Indian Journal of Science and Technology, Vol 11(5), DOI: 10.17485/ijst/2018/v11i5/119291, February 2018
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Review of Crash Prediction Models and their Applicability in Black Spot Identification to Improve Road Safety

Mohammad N. Al-Marafi* and Kathirgamalingam Somasundaraswaran

School of Civil Engineering and Surveying, University of Southern Queensland, Toowoomba, Queensland 4350, Australia; U1083416@usq.edu.au, Kathirgs@usq.edu.au

Abstract

Objective: This study aims to review of the development of crash prediction models, and their applications to analyse and identify black spots to improve road safety. **Methods:** Several modelling techniques have been reviewed in this study including, multiple linear regression, Poisson distribution, negative binomial, random effect technique, and multiple logistic regression models to identify their suitability to develop the crash prediction models. The studies related to the identification of black spots were also reviewed based on the type of crash data used in the identification process. **Result:** The reviewed documents highlight the shortcomings within the traditional crash prediction models (CPMs), as well as the demonstrated the flexibilities and effectiveness of the latest methods. Most suitable models can now be developed to represent the actual scenarios from several modelling techniques, where they provide a realistic and accurate prediction of crash frequency, for example, to determine if the location had a traffic safety problem compared to other locations with similar conditions and to identify the suitable measures to reduce crashes. **Application/Improvements:** The models identified in this research are already being used but the modelling approaches can be further modified to include the latest technical application on roads, available post-crash management system or safety culture which are commonly related road safety outcomes.

Keywords: Black Spots, Crash Prediction, Review, Road Crashes, Road Safety

1.Introduction

Worldwide, nearly 1.3 million people die annually, and about 20 to 50 million people sustain injuries as a result of road crashes, mostly in low and middle income countries[1]. This number of fatalities and injuries has a huge impact on the families affected, whose lives are often changed irrevocably due to these tragedies. Road crashes cannot be completely prevented, but through appropriate traffic engineering remedial actions and management approaches the crash rates can be reduced to certain acceptable societal limitations[2]. Proper verification of the cause of crashes can help to suggest preventive and corrective measures in terms of traffic control and road design at potential crash locations, so for this reason, systematic studies of traffic crashes are required to be undertaken. Traffic safety agencies in the past used various key measures such as the crash rate to traffic volume and/or the absolute crashes number at a location, to see whether that location had a traffic safety problem or not. Later on, the comparison with other locations of similar traffic and geometric conditions helped to improve the investigations. Thus, traditionally focuses have given on predicting the relation between the traffic crash frequency and other contributing variables which then propose strategies to alter the shape of traffic volume and road geometry. However, these methods tend to be subjective, short sighted, and an outdated view on road safety[3-5]. The main

*Author for correspondence

reason for this could be the availability of data for developing the models. However the systematic integration of crash data with safety policy and focuses on the capability of collected information to meet the requirements to develop the remedial actions helped the development of new modelling approaches.

The latest approaches proved that the appropriate relationships can be developed by using suitable statistical models to provide a realistic and accurate prediction of crash occurrences. In addition, these models have assisted road safety investigation in many ways including the identification of the factors contributing to crashes and estimating the potential crash frequency on a part of highway road network, for example, rural or urban segments and intersections. As a result, the techniques of developing crash prediction models and their applicability to identify the black spots locations have been improved and several shortcomings within these traditional methods have been removed. This study reviews the various types of crash prediction models to highlight the trends in modelling as well as their application in black spot identification to improve road safety. This review was conducted as a part of research study towards developing more sophisticated models towards breaking the barriers.

2.Crash Prediction Models

Road crashes are caused by a combination of many factors, including the roadway, the roadway environment, vehicles and road users' behaviour. Crash Prediction Models (CPMs) have been employed as useful tools by road engineers and planners to identify the reasons hence to propose remedial actions to improve road safety. Over the last two decades, substantial research has been conducted on the development of CPMs for the estimation of the predicted crashes or crash rates on road network[6-10]. In this process, several modelling techniques have been used in crash prediction models including, multiple linear regression, Poisson distribution, negative binomial, random effect technique, and multiple logistic regression models. A review of key studies that used these models are documented to provide below

2.1. Multiple Linear Regression Models

The multiple linear regressions modeling technique has been used in several fields including engineering, health, agriculture and economics, to model the relationship between two or more explanatory variables and an outcome variable by fitting a linear equation to observed data[11]. Detailed the creditability of the multiple linear regression models to describe relationships between continuous outcomes and explanatory variables. Although multiple linear regression models are used widely in road crash studies, they have limitations to describe adequately the random, non-negative, discrete, and typically sporadic events, which are all characteristics of road crashes[6,12-16] developed multiple linear regression models to investigate the effect of the roundabout geometry features on road crashes in urban and rural areas of Queensland, Australia. A total of 492 crashes and 100 roundabouts on urban and rural roads were studied. Data for this study on geometric condition, traffic volume, traffic control, and crashes were collected between 1986 and 1990. Three models were proposed to fit varying crash types (single vehicle, entering-circulating collisions, and approaching rear-end collisions). This study concluded with recommendations for the design and construction of roundabouts to minimize the number of crashes.

2.2. Poisson Distribution Models

Since crash occurrences are unavoidable, discrete and more likely random events, So many researchers argue that Poisson regression model appears to be more appropriate than multiple linear regression models[8]. Developed CPMS for urban intersections and road segments in Denmark by using Poisson distribution. The study found that additional explanatory variables including speed limit, road environment, parking facilities, number of minor side roads and number of exits per km proved to be important and significant explanatory variables for estimating the number of crashes in road segments. For intersections, however, the most significant variables in the models were those describing the traffic flow[16]. Stated that Poisson models have some limitations. One of these limitations is that the mean must equal the variance of road crash number (dependent variable). In most crash data, the variance value of the road crash number exceeds the mean value and, in such case, the data would be over-dispersed. Thus, some authors[13,14] recommend using alternative methods to solve the limitation of over-dispersion imposed through the Poisson regression model.

2.3. Negative Binomial Models

Negative Binomial (NB) regression method is considered as an alternative which does not require the equal mean

and variance assumption[17]. Attempted to define CPMs for two-lane rural road sections based on a combination of geometry, exposure, context and consistency variables related to the road safety performance. The roads considered were two-lane local rural roads, with a five-year crash analysis period to compensate for the low traffic flow and crash frequencies expected on local roads. The models suggested are also based on the generalized linear modelling (GLM) approach, assuming a negative binomial distribution error structure. In conclusion, three of the examined models were considered appropriate, based on practical considerations, statistical significance, and goodness of fit indicators[6]. Developed the generalized linear model (GLM) with negative binomial error structure to study road crashes on rural highway segments in Ghana. Data from 2005-2007 was collected from 76 segments with each segment ranging between 0.8 and 6.7 km. The results found that increased road segment length, increased traffic density, and decreased terrain tended to increase the probability of a crash. On the other hand speed, road curvature, and shoulder and road width were not found as statistically significant risk factors for road crashes.

It is also apparent that for some special purposes the Poisson distribution is still a fairly good representation in those cases where the negative binomial is suspected or known to apply. In a study by[18] lognormal, Poisson and negative binomial regression analyses were used to develop statistical models to investigate the relationship between road crashes and highway geometry, traffic control, and traffic volume variables for at-grade intersections. This study concluded that the Negative Binomial (NB) and lognormal distributions were more appropriate for the modeling of the crash frequency than the normal distribution. Similarly[19] modelled the relationship between road crashes and road geometry, traffic volumes, roadside hazards, road surfacing, cross-section and driveway density for two-lane rural roads in New Zealand. The results indicate that CPMs using a combination of models provide a good understanding how safety is affected by these variables.

2.4. Random Effect Models

The random effect technique assumes that road crash data is hierarchical in nature. The hierarchy in road crash data is proposed as follows: the lowest level of the hierarchy represents the crashes themselves, while the type of location on road network at which the crash occurred represents the higher-level hierarchy. In this type of model, the main assumption is that an association may exist among crashes occurring at the same location, so these crashes may share unobserved or unrecorded characteristics related to the location. Over the last decade, this hierarchical modeling technique has been gaining an increasing amount of attention in accounting for the multilevel data structure in crash prediction. These unobserved characteristics might include low pavement friction, poor pavement condition, poor reflectivity of road signs, and other similar factors[12]. The results from this technique may not be transferable to other data sets because the results are observation specific[14].

In[12] Developed a random-effects model to evaluate the relationship between the frequency of road crash and geometry, traffic and control characteristics, hence estimated the expected crash number at 52 signalized intersection in Malaysia. The results indicated that total approach volume, right-turn volume, uncontrolled left-turn lane, acceleration section, sight distance, median width, number of bus stops and bays, presence of surveillance camera, number of phases and signal control type significantly impact on the safety at the intersections. In the same study, the authors concluded that three of these variables including use of adaptive signal control, presence of bus bays and presence of an acceleration section tend to lower the crash frequency, while the remaining variables leading to increase it.

2.5Multiple Logistic Modeling

The multiple logistic regression technique is used to analyze only crash binary outcomes, meaning the value of the dependent variable ranges between 0 and 1. For example, this technique can be used to build a model to provide a measure of the probability of injury or non-injury crash outcomes. However, there are many studies in which crash outcomes are continuous (e.g., number of total crashes)[20]. Used this method to investigate the factors that significantly impact on intersection crashes involving injuries in Victoria, Australia, between 2000 and 2009. The results showed seven factors significantly related to the severity of intersection crashes, including speed zone, driver gender and age, time of day, seat belt usage, traffic control type, and crash type[21]. Used a binary logistic regression model to identify the main factors that effected on road crash severity. This study conducted in Kanas city con-

Table 1. Summary of models for analysing crash-frequency data

Model Type	Previous Research used or discussed this type	Advantages	Disadvantages
Multiple Logistic	Ackaah and Salifu[6], Chen et al.[20], Dissanayake and Roy[21], Kim et al.[13], Kutner et al.[11], and Montella et al.[14]	Suitable to study the effect of one variable while controlling for other variables[1]	used to analyze only crash binary outcomes
Multiple Linear	Arndt and Troutbeck[15], Chin and Quddus[12], Kim et al.[13], and Mustakim and Fujita[24]	Easy to estimate crash number	Unable to describe adequately the random, non-negative, discrete, and typically sporadic events.
Random Effects	Chin and Quddus[12] and Lord and Mannering[4]	Handle spatial correlation[2]	The results from this technique may not be transferable to other data sets because the results are observation specific
Poisson	Abdel-Aty and Radwan[16], Bauer and Harwood[18], Chin and Quddus[12], Greibe[8], andLord and Mannering[4]	Handle with unavoidable discrete and more likely random events	Cannot handle over- and under-dispersion(the mean must equal the variance of crash number)
Negative Binomial	Abdel-Aty and Radwan[16], Ackaah and Salifu[6], Bauer and Harwood[18], and Usman et al.[25]	Does not require the equal mean and variance assumption, able to describe adequately the random, non-negative, discrete, and typically sporadic events.	Cannot handle with small sample sizes

[1]In logistic regression the coefficients derived from the model (e.g., b1) indicate the change in the expected log odds relative to a one unit change in X1, holding all other predictors constant.
[2]Crashes occurring at the same location may share unobserved or unrecorded characteristics related to the location.

cluded that some of the significant variables which affect the probability of road crashes are asphalt road surface, speed, alcohol involvement, older driver, medical condition of the driver, daylight, type of vehicles, and fixed object types such as trees. However a few more detailed literature on CPMs including their advantages and disadvantages are summarised in Table 1.

3.Analysis in Black Spot Identification

In[22] Defined black spot as any location that has a higher predicted number of road crashes than other similar location as a result of local risk factors existing at the location. Identification of black spots, sometimes known as high-risk locations, hazardous road locations (HRL), hotspots, or crash-prone situations, is considered as the first step in the road crash reduction process. In general, the identification of black spots is divided into two main approaches based on the type of crash data used in the identification process.

3.1. Numerical Approach

The first approach depends on historical crash data. This method defines the black spot as the location which has a higher than average crash number, crash frequency (crash per year or crash per kilometer) or crash rate (crash per vehicle).

In[23-25] used crash frequency to identify the black spots whereas a road section is considered to be a black spot, from the crash frequency point of view, if: $A_j > A_c$, where:

$$A_c = F_{ave} + k_a \sqrt{\left(F_{ave} / L_j\right) - 0.5 / L_j} \quad (1)$$

Ac is a critical value for crash frequency, A_j is a number of crashes on segment j during a certain time period, L_j is a length of segment j, F_{ave} is the average crash frequency for all segments, and K_a is a constant that is selected for the significance test.

3.2. Model-Based Approach

The second approach is a model-based definition which depends on analyzing each site location by applying statistical models to identify black spots[26]. According to[27] the identification of hazardous locations represents a list of spots being prioritized for further researches of engineering which can distinguish road crash patterns, contributing factors and potential resolution. Furthermore, in these processes, cost-effective projects are often selected to get the best results from limited resources.

In[28] Investigated the possibility of using crash prediction models for the identification of black spots. The geometric and traffic characteristics of secondary rural roads in South Moravia were used in this study. The generalized linear model was used to determine the expected number of crashes for individual types of road segments. A critical road segment is defined as a segment where the observed number of crashes significantly exceeds the number of expected crashes on roads with similar geometric and traffic characteristics. The results indicated the possibility of using this method as an effective tool for road safety management.

In[29] Investigated the performance of three statistical models: Poisson lognormal, heterogeneous negative binomial, and traditional negative binomial model for ranking locations for road safety improvement. The authors compared these models for the identification of black spots based on the performance and practical implications. This study concluded that the choice of model assumptions and ranking criteria can lead to different lists of black spots.

In[24] Used 4 year's crash data from rural roadways to rank the black spots in Malaysia based on a crash point weightage formula as shown below:

$$CPW = X_1(0.6) + X_2(0.3) + X_3(0.8) + X_4(0.2) \quad (2)$$

In Equation (2), is the number of fatal, is the number of serious injury, is the number of slight injuries, and is the number of damage only. This study applied the Multiple Linear Regression method for developing a model which relates crash point weightage to rank the black spot locations.

In[30] stated that the best method to determine black spots is the expected crashes frequency, not the recorded crashes. At the same time, the combination of the recorded crashes number and the model estimate for that site is the best method to estimate the expected crashes frequency. A suitable technique to do this is to apply the empirical Bayes (EB) method[31]. Examined the ability to use the Sichel (SI) model in calculating empirical Bayes (EB) estimates. In order to accomplish the objective of this study, the SI model with a varying dispersion term and NB model were developed using the road crash data collected at 4-lane undivided rural highways in Texas. Results found that the selection of crash prediction model (i.e., the NB or SI model) will affect the value of the weighting adjustment factor used for calculating the EB outputs, and the determination of black spots by using the EB method can be different when the SI model is used. According to by calculating the weighted combination of the recorded and predicted crash frequency, the EB-adjustment technique is able to provide an expected crash frequency for a particular roadway segment or intersection.

4. Conclusions

This study comprehensively reviewed the simplified statistical models used for predicting road crashes and hence their application to identify the black spot locations to improve the road safety in urban and rural roads. Several studies show that the relationship between road crash frequency and explanatory variables are having a very good relationship. This finding has further led most road safety researchers to use statistical models in which the dependent variable is the crash frequency. The statistical methods such as Poisson and Negative binomial regression have traditionally been used as suitable techniques used for developing road crash models. This is due to the ability of these techniques to analyse data while preventing the possibility of having a negative integer crash value over some time period. At the same time, the selection of explanatory variables in most of the reviewed models has shown that the variables were included in the road crash models without an appropriate variable selection procedure. This means that the selection of the variables is done on a subjective basis (based on the availability of data) which might lead to biased results. So, the use of a variable selection procedure is useful to minimize such bias and misleading results.

Applying a suitable crash prediction model is very important to engineers and transportation planner, because it can help in identifying the black spot locations that require treatment and as well as ranking the hazardous locations by calculating Potential for Safety

Improvement (PSI). In general, the CPMs and observed crashes do not account for Regression to Mean effect (RTM) associated with crash data. RTM is the tendency of crash data to regress back to the mean[5]. For instance, a site may have high crashes at a given period and low crashes the next period without any road safety implementations. In addition, a high-risk site may have a certain period of randomly low frequency crashes and therefore be overlooked during road safety evaluation. The Empirical Bayes (EB) approach has been introduced by researchers as a means of solving the RTM problem. This approach identifies high crash locations (black spots) based on their Potential for Safety Improvement (PSI), calculated as the difference between predicted and expected crashes at the location.

This study indicates that the traditional methods have now been replaced by more advanced modelling techniques to support the analysis for developing innovative counter measures to improve road safety. The system of data collection has also been flexed to meet the systematic integration of the data with the road safety strategies and policies. However the future domain needs to break the barriers in providing additional information such as available advanced technology and communication, reliability of post-crash management system, and culture of road safety to the location as some of the key contributory factors for future studies.

5. Acknowledgment

The authors thank the Tafila Technical University for its financial assistance and the University of Southern Queensland for its support of this research.

6.References

1. World Health Organization. Global status report on road safety. Available from: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/. Date accessed: 12/11/2017.
2. Ganguly R, Gupta AK, Mishra M. Traffic Volume and Accident Studies on Nh-22 Between Solan and Shimla, India. European Scientific Journal. 2014 Sep; 2:248-54.
3. Hauer E. On exposure and accident rate. Traffic Engineering & Control. 1995; 36(3):134-8.
4. Lord D, Mannering F. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice. 2010 Feb; 44(5):291-305. Crossref.
5. Tegge RA, Jo JH, Ouyang Y. Springfield, Virginia: Development and application of safety performance functions for Illinois. 2010; p. 1-181.
6. Ackaah W, Salifu M. Crash prediction model for two-lane rural highways in the Ashanti region of Ghana. IATSS Research. 2011 Jan; 35(1):34-40.
7. Jadaan KS, Foudeh L, Al-Marafi MN, Msallam M. Modeling of Accidents Using Safety Performance Functions. Antalya, Turkey: International Conference on Agriculture, Environment and Biological Sciences (ICFAE'14). 2014; p. 73-7.
8. Greibe P. Accident prediction models for urban roads. Accident Analysis & Prevention. 2003 Mar; 35(2):273-85. Crossref.
9. Hauer E, Ng JC, Lovell J. Estimation of safety at signalized intersections. Transportation Research Record Journal. 1988; 1185:48-61.
10. Mountain L, Fawaz B, Jarrett D. Accident prediction models for roads with minor junctions. Accident Analysis & Prevention. 1996 Apr; 28(6):695-707. Crossref.
11. Kutner MH, Nachtsheim CJ, Neter J, Li W. New York, McGraw-Hill Irwin: Applied linear statistical models, 4th edn. 2005; p. 1-1415.
12. Chin HC, Quddus MA. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Analysis & Prevention. 2003 Mar; 35(2):253-9. Crossref.
13. Kim SH, Chung SB, Song KH, Chon KS. Development of an Accident Prediction Model using GLIM (Generalized Log-linear Model) and EB method: A case of Seoul. Journal of the Eastern Asia Society for Transportation Studies. 2005; 6:3669-82.
14. Montella A, Colantuoni L, Lamberti R. Crash prediction models for rural motorways. Transportation Research Record: Journal of the Transportation Research Board. 2008; 2083:180-9. Crossref.
15. Arndt O, Troutbeck RJ. Relationship between roundabout geometry and accident rates. Transportation Research Circular. 1998; 28:1-16.
16. Abdel-Aty MA, Radwan AE. Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention. 2000 Sep; 32(5):633-42. Crossref.
17. Cafiso S, Di Graziano A, Di Silvestro G, La Cava G, Persaud B. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. Accident Analysis & Prevention. 2010 Jul; 42(4):1072-9. Crossref. PMid:20441815.
18. Bauer KM, Harwood DW. Statistical Models of At-Grade Intersection Accidents Addendum, Report. Georgetown

Pike: Midwest Research Institute. 2000; p. 1-68. PMCid:PMC1572297.
19. Turner S, Singh R, Nates G. The next generation of rural road crash prediction models: final report. NZ Transport Agency Research Report. 2012; p. 1-98.
20. Chen H, Cao L, Logan DB. Analysis of risk factors affecting the severity of intersection crashes by logistic regression. Traffic Injury Prevention. 2012 Feb; 13(3):300-7. Crossref. PMid:22607253.
21. Dissanayake S, Roy U. Crash severity analysis of single vehicle run-off-road crashes. Journal of Transportation Technologies. 2014; 4(1):1-10. Crossref.
22. Elvik R. A survey of operational definitions of hazardous road locations in some European countries. Accident Analysis & Prevention. 2008 Nov; 40(6):1830-5. Crossref. PMid:19068283.
23. Kent S, Hans E. Black spot manual. General directorate of highways, traffic safety consultancy services, Ankara. 2001; p. 1-82.
24. Mustakim F, Fujita M. Development of accident predictive model for rural roadway. World Academy of Science, Engineering and Technology. 2011; 58(10):126-31.
25. Usman T, Fu L, Miranda-Moreno LF. Quantifying safety benefit of winter road maintenance: Accident frequency modeling. Accident Analysis & Prevention. 2010 Nov; 42(6):1878-87. Crossref. PMid:20728638.
26. AASHTO. Highway safety manual. Washington, D. C.: American Association of State Highway and Transportation Officials. 2010; p. 1-108.
27. Hauer E, Kononov J, Allery B, Griffith M. Screening the road network for sites with promise. Transportation Research Record: Journal of the Transportation Research Board. 2002; 1784:27-32.
28. Senk P, Ambros J, Pokorny P, Striegler R. Use of Accident Prediction Models in Identifying Hazardous Road Locations. Transactions on Transport Sciences. 2012 Dec; 5(4):223-32. Crossref.
29. Miranda-Moreno L, Fu L, Saccomanno F, Labbe A. Alternative risk models for ranking locations for safety improvement. Transportation Research Record: Journal of the Transportation Research Board. 2005; 1908:1-8.
30. Elvik R. State-of-the-art approaches to road accident black spot management and safety analysis of road networks. Norway: Institute of Transport Economics. 2007; p. 1-126.
31. Zou Y, Lord D, Zhang Y, Peng Y. Comparison of sichel and negative binomial models in estimating empirical bayes estimates. Transportation Research Record: Journal of the Transportation Research Board. 2013; 2392:11-21. Crossref.