"Effect of Time on Fatal Motor Vehicle Crashes" in the United States

Areej Altamimi
School of Computing
Clemson University
Clemson, SC
aaltami@clemson.edu

Farah Alshanik
School of Computing
Clemson University
Clemson, SC
falshan@clemson.edu

Devesh Kumar
Department of Civil Engineering
Clemson University
Clemson, SC
deveshk@clemson.edu

***Abstract*—Road Safety being of paramount importance, continuous efforts have been made to understand the factors which play significant role in crashes. These include road characteristics, human factors, and weather. Various studies have been done to assess the role these factors in road crashes. Recent studies have identified a new factor resulting in crashes-*Time.* However, there remains a dearth of detailed assessment to understand the association of time with fatal motor accidents in United States. The main purpose of this study is to establish trends in motor vehicle crashes associated with time. Time here, refer to different units of time i.e. hour, weekday, week, month, and year. Author analyzed the effect of time on fatal motor vehicle crashes. Based on the analysis, different machine learning algorithm were applied to predict and detect the probability of vehicle accidents. Python was used for developing model to predict the probability using machine learning algorithm (MLA). Based on the developed evaluation criteria to assess various MLAs, it was found that Gradient Boosting algorithm is best suited to predict the probability of crashes in given time.**

Keywords—time, fatal motor vehicle crashes, machine learning algorithm, gradient boosting.

I. Introduction

Road Safety has always been of prime importance in a highway system. In recent years, fatalities on U.S. highways have ranged between 40,000 and 43,000 per year [1]. Though it is lower than 55,000 per year experienced in 1970s, the former itself is not a low number and continues to represent staggering number. Each year road crashes results in severe injuries leading to temporary to permanent injuries such as long hospitalization, permanent injuries, inability to return to work or even loss of life.

Several factors such as behavioral factors, road characteristics, and environment contribute to road crashes. Behavioral factors in road accidents are difficult to study by traditional research methods for several reasons [2]. Road crashes are often unpredictable, so direct observation is practically impossible [3]. Many studies were conducted to analyze the factors contributing to vehicle crashes, using traditional methods that based only on descriptive analysis.

Recently, Machine Learning has been introduced as a valuable approach to establish relationships between various parameters. Machine learning is a technique wherein it studies the relationship in training/input data and make predictions based on established relationship. Advent of big data and its use in prediction using machine learning has been useful in many fields and the same can be applied in prediction of the most common risk factors for traffic accidents simultaneously, providing basis for minimizing these risk factors and ultimately reducing the frequency or severity of traffic accidents.

There have been several transportation and traffic accident related studies that assessed analysis of crash and traffic data such as work of Smeed published in 1949 who identified the number of fatally injured persons in an accident and compared those incident rates among various countries [4]. Similarly, Saha et al, in their work assessed the relationship between adverse weather conditions and fatal motor vehicle crashes in United States [5]. However, association between time and road crashes has not yet well developed and there remains a dearth of comprehensive assessment of association between them.

In this study, authors studied a new factor-time that has not been studied in detail in previous works. Author believes that a model can be established based on the study of trends of road crashes over the years/decades to associate effect of time with road crashes. Hence, author analyzed the effect of time on fatal vehicle crashes in US over a period from 2007 to 2015, followed by application of machine learning algorithms for training and prediction of probability of number of fatal motor vehicle crashes in given period (time).

A. *Objective*

Assessing the need for a comprehensive study, this study revolves around the objective of analyzing, using machine learning, effect of specific time (hour, day, week, month and year) on fatal motor vehicle crashes. In this study, the authors have pursued the following objectives:

1. To identify and understand role of various factors involved in road accidents using the available data on crashes in the United States;
2. To develop and evaluate various prediction models using machine learning, and assess the association of time and fatal road crashes;
3. To select best prediction model based on set evaluation criteria; and
4. To predict and evaluate occurrence of fatal motor vehicle crashes in a given period (time) based on Test Data.

B. *Hypothesis*

Various organization in US collects and maintains road accidents/crash data and make it available in public domain. Few of these organization are National Highway Traffic Safety Administration (NHTSA) and US Department of Transportation

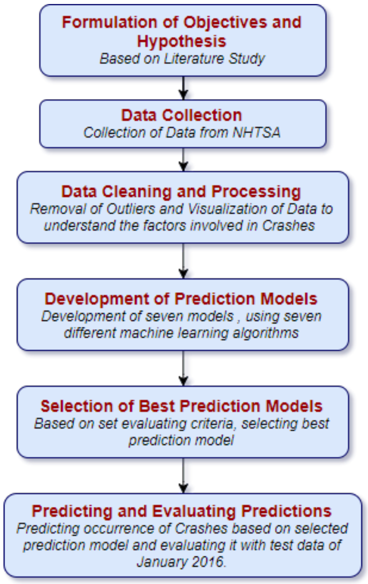
(DOT). NHTSA's Fatality Analysis Reporting System (FARS) is a census of fatal motor vehicle crashes. This study hypothesized that there exists a relationship between time and occurrence of fatal motor vehicle crash. Once a predictive model is developed using the available data on crashes, occurrence of crashes can be predicted in a given period of time. This shall provide opportunity to implement mitigation measures to eliminate factors which shall reduce number of fatal crashes and hence benefitting the economy.

II. Methodology

An analytical approach was developed for this study which started with data collection related to road crashes/accidents. The data was analyzed for all the states in US. For this, NHTSA's Fatality Analysis Reporting System (FARS) was used. FARS is a census of fatal motor vehicle crashes with a set of data files documenting all qualifying fatalities that occurred within the 50 States, the District of Columbia, and Puerto Rico since 1975. The crash data was collected for the year 2007 to 2016. Once the data was collected, it was cleaned and processed for removal of outliers. Visualization of collected data was also done to understand the data and its feature.

Crash data for the year 2007 to 2015 was used to train the machine learning algorithms and prediction (regression) models were developed. These were then, evaluated based on the set criteria of Root Mean Square Error (RMSE), and R Square values and the best model was selected. This model was used to predict the occurrence of fatal crashes for the month of January 2016, which was compared to the actual data of January 2016 (Test Data). This was followed by conclusion and recommendations. Fig I depict the methodology followed in this study.

FIGURE I. Study Methodology



Formulation of Objectives and Hypothesis
Based on Literature Study

Data Collection
Collection of Data from NHTSA

Data Cleaning and Processing
Removal of Outliers and Visualization of Data to understand the factors involved in Crashes

Development of Prediction Models
Development of seven models , using seven different machine learning algorithms

Selection of Best Prediction Models
Based on set evaluating criteria, selecting best prediction model

Predicting and Evaluating Predictions
Predicting occurrence of Crashes based on selected prediction model and evaluating it with test data of January 2016.

III. Data Collection and Processing

A. Fatality Analysis Reporting System (FARS)

The Fatality Analysis Reporting System (FARS), which became operational in 1975, contains data on a census of fatal traffic crashes within the 50 States, the District of Columbia, and Puerto Rico. For a crash data, to be included in FARS, it must involve a motor vehicle traveling on a traffic-way customarily open to the public, and must result in the death of an occupant of a vehicle or a non-occupant within 30 days (720 hours) of the crash.

FARS is directed by the National Center for Statistics and Analysis (NCSA), which is a component of NHTSA. NHTSA has a cooperative agreement with an agency in each State's government to provide information on all qualifying fatal crashes in the State. FARS data are obtained from various States' documents [6], such as:

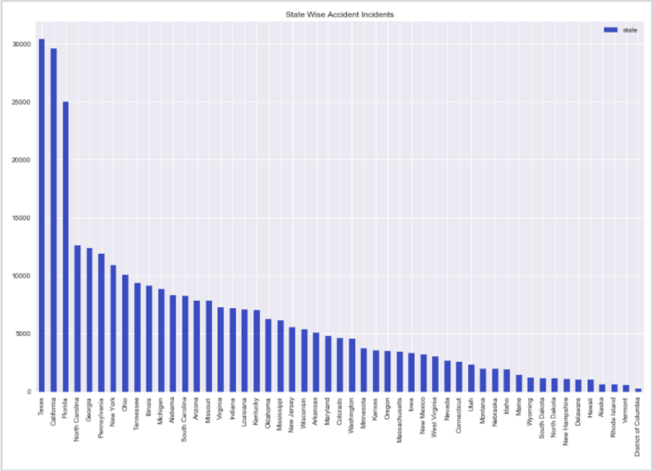
- Police Crash Reports
- Death Certificates
- State Vehicle Registration Files
- Coroner/Medical Examiner Reports
- State Driver Licensing Files
- State Highway Department Data
- Emergency Medical Service Reports
- Vital Statistics and other State Records

The data is coded automatically checked when entered for acceptable range values and for consistency. enabling the analyst to make corrections immediately. Several programs continually monitor and improve the completeness and accuracy of the data [7]. Annual FARS data files are available for 1975 through 2016. These files were downloaded and processed to develop prediction models.

B. Data Cleaning and Visualization

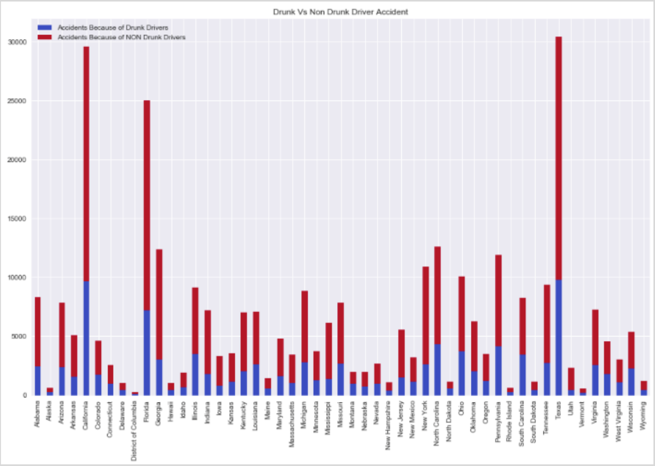
The crash data was collected for the year 2007 to 2016 and it was cleaned for removal of outliers and anomalies. This data contained few records with month value greater than 12, days values greater than 31. Such kind of data records were removed in data cleaning process. This cleaned data was then plotted to understand the trend of accidents happening all over the country.

FIGURE II. Statewise Car Accidents Incidents



For example, number of car accidents were plotted for all states in the US to understand which states are having more number of car accidents. This is presented in Fig. II. Similarly, to analyze the effect of drunk driving on accidents, drunk vs non-drunk driving accidents data was plotted. This is represented in Fig. III. It can be observed from Fig. III, that beside drunk driving, there are few other factors which contribute to road accidents with drunk driving being the most common cause.

FIGURE III. Drunk vs non-drunk driver accident



After cleaning and visualization, the data for the year 2007 to 2015 was used for developing prediction models using seven machine learning algorithms. These models are discussed in the following sections.

IV. Development of Prediction Models

Once the data classification and trend assessment is done, the authors, proceeded with model development. It was decided to develop seven various types of models using Python coding. These are as follows:

1. Multiple linear Regression;
2. Lasso Regression;
3. Elastic Net;
4. Kernel Ridge;
5. Gradient Boosting;
6. Xg Boost; and
7. Light Gradient Boosting.

These models are discussed in the following sections,

1) Multiple Linear Regression

Multiple regression is simply an extension of normal linear regression. It is used when we want to predict the value of a variable based on the value of two or more variables. In this case, the analysis is known as multiple linear regression. The main idea of multiple linear regression method is to build correlation analysis between dependent (crashes) and independent variables (Time, day, etc) [8]. The basic function is of the following form:

$$\boldsymbol{Y = C_0 + C_1X_1 + C_2X_2 + C_3X_3 + C_4X_4 + C_5X_5}$$

Where, Y= Dependent Variable

C_0= Regression Constant

C_1, C_2, C_3, C_4, C_5 =Regression Coefficients of respective 5 dependent variables.

X_1, X_2, X_3, X_4, X_5= Independent Variables

The dependent variable in above example is crash number, and independent variables are hour, day, week, month, and year.

2) Lasso Regression

In statistics and machine learning, lasso (least absolute shrinkage and selection operator) (also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces.

3) Elastic Net Regression

Estimates Lasso and Elastic-Net regression models on a manually generated sparse signal corrupted with an additive noise. Estimated coefficients are compared with the ground-truth.

4) Kernel Ridge Regression

Kernel ridge regression, in machine learning, combines Ridge Regression (linear least squares with l_2-norm regularization) with the kernel trick. It thus learns a linear function in the space induced by the respective kernel and the data. For non-linear kernels, this corresponds to a non-linear function in the original space.

5) Gradient Boosting Regression

GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage, a regression tree is fit on the negative gradient of the given loss function.

6) Xg Boost Regression

Xg Boost is short for “Extreme Gradient Boosting”. Xg Boost is used for supervised learning problems, where we use the training data (with multiple features) X_i to predict a target variable Y_i.

7) Light Gradient Boosting Regression

Light Gradient Boosting (LGB) is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking, classification and many other machine learning tasks. Since it is based on decision tree algorithms, it splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf-wise.

So, when growing on the same leaf in LGB, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy which can rarely be achieved by any of the existing boosting algorithms. Also, it is surprisingly very fast, hence the word ‘Light’.

V. Analysis and Findings

Once the regression models were developed, a regression evaluation matrix was developed to evaluate each model. It included three common evaluation matrices for regression analysis. These are:

a) Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is the mean of the absolute value of the errors. This is given by following formula:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

b) Mean Squared Error (MSE)

Mean Squared Error (MSE) is the mean of the squared errors. This is given by following formula:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

c) Root Mean Squared Error (RMSE)

Root Mean Squared Error (MSE) is the square root of the mean of the squared errors. This is given by following formula:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The matrices were used due to following reason:

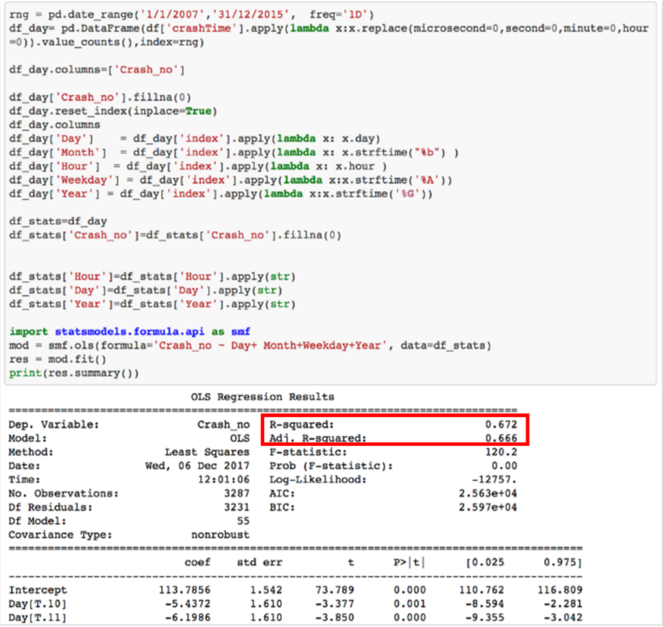
- MAE is the easiest to understand, because it's the average error;
- MSE is more popular than MAE, because MSE "punishes" larger errors, which tends to be useful in the real world;
- RMSE is even more popular than MSE, because RMSE is interpretable in the "y" units.

These functions are loss functions, because these are to be minimized. At first, hour, day, week, month and year were selected as independent variable to predict the value of dependent variable "crash number". However, R square value for this model was around 20% which suggests that hourly intervals are not a good variable for developing regression model.

Hence, "hour" as independent variable was dropped and daily intervals were selected as minimum period. The R square value obtained from this model (refer Fig IV) was around 67% which is considerably good because only few features of data and real-world data is being used. Thus, the final model includes four independent variable which excludes "hour". The comparison of all seven model (for four variables) based on above three matrices is presented in Fig. V.

Based on the above evaluation, it can be observed that Gradient Boosting has least RMSE value and maximum R square value. Evaluation based on above three matrices was further substantiated by plotting the prediction of crashes as per the test data including residual histogram. Fig. VI and VII shows an example for plot (prediction vs test data plot) and residual histogram each in case of Gradient Boosting.

FIGURE IV. OLS REGRESSION RESULTS



```
rng = pd.date_range('1/1/2007','31/12/2015',  freq='1D')
df_day= pd.DataFrame(df['crashTime'].apply(lambda x:x.replace(microsecond=0,second=0,minute=0,hour
=0)).value_counts(),index=rng)

df_day.columns=['Crash_no']

df_day['Crash_no'].fillna(0)
df_day.reset_index(inplace=True)
df_day.columns
df_day['Day']     = df_day['index'].apply(lambda x: x.day)
df_day['Month']  = df_day['index'].apply(lambda x: x.strftime("%b") )
df_day['Hour']  = df_day['index'].apply(lambda x: x.hour )
df_day['Weekday'] = df_day['index'].apply(lambda x:x.strftime('%A'))
df_day['Year'] = df_day['index'].apply(lambda x:x.strftime('%G'))

df_stats=df_day
df_stats['Crash_no']=df_stats['Crash_no'].fillna(0)


df_stats['Hour']=df_stats['Hour'].apply(str)
df_stats['Day']=df_stats['Day'].apply(str)
df_stats['Year']=df_stats['Year'].apply(str)

import statsmodels.formula.api as smf
mod = smf.ols(formula='Crash_no ~ Day+ Month+Weekday+Year', data=df_stats)
res = mod.fit()
print(res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Crash_no   R-squared:                       0.672
Model:                            OLS   Adj. R-squared:                  0.666
Method:                 Least Squares   F-statistic:                     120.2
Date:                Wed, 06 Dec 2017   Prob (F-statistic):               0.00
Time:                        12:01:06   Log-Likelihood:                -12757.
No. Observations:                3287   AIC:                         2.563e+04
Df Residuals:                    3231   BIC:                         2.597e+04
Df Model:                          55
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept       113.7856      1.542     73.789      0.000     110.762     116.809
Day[T.10]        -5.4372      1.610     -3.377      0.001      -8.594      -2.281
Day[T.11]        -6.1986      1.610     -3.850      0.000      -9.355      -3.042
```

FIGURE V. SCORING OF MODELS BASED ON DEVELOPED MATRICES

S. No.	Name	RMSE		R^2	
		Mean	Std. Dev	Mean	Std. Dev
1	Multiple Linear Regression	0.576845	0.0231366	0.666128	0.017321
2	Lasso Regression	0.576736	0.0241598	0.666237	0.018469
3	Elastic Regression	0.576768	0.0241771	0.666204	0.018403
4	Kernel Ridge Regression	0.578823	0.0249323	0.663837	0.018815
5	Gradient Boosting	0.557086	0.0253957	0.688286	0.023956
6	Xg Boost	0.563610	0.0247817	0.680842	0.025267
7	Light Gradient Boosting	0.572354	0.0258934	0.671150	0.021999

FIGURE VI. PREDICTION VS TEST DATA (GRADIENT BOOSTING)

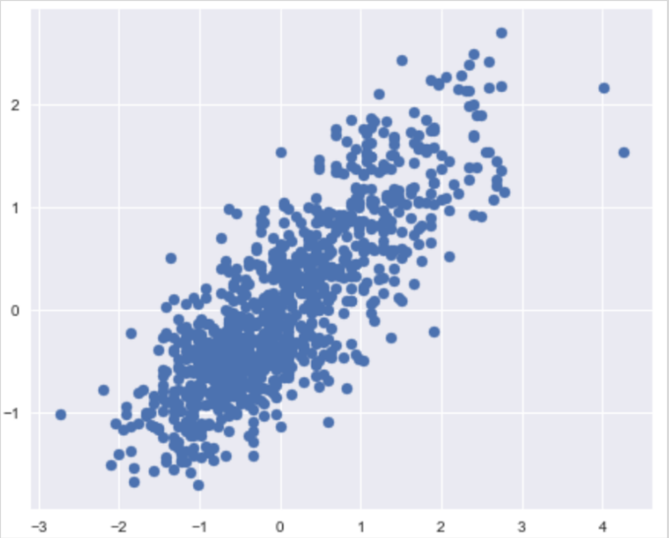
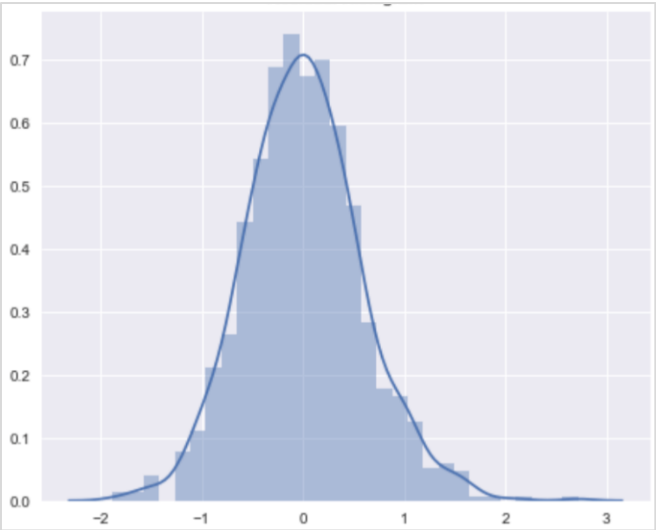


FIGURE VII. RESIDUAL HISTOGRAM (GRADIENT BOOSTING)

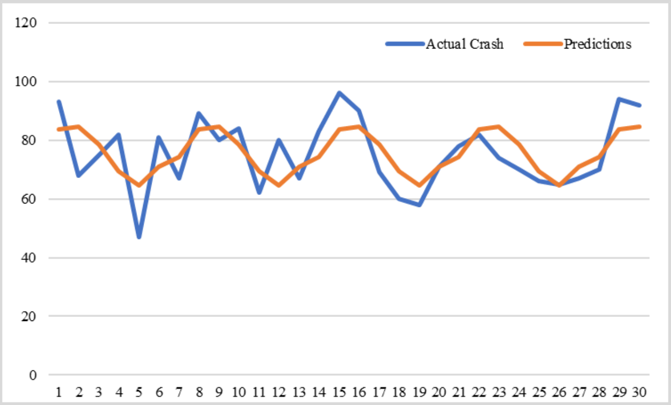


Gradient Boosting regression model was found to be the best among all seven. Hence, this model was selected for further prediction and evaluation with test data of year 2016.

VI. Evaluation of Selected Prediction Model

The above model was then used to predict number of crashes in 2016. The prediction values as presented in Fig VIII were found quite close to the actual value. Hence, it can be inferred that four variables (day, week, month and year) can be used to develop a prediction model.

FIGURE VIII. Evaluation of Prediction Results



VII. Conclusion

This study started with the hypothesis that variables such as hour, day, week, month and year can be used to develop prediction model to predict number of crashes in the given period. After analyzing all the prediction models in this study, it can be concluded that hourly interval is not a good fit for predicting number of crashes. However, day, week, month and year are a good fit for predicting number of crashes.

Considering the fact that real-world data was used in the study and only selected features were used to develop prediction model, the prediction results are quite acceptable. This also provides an opportunity to increase the scope of this study by including more features into the model so that prediction results can reach up to actual number of crashes.

References

[1] R. P. Roess, E. S. Prassas and W. R. McShane, Traffic Engineering 4th Edition, Prentice Hall, 2011.

[2] D. D. Clarke, R. Forsyth and R. Wright, "Behavioural factors in accidents at road junctions: the use of a genetic algorithm to extract descriptive rules from police case files," *Accident Analysis & Prevention,* vol. 30, pp. 223-234, 1988.

[3] D. D. Clarke, R. Forsyth and R. Wright, "Junction road accidents during cross-flow turns: a sequence analysis of police case files," *Accident Analysis & Prevention,* vol. 31, pp. 31-43, 1999.

[4] Smeed, "Some Statistical Aspects of Road Safety Research," *Journal of Royal Statistical Society Series,* vol. A 112, pp. 1-34, 1949.

[5] S. Saha, P. Schramm, A. Nolan and J. Hess, "Adverse weather conditions and fatal motor vehicle crashes in the United States, 1994-2012," *Environmental Health,* vol. 15, no. 104, p. 1, 2016.

[6] N. H. T. S. Administration, "Fatality Analysis Reporting System (FARS)," US Department of Transportation, October 2017. [Online]. Available: ftp://ftp.nhtsa.dot.gov/fars/. [Accessed 15 September 2017].

[7] N. H. T. S. Administration, "Fatality Analysis Reporting System (FARS)," October 2017. [Online]. Available: ftp://ftp.nhtsa.dot.gov/fars/FARS-DOC/. [Accessed 15 September 2017].

[8] H. Gupta and S. Rokade, "Development of Crash Prediction Model Using Multiple Regression Analysis," *Technical Research Organization India,* vol. 4, no. 6, pp. 82-86, 2017.