ZERO INFLATED NEGATIVE BINOMIAL FOR MODELING HEAVY VEHICLE CRASH RATE ON INDIAN RURAL HIGHWAY

A.K.Sharma[1], V.S.Landge[2]

[1]Department of Civil Engineering, RCOEM, Nagpur, India
[2] Department of Civil Engineering, VNIT, Nagpur, India

ABSTRACT

Poisson regression and negative binomial regression have been widely used to model the road crashes and to predict crash frequency. Zero inflated models have been shown to be a powerful tool to predict crash frequency when crash data are characterized by preponderance of zero. This paper presents the research work aiming to correlate the road traffic crash rate with road geometry and traffic characteristics for crashes involving heavy vehicles on national highway number 6(NH-6), one of the busy rural roads in central India. Stochastic regression models were developed using crash data collected during 2005-09 over a stretch of 100 km of road length. Zero Inflated Negative Binomial (ZINB) regression method has been used to model the occurrence of road traffic crashes. The Akaike Information Criterion (AIC) has been used to measure the relative goodness of fit. The independent variables selected in this study were shoulder width (SW), lane width (LW), access density (AD), spot speed (SS) and annual average daily traffic (AADT).The model demonstrates that access density, lane width and shoulder width are important parameters affecting the traffic safety of the selected highway.

KEYWORDS: *Access density, Crash rate, Lane width, Shoulder width, Zero inflated.*

I. INTRODUCTION

Heavy commercial vehicles are a very common sight on Indian Rural Highways. They are the integral part of transportation of passenger and freight throughout India. Statistics indicates that trucks, buses and other articulated vehicles accounted for the highest share in total road accidents (31.3%) as well as fatal accidents (39.7%) and persons killed(39%) in the year 2009[1]. Table 1 shows the total traffic accidents and percentage share of different vehicles.

Traffic accidents have significant impact on the society. It results in enormous costs in terms of lost productivity and property damage. Efforts are required to have better understanding of the factors that influence accident. The knowledge about the relationship between the accident and the factors responsible is incomplete. The causes of road accidents are always complicated due to presence of many factors such as, roadway geometric design, traffic characteristics, human factor, weather condition etc. Most of the studies so far have focused on some risk factors such as drinking and driving, restraint systems and tried to determine their relationship with accident rate. Previous research has shown that accidents involving trucks have a likelihood of producing a severe injury or fatality. However the relative impacts of various factors (Roadway geometry, traffic characteristics, etc.) have not been quantified. The objective of this research is to develop accident prediction model (APM) for heavy vehicle accidents and quantify the factors responsible for the same.

The paper begins with statement of objectives and methodological approach adopted for the study. This follows a brief description of study area and scope, data collection and analysis, variables used in the model , modeling techniques and model selection criterion. The paper concludes with results and discussion and direction for future research.

Table 1. Total Traffic Accidents and Percentage Share of Different Vehicles

	2-Wheelers	Auto-rickshaw	Cars	Trucks, buses and other articulated vehicles	Other Motor Vehicles	Other Vehicles/ Objects
Accidents	22.4	6.9	20.6	31.3	10.9	7.9
Fatal Accidents	17.8	4.3	17.1	39.7	11.4	9.7
Persons Killed	15.7	4	17.5	39	13.8	10
Persons Injured	20.2	7.7	20.3	32.8	11.4	7.6

II. OBJECTIVE

The specific objectives of studies are:

- Development of Correlation between road accidents, volume of heavy vehicles, geometric design parameters of highway along with traffic operating characteristics.
- Evolving engineering remedial measures for improving safety on the selected stretch.
- Practical recommendations for improving traffic safety on the said highway.

III. METHODOLOGY

Methodology adopted for the study is as specified below:

- Identification of study area and scope
- Crash data collection from the law enforcement agency and insurance companies.
- Road geometric and Traffic parameter data from field studies
- Selection of variables and modeling methods
- Development of accident prediction models
- Testing of models, interpretation of results and remedial measures suggestions.

IV. STUDY AREA AND SCOPE

The study area chosen is National Highway No.6 commonly refer to as NH-6 or G.E. Road (Great Eastern Road),which is a one of the most busy national highways in India. It's a connecting corridor to major states of India namely Gujarat, Maharashtra, Chhatisgarh, Orissa, Jharkhand and West Bengal. The Highway passes through the cities of Surat, Dhule, Nagpur, Raipur, Sambalpur, Kolkata. NH-6 eventually will be one of the important links of Asian Highway network (AH-46).

The scope of present study is limited to road section passing through central Indian state of Maharashtra. Maharashtra, one of the most advanced states of India has high traffic accident rate. Most of these accidents occur on the National highways. The highway under consideration has very high rate of accidents. Many of these accidents are fatal and involvements of heavy vehicles in such accidents are in large proportion. For the present study the geometric parameters like, Shoulder width (SW), Lane Width (LW) and access density (AD) and traffic parameters like, Spot Speed(SS), Annual Average Daily Traffic Volume (AADT) are taken into consideration.

V. DATA COLLECTION AND ANALYSIS

Accident data was collected from the police stations and insurance companies. Road geometry and traffic data was collected through field studies and traffic count survey. For the purpose of collecting road geometry and traffic data, the road was divided into segments of similar characteristics ranging from 0.2 to 0.6km length for curve portion and 1.0 to 2.2km for straight portion. This study deals with the data collected for straight portions.

National Highway No. 6 experiences the crash rate as high as 1.62 accidents per year per km. It has a very high rate of fatality 0.38 death /km/year .The highway share heavy vehicles, passenger cars, two wheelers, animal drawn carts, cattle, and pedestrians. Heavy vehicles are involved in 78% of the accidents, passenger cars are involved in 48% of accidents, two wheelers are involved in 62% of accidents and pedestrians are involved in 21% of accidents.

The preliminary analyses of the data are given from fig.1 to fig.5. Fig.1 shows a positive relation between dependent variable and access density at most of the locations. It suggests that more access points to the main road will give rise to more traffic accidents. Fig.2 which gives relation of AADT with the dependent variable suggests that accident rates are more when AADT variation is from 8000 to 11500. Fig.3 gives a relation of, shoulder width with accident rate, which suggest that more deficient shoulders gives rise to more accidents. Similarly relationships of other variables are presented graphically in different figures.

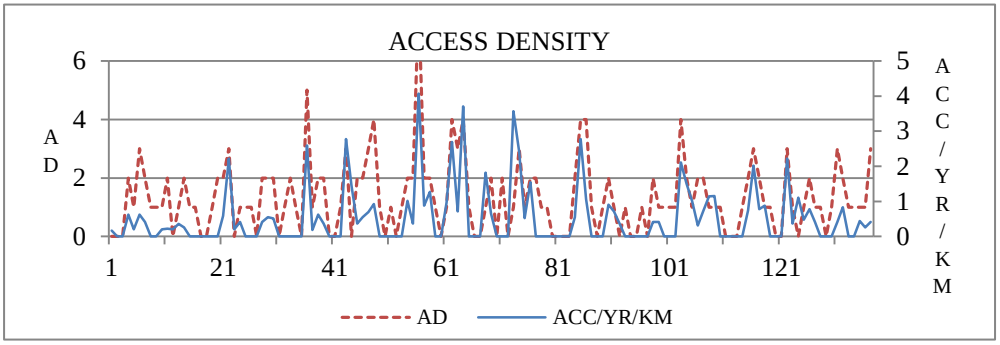

Figure 1. Access Density vs crasht Rate

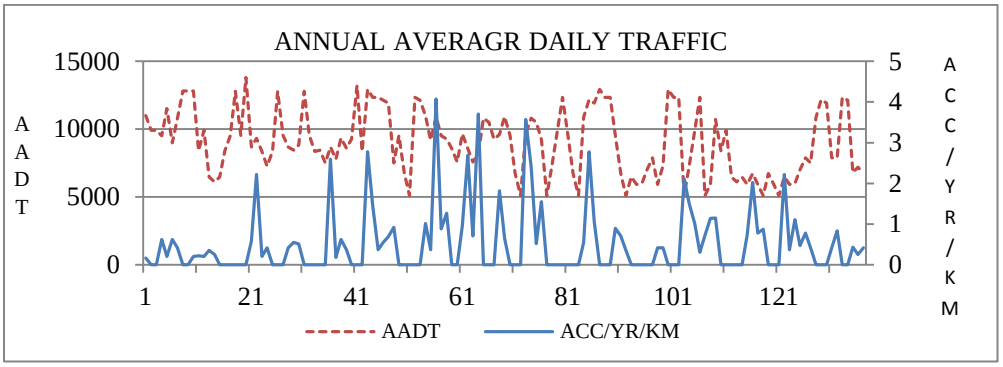

Figure 2. AADT vs Crash Rate

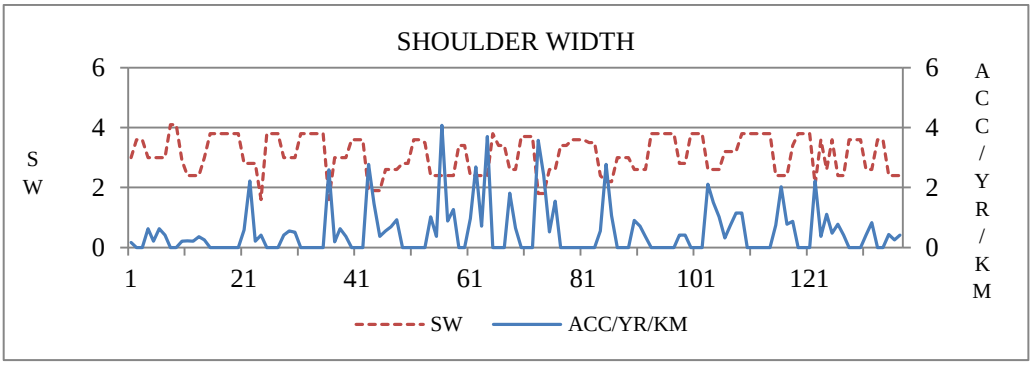

Figure 3. Shoulder Width vs Crash Rate

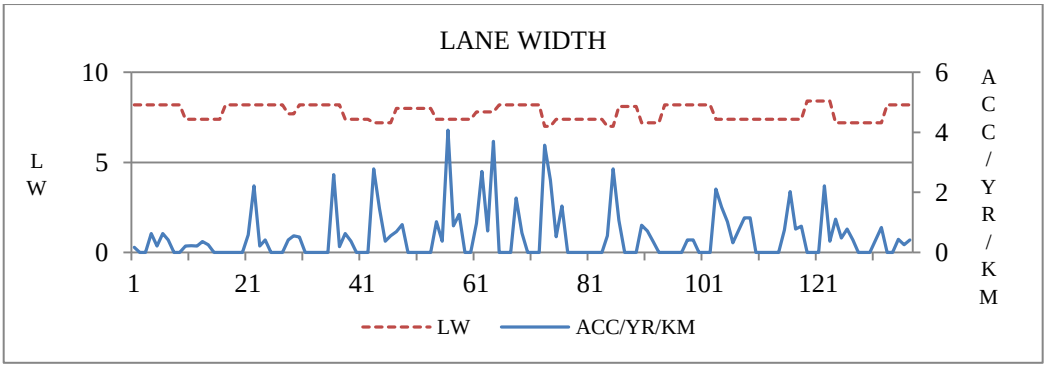

Figure 4. Lane Width vs Crash Rate

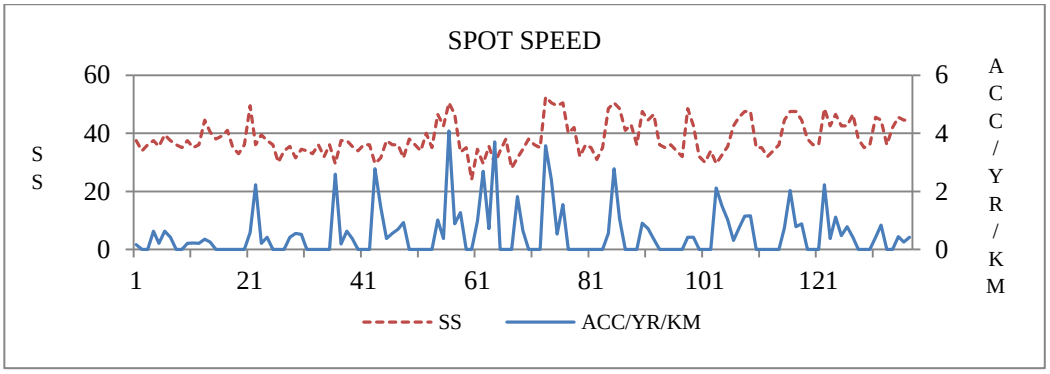

Figure 5. Spot speed vs Crash Rate

VI. VARIABLES IN THE MODEL

After preliminary analysis of the data, following parameters were selected as independent variables.

i. Shoulder width

Shoulder provides an area along the highway for vehicle to stop, particularly during emergency. Slow moving vehicles, pedestrians can use the shoulder and keep the carriageway free for heavy and fast moving vehicles. A report by Zegeer et al. [2] (1981) on the "Effect of lane and shoulder widths on accident reduction on rural two-lane roads" indicated that a paved shoulder widening of 2 feet per side reduces accidents by 16%. Shoulder width has been a parameter with significant influence on safe operations of traffic and hence selected as a variable. The area under consideration shows a wide variation in the shoulder width from 1.2 to 3.10 m.

ii. Lane width

Traffic flow tends to be restricted when lane width reduces. This is because vehicles have to travel closer together in lateral direction. Lane width, therefore, is treated as an important parameter. It has been found that accident rate reduces as lane width increases.

iii. Access density

In urban and suburban areas, the rapid growth of the local economy has steadily increased the demand for access points along multilane highways. The availability of access is necessary to commercial or residential developments, usually at the expense of traffic operations and the safety of local highway systems. To achieve a good coordination of these two aspects, compromises are often required to be made between accessibility and mobility or capacity and safety. On the study area access density ranging from 0 to as much as 8 per km of road length is observed. A close look at the accident pattern occurring near the access points and segments of high access density, point to involvement of this variable. In the view of above access density is an important variable governing safety on the said highway and has been selected as one of the independent variable.

iv. Traffic volume

Traffic volume is believed to have considerable impact on the crash rate [3][4]. For this study annual average daily traffic (AADT) is used as a parameter to indicate traffic volume. Traffic volume, converted in to passenger car unit (pcu), for various segments was collected and included in the model.

v. Spot speed

Speed and travel time are the most common indicator of the performance of a traffic facility. Spot speed is one of the major parameter that is used as an indicator of traffic performance. Spot speed of a location has considerable impact on the traffic safety of the area. The data collected shows a wide variation in the spot speed from 25kmph to 60kmph.

6.1 VARIABLE SELECTION

Various combinations of the variables selected were tried and the best combinations are given in Table 2. Crash rate per year per km was chosen as dependent variable. Crash rate is defined as

$$CRi = TAi / Li / NY \quad (1)$$

Where , CR_i = Crash Rate on segment i
TA_i = Total accidents on segment i
L_i = Length in Km. of segment i
NY = Number of Years

Table 2. Variables in the Model

Model No.	Independent Variables	Dependent Variable
1	AD, SW, LW, AADT, SS	Accidents/Yr/Km(Crash Rate)
2	AD, SW, LW ,SS	
3	AD, SW, LW	
4	SW ,LW ,SS	

VII. MODELING OPTIONS AVAILABLE

Various modeling techniques have been tried to model accidents aiming for accuracy. However the suitability of the model depends on the data quality and is location specific. The model building methodology is selected based on the availability of data and accuracy of the data. Various modeling techniques popularly employed are as discussed below:

7.1 DETERMINISTIC MODELS

Deterministic models which are widely used are not considered to be suitable for an arbitrary and sporadic event like traffic crashes. Much of the early work in the empirical analysis of accident data was done with the use of multiple linear regression models. As the literature has repeatedly pointed out, these models suffer from several methodological limitations and practical inconsistencies in the case of accident modeling. To overcome these limitations, researchers turned to stochastic models.

7.2 STOCHASTIC MODELS

Stochastic models are logical alternative for events that occur randomly and independently over time. Unlike deterministic models stochastic models assume accident as random event. Early in 1989; Okamoto et al.[5] suggested that the occurrence of traffic crashes follows stochastic distribution. In 1990, Garber et al. [6] developed several models to describe the occurrence of crashes in using stochastic modeling techniques, like Poisson regression (PR) [7] and negative binomial regression (NBR). Various studies further examined the goodness-of-fit of these regression models. More stochastic models were also proposed other than Poisson regression models and negative binomial regression model, which included Zero Inflated Poisson Regression Model and Zero Inflated Negative

Binomial Regression Models. This paper presents models build using Zero Inflated Negative binomial Regression.

7.3 MODELING METHOD

Crash being non negative, sporadic and discrete, use of deterministic models weakens the analysis. Stochastic modeling methods[8][9] overcome this limitation and are a better option for random events like accidents .Zero Inflated Negative Binomial Model(ZINB) was selected to model accidents. ZINB models have been principally applied when crash data are characterized by a preponderance of zeros.

7.4 ZERO INFLATED NEGATIVE BINOMIAL REGRESSION MODELS

One of the problems that plagues Poisson regression model is over dispersion. Negative binomial regression was developed to be an improvement on the Poisson regression process. The negative binomial regression model allows for over dispersion in the model and can be used to quantify various parameters more effectively. Transportation safety analysts have typically justified the use of Zero Inflated (ZI) models because of the improved statistical fit compared to traditional Poisson and NB models.
Zero Inflated regression models are two regime models. First probability model governs whether a count number is zero or positive number, called as inflated model. Then the positive part of the distribution is described by suitable stochastic distribution, called as base model.
In Miaou's[10] study, the negative binomial regression model was of the form;

$$y_i = 0,1,2 \dots \dots with\ probability\ \frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha*\lambda_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha*\lambda_i}{1+\alpha*\lambda_i}\right)^{y_i} \quad (2)$$

$$\lambda_i = e^{\beta_i x_i} \quad (3)$$

$$where\ x_i\ is\ i^{th}\ covariate\ and\ \beta_i\ is\ the\ regression\ coefficient$$

For Zero Inflated Negative Binomial regression

$$y_i = 0\ with\ probability\ \ p_0 + \left(\frac{1}{1+\alpha*\lambda_i}\right)^{\frac{1}{\alpha}} \quad (4)$$

$$y_i = 1,2 \dots \dots with\ probability\ \ (1-p_0) * \frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha*\lambda_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha*\lambda_i}{1+\alpha*\lambda_i}\right)^{y_i} \quad (5)$$

where p_0 can be represented by probability model incorporating the effects of covariates, such as logit model.

$$p_0 = \frac{e^{r'w_i}}{1+e^{r'w_i}} \quad (6)$$

r' is the coefficient matrix and w_i is the i^{th} covariate
$\Gamma(.)$ is Gamma function; and α is the rate of over dispersion.

7.5 MODEL SELECTION CRITERIA

Maximum likelihood estimation method has been employed widely in estimating Poisson, negative binomial [11] and zero inflated regression models. According to definition of maximum likelihood estimation method the estimated parameters are best when the maximum value of likelihood is obtained. Akaike Information Criterion (AIC) [12] was used to judge the performance of the model. Smaller the AIC value, the better the model.

$$AIC = -2\text{Log L} + 2K \quad (7)$$

Where Log L is the log likelihood; K is the number of estimated parameters.

VIII. RESULTS AND DISCUSSION

The results obtained after analysis of data, using SPSS software, are shown in Table 3 and Table 4. Table 3 shows the parameter estimates for the various variables used in the base model and table 4 shows the parameter estimates for the various variables used in inflate model. The final model is selected on the basis of Akaike Information criteria (AIC).
In order to see the performance of the model the coefficient of variables has to be examined. Models with logical algebraic signs of the variables were selected. Along with the strong statistical tools, proper engineering judgments are required to decide upon the selection of final model. For example, a negative sign of shoulder width suggest that more is the shoulder width less will be the accident and it is logically acceptable.

8.1 FINAL MODEL

Base Model Selected

$$\text{Acc/Yr/Km} = 2.913 + 0.326 * \text{AD} - 0.478 * \text{SW} - 0.173 * \text{LW} \quad (8)$$

$$\lambda = e^{2.913+0.326*\text{AD}-0.478*\text{SW}-0.173*\text{LW}} \quad (9)$$

Inflate Model

$$p_0 = \frac{e^{-7.001-1.475*\text{AD}+2.719*\text{SW}+1.371*\text{LW}-0.285*\text{SS}}}{1+e^{-7.001-1.475*\text{AD}+2.719*\text{SW}+1.371*\text{LW}-0.285*\text{SS}}} \quad (10)$$

$$\text{Predicted Frequency for } (y=0) = \varepsilon * \left[p_0 + \left(\frac{1}{1+\alpha*\lambda_i}\right)^{\frac{1}{\alpha}} \right] \quad (11)$$

$$\text{Predicted Frequency for } (y_i = 1,2 \ldots)$$

$$= \varepsilon * \left[(1-p_0) * \frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)} \left(\frac{1}{1+\alpha*\lambda_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha*\lambda_i}{1+\alpha*\lambda_i}\right)^{y_i} \right] \quad (12)$$

where ε is exposure term = total length of study area in kilometers
α is over dispersion parameter =0.711 (for present data collected)

Table 3. Parameter estimates for base model (Using SPSS)

Variable	Model 1	Model 2	Model 3	Model 4
Intercept	3.074	3.077	2.913	4.159
AD	0.327	0.327	0.326	
SW	-0.485	-0.485	-0.478	-0.844
LW	-0.179	-0.179	-0.173	-0.130
AADT	$3.08*10^{-7}$			
SS	-0.002	-0.002		$-7.22*10^{-5}$
AIC	217.18	214.96	212.86	263.41

Table 4. Parameter estimates for inflate model (Using SPSS)

Inflate Variable	AD	SW	LW	SS	Intercept
Coefficient	-1.475	2.719	1.371	-0.285	-7.001

8.2 MODEL TESTING

The model finally selected is tested for the accident data collected for the year 2010. The testing result is given in Table 5 and Figure 6

Table 5. Predicted and observed Frequency

Accident Frequency	Predicted Frequency	Observed Frequency	R-Square
0	64.32	55	
1	17.38	26	
2	11.40	4	
3	7.00	3	
4	4.25	3	
5	2.51	1	
6	1.44	3	**0.954**
7	0.824	0	
8	0.480	1	
9	0.274	0	
10	0.137	0	

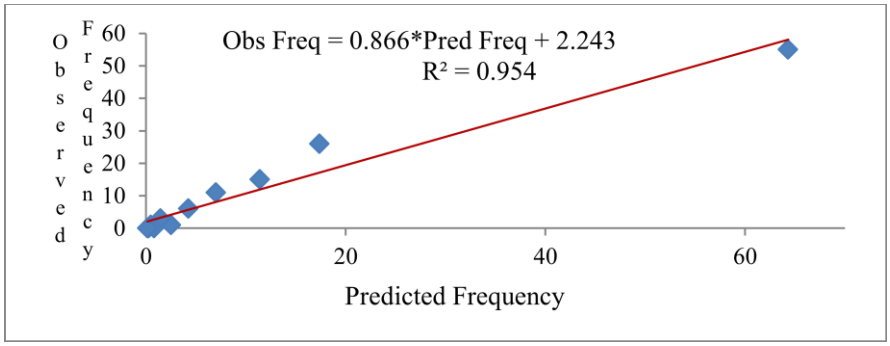

Figure 6. Observed vs Predicted Frequency

IX. CONCLUSIONS AND REMEDIES

Zero inflated negative binomial regression was proposed to establish an empirical relationship between heavy vehicle traffic accidents and highway geometric and traffic parameters. The result of this study can eventually be employed to identify the locations with certain numbers of accident frequencies under different geometric and traffic conditions for national highway number-6 (NH-6) in India. The results obtained here provide valuable insight into the underlying relationship between risk factors and vehicle accidents. The model selected may be used to develop strategies for enhancing safety with optimum use of resources. The effects of various variables can also be quantified from the parameter estimates of the models.

Heavy Vehicle safety is greatly influenced by number of access points per unit length of the road. Each additional access point per kilometer of road length may increase accident rate by 60%.Shoulder width is also a highly influential factor affecting heavy vehicle safety. Increasing shoulder width by 0.25m on either side of the carriageway may reduce the accident rate by 40%. Similarly increase in lane width by 1m, may reduce the accident rate by 30%.

10.1 REMEDIES

- Access points depends on number of villages and land use of the area, so it is not possible to reduce number of access points, but it is suggested that the access point to the main highways should be properly designed with the help of auxiliary lanes so as to have sufficient sight distance and safe entry to main route.
- Shoulder is a recovery area for the driver's error. The minimum width of the shoulder on both sides of carriage way should be at least equal to width of a regular commercial vehicle using the facility, so that they can be parked there in case of emergency without disturbing the traffic flow of the main carriageway.
- Sufficient width of carriageway should be provided, depending on traffic volume. A width of 7.5 m (two lane) is not at all sufficient for two lane undivided facility, considering present situations, a minimum width of 10.5 m (three lanes) is suggested for such road infrastructures.

X. SCOPE FOR FUTURE WORK

The regression method may be used to model traffic accident for different set of crash data collected from different sections of the same or some other highway. The method may be compared with other methods like ANN. Future work might also focus on improving the prediction performance of the ZINB models.

REFERENCES

[1]. Road accident in India (2009)" (Ministry of Road Transport and Highway).
[2]. Zegeer, C. V., Deen, R. C. & Mayes, J. G. (1981) "Effect of lane and shoulder widths on accident reduction on rural two-lane roads". TRR 806 .pp 33-43.
[3]. Karlaftis M.G. & Golias, I.(2002) , "Effect of road geometry and traffic volumes on rural roadway accident rates", Accident Analysis and Prevention vol.34, Issue 3, pp. 357-365.
[4]. Williams Ackaah, Mohammed Salifu (2011), "Crash prediction model for two lane rural highways in Ashanti region of Ghana", International Association of Traffic and Safety Sciences(IATSS) Research, http://dx.doi.org/10.1016/j.iatssr.2011.02.001
[5]. Okamoto (1989), "A Method to cope up with random errors of observed accident rates" Safety literature, Vol.4, page 317-332.
[6]. Garber, N.J., Wu, L(1989), "Stochastic models relating crash probabilities with geometric and corresponding traffic characteristics data", Research report No UVACTS-5-15- 74 Center for transportation studies at the University of Virginia.
[7]. Jovanis, P. & Chang, H(1986)., "Modeling the relationship of accidents to miles traveled", Transportation Research Record 1068, pp 42-51.
[8]. Poach, M. & Mannering, F.(1996), "Negative binomial analysis of intersection accident frequencies" ASCE's Journal of Transportation Engineering Vol.122 No. 2, pp. 105-113.
[9]. Ziad Sawalha & Tarek Sayed (2003) "Statistical Issues In Traffic Accident Modeling" TRB annual meeting (CD-ROM).
[10]. Miaou, S.P.(1994) "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions", Accident Analysis and Prevention vol. 26, issue 4, pp. 471-482
[11]. 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).
[12]. Sharma A.K., Landge V.S.(2012),"Pedestrian Accident Prediction Model For Rural Road", International Journal of Science and Advanced Technology Volume 2, No 8 .

AUTHORS

A. K. Sharma ,born on 28-12-64 ,graduated in civil Engineering from NIT Raipur(Chattisgarh,India) in 1987. He completed his post-graduate in Highway and Traffic Engineering from IIT Kharagpur (India) in 1989. Presently working at Shri Ramdeobaba College of Engineering and Management, Nagpur , as Associate professor in Civil Engineering Department. His main field of work has been pavement and traffic

engineering. He has guided many undergradute and post graduate dissertations. Presently he is persuing his Doctoral research in the field of traffic safety.

V. S. Landge, born on 02-10-68, graduated from RKNEC, Nagpur (India) in 1991. He completed his post graduate studies in Transportation Engineering from BITS Pillani (Rajasthan, India) in 1993. He completed his Doctoral degree from IIT Roorkee (India) in 2006. Presently working at VNIT, Nagpur as Associate professor in Civil Engineering Department. His main field of work has been traffic engineering particularly in the field of traffic Safety. He is a member of the State Technical Agency for Pradhan Mantri Gram Sadak Yojna for Vidarbha Region in India He has guided many undergraduate and post graduate dissertations.