Predicting Traffic Accidents using machine Learning Algorithms

S.Anil Kumar[1], A.Umadevi[2], K.Jyothirmai[3], D.Venkatareddy[4], K.Karthik[5]

[1]Assistant Professor, CSE, Tirumala Engineering College, Narasaraopet, A.P., India
[2,3,4,5] B. Tech Students, CSE, Tirumala Engineering College, Narasaraopet, A.P., India

Abstract: Models of accident prediction (APMs) are extremely important instruments for estimating the number of expected accidents in such entities as intersections and street segments. These estimates are typically used in the identification and assessment of sites for possible safety treatment. An APM is, essentially, a mathematical equation that describes the overall site incident incidence as a result of traffic flow and other site features. At the same period, the reliability of an APM is strengthened when the APM is based on data as long as possible, particularly if data for the same years is used in the safety review of a location. This paper covered as many articles as practicable and numerous work holes and a future study possibility were suggested in this field. Several models such as multiple linear regressions, Poisson regression, Conway-Maxwell Poisson regression modeling, artificial neural networks and furious logics have been discussed in this paper.

Keywords: Traffic Accidents Prediction, Multiple Linear Regression, Poisson Regression, Literature Review, Artificial Neural Networks.

I. INTRODUCTION

Road accidents remain a significant worldwide problem, both from the public health point of view and from the socio-economic point of view [1]. Cars, vans, boats, bicycles, tourists, animal species, taxis and other types of travelers worldwide share highways, which in many countries lead to economic and social development [2]. Yet every year, many vehicles take part in accidents that cause millions of deaths and injuries. Approximately 1, 25 million people are killed every year in motor vehicle crashes and about 50 million more are hurt. According to current trends, by 2030 around two million people are expected to die each year in motor vehicle crashes [2].

Today, road accidents have become the ninth most serious cause of death in the world and by 2020 fatal accidents will likely rise to third place without new initiatives aimed at improving road safety[3]. The numbers of road accidents in developed countries have declined since the 1960s because of positive measures such as protective laws on seat belts, regulation of speed limits, warnings to the dangers of mixing alcohol with transportation and a better construction and use of roads and cars. In the United States for example, road deaths decreased by roughly 25,0% between 2005 and 2014, and the number of people hurt declined by 13,0% between 2005 and 2014[3]. In Canada, the number of accidents on roads decreased by approximately 62% from 1990 to 2014, and the number of injuries fell by approximately 68% over the same period [2].

In the developing countries, however, traffic accidents have decreased from 1990 to 2014; for example, the number of road accidents within Malaysia has risen by 44 percent and in China by around 43 percent. Developing countries face a huge responsibility globally, responsible for 85% of annual deaths and 90% of increased years of life in disabilities. Less than one third of all roads going by involve people between the ages of 15 and 44; in their productive years. In fact, this age group's illness responsibility accounts for nearly 60% of all modified life years. The expenses and effects of these damages are significant. Three-quarters of all disadvantaged families who lost their position in a car accident reported a drop in their living standards and about 61 percent reported borrowing cash to cover their costs as a result of their tragedy. While transport authorities also attempt to identify the most dangerous road sites and make great strides in protective measures, for example lighting and policy enforcement, the number of road incidents annually has still not decreased significantly. In 2015, for example, there were 3, 5092 traffic fatalities reported in the US, a rise of 7.2% compared to the previous year [5, 6]. Around 2014 and 2015 the fatality rate for every 100 million vehicle miles traveled (VMT) rose by 3.7 percent. Every month, except in November, the number of fatalities in motor vehicles increased per month from 2014 to 2015; the largest rises occurred in July and September [6].

II. RELATED WORD

We address the related work in this segment. Next, numerous studies were conducted to examine the origin and nature of road accidents. Li et al. [5] used data mining algorithms such as the Multiple Linear algorithm, the classification method of the Naïve Bayses and the K-mean data traffic clustering algorithm. The research was carried out by the scientists, including temperature, light or surface conditions. However, an analysis of closely related attributes such as the type of accident or the type of road is required. Chong et al. [3] used algorithms in machine learning to predict the extent of injuries in traffic accidents. We called neural networks educated in deep learning, help vector machines, decision-making structures, and the concomitant hybrid paradigm of decision-making and neural networks. In comparison, the decision-tree used to evaluate and develop the N5 National Highway traffic accident statistics in Bangladesh [8] Oña et al. [6] provided the study of the rural road accidents in Spain by means of Bayesian Networks to identify traffic accidents according to their injury seriousness. In [10] a model for forecasting the frequency of traffic collisions in Abu Dhabi was suggested. Abellán et al. [1] proposed an effective way to remove rules from the decision tree in order to extract important relationships between variables. And the proposed approach was used to collect relevant rules from Spain's rural traffic accident records. In [7], authors analyzed the seriousness of a traffic accident on Slovenian highways with an algorithm for the classification and regression tree. Castro and Kim [2] use the Bayesian network, decision-making tree and artificial neural networks to detect car accidents in the UK.

There are also several articles documenting Korea's traffic accidents. A Structural Equation Model (SEM) [4] is recommended to compare the traffic accident variables. This uses crash details from roads in Korea and measures the association between exogenous factors and the severity of the traffic accident. Sohn et al. [9] used data mining techniques to estimate and classify the seriousness of road accident types in Korea. We used the neural network, the logistic regression and the decision tree to pick a number of factors and to define classified modes for seriousness of the incident.

III. CLASSIFICATION OF ACCIDENT PRIDICTION MODELS

The following subsection presents a number of mostly used models.

A. Multiple Linear Regressions

Early simulations of equations based on a simple multiple linear regression method, considering errors normally distributed. The general form of the linear crash prediction model is as follows:

$$Y/\Theta \sim Dist(\Theta) with\ \Theta = f(x\beta, \varepsilon) \qquad (1)$$

Where,
Y: the dependent variable (i.e. crash frequency),
Θ: the crash dataset, Dist (θ): the model distribution,
X: a vector representing different independent variables (i.e. risk factors),
β: a vector of regression coefficients,
f (.): link function that relates X and Y together,
ε: the disturbance or error terms of the model.

B. Poisson Regression

Although several linear regression models have typically been associated, it has been observed that crash incidents are often better suited to a Poisson circulation. The simulation of incident data as continuous data by using an ordinary lowest square regression is a recurrent problem [7]. This is false since regression models can generate expected values that are incompatible with continuous data processing and can also project negative values. Furthermore, several distributions of crash data with multiple findings in the data set with a value of zero [8] are significantly biased. The high number of zeroes in the data set prevents a skewed distribution from turning into a normal distribution that is a requirement for normal distribution [9]. The use of a poisson distribution or one of its variations is an option.

Poisson distributions have several advantages over normal distribution, including skewed, discrete and non-negative number limits to predicted values [7]. In order to investigate the interaction between risk factors and forecasting of incidents in traffic [10, 11, 12, 13, and 14], common linear simulation versions of Poisson models were introduced. A wide range of transport count results, including accident frequency, were used for the regression of Poisson. A Poisson regression model is identical with two variations to a regular linear regression. In the first place, it assumes that the errors follow a Poisson circulation (not normal). Third, it modeless the normal log of the answer variable in(Y) rather than the simulation of the response variable Y as a linear function of the regression coefficients [7]. The configuration of Poisson is as follows:

$$p(n_i) = \frac{\lambda t EXP(-\lambda t)}{n!} \qquad (2)$$

Where,
P (ni): The probability of n collisions in the section I of the road ni: number of observations per period (e.g. year), μi: expected frequency of collision on the street segment I for each period (i.e. average distribution). One hypothesis of the Poisson models is that the mean and variance is equal, an assumption sometimes violated [14]. This can be achieved using a conditional function, if there is a small difference, or using a negative binomial regression model, if there are significant differences [15].

C. Conway-Maxwell Poisson Regression Models

The Conway-Maxwell Poisson model was recently tested for safety issues but is relatively limited in its use in the field of incident incidence modeling [6]. Generalized additive models have been investigated since they can provide the explanatory variables with smoothing functions [6]. The Conway-Maxwell

Poisson process is a Poisson circulation theory, which can accommodate details on under-dispersed and over-dispersed incidents. The main favor of this model is to deal with the under-dispersion of crash data that cannot be modeled on the Poisson model or the Negative Binomial model. The low mean and small sample size of the under-dispersed accident data may however influence the estimated parameters and the frequency of use of accidents were therefore limited [7]. However, it can become very difficult in practice to approximate such models because they need further parameters, an issue which has possibly hampered their application to the estimation of accident rates [7, 8].

D. Artificial Neural Networks and Fuzzy Logic models

The interaction of dependent variables with the corresponding independent variables in accident modeling cannot properly be represented by a linear method, as well as by non-linear approximates such as furious logic and neural networks. Artificial Neural Networks (ANNs) are a collection of artificial intelligence devices that can be used for problems of inference and classification. ANNs can model extremely complex non-linear functions at high levels of precision by means of learning processes identical to the cognitive system modeling technique in the human brain. The network composition comprises of input layers, cached layers and output layers [10, 11]. Such models can be trained to approximate some non-linear behavior in a supervised learning process to the necessary degree of accuracy using a learning algorithm (e.g. back propagation). In contrast with statistical models, ANNs have some advantages [11]. For example, regression models require a predetermined relationship or functional shape of the dependent variable (crash frequency) with independent explanatory variables which can be estimated by some statistical approaches, while ANNs don't require these functional forms and can be easily analyzed [13]. In comparison, the ANNs vary from statistical models by serving as black boxes and do not view parameter estimates [14]. In recent years, Fuzzy logic systems are shown to have significant capacity estimation injuries [13, 14, and 15]. Fuzz's logic scheme is defined as the non-linear map of an input data set to a scalar output data. The first stage of the process (so-called "fuzzification"), consists in the compilation of a narrow set of input data, which is transformed into a bunch of fuzzy inputs using fluent linguistic variables. [8]. Following this, an approximation is generated based on a set of fuzzy rules and the resulting fuzzy output is mapped with the membership functions to a narrow output in the defuzzification stage[8, 9].

IV. PROPOSED SCHEMA

In this segment we develop and describe a large data model for the study of traffic accidents step by step. Figure 1 shows a model for analysis of accident factors in traffic. The first step is to define the problem and a consistent research approach. We define the problem in this paper as an analysis of the causes and rules of association of road accidents in Korea. The second step is to collect data for the study of big data. For this analysis we use traffic mortality data from 2015-206 in Korea. Next, in the third step the characteristics of the data collected are calculated and preprocessed. We analyze the data in the fourth step using the statistical method and Multiple Linear algorithm. In the last stage, we draw conclusions based on the results evaluated.

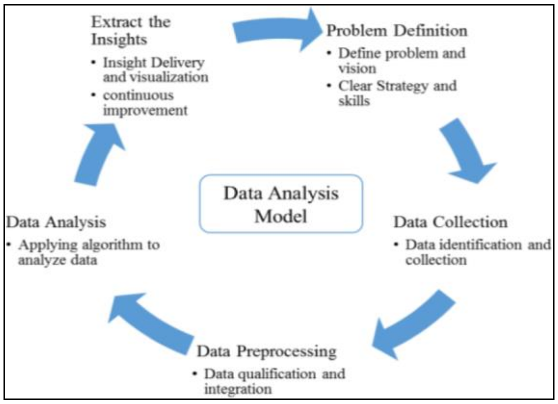
Extract the Insights
• Insight Delivery and visualization
• continuous improvement
Problem Definition
• Define problem and vision
• Clear Strategy and skills
Data Analysis Model
Data Analysis
• Applying algorithm to analyze data
Data Collection
• Data identification and collection
Data Preprocessing
• Data qualification and integration

Figure 1: Proposed architecture

A. Data Cleaning

In this section, we describe the data cleaning phase. First, we describe the data used in the analysis, and explain the preprocessing step of the collected data.

B. Data Collection

The purpose of this study is to examine driving accident factors in Korea. For research, we use the traffic accident reports collected by the government in 2015 and 206. The data collected includes 23 attributes such as accident date and time, accident type, death and injury numbers, etc. and data are supplied in a csv format. The key characteristics of the data collected are listed in Table 1.

Table 1. Main attributes of the collected data

Attribute	Description
The date and time of accident	The year, month, day and time of the accident
Time slot	Time of accident (day or night)
Day of the week	Days of accidents (Mon., Tues., Sun.)
Number of deaths	The number of people who died due to traffic accidents
Number of seriously injured people	The number of injured people (More than 3 weeks of treatment is required)
Number of slightly injured people	The number of injured people (More than 5 days and less than 3 weeks of treatment is required)
Number of injured people	The number of injured people (Less than 5 days of treatment is required)
Area	Traffic accident area (city, country, district)
Type of accident	Types of accidents (Car-Car, Car-Human accidents, etc.)
Type of accident: subcategory	Type of behavior at the time of the accident (frontal collision, side collision, walking, etc.)
Road type	Type of accident road (highway, underpass, etc.)
Accident vehicle	Type of accident vehicle of the biggest fault or the smallest damage among the traffic accident (Truck, car, etc.)
Victim vehicle	Type of accident vehicle of the fewest faults or the biggest damage among the traffic accident (car, passenger, etc.)

C. Data Preprocessing

In this portion, the method of preprocessing the data collected for analysis is defined. Since the currently available data provide information only for the occurrence of a traffic accident, the scale of the accident could not be determined. Therefore we are creating an' severity' variable in this paper, which measures the scale of an injury so as to quantify the scale of the accident and to easily determine the degree of accident. The' severity' term is used to calculate the incident rate. We weighed for this purpose the variables ' Number of deaths,'' Number of seriously injured persons,'' Number of slightly injured persons' and' Number of injured persons.'

Therefore, we establish a vector' Accident scale' in order to better understand' Severity.' It derives from the attribute ' Large-scale Unfelt' when the severity is higher than 50. If it is higher than 20, it derives from the' Middle accident' and if it is less than 20, it derive from the' Small-scale Unfelt.'

V. THE RESULT OF DATA ANALYSIS

We evaluate the data and explain the findings in this section. First of all, we performed mathematical research focused on the pre-processed data attributes. The results of the study of the traffic incident form and the most crash happened during the crossing as seen in Figure 2.

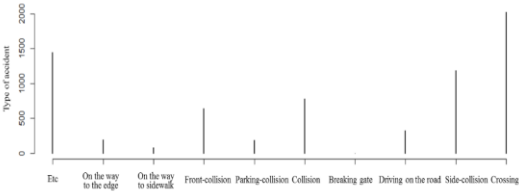

Figure. 2. Traffic accident analysis by type of accident.

Table 2 shows the results of the' Rate of Incident' traffic accident review extracted from segment 4.1.2. Table 2 reveals that there is more small-scale injury than large-scale crash. Large-scale collisions are the least incidents, but because of the number of deaths and casualties, they should be vigilant. The main reason for traffic accidents is also considered to vary depending on the scale of the accident. Therefore we evaluate the traffic accident according to the severity of the incident in this article.

Table 2. The result of traffic accident analysis by 'Scale of accident'.

	Small-scale Accident	Medium-scale Accident	Large-scale Accident
Number of cases	6,738	103	15

A. Large-scale Accident Analysis

In this section, we analyze the major accident that is more than 50. This incident is a significant event and does not occur frequently. We set the minimum support for the experiment to 0.4 and the reliability to 0.8. The more common associate rules indicate that the accident time is during the day and the cars are the type of victim. This means that many large-scale accidents occur during the daytime because the number of passengers in the vehicle during the day is longer than at night. And there are many major accidents where' ongoing crash' is a type of accident. Furthermore, if people do not comply with the safety obligation, many major accidents occur.

B. Medium-scale Accident Analysis

They discuss in this segment the medium-sized crash with a frequency greater than 20 and less than 50 accidents. First, a medium-scale accident occurred more during the day than during the night. But it only happened 1, 6 times during the day rather than at night, so there is no big difference. This means that medium-sized accidents often occur irrespective of time of the day. Therefore, all crash and survivor vehicles are' passenger cars,' which implies that the number of passengers is less than the amount of passengers in the car but the number of accidents is greater.

C. Small-scale Accident Analysis

In this section, we analyze the small accident that is less than 20. Small-scale injuries are most likely because the number of victims and fatalities is smaller than other incidents. For small-scale collisions the odd argument is that it occurs more during the night than during the day. Although this is a slight difference, it means that small-scale incidents mostly happen at night, while other accidents happen more in daytime. Additionally, the perpetrator is presumably a' federal' rather than a car. It indicates that many accidents occur between car and pedestrians rather than between cars in the event of a small-scale collision. Table 3 shows some of the results from large-scale accident data in order to find association rules. The law with stronger support and trust is the more related norm.

Table 3. The result of traffic accident analysis

lhs	**rhs**	**support**	**confidence**
{daytime}	{Large-scale accident}	0.933	1
{Accident vehicle: Van}	{Large-scale accident}	0.6	1
{Driving on the road}	{Large-scale accident}	0.533	1

{daytime, do not following the Safety driving obligation}	{Large-scale accident}	0.533	1
{Cross road}	{Large-scale accident}	0.4	1
{daytime, Victim vehicle: Passenger car}	{Large-scale accident}	0.4	0.857

VI. CONCLUSIONS

The prediction models of traffic accidents are extremely important instruments used by road safety initiatives, government authorities, the military, health departments, road safety educational institutions, cars and drivers ' instruction. These can be used to forecast the likelihood of collisions and the factors contributing to the transportation policies. The World Health Organization (WHO) estimates that traffic accidents are listed as the world's ninth-largest cause of death, and are the world's leading cause of death from 1-29 years. Every year, approximately 1.25 million people are killed and about 50 million more wounded by road accidents. Despite this phenomenon, the paper proposed various kinds of models for forecasting incidents in transport to better understand the techniques and their leading risk factors for accidents

REFERENCES

[1] .Leden, L. (2002). Pedestrian Risk decreases with pedestrian flow. A case study based on data from signalized intersections in Hamilton, Ontario – Accident Analysis and Prevention, Vol.
34, p. 457-464.

[2] World Health Organization (2015) Global Status Report on Road Safety

[3] NHTSA—National Center for Statistics and Analysis (NCSA) (2016) NHTSA Studies Vehicle Safety and Driving Behavior to Reduce Vehicle Crashes. http://www.nhtsa.gov/NCSA

[4] Beirness, D.J. and Beasley, E. (2011) A Comparison of Drugand Alcohol-Involved Motor Vehicle Driver Fatalities. Canadian Centre on Substance Abuse, Ottawa.

[5] World Bank (2015) The World Bank-Transport for Development. http://blogs.worldbank.org/transport/whyvehicle-

[6] NHTSA—National Center for Statistics and Analysis (NCSA)
(2016) NHTSA.

[7] Glenberg, A. (1996) Learning from Data: An Introduction to Statistical Reasoning. 2nd Edition, Lawrence Erlbaum Associates, Mahwah.

[8] Gelman, A. and Hill, J. (2007) Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, London.

[9] Kim, D.G., Lee, Y., Washington, S. and Choi, K. (2007) Modeling Crash Outcome Probabilities at Rural Intersections: Application of Hierarchical Binomial Logistic Models. Accident Analysis and Prevention, 39, 125-134.

[10] Abdulhafedh, A. (2016) Crash Frequency Analysis. Journal of Transportation Technologies, 6, 169-180.

[11] Blincoe, J., Miller, R., Zaloshnja, E. and Lawrence, A. (2015) The Economic and Societal Impact of Motor Vehicle Crashes, 2010. National Highway Traffic, Washington DC.

[12] Park, S. and Lord, D. (2007) Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. Transportation Research Record, 2019, 1-6. https://doi.org/10.3141/2019-01

[13] Ma, J., Kockelman, K.M. and Damien, P. (2008) A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, Using Bayesian Methods. Accident Analysis and Prevention, 40, 964-975.

[14] El-Basyouny, K. and Sayed, T. (2009) Collision Prediction Models Using Multivariate Poisson- Lognormal Regression. Accident Analysis and Prevention, 41, 820- 828.

[15] Hilbe, J. (2007) Negative Binomial Regression. Cambridge University Press, London.