CLASSIFICATION OF TRAFFIC ACCIDENT PREDICTION MODELS: A REVIEW PAPER

[1]ALI AHMED MOHAMMED, [2]KAMARUDIN BIN AMBAK, [3]AHMED MANCY MOSA, [4]DEPRIZON SYAMSUNUR

Faculty of Civil and Environmental Engineering, University Tun Hussein Onn Malaysia(UTHM), Smart Driving Research Center (SDRC), 86400 Parit Raja, Batu Pahat, Johor, Malaysia. Faculty of Civil and Environmental Engineering, University Tun Hussein Onn Malaysia(UTHM), Smart Driving Research Center (SDRC), 86400 Parit Raja, Batu Pahat, Johor, Malaysia. Civil Engineering Department , Al-Mansour University College, Baghdad , Iraq. Faculty of Civil Engineering Department UCSI University, Jalan Menara Gading 1, Taman Connaught, 56000 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur.
E-mail : hf160001@siswa.uthm.edu.my, kamardin@uthm.edu.my

Abstract - Accident prediction models (APMs) are extremely important tools for estimating the expected number of accidents on entities such as intersections and street segments. These estimates are typically used in the identification of sites for possible safety treatment and in the evaluation of such treatments. An APM is, in principally, a mathematical equation that expresses the average accident frequency of a site as a function of traffic flow and other site characteristics. whilst, the credibility of an APM is enhanced if the APM based on data as many years as possible especially if data for those same years are utilized in the safety analysis of a site. This paper covered a review as many papers as possible and various gaps in research along with a future possibility of study in this area have been indicated. Several models were discussed in this paper such as multiple linear regressions, Poisson regression, Conway-Maxwell Poisson regression models, artificial neural networks and fuzzy logics.

Index Terms - Traffic Accidents Prediction, Multiple Linear Regression, Poisson Regression, Literature Review, Artificial Neural Networks.

I. INTRODUCTION

Road traffic accidents continue to be a major problem in the world, both from the public health and socio-economic perspectives [1]. Throughout the world, cars, buses, trucks, motorcycles, pedestrians, animals, taxis and other categories of travelers, share the roadways, contributing to economic and social development in many countries [2]. Yet each year, many vehicles are involved in accidents that are responsible for millions of deaths and injuries. Globally, every year, about 1.25 million people are killed in motor vehicle crashes and approximately 50 million more are injured. Following current trends, about two million people could be expected to be killed in motor vehicle crashes each year by 2030 [2]. Currently, road accidents are ranked as the ninth most serious cause of death in the world, and without new initiatives to enhance road safety, fatal accidents will likely rise to the third place by the year 2020 [3]. In developed countries, road traffic accidents rates have decreased since the 1960s because of successful interventions such as seat belt safety laws, enforcement of speed limits, warnings about the dangers of mixing alcohol consumption with driving, and safer design and use of roads and vehicles. For example, road traffic fatalities have declined by about 25.0 percent in the United States from 2005 to 2014 and the number of people injured has decreased 13.0 percent from 2005 to 2014 [3]. In Canada, the number of road traffic accidents has declined by about 62% from 1990 to 2014, and the number of injuries has declined by about 68% during the same period [2]. However, traffic accidents have increased in developing countries from 1990 to 2014; for example, number of road traffic accidents increased by 44% in Malaysia and by about 43% in China.

Developing countries bear an extensive offer of the burden, accounting for 85% of annual deaths and 90% of the disability-adjusted life years. More than one-portion of all street movement passing's all around include individuals ages 15 to 44; amid their most beneficial gaining years. Moreover, the disability burden for this age group accounts for about 60% of all disability-adjusted life years. The costs and consequences of these losses are significant. Three-quarters of every poor family who lost a part in a traffic accident reported a decrease in their standard of living, and around 61% revealed borrowing cash to cover costs following their misfortune [4]. Although transportation agencies often try to identify the most hazardous road sites, and put great efforts into preventive measures, such as illumination and policy enforcement, the annual number of traffic accidents has not yet significantly decreased. For instance, 35092 traffic fatalities were recorded in United States of America during 2015; an increase of 7.2% compared to the previous year [5, 6]. The fatality rate per 100 million vehicle miles traveled (VMT) increased by 3.7% between 2014 and 2015. Thirty-Five States had more motor vehicle fatalities in 2015 than in 2014 Except November, every month

witnessed increases in fatalities from 2014 to 2015; the highest increases occurred in July and September [6].

II. CLASSIFICATION OF ACCIDENT PRIDICTION MODELS

The following subsection presents a number of mostly used models.

A. Multiple Linear Regression

Early models of traffic accident models based on the simple multiple linear regression approach assuming normally distributed errors. The general form of the linear crash prediction model can be expressed as follows:

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

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 multiple linear regression models have been generally connected, it has been discovered that crash event can frequently be better fitted with a Poisson circulation. One frequent pitfall is to model accidents data as continuous data by applying an ordinary least square regression [7]. This approach is inappropriate because regression models can produce predicted values that are non-integers and can also predict values that are negative, both of which are inconsistent with continuous data modeling. In addition, many distributions of crash data are positively skewed with many observations in the data set having a value of zero [8]. The high number of zeros in the data set prevents the transformation of a skewed distribution into a normal one which is a requirement of normal distribution [9]. An alternative is to use a Poisson distribution or one of its variants. Poisson distributions have a number of advantages over an ordinary normal distribution including a skew, discrete distribution, and the restriction of predicted values to non-negative numbers [7]. Whereby, generalized linear modeling variates of the Poisson regression model have been proposed to explore the relationship between the risk factors and traffic accident modeling [10, 11, 12, 13, 14]. Poisson regression has been applied to a wide range of transportation count data, including crash frequency. A Poisson regression model is similar to an ordinary linear regression with two exceptions. To begin with, it assumes that the errors follow a Poisson (not normal) circulation. Second, rather than modeling the response variable Y as a linear function of the regression coefficients, it models the natural log of the response variable ln(Y) as a linear function of the coefficients [7]. The Poisson model can be expressed as follows:

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

where,

P (ni): the probability of n crashes occurring on a highway segment i .
ni : the number of observations per time period (such as a year),
λi: the expected crash frequency on road segment i per time period (i.e. the mean of distribution) which can be estimated as follows:

$$\lambda i = EXP(\beta Xi) \quad (3)$$

where
Xi: a vector of the independent variables (i.e. risk factors),
β: a vector of the estimates (coefficients) of the independent variables Xi.
This model is estimable by standard maximum likelihood methods, with the log likelihood (LL) function given as:

$$ll(\beta) = \sum_1^n [-EXP(\beta Xi) + n(\beta Xi) - Ln(n!)] \quad (4)$$

One assumption of Poisson models is that the mean and the variance are equal, an assumption that is sometimes violated [14]. This can be dealt with by using a dispersion parameter if the difference is small, or by using a negative binomial regression model if the difference is large [15].

C. Conway-Maxwell Poisson Regression Models

The Conway-Maxwell Poisson model has been recently investigated with respect to transportation safety issues, but its application in accidents frequency modeling has been rather limited [16]. Generalized additive models have been explored given that they can provide smoothing functions for the explanatory variables [16]. The Conway-Maxwell Poisson circulation is a speculation of the Poisson circulation that can handle both under-dispersed and over-dispersed accidents information. The fundamental preferred of this model is to handle the under-dispersion in crash data that cannot be modeled by the Poisson model or the Negative Binomial model. However, the low sample-mean and small sample size of the under-dispersed accidents data can influence the estimated parameters, and therefore, it has been limited in the application of accidents frequency [17]. Moreover, in practice, the estimation of these models can become very difficult as they require more parameters, a problem that has likely impeded their application to accidents frequency prediction [17, 18].

D. Artificial Neural Networks and Fuzzy Logic models

A linear function may not sufficiently expound the relationship between the dependent variables and the associated independent variables in accidents modeling, non-linear approximators such as fuzzy logic and neural networks have also been explored. Artificial Neural Networks (ANNs) are a class of computational intelligence tools that can be used for prediction and classification problems. ANNs can model very complex non-linear functions to high reliability levels using a process of learning that is similar to the learning procedure of the cognitive system in the human brain. The network body is composed of input layers, hidden layers, and output layers [10, 11]. These models can be trained to approximate any nonlinear function to a required degree of accuracy using a learning algorithm (such as back propagation) that would give the desired output, in a supervised learning process. ANNs have some advantages over the statistical models [11]. For instance, regression models need a predefined relationship or functional form between the dependent variable (crash frequency) and the independent explanatory variables that can be estimated by some statistical approaches whereas the ANNs do not require the establishment of these functional forms and can be easily applied in the analysis [13]. Furthermore, the ANNs differ from the statistical models in that they behave as black-boxes and do not provide interpretation for the parameter estimates [14]. Fuzzy logic applications have increasingly been proven to have some significant accidents predicting capability in recent years [13, 14, 15, and 16]. Fuzzy logic system is defined as the nonlinear mapping of an input data set to a scalar output data, and the first step of the process (known as fuzzification) consists of gathering a crisp set of input data that will be changed to a fuzzy set utilizing fuzzy linguistic variables, fuzzy linguistic terms, and enrollment capacities [18]. After that, an inference is made based on a set of fuzzy rules, and then, the resulting fuzzy output is mapped to a crisp output using the membership functions, in the defuzzification step [18, 19].

III. ADVANTAGES AND LIMITATIONS OF USED MODELS

Table 1 abstracts the advantages and limitations of regression models

Table I: Advantages and Limitations of each model

Model type	Advantages	Limitations
A. Multiple Linear Regression	❖ ability to determine the relative influence of one or more predictor variables to the criterion value. ❖ ability to identify outliers, or anomalies.	❖ Disadvantages: Outputs of regression can lie outside of the range [0,1]. ❖ It has limitations in the shapes that linear models can assume over long ranges. ❖ The extrapolation properties will be possibly poor. ❖ It is very sensitive to outliers It often gives optimal estimates of the unknown parameters.
Poisson	Most basic model; simple to estimate	negatively influenced by the low sample-mean and small sample size bias; Cannot handle over- and under-dispersion
Conway Maxwell Poisson	be able to handle under- and over dispersion or combination of both using a variable dispersion parameter	No multivariate extensions available to date; Could be negatively influenced by the low sample-mean and small sample size bias
Artificial Neural Networks and Fuzzy Logic models	❖ First check really ANN is required to model the given problem. ❖ The model need not assume any model structure before starting the ANN model. ❖ It can be used for non-linear problems. ❖ It is a non-parametric method, thus eliminates the error in parameter estimation.	It cannot extrapolate the results

CONCLUSIONS

Traffic accidents prediction models are extremely important tools in road safety programs used by transportation agencies, police, health departments, education institutions that oversee road safety, vehicles, and the driver's education. They can be used to predict both the frequency of accidents occurrence and the contributing factors that could then be addressed by transportation policies. According to the world health organization (WHO), road crashes are ranked as the ninth most serious cause of death in the world, and present the world's leading cause of death for individuals between the ages of one and twenty-nine. Each year, traffic accidents are responsible for killing about 1.25 million people and injuring approximately 50 million more. Given such trend, this paper presented different types of traffic accidents prediction models to gain a better understanding of the techniques used to predict accidents and their contributing risk factors

ACKNOWLEDGMENTS

The author of this work would like to express their Serious thanks and appreciation for UTHM laboratory technician Smart Driving Research Center (SDRC) for their valuable support.

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 2015. http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/.

[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 Drug- and 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/why-vehicle- safety -matters -crash-related-deaths ?cid=EXT_WBBlogSocialShare_D_EXT.

[6] NHTSA—National Center for Statistics and Analysis (NCSA) (2016) NHTSA. https://www.nhtsa.gov/press-releases/us-dot-announces-steep-increase-roadway-deaths-based-2015-early-estimates

[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.https://doi.org/10.1017/CBO9780511811852.

[16] Geedipally, R., Lord, D. and Dhavala, S. (2012) The Negative-Binomial Lindley Generalized Linear Model: Characteristics and Application Using Crash Data. Accident Analysis and Prevention, 45, 258-265.

[17] Lord, D. (2006) Modeling Motor Vehicle Crashes Using Poisson-Gamma Models: Examining the Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter. Accident Analysis and Prevention, 46, 751-766.

[18] Li, X., Lord, D., Zhang, Y. and Xie, Y. (2009) Predicting Motor Vehicle Crashes Using Support Vector Machine Models. Accident Analysis and Prevention, 40, 1611- 1618.

[19] Meng, H., Zheng, L. and Qing, M. (2009) Traffic Accidents Prediction and Prominent Influencing Factors Analysis Based on Fuzzy Logic. Accident Analysis and Prevention, 9, 87-92.

★★★