* predictor variables; β is a *p x 1* column vector of the fixed-effects regression coefficients; $\mathbf{Z}$ is an $(N \times r)$ design matrix for the *r* random effects (the random compliment to the fixed $\mathbf{X}$); $\boldsymbol{\gamma}$ is a $(r \times 1)$ vector of random effects (the random compliment to the fixed β) which are assumed to be normally distributed with mean 0 and variance matrix $\mathbf{G}$ (of random effects); $\boldsymbol{\varepsilon}$ is an *(N x 1)* column vector of the residuals, that part of λ which is not explained by the model; variance matrix of repeated measures is R and N is total number of observations. NB is the conditional target probability distribution. If there are no random effects, the model reduces to a regular negative binomial model. For a detailed discussion about RENB, readers are referred to Shankar et al. (1998) and Chin and Quddus (2003).

3.3. Data set and methodology

The data used for this study was collected over a number of non-urban sections of Roads in Haryana namely National Highways NH-1, NH-65, NH-73A and NH-248A and State Highways SH-11, SH-18 and SH-20 and MDR-137. The data for road geometry and road environment was collected through field visits along various highways. Traffic volume data for National Highway sections was collected from Toll plazas and from various Detailed Project Reports prepared by National Highway Authority of India (NHAI). Traffic volume for State Highway sections was obtained from traffic registers of Haryana Public Works Department. The traffic volume data for MDR-137 was collected from toll plaza. The non-available Traffic data for some of the sections was collected by 16–24 h traffic count on National Highway sections and by 12–16 h count on State Highway sections. The data of spot speeds was collected by using the radar gun.

On the basis of collected data, road sections having similar traffic, road width, median width and shoulder width were identified. Total road length of various highways covered in this study was more than 250 km and it was divided into 68 uniform sections of varying length. NH-1 was subdivided into 21 sections, NH-65 into 6 sections, NH-73A into 7 sections, NH-248A into 3 sections, SH-20 into 9 sections, SH-18 into 8 sections, SH-11 into 7 sections and

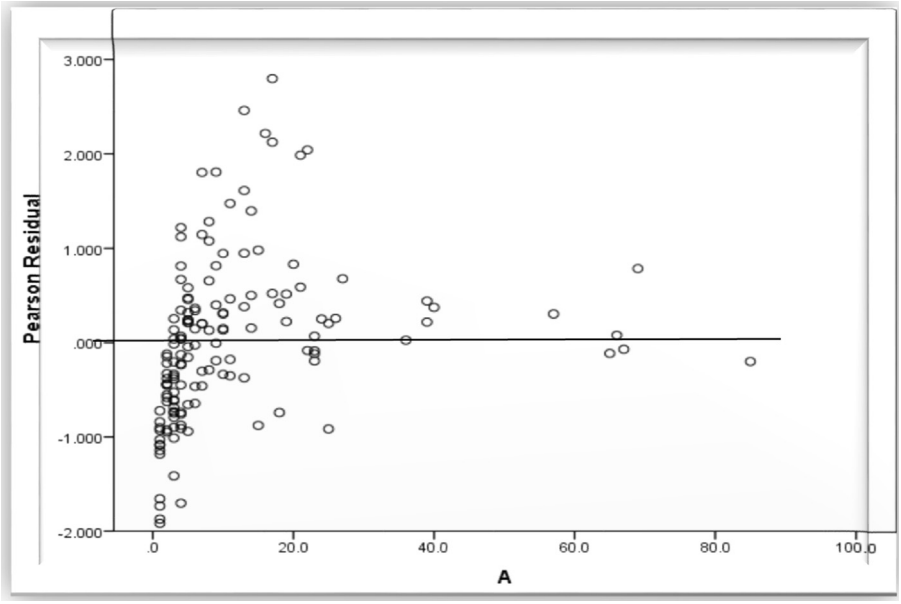

Fig. 2. Plot of Pearson residuals for RENB2 model with train dataset (Where A stands for actual accident frequencies).

MDR-137 into 7 sections. Road accidents data for a period of 2–6 years (ranging between year 2007–2014) on different sections was collected from police records. 85.7% of this data belonged to years 2009-13. The accident data included all types of accidents reported to police i.e. fatal, serious and minor injury accidents and property damage only accidents. This resulted in a data set comprising a total of 222 samples. The summary statistics for all 16 variables collected from various sources as described above is provided in Table 1.

Based on the literature review in Section 2.2, fifteen input variables namely annual average daily traffic (AADT), section length (L), carriage way width (RW), paved shoulder width (PSW), median width (MW), 98th percentile Speed (98th_speed), standard deviation of speed (STD), proportion of trucks (Truck_perc) and cars (Cars_perc) in the traffic, number of Minor access points (M_Acc), driveways and commercial establishments (DW), median openings (MO), horizontal curves (HC), narrow bridges and culverts (BC) and length of service road along the highway (SR) were used as independent (input) variables whereas accident frequency (A) represented as Accidents/year was chosen as dependent variable.

Random division (train/test approach) is one of the most commonly used approaches to test the effectiveness of a machine learning algorithm like M5 model tree. This approach involves in dividing the total dataset randomly into two parts. In the present study, a total of 148 samples were used for training (creating the model) and remaining 74 samples for testing models with different modeling approaches. For M5 model tree and RENB model Weka (Witten and Frank, 2005) and SPSS (IBM, 2013) software were used respectively. Both models are validated using test data in terms of three different statistical measures namely, correlation coefficient (CC), Root mean square error (RMSE) and mean absolute error (MAE).

98^{th}_ speed <= 85.5 :
| STD <= 14.425 :
| | trucks_perc <= 36.535 : LM1
| | trucks_perc > 36.535 : LM2
| STD > 14.425 :
| | M_ACC <= 8.5 :
| | | HC <= 2.5 :
| | | | B/C <= 0.5 : LM3
| | | | B/C > 0.5 : LM4
| | | HC > 2.5 :
| | | | L<= 3.725 : LM5
| | | | L> 3.725 : LM6
| | M_ACC > 8.5 :
| | | trucks_perc <= 51.369 : LM7
| | | trucks_perc > 51.369 : LM8
98^{th}_ speed > 85.5 :
| SR <= 0.99 :
| | M_ACC <= 5.5 : LM9
| | M_ACC > 5.5 : LM10
| SR > 0.99 :
| | SR <= 5.05 :
| | | DW <= 24 : LM11
| | | DW > 24 : LM12
| | SR > 5.05 : LM13

Fig. 3. M5 Model tree with train dataset.

4. Results

Table 2 provides the correlation coefficient (CC), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values using both M5 model tree and RENB regression model in predicting accident frequencies using test dataset. In comparison to RENB2 model, results from Table 2 indicate a seemingly improved performance by M5 model tree for the used dataset but the differences in predicted values are not significant as suggested by student's *t*-test ($P_{\alpha=0.05} = 0.445$, *and* $t = -0.141$).

To develop RENB model, first, the correlation between the pairs of independent variables was checked. Variable pairs lnAADT and RW and lnAADT and MW, found strongly correlated (Pearson's correlation higher than ±0.85 and $p < 0.05$ indicating statistically

LM1
A = 0.0176 * AADT + 0.05 * L - 0.3854 * PSW+ 0.1138 * MW+ 0.1438 * M_ACC+ 0.4213 * SR+ 0.0476 * trucks_perc + 0.045 * cars_perc+ 0.033 * DW + 0.0273 * 98th _speed + 0.2144 * STD- 5.2185
LM2
A = 0.0176 * AADT+ 0.05 * L - 0.3854 * PSW+ 0.1138 * MW+ 0.1438 * M_ACC+ 0.4213 * SR+ 0.0675 * trucks_perc + 0.0523 * cars_perc+ 0.033 * DW+ 0.0273 * 98th _speed + 0.2144 * STD - 5.5808
LM3
A = 0.0409 * AADT+ 0.1562 * L - 0.5171 * PSW+ 0.1138 * MW+ 0.1433 * M_ACC+ 0.4213 * SR- 0.0438 * trucks_perc+ 0.0306 * cars_perc+ 0.033 * DW- 0.2487 * B/C+ 0.0538 * 98th _speed + 0.0809 * STD - 0.4814
LM4
A = 0.0409 *A ADT+ 0.1562 * L - 0.5693 * PSW+ 0.1138 * MW+ 0.1433 * M_ACC+ 0.4213 * SR - 0.0438 * trucks_perc + 0.0306 * cars_perc+ 0.033 * DW- 0.1973 * BC+ 0.0538 * 98th _speed + 0.0809 * STD - 0.9903
LM5
A =0.1461 * AADT + 0.3257 * L - 0.3773 * PSW + 0.1138 * MW + 0.1433 * M_ACC+ 0.4213 * SR- 0.2192 * trucks_perc + 0.0306 * cars_perc+ 0.033 * DW - 0.1116 * BC+ 0.0538 * 98th _speed + 0.0809 * STD + 2.8131
LM6
A =0.0993 * AADT+ 0.3199 * L - 0.3773 * PSW+ 0.1138 * MW+ 0.1433 * M_ACC+ 0.4213 * SR - 0.1377 * trucks_perc + 0.0306 * cars_perc+ 0.033 * DW - 0.1116 * BC+ 0.0538 * 98th _speed + 0.0809 * STD + 0.6912
LM7
A = 0.0676 * AADT- 0.0816 * L - 1.4575 * PSW+ 0.1138 * MW+ 0.1695 * M_ACC+ 0.4213 * SR+ 0.0614 * trucks_perc + 0.1077 * cars_perc+ 0.033 * DW+ 0.1404 * 98th _speed + 0.0809 * STD - 9.0128
LM8
A = 0.0276 * AADT - 0.1731 * L - 1.1888 * PSW+ 0.1138 * MW+ 0.1695 * M_ACC+ 0.4213 * SR+ 0.1118 * trucks_perc + 0.1559 * cars_perc+ 0.033 * DW+ 0.2763 * 98th _speed + 0.0809 * STD - 19.8178
LM9
A = 0.0602 * AADT + 0.1034 * L - 0.4064 * PSW+ 0.0065 * MW+ 0.2771 * M_ACC+ 2.0051 * SR - 0.1916 * trucks_perc - 0.0794 * cars_perc+ 0.2745 * DW+ 0.0565 *98th _speed + 5.7275
LM10
A = 0.066 * AADT+ 0.1034 * L - 1.4055 * PSW- 0.2739 * MW + 0.3176 * M_ACC+ 2.0051 * SR - 0.1916 * trucks_perc - 0.0794 * cars_perc + 0.2745 * DW+ 0.0565 * 98th _speed + 9.184
LM11
A = 0.0908 * AADT+ 0.1034 * L - 0.9838 * PSW + 0.2354 * MW+ 0.1679 * M_ACC + 2.7487 * SR- 0.2468 * trucks_perc - 0.0816 * cars_perc+ 0.5266 * DW+ 0.0565 * 98th _speed + 10.2665
LM12
A = 0.0935 * AADT + 0.1034 * L - 0.9838 * PSW+ 0.2354 * MW+ 0.1679 * M_ACC+ 2.7487 * SR - 0.2468 * trucks_perc - 0.0816 * cars_perc + 0.5505 * DW+ 0.0565 * 98th _speed + 10.5222
LM13
A =0.0949 *A ADT + 0.1034 * L - 0.9838 * PSW + 0.2354 * MW+ 0.1679 * M_ACC + 3.0539 * SR - 0.2687 * trucks_perc - 0.0816 * cars_perc + 0.4959 * DW+ 0.0565 * 98th _speed + 16.2309

Fig. 4. Accident prediction equations by M5 Model tree with train dataset.

significant correlation at 95% confidence level), were discarded and only lnAADT was used during model development. A full Poisson model (taking all input variables) was then developed and checked for goodness of fit. For full Poisson model, the Pearson χ^2 value divided by degrees of freedom was 2.939 indicating over-dispersion and a fit case for the use of NB model. To account for temporal and spatial variability in the dataset used (multi-year data for sections of 8 highways belonging to the same state) RENB models (Shankar et al., 1998) were used. In the SPSS Generalised linear mix model specification, model structure was defined by selecting HID (Highway ID, 1–8) and RID (section ID, 1–19) as subject and year as repeated measure. The variables whose coefficients were found significant ($p < 0.05$), improved the finite sample corrected Akaike Information Criterion (AICC) and Bayesian Information Criterion (BIC) statistics were included in the final model (Vogt and Bared, 1998; Sawaldha and Sayed, 2003).

Three models developed with different specifications are reported in Table 3: FENB model, considering all observations as independent (Column 3, Table 3); RENB1 model, considering only serial correlation effects (Column 4, Table 3) and RENB2 model, considering both location and time effects with subject specification HID*RID (Column 5, Table 3). Fourth model (RENB3 model) was also developed considering both location and time effects and two subjects HID and HID*RID. Convergence problem occurred in RENB3 model and the results were only marginally different from the RENB2 model in terms of AICC and BIC statistics. The variance of HID was also found insignificant. Therefore its results are not being reported here. The comparative statistics of FENB, RENB1 and RENB2 models (Table 3) indicate the presence of both spatial and temporal correlation in the data. The detailed results for the RENB2 Model (Table 3, column 5) are reported in Table 4–6 and Fig. 2.

Table 4
Model Specifications.

Sr. No.	Description	Effects	Number
1	Covariance Parameters	Residual effects	8
		Random effects	1
2	Design matrix	Fixed effects	7
		Random effects	1[a]
3	Common subjects		63

Common subjects are based on subject specifications for random and residual effects and are used to chunk the data for better performance.
[a] This is the number of columns per common subject.

Fig. 3 provides the output of M5 model tree and the model coefficients for accident prediction equations at terminal nodes are given in Fig. 4. These simple linear equations at nodes (Figs. 3 and 4), can easily be used to predict accident frequencies for the dataset with in the given range as provided in Table 1.

The predictions by both M5 model tree and RENB2 model for test data were compared with the actual accident frequency values and plotted in Figs. 5. Results from Table 2 and Figs. 5 indicate that the performance of M5 model tree based approach is comparable to that achieved by RENB model.

5. Sensitivity analysis using M5 model tree

This section discusses the effect of the four significant input parameters (AADT, M_ACC, SR and STD) on the accident frequencies, using M5 model tree. This is achieved by testing the model created by M5 model tree using training data (148 samples) with hypothetical test datasets. The hypothetical test dataset is created by varying only one input parameter while keeping all other input parameters constant of one sample taken from the test data. To study the effect

Table 5
Model coefficients for fixed effects in RENB2 model.

Sr. No.	Model term	Coefficient	Std. Error	t	Sig.	Exp(coefficient)	95% confidence interval for Exp(coefficient)	
							Lower	Upper
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	Intercept	−0.643	0.393	−1.634	0.109	0.526	0.239	1.159
2	ln AADT	0.240	0.093	2.567	0.013	1.271	1.054	1.533
3	ln L	0.143	0.078	1.837	0.074	1.153	0.986	1.350
4	M_Acc	0.038	0.006	5.923	0.000	1.039	1.025	1.054
5	PSW	−0.141	0.036	−3.871	0.000	0.868	0.807	0.934
6	SR	0.105	0.011	9.308	0.001	1.110	1.077	1.145
7	STD	0.085	0.016	5.170	0.000	1.088	1.053	1.125
	Dispersion parameter	0.174						

Table 6
Covariance parameter estimates for Random and Residual effects in RENB2 model.

Sr. No.	Covariance Parameter	Estimate	Std. Error	Z	Sig.	95% confidence interval for Exp(coefficient)	
						Lower	Upper
A	Residual effect						
1	Var(YEAR = 2007)	0.109	0.280	0.388	0.698	0.001	16.923
2	Var(YEAR = 2008)	3.038	1.403	2.164	0.030	1.228	7.513
3	Var(YEAR = 2009)	0.415	0.251	1.652	0.099	0.127	1.360
4	Var(YEAR = 2010)	0.324	0.144	2.256	0.024	0.136	0.772
5	Var(YEAR = 2011)	0.941	0.279	3.373	0.001	0.526	1.681
6	Var(YEAR = 2012)	0.688	0.255	2.698	0.007	0.333	1.423
7	Var(YEAR = 2013)	0.587	0.276	2.113	0.035	0.232	1.485
8	Var(YEAR = 2014)	0.291	0.383	0.758	0.448	0.022	3.854
B	Random effect	0.133	0.049	2.702	0.007	0.065	0.276

Covariance Structure: Unstructured.
Subject specification: HID*RID.
The covariance structure is changed to scaled identity because the random effect has only one level.

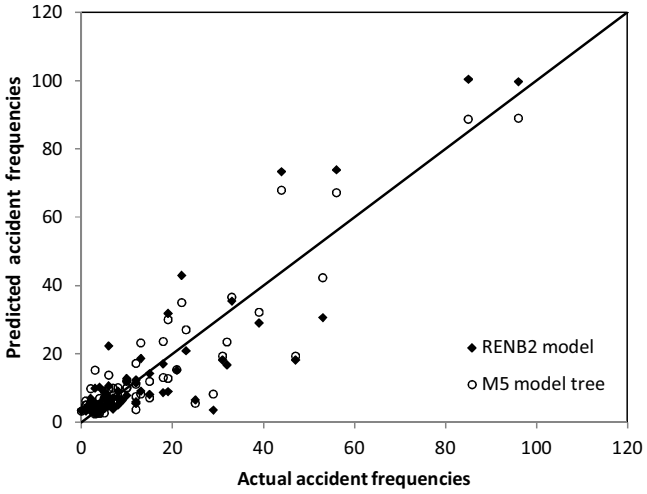

Fig. 5. Plot between actual and predicted values for M5 model tree and RENB2 Model with test dataset.

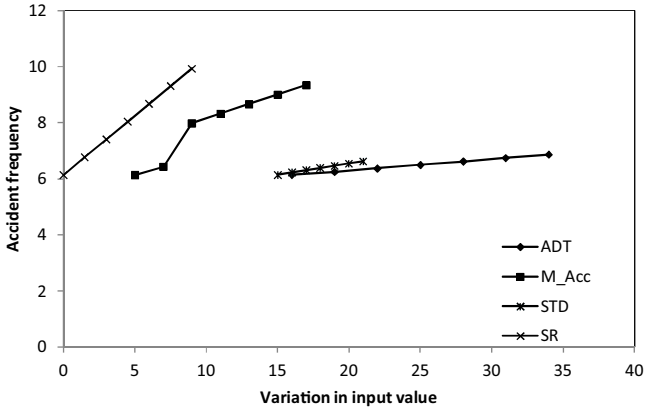

Fig. 6. Variation of Accidents frequency with increasing value of different input parameters.

of the change in STD, SR, M_ACC and AADT on accident frequencies, STD was increased by a value 1, SR value by 1.5, M_ACC by a value 2, and AADT by a value 3, while other input parameters were kept constant. This way seven datasets having different values of each of the STD, SR, M_ACC and AADT were created and used for further processing for sensitivity analysis.

Results plotted in Figs. 6 suggest that M5 model tree was able to create the conditions similar to physical modeling in justifying that accidents increase with increasing value of STD, SR, M_ACC and AADT. Thus, M5 model based predictive approach is capable of generalisation well in predicting accident frequencies within the range of the input parameters used in this study.

6. Discussion

6.1. Results of M5 model tree

By observing the regression tree (Fig. 3) by M5 model tree approach and the linear predictor equations LM 1–13 at terminal nodes (Fig. 4), it is clear that out of the 15 explanatory variables, M5 model considered three variables, carriageway width, median openings, number of narrow bridges and culverts as insignificant. Other variables were considered significant either in the division of data, like variables speed and standard deviation of speed, number of minor accesses, length of service road along highway and proportion of trucks; or in linear predictor equations, like variables traffic volume, section length and paved shoulder width etc. The first division of the data is based on 98th percentile speed, suggesting it to be the most important input variable in the design of M5 model tree.

After 98th percentile speed, the splitting criteria used by M5 model tree (Fig. 3) suggests STD (when 98th percentile speed <=85.5) and SR (when 98th percentile speed >85.5) are other two most important variables. In India, the design speed of National highways is generally kept 100 kmph and that of state highways 80 kmph. The first division, therefore, indicates that the equations LM 1–8 can be used for state highways and equations LM 9–13 for national highways. Thus, in the case of state highways, M5 model tree considers standard deviation of speed, number of minor accesses, proportion of trucks and number of curves as the significant input variables for the division of data. On the other hand on national highways service road length, number of minor access and number of driveways and commercial establishments along the highway were considered significant.

Every linear model in Fig. 4 is applicable to a set of conditions defined by model tree (Fig. 3). For example when speed on a particular section is less than 85.5 kmph, the standard deviation of speed is below 14.425 kmph and truck percentage is below 36.535%, accident frequency will be given by linear model LM1.

Speed and its variance both were found positively associated with accidents. These results are in accordance with studies reported in reported in the literature (Quimby et al., 1999; Taylor et al., 2002). Effect of speed on accident frequencies was maximum on sections with a large percentage of trucks and large number of minor accesses. The standard deviation of speed was found having significant effect only on low speed roads and not on national highways. Similar to the study reported by Ackaah and Salifu (2011), minor accesses on highway were found positively associated with accidents.

Exposure variables, traffic volume and section length were found positively associated with accident frequency. Fig. 4 indicates that on state highways the effect of traffic volume increases with increasing value of standard deviation of speed thus suggesting that on two roads carrying same traffic volume, accidents will be more on the road carrying mixed traffic. The effect of traffic volume on accident frequencies was also found associated with number of curves and minor accesses on state highways and with length of service road and number of minor accesses and drive ways on national highways. Accident frequencies also increase with the increase in exposure in terms of distance travelled. This exposure effect is comparatively high when traffic is mixed (standard deviation is more) and the section has a large number of curves.

The linear models in Fig. 4 indicate that the accident frequency increases with increasing length of service road. This apparently seems working against the common assumption as the service roads generally run parallel to the highway and separated from the highway by raised curb. Service roads collect the local traffic and provide entry at limited points thus preventing the wrong side entries to the highways leading to reduction in accidents. But positive impact of service road length on accident frequencies in the present study can be a result of dysfunctional service roads with no control of access in the study area. Along many sections, service road is provided only on one side of a divided highway. At many locations, entry/exit from service road to the main highway are provided without consideration of local needs and the traffic has to travel some distance on the wrong side to find a median opening. At many places service roads are ill maintained, used for parking and encroached upon for commercial activities like eateries, petrol pumps and shops all along the service roads which lead to wrong side traffic movement. Along many sections, no raised curbs are provided between service road and highway resulting in uncontrolled access from service road to highway. In this way service roads increase conflict points along the highway without any warning, resulting in increased accident frequencies along sections with longer service roads.

Paved shoulder width was found negatively associated with accidents (Fig. 4). Similar results have been reported in the literature for multilane roads (Hauer et al., 2004; Dinu and Veeraragavan, 2011). On state highways, effect of increase in paved shoulder width is more when speed variance is large and minor accesses are in large number. On national highways the paved shoulders are more effective when service roads are of short length and minor accesses are more in number. This is probably because of the reason that the paved shoulders act as acceleration or deceleration lane for diverging or merging traffic from minor access. Provision of the paved shoulder was found comparatively less effective in reducing accidents on roads with a large number of curves.

Median width was found to be positively associated with the accident frequencies. These results are in line with findings of Hauer (2000) and Elvik and Vaa (2004). Median width increases the number of all accident types, possibly due to the hindrance during overtaking manoeuvres, the presence of a new hazard in the carriageway (Elvik and Vaa, 2004) and due to its association with increased speeds (Hauer, 2000), particularly during night time. The contribution of median width was found minimum when service road length was short and minor accesses were less, i.e. on road sections with fewer conflicts.

The proportion of cars in total traffic was found positively associated with accidents on low-speed roads but have negative impact on high-speed roads. Although the results partially confirm the findings of Dinu and Veeraragavan (2011), which suggest the proportion of cars to be negatively correlated with accidents, but the results seem reasonable as increase in the proportion of cars on low-speed roads increases speed variance but on high-speed roads it makes traffic more homogeneous. On low-speed roads, the effect of increase in the proportion of cars was highest when speed variance, percentage of trucks and number of minor accesses were larger. The proportion of trucks in total traffic was found positively associated with accidents when speed and standard deviation both were less. The additive effect was comparatively high when proportion of trucks was above 51%, speed variance was more than 14.425 kmph, and number of minor accesses was more than eight. On high as well as low-speed roads with less number of minor accesses the proportion of trucks was found negatively associated with accident frequencies. Number of driveways and commercial establishments along the highway was found positively associated with accident frequency. This finding is similar to the results reported by Dinu and Veeraragavan (2011) but this association was significant only on high-speed road sections.

6.2. Results of RENB model

Results from Table 3 indicate that RENB2 model with location and time specific random effects shows considerable improvement in comparison with FENB and RENB1 models in terms of AICC, BIC statistics and residual effects. The standard error of coefficient estimates is generally higher for RENB2 model which is in accordance with findings of Shankar et al. (1998). This clearly indicates the presence of unobserved heterogeneity in the data. The RENB2 model considers only six input variables: Traffic volume, length of section, paved shoulder width, number of minor accesses, length of service road and standard deviation of speed having significant effect on accidents in comparison to other variables. Other variables were not found significantly affecting accident frequencies. Out of six significant variables, only PSW was found negatively associated with accident frequencies. The column 7 in Table 5 indicates the incidence rates of the variables. These values indicate the corresponding change in accident frequency for a change of one unit in the value of the variable keeping all other factors including random effects constant. The values lying above one indicate positive association while those lying below one show the negative association

of a variable with accidents. The values close to one indicates that the variable has no significant effect on accident frequency.

Table 6 shows variance estimates of random and residual effects. Within section random effects show a significant variance of 0.133. Significant residual effects were observed for years 2010–2013. The estimated variance of residual effects for years 2007, 2009 and 2014 are less significant as compared to other years (Table 6, column 6) and significant but quite large for the year 2008 (Table 6, column 3). This is probably due to non-availability of data for these years for many road sections. The Pearson residuals plotted in Fig. 2 for the final RENB model indicate a considerably significant statistical fit.

6.3. Comparison of M5 and RENB model results

The variables Minor access, length of service road, standard deviation of speed, traffic volume, length of section and paved shoulder width were found to be significant input variables in both the models. Speed was considered the most significant variable by M5 model tree but it was not considered significant in RENB2 model. M5 model also suggests different sets of significant variables for state and national highways. For example service road length and number of driveways was considered significant for high-speed national highways. Similarly, number of curves and standard deviation of speed were considered significant for state highways only. RENB model provides no such division unless different models are developed for different road types.

On the basis of discussion in Sections 6.1 and 6.2, it can be concluded that M5 model tree based regression approach is simple and easy to interpret and its results are comparable to RENB model. M5 model allows change of sign and variation in the values of the coefficients across observations, resulting in a better insight of the effect of an explanatory variable or group of variables over accident frequency.

The M5 model tree provides both theoretical and applied advantages over FENB/RENB model. Theoretically, it does not require a predetermined functional form of the model. Neither has it required any prior assumption related to the effect of risk factors on accidents. From the practical point of view, M5 models are capable of graphically displaying the results; thus make the results easily understandable. The model tree outcome can be structured as a sequence of "if-then" rules. By using these rules a practicing engineer can trace a path down the tree to a terminal node where he finds a simple linear equation to predict accidents. The number of rules provided by M5 model tree is the function of the complexity of the modeling problem and thus may be considered as one of its drawbacks in case number of rules becomes too large.

7. Conclusions

This paper investigates the potential of M5 model tree and conventionally used FENB/RENB regression models for predicting the accident frequencies and identification of key risk factors for road accidents using the field dataset collected on non-urban sections of highways in India (Haryana). A major conclusion from this study is that both models perform comparably well. Availability of simple linear relations in M5 model tree approach is a major advantage as that may be useful in predicting the accident frequencies within the given input data range and provide a good insight about the changing effect of variables over accidents across observations. Sensitivity analysis also suggests that M5 model tree algorithm is successful in modeling the physical process of accident occurrence on rural highways.

The two well-performing modeling approaches lead to similar conclusions regarding potential risk factors. Accidents were found increasing with increase in traffic volume, length of section, standard deviation of speed and service road length and number of minor access whereas accident frequency decreased with an increase in paved shoulder width. The results also suggest that to improve safety on highway sections in India minor access road junctions need to be improved and service roads are to be made functional and access controlled. The significant impact of standard deviation of speed on increase in accident frequencies (in comparison to speed itself) is another area of concern reflecting mix traffic conditions on Indian highways.

In spite of the encouraging performance by M5 model tree in the present study, further investigations are needed to judge the effectiveness of this approach in predicting road accidents as artificial intelligence-based approaches are data dependent and their output may change depending on the dataset, as well as number of data available for training.

References

Ackaah, W., Salifu, M., 2011. Crash Prediction Model for two-lane rural Highways in the Ashanti region of Ghana. IATSS Res. 35 (1), 34–40.

Agresti, A., Booth, J.G., Caffo, B., 2000. Random-effects modeling of categorical response data. Sociol. Method. 30, 27–80.

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

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1994. Classification and Regression Trees. CRC Press.

Cafiso, S., Graziano, A.D., Silvestro, G.D., Cava, G.L., Persaud, B., 2010. Development of comprehensive accident models for two lane rural highways using exposure, geometry, consistency and context variables. Accid. Anal. Prev. 42, 1072–1079.

Chang, L.Y., Chen, W.C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Saf. Res. 36 (4), 365–375.

Chang, Li-Yen., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Saf. Sci. 43, 541–557, http://dx.doi.org/10.1016/j.ssci.2005.04.004.

Chikkakrishna, N.K., Parida, M., Jain, S.S., 2013. Crash prediction for multilane highway stretch in India. Proc. East. Asia Soc. Transp. Stud. 9.

Chin, H.C., Quddus, M.A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accid. Anal. Prev. 35 (2), 253–259.

Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.L., 2002. The Analysis of Longitudinal Data, 2nd ed. Oxford University Press, Oxford.

Dinu, R.R., Veeraragavan, A., 2011. Random parameter models for accident prediction on two-lane undivided highways in India. J. Saf. Res. 42, 39–42.

Divakaran, L.M., Sreelatha, T., 2013. Accident prediction models for urban unsignalised intersections. Int. J. Innovative Res. Sci. Eng. Technol. 2 (1), Proceedings of International Conference on Energy and Environment-2013 (ICEE 2013).

Eenink, R., Reurings M., Elvik R., Cardoso J., Wichert S. and Stefan C. (2008). Accident Prediction Models and Road Safety Impact Assessment: Recommendations for using these tools. Sixth Framework Programme, 2008. www.ripcord-iserest.com, RIPCORD-iSEREST-Deliverable-D2. doc.

Elvik, R., Vaa, T., 2004. Handbook of Road Safety Measures. Elsevier, Oxford, United Kingdom.

Elvik, R., 2011. Assessing causality in multivariate accident models. Accid. Anal. Prev. 43, 253–264.

Emerson, D., Nayak, R., Weligamage, J., 2011. Using data mining to predict road crash count with a focus on skid resistance values. In: 3rd International Road Surface Friction Conference, 15–18 May 2011, Gold Coast, Queensland, Australia, in press.

Fletcher, J. P., Baguley, C, J, Sexton, B., Done, S. (2006) Road Accident Modeling for Highway Development and Management in Developing Countries. Main Report: Trials in India and Tanzania. Project Report No: PPR095, DFID.

Hauer, E., 2010. On prediction in road safety. Saf. Sci., http://dx.doi.org/10.1016/j.ssci.2010.03.003.

Hauer, E., Council, F.M., Mohammedshah, Y., 2004. Safety models for urban four-lane undivided road segments. TRB 1897, 96–105.

Hauer, E., 2000. The Median and Safety NCHRP 17–19(4). Transportation Research Board.

IBM Corp. Released, 2013. IBM SPSS Statistics for Windows, Version 22.0. IBM Corp., Armonk, NY.

Jacob, A., Anjaneyulu, M.V.L.R., 2013. Development of crash prediction models for two-lane rural highways using regression analysis. Highway Res. J. 6 (1), IRC, New Delhi.

Jacobs, G.D., Aeron-Thomas, A., Astrop, A., 2000. Estimating Global Road Fatalities, TRL Report 445. Transport Research Laboratory Ltd., Crowthome.

Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accid. Anal. Prev. 34 (3), 357–365.

Landge, V.S., Jain, S.S., Parida, M., 2006. Modeling traffic accidents on two lane rural highways under mixed traffic conditions. 87th Annual Meeting of Transportation Research Board CD Rom.
Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using Support Vector Machine models. Accid. Anal. Prev. 40, 1611–1618.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. A 44, 291–305.
Lord, D., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transp. Res. Rec. J. Transp. Res. Board 1717 (1), 102–108.
Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37 (1), 35–46.
Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., Ivan, J.N., vanSchalkwyk, I., Lyon, C., Jonsson, T., 2008. Methodology for Estimating the Safety Performance of Multilane Rural Highways. NCHRP Web-Only Document 126. National Cooperation Highway Research Program, Washington, DC, http://onlinepubs.trb.org/onlinepubs/nchrp/nchrpw126.pdf (accessed 03.06.10).
Lord, D., 2006. Modeling motor vehicle crashes using Poisson-Gamma Models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accid. Anal. Prev. 38, 751–766.
Maher, M., Mountain, L., 2009. The sensitivity of estimates of regression to the mean. Accid. Anal. Prev. 41 (4), 861–868.
Malyshkina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accid. Anal. Prev. 41 (2), 217–226.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.
Miaou, S.P., 1996. Measuring the Goodness of Fit of Accident Prediction Models. Publication FHWA-RD-96-040, FHWA.
Ministry of Road Transport and Highways, India. Sundar Committee Report. (2007).
Ministry of Road Transport and Highways, India. Road Accidents in India, (2013).
Ministry of Surface Transport, Government of India, New Delhi. Evaluation of capacity augmentation projects of National Highways and State Highways. Final Report. (2000).
Mohan, D., Tsimhoni, O., Sivak, M., Flannagan, M.J., 2009. Road safety in India: challenges and opportunities. UMTRI 1.
Mohan, D., 2002. Traffic safety and health in indian cities. J. Transp. Infrastruct. 9 (1).
National Highway Authority of India. (2014). http://www.nhai.org/roadnetwork.htm.
National Informatics Centre (NIC), 2014. https://data.gov.in/catalog/stateut-wise-distribution-accidental-deaths-natural-causes. In: State-Wise Distribution of Accidental Deaths by Unnatural Causes During. Department of Electronics & Information Technology, Ministry of Communications & Information Technology, Government of India https://data.gov.in/catalog/stateut-wise-distribution-accidental-deaths-natural-causes.
Persaud, B.N., 2001. Statistical Methods in Highway Safety Analysis: A Synthesis of Highway Practice. NCHRP Report 295. Transportation Research Board, Washington, D.C (nchrp_syn_295).
Quimby, A., Maycock, G., Palmer, C., Buttress, S., 1999. The Factor That Influence a Driver's Choice of Speed – A Questionnaire Study, TRL Report 325. Transport Research Laboratory, Crowthorne, Berkshire.
Quinlan, J.R., 1992. Learning with continuous classes. In: Proceedings of Australian Joint Conference on Artificial Intelligence, World Scientific Press: Singapore, pp. 343–348.
Riviere, C., Lauret, P., Ramsamy, J.F.M., Page, Y.A., 2006. Bayesian neural network approach to estimating the energy equivalent speed. Accid. Anal. Prev. 38 (2), 248–259.
Robert, V.R., Veeraragavan, A., 2007. Accident prediction factors for rural highway segments in developing countries Transportation Research Board. 86th Annual Meeting Compendium of Papers CD–ROM.
Rokade, S., Singh, K., Katiyar, S.K., Gupta, S., 2010. "Development of accident prediction model". Int. J. Adv. Eng. Technol. 1 (3), 25–40, E-ISSN 0976-3945.
Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-Injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676.
Sawaldha, Z., Sayed, T., 2003. Statistical issues in traffic accident modeling. In: 82nd Annual Meeting of the Transportation Research Board, TRB.
Shaheem, S., Das Gupta, G.C., 2005. Impact of road development on road safety – a case study of Aluva-Cherthala section of NH-47. J. Indian Roads Congress 66, 615–639.
Shaheem, S., Mohammed, K.M.S., Rajeevan, 2006. Evaluation of cost effectiveness of improvements of accident prone locations on NH-47 in Kerala state. Indian Highways 34 (2006), 35–46.
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.
Sharma, A.K., Landge, V.S., 2012. Pedestrian accident models for rural roads. Int. J. Sci. Adv. Technol. 2 (8), 66–73.
Sharma, A.K., Landge, V.S., 2013. Zero Inflated Negative Binomial for modeling heavy vehicle crash rate on Indian rural highway. Int. J. Adv. Eng. Technol. 5 (2), 292–301, ISSN. 2231-1963.
Sharma, A.K., Landge, V.S., Deshpande, N.V., 2013. Modeling motorcycle accidents on rural highway. Int. J. Chem. Environ. Biol. Sci. 1 (2), 313–317.
Sharma, A.K., Landge, S., Rao, T.K., 2014. Access density and standard deviation of speed as contributing factors for accident of high speed cars. International Conference on Quality Up-gradation in Engineering Science and Technology. (ICQUEST-2014), International Journal of Computer Applications (0975-8887).
Shugan, S.M., 2006. Editorial: errors in the variables, unobserved heterogeneity, and other ways of hiding statistical error. Mark. Sci. 25 (3), 203–216.
Sikka, S., 2014. Prediction of road accidents in delhi using back propagation neural network model. Int. J. Comput. Sci. Eng. Technol. 5 (8), ISSN: 2229-3345.
Srinivas, C., Dinu, R.R., Veeraragavan, A., 2007. Application of poisson and negative binomial regression for modeling road accidents under mixed traffic conditions. Transportation Research Board 86th Annual Meeting No. 07-1617.
Taylor, M.C., Baruya, A., Kennedy, J.V., 2002. The Relationship Between Speed and Accidents on Rural Single-carriageway Roads, TRL Report TRL511. Transport Research Laboratory, Crowthorne, Berkshire.
Varghese, M., Mohan, D., 1991. Transportation injuries in rural haryana, north India. In: Proceedings International Conference on Traffic Safety, Macmillan India Ltd., Delhi, pp. 326–329.
Vogt, A., Bared, J.G., 1998. Accident Models for Two-lane Rural Roads: Segments and Intersections. Publication No. FHWA-RD-98-133.
Wang, Y., Witten, I.H., 1997. Induction of model trees for predicting continuous classes. In: Proc European Conference on Machine Learning Prague, Czech Republic, pp. 128–137.
Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2010. Statistical and Econometric Methods for Transportation Data Analysis, second edition. Chapman Hall/CRC, Boca Raton, FL.
Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco.
Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian Neural Network models: an empirical analysis. Accid. Anal. Prev. 39, 922–933.
Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Anal. Methods Accid. Res. 1, 39–52.