Tree model based-predictive modelling of road accidents on different road stretches of Haryana.

Dr. Baswaraj Patel[a]

[a] Assistant Professor, Department of Civil Engineering, DCRUST, Murthal, Sonepat, Haryana.

Abstract

By creating pedestrian accident frequency prediction models, the study aims to identify several factors that influence the frequency of pedestrian accidents on Haryana's non-urban road lengths. Comparative performance of different pedestrian accident predictive models is also discussed in this study. To get pedestrian accident frequency, the different types of accidents was aggregated per year for all the road sections under study. The resulted data contained a total of 268 distinct samples out of which 178 samples were utilized for training and rest of 90 samples were adapted for validation of prediction model. As per available data and literature review, 14 input significant variables namely average daily traffic (ADT), carriageway width (CW), section length (SL), number of horizontal/ vertical curves (HVC), number of Minor access points (MA), 98th percentile Speed (98ps), median openings (MO), Shoulder width (SW), length of service road on the highway section(LSR), median width (MW), and number of bridges and culverts (BC) were used for prediction of pedestrian accident frequencies (A). Different approaches have been suggested to validate and compare the performances of used models. In case of accident frequency prediction, three statistical measures viz. CC, RMSE and MAE were used. Keeping in view the robust performance of non-parametric models in pedestrian accident prediction as reported in literature, M5 pruned model and REP model were used. REP regression model was developed using IBM SPSS software whereas WEKA software was used for M5 model.

Introduction

Pedestrians have become more vulnerable to traffic accidents as vehicular traffic continues to expand on roadways around the world, particularly in developing nations like India where traffic restrictions are not enforced efficiently [1]. In India, real-time accident data is often unavailable; making it much more difficult to pinpoint the actual causes of accidents. It is difficult to develop relationship among various causal factors for the purpose of pedestrian road safety. Different factors were examined, such as vehicle speed, motorized vehicle type, road environmental and pedestrian characteristics. Accident Prediction Models (APMs) have been utilized as effective tools for road accidents frequency [2]. The factors chosen for modelling in the past have comprised of road and traffic characteristics including geometric design factors and road environmental factors [3].

Literature Review Poisson Model

An accident prediction model (APM) is a mathematical expression that describes the relation between accident frequency or accident severity and variables (road length, medians, traffic, speed etc.). The parameter of the final model can vary for different types of roads and countries due to differences in road features, road user behavior, type of vehicles and road environment [4, 5].

Generalized linear model (GLM) is the most popular approach for the development of accident prediction models [6, 7]. Poisson regression model, a family of GLMs, is broadly used in accident prediction modeling due to count and discrete properties of accident data. An accident is rare and random incident and accident frequency has real and discrete value. Due to distinct non-negative value of accident frequency, the probability of pedestrian accident can be better characterized by Poisson distribution. In the present study, Poisson distribution is used for developing accident frequency mode analysing the dataset collected, as mean and variance are approximately equal for the dataset collected. Poisson regression model assumes dependency of variable h_i and the different crashes counting in i^{th} highway segments for a fixed time, followed by Poisson distribution having the parameter μ_i which is projected accident frequency for i^{th} highway section during a time period.

$$Q(h_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \qquad (1)$$

here, $Q(h_i)$ is the probability of h accidents occurring at i^{th} highway section in a fixed time. The predictable frequency of accidents presumed as a descriptive variables function,

$$\mu_i = \exp(m_o + m_1{*}X_1 + m_2{*}X_2 + \ldots\ldots + m_p{*}X_p) \qquad (2)$$

Where, $X_1, X_2, \ldots\ldots\ldots, X_p$ are the descriptive variables that comprises of traffic and road characteristics at i^{th} section of highway, coefficients for different explanatory variables in the prediction model are m_0, m_1, m_2, ……., m_p of the model, which are calculated using maximum likelihood methods. Model coefficients are the predicted regression coefficients for the accident prediction model.

Fixed effect Poisson model (FEP model) does not consider location-specific effects over time for collected accident data for number of years. Therefore, random effect Poisson model (REP) was selected in the study to address random location and time effects in any location group. Each accident observation for year *t* for *i*th location group was considered as independent observation producing N x T independent observations where, N = number of location groups and T = number of years for which data was recorded. Mannering et al. (1998) [8] and Quddus (2003) [9] mentioned that with random location specific and temporal effects in any location group, the observation dataset must be modelled considering penal matrix having 'N' number of location groups and 'T' number of years. For REP model statistical software IBM SPSS was used. The model having minimum values of Akaike information criteria, Bayesian information criteria and log likelihood was selected as final model in the study.

Study area and data collection

The descriptive statistics of these variables used for frequency and severity modeling are mentioned in Table 1. Model Variables

Table 1: For accident frequency and severity modelling

Sr. No.	**Variable**	**Designation**	**Minimum**	**Maximum**	**Mean**	**Standard. Deviation**
1	(Dependent variable) Accident Frequency	A	1	16	6.46	2.65
2	Average Daily Traffic (1000 PCU/day)	ADT	6.39	93.75	32.66	26.10
3	Section length (km)	SL	0.40	13.20	4.89	3.31
4	Carriageway width (m)	CW	5.5	28.0	12.21	7.14
5	Shoulder width (m)	SW	0	6.0	2.39	1.43
6	Median width (m)	MW	0	8.0	1.60	2.16
7	Number of minor access	MA	0	29	10.24	6.79
8	Number of hor./ Vertical curves	HVC	0	12	4.76	4.75
9	Number of Median openings	MO	0	15	3.13	4.34
10	Length of service road (km)	SR	0	18.20	1.84	4.17
11	Percentage of trucks in ADT	TP	11.00	64.08	37.88	15.10
12	Percentage of cars in ADT	CP	10.45	65.00	33.05	13.57
13	Driveways and commercial units (numbers)	DW	0	15	5.09	3.48
14	98th percentile speed (KMPH)	98ps	61	109	81.42	11.87
15	Bridge and Culverts (numbers)	BC	0	6	2.14	1.95

Prediction of pedestrian accident frequency

The dataset used in the study was gathered from different road sections of seven highways passing from Sonipat, Haryana (India). The road geometry and environment data was gathered from state agencies and through field visits. This comprised of dimensions of carriageway width, width of median, paved shoulder width, and length of service road along the section. Other measurements included number of narrow bridges, median openings, minor access roads, drive ways and commercial establishments in a road section. Various government entities provided traffic volume statistics for various roadways. For road sections where traffic data was not available, it was collected by 12 hour traffic count. The spot speed data was also obtained on various road sections under study. The data collected to model accident frequency consisted of total of 1875 accidents. The non-parametric model (M5 Tree) in accident prediction as reported in the literature was used along with REP model.

The dataset used for modelling accident frequency is summarised in Table 1 and 2. The research compares the outcomes of several models used in this study. In this research, out of 268 datasets, a total of 178 datasets were used for creating model (training dataset) and rest of the datasets were used for testing the developed prediction model. All the prediction models are evaluated using three statistical measures namely, correlation coefficient, Root mean square error and mean absolute error using test dataset.

Results of random effect Poisson (REP) model

In REP model, for temporal variability data was collected across the years, whereas for spatial variability data for different sections from seven highways was used [8]. REP model in SPSS was developed by considering highway and Road section ID as subject measures and year as repeated measure. Only those explanatory variables which improved the χ^2statistics and whose coefficients were found significant at 95% confidence level were included in the final prediction model [10].

Table 2: Parameter estimates of REP model

Sr. No.	Model characteristics	REP Model	
1	2	3	4
	Parameter estimates	Coeff.	Exp (Coeff.)
1	(Intercept)	-2.969 (0.943)*	0.051
2	In ADT	0.255 (0.098)	1.290
3	CW	0.044 (0.011)	1.045
4	SW	-0.169 (0.041)	0.845
5	SR	0.025 (0.010)	1.025
6	TP	0.012 (0.004)	1.012
7	DW	0.021 (0.007)	1.021
8	MW	0.084 (0.026)	1.088
9	98th percentile speed	0.028 (0.006)	1.028

*Values in parenthesis are estimates of standard error.

Table 2 provides both the coefficient estimates (column 3) and the exponential coefficients (column 4) of the Poisson regression. The exponential coefficient for carriageway width is 1.045 in Table 2, column 4, means that the number of pedestrian accidents will be 1.045 times greater (i.e. 4.5 % higher) for each extra 1 m width added.

The log of ADT is statistically significant and directly correlated with pedestrian accident frequency concluded from positive coefficient in Table 2. It can also be concluded from exponential coefficient in Table 2, column 4 that keeping all other effects constant, every 1% increment in the log of ADT may raise pedestrian accident frequency by 29%.

Numerous studies testified that the likelihood of pedestrian accidents have positive correlation with traffic [11-13].

Table 2 shows that median width is associated with pedestrian accident frequency. It can also be implied from Table 2, column 4 that 1 m increment in the median width may raise pedestrian accident frequency by 8.8%. These results were found similar to the outcomes of Baruya (1998) [14]; Reurings and Janssen (2006) , Gomes (2013) [15, 16]; Venkataraman et al. (2011) [17] and Abdel-Aty and Radwan (2000) [18] about contribution of median in accident frequency. The results also conclude that carriageway width is positively correlated with the pedestrian accident frequency. The partial effect for carriageway width with all other variables being equal as concluded from Table 2, column 4 that 1 metre increase in the carriageway width increases accidents involving pedestrians by 4.5 %. Various researches conducted in the past had also reported positive correlation between carriageways width and pedestrian road accidents [19]; [20]. Hauer et al. (2004) [21] established that wider carriageway in the developed countries has a higher chance of pedestrian road accidents.

Shoulder width was negatively linked with pedestrian accident frequency. It can imply from Table 2, column 4 that 1 m increment in the shoulder width reduced pedestrian accident frequency by 15.5 %. It was concluded in the study that truck percentage was positively linked with pedestrian accidents on roads. The effect was relatively high when truck proportion in the traffic was above 40 - 45% .The study shows found positively associated with pedestrian accidents. It can be implied from the present study that 1 unit increment in the speed may increase pedestrian accident frequency by 2.8 %. Similar observations have also been reported that for every 1 mph increase in the speed, accidents were increased by 2 to 7 percent (TRL, 2000). This result was also in accordance with the studies reported by Quimbly et al. (1999) [22] and Taylor et al. (2002) [23].

Results and Discussion of M5 pruned tree model

Figure 1 shows M5 pruned tree results and the accident prediction equations at end nodes are provided in Figure 3. Although, the prediction of pedestrian accident frequencies by FEP/ REP regression model are better, the advantages of M5 model tree are (i) there is no requirement of predefined underlying relationships between accidents and input variables and (ii) the availability of simple linear equation at nodes i.e. LM1, LM2……, LM 9 (Figure 2), which can easily be utilised for prediction of accident frequencies with in the given range of training dataset use to create the model.

```
CW <= 17.5 :
|  SW <= 2.25 : LM1
|  SW >  2.25 :
|  |  DW <= 13.5 : LM2
|  |  DW >  13.5 :
|  |  |  MA <= 13.5 : LM3
|  |  |  MA >  13.5 : LM4
CW >  17.5 :
|  Ln SL <= 2.128 :
|  |  ln ADT <= 11.445 : LM5
| |  ln ADT >  11.445 :
| | |  BC <= 1 :
| | | |  MW <= 3.875 : LM6
| | | |  MW >  3.875 : LM7
| | |  BC >  1 : LM8
|  Ln SL >  2.128 : LM9
```

Figure 1: M5 pruned model tree with train Data set

LM num: 1

A/yr = 0.4503 * Ln SL + 0.174 * CW - 0.7408 * SW - 0.1219 * HVC - 0.0824 * MO + 0.019 * CP + 0.1667 * DW - 0.1834 * BC + 0.0927 * 98ps - 0.8636

LM num: 2

A/yr = 0.474 * ln ADT + 0.2629 * Ln SL + 0.0248 * CW - 0.2193 * SW - 0.093 * MW - 0.0404 * MO + 0.0426 * TP + 0.0464 * CP + 0.0613 * DW - 0.3523 * BC + 0.0865 * 98ps - 10.2441

LM num: 3

A/yr = 0.4148 * ln ADT + 0.244 * Ln SL + 0.0248 * CW - 0.2193 * SW - 0.04 * MW - 0.0404 * MO + 0.0373 * TP + 0.0406 * CP + 0.0613 * DW - 0.3185 * BC + 0.0865 * 98ps - 8.7569

LM num: 4

A/yr = 0.4148 * ln ADT + 0.244 * Ln SL + 0.0248 * CW - 0.2193 * SW - 0.0451 * MW - 0.0404 * MO + 0.0373 * TP + 0.0406 * CP + 0.0613 * DW - 0.3185 * BC + 0.0865 * 98ps - 8.6628

LM num: 5

A/yr = 0.4148 * ln ADT + 0.244 * Ln SL + 0.0248 * CW - 0.2193 * SW - 0.093 * MW - 0.0404 * MO + 0.0373 * TP + 0.0406 * CP + 0.0613 * DW - 0.3185 * BC + 0.0865 * 98ps - 8.7156

LM num: 6

A/yr = 0.1124 * Ln SL + 0.0248 * CW - 0.2193 * SW - 0.1269 * MW - 0.0404 * MO + 0.1275 * DW - 0.0823 * BC + 0.1094 * 98ps - 2.6111

LM num: 7

A = 0.0908 * ln ADT+ 0.1034 * ln SL - 0.9838 * SW + 0.2354 * MW+ 0.1679 * MA + 2.7487 * SR- 0.2468 * TP - 0.0816 * CP + 0.5266 * DW+ 0.0565 * 98ps + 10.2665

LM num: 8

A = 0.0935 * ln ADT + 0.1034 * ln SL - 0.9838 * SW+ 0.2354 * MW+ 0.1679 * MA + 2.7487 * SR - 0.2468 * TP - 0.0816 * CP + 0.5505 * DW + 0.0565 * 98ps + 10.5222

LM num: 9

A =0.0949 * ln ADT + 0.1034 * ln SL - 0.9838 * SW + 0.2354 * MW+ 0.1679 * MA + 3.0539 * SR - 0.2687 * TP - 0.0816 * CP + 0.4959 * DW + 0.0565 * 98ps + 16.2309

Number of Rules : 9

Figure 2: Prediction equations developed by M5 pruned model tree

Figures 1 and 2 demonstrate that M5 model tree evaluated carriageway width, shoulder width, driveways and commercial units, section length, number of minor accesses, median width, number of narrow bridges and culverts, traffic volume and number of minor accesses as significant factors out of the 14 explanatory factors.

As carriageway width is the most important variable in the M5 model tree, it is recommended as the primary partition of the data. After carriageway width, the splitting criterion suggests shoulder width (when CW <= 17.5) and ln SL (when CW > 17.5) are the two most important variables in M5 model tree as shown in Figure 2. The interpretation of results of M5 model tree is quite simple. All the linear models in Figure 3 are appropriate for a set of conditions defined by corresponding model (LM1, LM2,….., LMn). For instance, when carriageway width on a road section under study is less than 17.5 m, shoulder width is below 2.25, pedestrian accident frequency will be ascertained by LM1. The major findings from Figures 1 and 2 are summarised in the following paragraphs.

The results in M5 pruned tree based regression are relatively simple to interpret as compared to the REP model approach. Though the unobserved heterogeneity is accounted for in the REP model, the direction of the effect of an explanatory variable is fixed. On the other hand, M5 pruned model allows change of sign and value of the coefficient across the observed values. Therefore, M5 pruned model provides a better understanding of the effect of independent variables on pedestrian accident frequency.

The variables such as minor access, length of service road, speed, traffic volume on the road, length of section and shoulder width were found to be significant variables in both REP and M5 pruned tree models, whereas speed was found significant only in REP model.

The M5 pruned model tree provides not only theoretical but also application advantages over FEP/ REP model. It does not have any prerequisite requirement like predetermined functional form of the model, which is related to the effect of causal factors on pedestrian accidents. It is as such obvious that in the case with highway safety problems, the errors in coefficient estimates by FEP model may be underestimated leading to erroneous interpretation of significance of the explanatory variables. Although REP model accounts for unobserved heterogeneity and estimates the variance of random effects but it requires full data set of all sections and years. The estimation process is quite cumbersome and the results may not be transferrable to other dataset. In case of M5 pruned model tree, the presence of correlation between explanatory variables is not of much concern. Further, presence of outliers can adversely affect the estimation of accident frequencies by FEP/REP regression model, and need to be identified and deleted from the data in advance, whereas in case of M5 pruned model tree regression, outliers are isolated into a node and do not contribute in splitting.

In M5 pruned models results are graphically displayed in easily understandable trees as shown in Figure 2. These model tree results may be turned into a series of "if-then" rules. These principles may be used to trace a route down the tree to a terminal node, where a simple linear equation (LM 1, LM 2, ----, LM n) can be found to predict pedestrian accidents.

Comparative statistics

Table 3 provides the comparison of three statistical measures i.e. CC, RMSE and MAE values for M5 and REP regression models to predict accident frequencies. Results from Table 3 indicate that for the used dataset REP model best predicts the accident frequencies as compared to M5 pruned tree model.

Table 3: Comparative table of various performance indicators of different models

Model	CC	RMSE value	MAE value
M5 model tree	0.7528	1.7314	1.3734
REP model	0.914	1.572	0.835

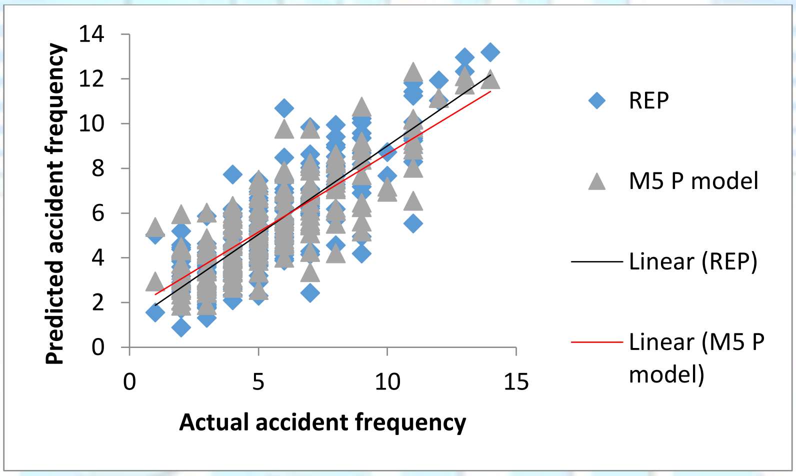

Figure 3: Actual vs predicted accident by different modelling approaches for test dataset.

Conclusions of the study

The M5 model tree and REP model pedestrian accident frequency predict for test data were compared and displayed against the actual accident frequency values. The figure indicates that the scatter plot of REP model is closer to the best fit line as compared to those of M5 pruned. The R^2 values for M5 pruned model and REP model scatter plots were found to be 0.566 and 0.835 respectively, therefore, recommending a better fit by REP model for test data in comparison to M5 pruned . REP model performance was found to be the best.

The variables median width, traffic volume, carriageway width, speed, driveways, length of service road, truck percentage in the traffic and shoulder width were input variables affecting accident frequencies for random effect Poisson model. Accidents were increased when median width, length of service road, speed, traffic volume, road width, driveways and truck percentage in the traffic increased. Result shows with increase in shoulder width accident frequency were found decreasing. By keeping all other effects constant REP model shows, 15.5 percent accidents reduce by small increase of one meter shoulder width. The increase of speed by 1 kmph increases accident frequency by 2.8 percent. By increase of 1 km in service road length accidents increase by 2.5 percent. It was also found that one unit increase in road width, drive way; median width and percentage of trucks in the traffic will lead to increase the pedestrian accident by 4.5%, 2.1%, 8.8% and 1.2% respectively.

References

1. Peden, M., et al., *World report on road traffic injury prevention*. 2004: World Health Organization.
2. La Torre, F., et al., *Development of an accident prediction model for Italian freeways.* Accident Analysis & Prevention, 2019. **124**: p. 1-11.
3. Caliendo, C., M. Guida, and A. Parisi, *A crash-prediction model for multilane roads.* Accident Analysis & Prevention, 2007. **39**(4): p. 657-670.
4. Eenink, R., *Accident prediction models and road safety impact assessment.* Ripcord Iserest, 2008.
5. Eenink, R., et al., *Accident prediction models and road safety impact assessment: recommendations for using these tools.* Institute for Road Safety Research, Leidschendam, 2008.
6. Chikkakrishna, N.K., M. Parida, and S.S. Jain. *Crash prediction for multilane highway stretch in India*. in *Proceedings of the Eastern Asia Society for Transportation Studies*. 2013.
7. Zou, Y., Y. Zhang, and D. Lord, *Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models.* Analytic methods in accident research, 2014. **1**: p. 39-52.
8. Shankar, V., et al., *Evaluating median crossover likelihoods with clustered accident counts: An empirical inquiry using the random effects negative binomial model.* Transportation Research Record: Journal of the Transportation Research Board, 1998(1635): p. 44-48.
9. Chin, H.C. and M.A. Quddus, *Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections.* Accident Analysis & Prevention, 2003. **35**(2): p. 253-259.
10. Sawalha, Z. and T. Sayed. *Statistical issues in traffic accident modeling*. in *82 nd Annual Meeting of the Transportation Research Board*. 2003.
11. Elvik, R., *Assessing causality in multivariate accident models.* Accident Analysis & Prevention, 2011. **43**(1): p. 253-264.
12. Ukkusuri, S., et al., *The role of built environment on pedestrian crash frequency.* Safety science, 2012. **50**(4): p. 1141-1151.
13. Leden, L., *Pedestrian risk decrease with pedestrian flow. A case study based on data from signalized intersections in Hamilton, Ontario.* Accident Analysis & Prevention, 2002. **34**(4): p. 457-464.

14. Baruya, A., *MASTER: Speed accident relationship on single carriageway roads of UK.* 1998.
15. Reurings, M. and T. Janssen, *Accident prediction models for urban and rural carriageways*. 2006, R-2006-14, SWOV.
16. Gomes, S.V., *The influence of the infrastructure characteristics in urban road accidents occurrence.* Accident Analysis & Prevention, 2013. **60**: p. 289-297.
17. Venkataraman, N.S., et al., *Model of relationship between interstate crash occurrence and geometrics: exploratory insights from random parameter negative binomial approach.* Transportation Research Record, 2011. **2236**(1): p. 41-48.
18. Abdel-Aty, M.A. and A.E. Radwan, *Modeling traffic accident occurrence and involvement.* Accident Analysis & Prevention, 2000. **32**(5): p. 633-642.
19. Choueiri, E.M., et al., *Safety aspects of individual design elements and their interactions on two-lane highways: international perspective.* Transportation Research Record, 1994(1445).
20. Zajac, S.S. and J.N. Ivan, *Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural Connecticut.* Accident Analysis & Prevention, 2003. **35**(3): p. 369-379.
21. Hauer, E., F. Council, and Y. Mohammedshah, *Safety models for urban four-lane undivided road segments.* Transportation Research Record: Journal of the Transportation Research Board, 2004(1897): p. 96-105.
22. Quimby, A., et al., *The Factors the Influence a Driver's Choice of Speed: A Questionnaire Study*. 1999: Citeseer.
23. Taylor, M.C., A. Baruya, and J.V. Kennedy, *The relationship between speed and accidents on rural single-carriageway roads*. Vol. 511. 2002: TRL.