ISSN (Print) : 2319-8613
ISSN (Online) : 0975-4024

Accident data analysis using Statistical methods –A case study of Indian Highway

Rahul Badgujar[1], Priyam Mishra[2], Mayank Chandra[3],
Sayali Sandbhor[4] , Humera Khanum[5]

[1,2,3] Undergraduate scholars, Department of Civil Engineering,
Symbiosis Institute of Technology, SIU, Pune, India.
[4,5] Assistant Professor, Department of Civil Engineering,
Symbiosis Institute of Technology, SIU, Pune, India.
[1]rahul.badgujar@sitpune.edu.in, [2]priyam.mishra@sitpune.edu.in, [3]mayank.chandra@sitpune.edu.in
[4]sayali.sandbhor@sitpune.edu.in, [5] humera.khanum@sitpune.edu.in

***Abstract*-Road accidents are one of the major causes of death worldwide as per Global safety report more than 1.5 million people are killed in road accidents every year throughout the world[5] The probability of occurrence of accident depends on numerous factors like roadway condition, geometrics of road, vehicle, pavement condition and weather condition, etc[10] each factor contributes its own share towards occurrence of accidents and there can be many more factors which are situation specific. To ascertain the effect of various parameters an accident occurrence, data of road accidents for a stretch of 101 Kms of an Indian national highway was collected for past 3 years. The analysis of the data using regression technique enables to predict the occurrence of accidents for a certain situation and similar application can be carried out on any stretch to get regression equation of similar type.**

Keywords: Nature of Accidents, Cause of Accidents, Road Feature, Classification of Accident, Road Safety.

I. INTRODUCTION

In today's world road and transport has become an integral part of growth and development of a nation. Everybody is a road user in one or other shape. The present transport system has minimized the distance but it has on the other hand increased the life risk. Every road crashed result in loss of lakh of lives and serious to injuries to corers of people. India has a total of about 2 million kilometers of roads out of which 960,000 km are surfaced road and about 1 million km of roads in India are of poor quality[11]. Rural areas have unsurfaced roads &urban areas have high severity of congestion. These are some of the major problem face on any Indian roads. This is very serious situation and requires proper attention with the use of some statistical methods. In this study analysis of data is done using regression analysis. In today's world where the number of road commuters is increasing drastically, its demanding more & more safer roads to have accident free roads. A lot of initiative are being taken up by the government to tackle this issue but needs a little more research attention.

A.Introduction to case study

As we experience increase in number of vehicles on road simultaneously road accidents are also increasing in same manner. Road accidents are one of the biggest killers in India. Statistics suggests that one person dies in a road accident in India every four minutes [12]. In spite of these numbers being so high, not much effort is made to make roads safer.

National highway (NH9) is a major East – West highway in India &passes through almost 7 states in present case study analysis data pertaining to accidents on a 101 kms stretch on this highway from Pune to Solapur cities in Maharashtra state of India. Data for 3 years is taken into consideration for analysis. This case study is an effort to make roads safer for road user by using a prediction model developed through regression analysis [6].

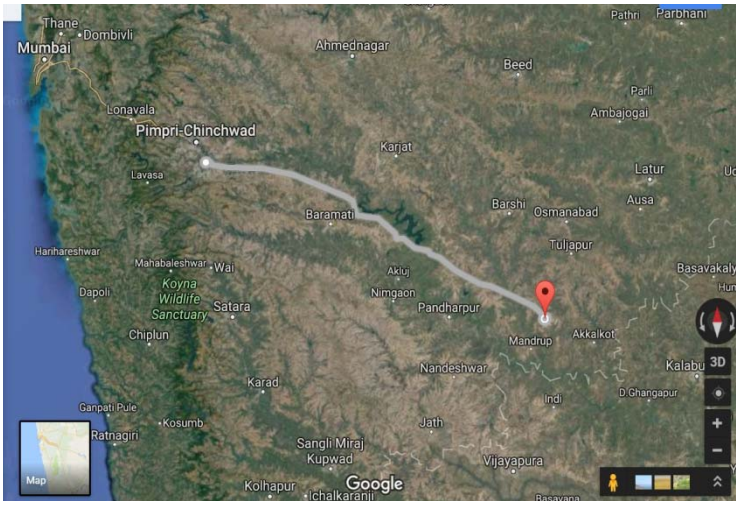

Fig. 1. Study Stretch

II. RESEARCH METHODOLOGY

Methodology for the research work is mainly divided into three parts:-

- Analysis based on location – analysis for identifying stretch points with higher number of accidents.
- Analysis based on time – analysis for identifying the hours having highest number of accidents occurring on stretch.
- Regression analysis – for obtaining prediction equation by using parameters available in accidental data.

I. *Analysis based on location*

For location based analysis stretch is divided into 11 sections for equal length and further divided into right and left lane for analysis. Data is sorted for both lanes and respective accident data for each section is recorded as shown in table 1.

Table 1. Location Wise Distribution of Stretch

Section	Accident Location(KM)
1	150-159
2	160-169
3	170-179
4	180-189
5	190-199
6	200-209
7	210-219
8	220-229
9	230-239
10	240-249
11	250~

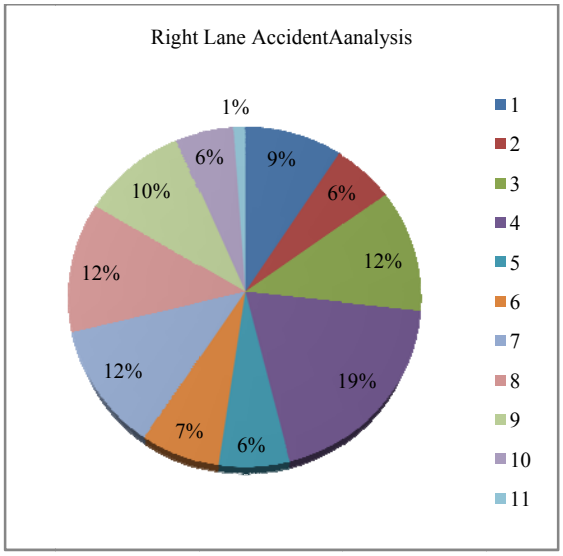

Figure 2: Right Lane analysis

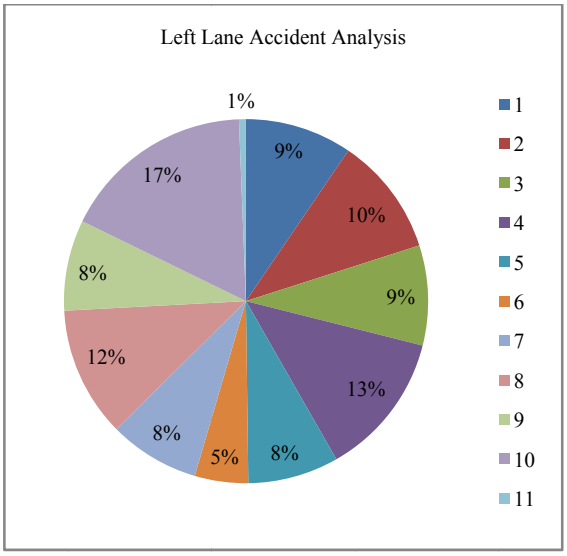

Figure 3: Left Lane analysis

Figure 1 shows % accidents on right lane section wise &Figure 2 shows % accidents on left lane section wise.

II. *ANALYSIS BASED ON TIME*

For analysis based on time the study stretch is divided into 8 time intervals of 3hours each and, corresponding number of accident are found out. As available data is for 3 years,year wise sorting is done for in-depth study of accident data and for understanding the pattern for accident data on basis of time represented as 1T to 8T on the basis of time intervals. [1][7][9]

Table 2. Accident Distribution Based on Time

Section	Time of Accident	No. of Affected Persons	Fatal	Grevious	Minor	Injury
1T	00:00-02:59	102	8	16	54	24
2T	03:00-05:59	52	10	21	64	13
3T	06:00-08:59	56	8	22	84	15
4T	09:00-11:59	139	7	27	74	18
5T	12:00-14:59	126	11	28	70	17
6T	15:00-17:59	162	16	34	82	30
7T	18:00-20:59	181	14	36	86	38
8T	21:00-23:59	160	12	38	72	40

Table 2 shows timewise accident distribution & gives detailed numbers regardingc classification of accident.

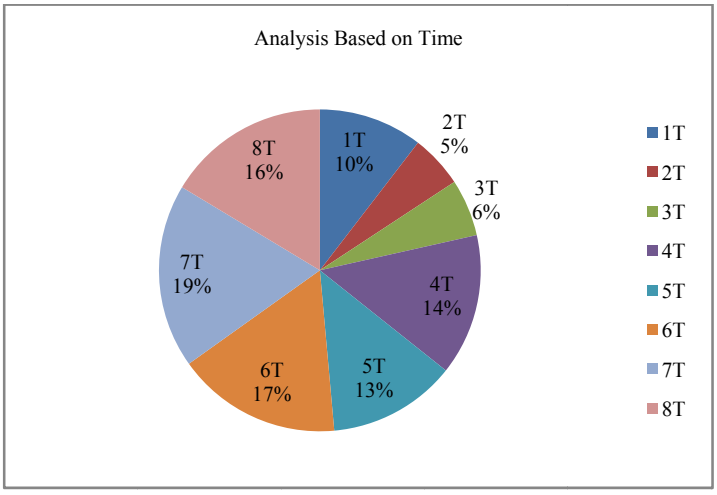

Figure 4: Analysis based on time

Figure 4 shows % distribution on basis of time. The available data has been processed to apply regression analysis and predict the accident possibility. Following section explains the same in detail.

III. REGRESSION ANALYSIS

The regression model is developed for predicting accident and fatalities, considering factors contributing to occurrence of accidents as independent variables and accident as dependent variable using regression equation.

A. Introduction

The form for linear regression models developed will be in following form:

$$y = mx + c$$

Where,

y= accident to be predicted.

x=factor contributing to occurrence of accident.

m & c = slope and coefficient.

For multiple regressions if there are n predictor variables, then the regression equation model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + c$$

The $x_1, x_2, \ldots x_n$ represent the n predictor variables. Those parameters are the same as before, β_0 is the constant, β_1 is the coefficient on the first predictor variable, β2 is the coefficient on the second predictor variable, and so on. c is the error term or the residual that can't be explained by the model.

Microsoft Excel was used to perform the regression analysis on available data.The df(Regression) is one less than the number of parameters being estimated. There are n predictor variables and so there are n parameters for the coefficients on those variables. There is always one additional parameter for the constant so there are $n + 1$ parameters. But the df is one less than the number of parameters, so there are $n + 1 - 1 = n$ degrees of freedom. That is, the df(Regression) = # of predictor variables.[8]

The df (Residual) is the sample size minus the number of parameters being estimated, so it becomes df (Residual) =$m - (n + 1)$ or df (Residual) =$m - n - 1$. It is easier just to use subtraction once you know the total and the regression degrees of freedom. Thedf(Total) is still one less than the sample size as it was before. df(Total) =$m - 1$. A variance is a variation divided by degrees of freedom, that is MS = SS / df. The F test statistic is the ratio of two sample variances with the denominator always being the error variance. So F = MS(Regression) / MS(Residual).

The null hypothesis claims that there is no significant correlation at all. That is, all of the coefficients are zero and none of the variables belong in the model.The alternative hypothesis is not that every variable belongs in the model but that at least one of the variables belongs in the model. If p-value is 0.000, we must conclude that there is no correlation at all and have a good model for prediction.If the coefficient is zero, then that variable drops out of the model and it doesn't contribute significantly to the model. If the p-value < 0.05, we'll reject the null hypothesis and retain it otherwise.[13][2][3]

B. Steps for Regression Analysis

Enter the data into the spreadsheet that you are evaluating. You should have at least two columns of numbers that will be representing your Input Y Range and your Input X Range. Input Y represents the dependent variable while Input X is your independent variable. Open the Regression Analysis tool. Define your Input Y Range. Repeat the previous step for the Input X Range. Click OK. The summary of your regression output will appear where designated.

1) *Identification of parameters:* Total 7 parameters have been identified which are further subdivided into various categories. Dependent and independent variables have been classified (Table 3). Dependent and Independent variables[4]for analysis:-

- Y (dependent variable)

1. Classification of accident
2. Nature of accident

- X (independent variable)

1. Road feature
2. Intersection type and control
3. Road condition
4. Causes
5. Weather condition

Table 3. Parameters

Parameters	Categories	Code
Nature of Accident (A)	Overturning	A1
	Head on Collision	A2
	Rear End Collision	A3
	Collision Brush	A4
	Right Turn on Collision	A5
	Skidding	A7
	Other	A8
Classification of accident (B)	Fatal	B1
	Grievous	B2
	Minor Injury	B3
	Non Injury	B4
Causes (C)	Drunken	C1
	Over Speeding	C2
	Vehicle Out of Control	C3
	Fault of Driver	C4
	Defect in Vehicle	C5
Road Feature (D)	Single Lane	D1
	Two Lane	D2
	Three Lane or More Without Central Divider	D3
	Four Lane or More With Central Divider	D4
Road Conditions (E)	Straight Road	E1
	Slight Curve	E2
	Sharpe Curve	E3
	Flat Road	E4
	Gentle Slope	E5
	Steep Slope	E6
	Hump	E7
	Dip	E8
Intersection Type and Control (F)	T Junction	F1
	Y Junction	F2
	Four Arm Junction	F3
	Staggered Junction	F4
	Junction with More Than Four Arms	F5
	Roundabout Junction	F6
	Manned Rail crossing	F7
	Unmanned Rail Crossing	F8
Weather Conditions (G)	Fine	G1
	Mist /Fog	G2
	Cloud	G3
	Light Rain	G4
	Hail	G5

	Snow	G6
	Strong Wind	G7
	Dust Storm	G8
	Very Hot	G9
	Very Cold	G10
	Other Extraordinary Weather Conditions	G11

For regression analysis following parameters are choose on basis of correlation and regression is applied using Microsoft Excel (2007).

C. Application of regression analysis

y=Classification of accident, x_1 = Road Feature, x_2 = Road Condition, x_3= Intersection Type and Control, x_4 = Weather Condition

Table 4 below shows the results of application of regression analysis.

Table 4. Summary Output

Regression	**Statistics**
Multiple R	0.162933
R Square	0.265473
Adjusted R Square	0.018821
Standard Error	0.824064
Observations	634

	Coefficients
Intercept	2.460882533
x_1	-0.003573306
x_2	0.016035055
x_3	-0.017147712
x_4	0.009261565

	df	SS	MS	SF	Significance F
Regression	4	11.66725968	2.333452	3.436185	0.00451679
Residual	630	427.821734	0.679082		
Total	634	439.4889937			

From regression analysis output we obtain coefficients for parameters and obtained equation is:

$$y = (-0.00357x_1 + 0.016035x_2 - 0.01715x_3 + 0.009262x_4) + 2.460883$$

D. Prediction

Using the equation form from regression analysis of accidental data prediction is made for different independent variables along the stretch so as to find classification of accident that can occur. Validation of prediction model is done using available accident data. Error in respective model is indentified and their validity is checked (Table 5).

Table 5. Prediction for classification of accident

Predicted Value	**Classification of Accident (Actual Value)**	**Difference**	**Accuracy (%)**
2.856232	3	-0.143768	95.2077333
2.872267	3	-0.127733	95.7422333
0.571827	1	-0.428173	57.1827
4.091291	4	0.091291	98.1808
3.030221	3	0.030221	99.6101
2.856232	3	-0.143768	95.2077333

IV. OBSERVATIONS

Analysis based on accident location predicts 180-189km on right lane have highest no of accidents on study stretch. Analysis based on accident location predicts 240-249km on left lane have highest no of accidents on study stretch. Analysis based on time predicts 18:00-20:59 hrs have highest no of accidents on study stretch. Using prediction equation and proper parameters, prediction for classification of accident that can take happen on stretch can be made. For prediction purpose, random 60 values from available accident data is chosen for input variables and prediction is made using equation. Predicted values from regression equation were compared with available accident data and found that prediction model for classification of accident predicts 66% values with an error of approx 10%, 12% values with an error of 20% and 22% values with an error higher than 30%.Using field details of stretch and prediction model potential points for accidents can be predicted.

V. CONCLUSION

Severity of accident can be reduced by applying prediction model with proper input of parameters. The likelihood of accidents on the study stretch can be reduced. The need of costly remedial work can be reduced. The total cost of stretch safety for community, including accidents, disruption and trauma is minimized. Safety provisions needed for KM180-189 on right lane and on KM240-249 on left lane as these have highest no. of accidents. Lighting provisions must be improved for 18:00-20:59 hrs on study stretch.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to our guide Ms. Sayali Sandbhor for providing her valuable guidance, comments and suggestions throughout the course of the project. We would also like to thank Mrs. Humera Khanum for introducing us the topic and providing data.

We are really grateful towards our college for providing us all the necessary permissions and details we required. The faculty was also very helpful for any of our doubts related to project.

REFERENCES

[1] Accident Analysis on NH-18 by Using Regression Model and its Preventive Measures VeluruSailaja1 , Dr. S. Siddi Raju2
[2] Road Accidents Study Based On Regression Model: A Case Study of Ahmedabad City ManishaMinesh Desai
[3] Accident Analysis and Prediction of Model on National Highways Rakesh Kumar Singh & S.K.Suman
[4] Analysis of relationship between road safety and road design parameters of four lane National Highway in India Ravi Shenker1,Arti Chowksey2 and HarAmrit Singh Sandhu3
[5] http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/
[6] Wikipedia
[7] Transport Research Wing of the Ministry of Road Transport and Highways
[8] SAS
[9] Big sky
[10] AASHTO
[11] www.badroadsinindia.com
[12] http://timesofindia.indiatimes.com/topic/Road-accident
[13] https://people.richland.edu/james/ictcm/2004/multiple.html

AUTHOR PROFILE

Author1 Rahul Badgujar Undergraduate scholar, Department of Civil Engineering, Symbiosis Institute of Technology, SIU, Pune, India

Author2 PriyamMishra Undergraduate scholar, Department of Civil Engineering, Symbiosis Institute of Technology, SIU, Pune, India

Author3 Mayank Chandra Undergraduate scholar, Department of Civil Engineering, Symbiosis Institute of Technology, SIU, Pune, India

Author4 Sayali Sandbhor Assistant Professor, Department of Civil Engineering, Symbiosis Institute of Technology, SIU, Pune, India

Author5 Humera Khanum Assistant Professor, Department of Civil Engineering, Symbiosis Institute of Technology, SIU, Pune, India