СРПСКО ДРУШТВО ЗА ПУТЕВЕ

Пут и саобраћај
Journal of Road and Traffic Engineering

www.putisaobracaj.rs

Road Crash Prediction Models: A Review of Methods and Applications

Tilahun Mintie Wubie [a, b], Abeje Tilahun Fetene[c]

[a] *Faculty of Civil and Water Resources Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia*
[b] *Department of Civil Engineering, Haramaya Institute of Technology, Haramaya University, Dire Dawa, Ethiopia*
[c] *School of Civil, Hydraulic and water resource Engineering, university of Gondar, Ethiopia*

ARTICLE INFO

DOI: 10.31075/PIS.70.04.02
Professional paper
Received: 26.11.2024.
Accepted: 22.12.2024.
Corresponding author: e-mail:
tilahunmintie97@gmail.com

ORCID ID
Tilahun Mintie Wubie: 0000-0002-9798-7280
Abeje Tilahun Fetene:0000-0003-4822-5180

Keywords:
Crash prediction models
Black spot analysis
Statistical crash modeling
Advanced Crash Modeling

ABSTRACT

Road traffic crashes are still the major road safety problem in the world causing for the deaths of more than 1 million people each year, although the problem is more serious in low- and middle-income countries. Therefore, road crash prediction models play an important role in road safety management in determining both the predicted crash frequency and the contributing factors that could then be addressed by transport policies. Many types of statistical crash prediction models have been proposed to estimate predicted crash frequencies in road networks, ranging from basic Poisson and negative binomial models to more complicated models, such as zero-inflated and Conway-Maxwell Poisson regression models. However, little effort has been made to assess the performance and practical implications of these models when they are used to identify black spot locations. The study aims to critically summarise the global experience on the development and application of CPMs to analyse and identify black spots for road safety improvements. To achieve the objective, several crash modelling techniques have been reviewed. The study also reviewed data and methodological issues in the development of crash prediction models including data collection methods, network segmentation, and selection of explanatory variables and the application of crash prediction models in black spot identification for road safety improvements. The study identified the limitations of the most traditional crash modelling techniques and examined the flexibilities and effectiveness of the latest crash modelling techniques.

1. Introduction

Road traffic crashes are a major issue that leads to an increase in property losses, injuries, and fatalities. They can provide serious health, financial, and developmental difficulties for drivers. As of 2019, road traffic crashes rank as the 12th most common cause of death overall and the leading cause of mortality for children and youth (ages 5 to 29). As compared to the 1.25 million road traffic deaths in 2010, there were an expected 1.19 million deaths in 2021, that means a 5% of traffic accident was decreased. The European Region has had a 36% fall in deaths, the biggest drop since 2010. The number of deaths has stayed constant in the Region of the Americas, with a 16 percent drop reported in the Western Pacific Region, and a 2 percent decline in the South-East Asia Region. However, in 66 countries. There was a rise of traffic accident; 28 of these countries are in the African Region, which has seen a 17% rise in the number of deaths since 2010. Additionally, over 92% of road traffic deaths and injuries occur in low- and middle-income nations, while no low-income country has seen a decrease in the number of road traffic fatalities since 2020 (WHO,2023). Because of successful interventions such as seat belt safety regulations, enforcement of speed limits, warnings about the dangers of mixing alcohol and driving, and safer construction and use of roads and cars, road traffic death rates in developed countries have dropped since the 1960s. For example, in the United States, road traffic fatalities have decreased by approximately 25.5% from 2005 to 2014, while the number of people injured has decreased by 13.0 percent (Abdulhafedh, 2017). This high number of deaths and injuries in LMIC has a significant impact on the families involved, whose lives are frequently irreversibly altered as a result of these catastrophes. Although road crashes cannot be completely avoided, they can be reduced to acceptable societal limits with suitable corrective activities and management approaches in traffic engineering (Ganguly et al., 2014).

Decision-making for road safety interventions is a complex procedure which involves a number of actors (experts, politicians, public, etc.) and issues (environmental, mobility, economical) that compete for a limited resources available. New knowledge and instruments are required for the detection of unsafe road sections and the calculation of the effects of safety (design) measures to allow efficient resource allocation. Crash prediction models (CPMs) can be used to identify relative dangerous road sections in a network and to estimate the effects of putting safety measures in place on those sections (Wegman, 2014). CPMs are used for a variety of purposes; most frequently to estimate the predicted accident frequencies from various roadway entities (highways, intersections, interstates, etc.) and also to identify geometric, environmental, and operational factors that are associated with the occurrence of accidents (Yannis et al., 2017). According to Ambros et al. (2018), they can be used in:

- Investigating and comparing different combinations of individual risk factors that make certain roads unsafe;
- Network safety screening: (identification of hazardous locations);
- Assessing safety of contemplated (re)constructions or safety treatments;
- Economic analysis of project (costs vs. safety benefits).

Several research studies have been conducted on the development of CPMs to estimate crash rates on the road. In this regard, multiple linear regression, Poisson distribution, negative binomial, Poisson log normal, zero inflated Poisson and negative binomial, Conway-Maxwell Poisson, multiple logistic regression and artificial neural network models were used in crash prediction models by different researchers.

Previously, road traffic crash analysis models were basically based on simple multiple linear regression methods, assuming that errors are normally distributed. However, researchers soon discovered that crash occurrence could be better fitted with a Poisson distribution. Hence, a Poisson regression and negative binomial models were soon adopted over conventional multiple linear regression techniques due to their ability to describe adequately the random, non-negative and discrete nature of crash data (Abdulhafedh, 2016, Basu and Saha, 2017). However, Poisson and negative binomial regression techniques cannot handle over dispersion which is a major statistical problem in road crash data. To address this problem, different advanced modeling techniques have been proposed by researchers as discussed in the following sections.

The AASHTO (2014) highway safety manual recommends the application of Safety Performance Functions (SPF) in predicting crash frequencies on different types of roads with various road environment conditions. SPFs allow calibration for the local conditions at the time of application. The HSM provides procedures for the calibration process and incorporates empirical Bayes (EB) method for the regression of the SPF model. However, the HSM is appropriate for homogeneous road sections in terms of traffic volumes, roadway design characteristics, and traffic control features.

The objective of this paper is to critically summarise the global experience in the development and application of CPMs. To achieve this objective, the literatures that have been done on CPMs were searched based on the following criteria:

- Sources: Books, Journal articles, conference papers and reports of agencies;
- Keywords: crash prediction, accident prediction, safety performance functions, road safety evaluation tools;
- Time frame: No time limit

2. Methodological Issues in Developing CPMs

Properties of road crash data and methodological difficulties have been widely discussed in recent years by researchers in the field of road traffic safety. Crash data have been shown to be sources of errors in crash modeling, resulting in erroneous predictions and incorrect conclusions about the components that cause a road crash (Mekonnen and Sipos, 2022). Therefore, care should be taken in the data collection, network segmentation, selection of explanatory variables, and selection of modelling techniques before the development of CPMs.

2.1 Data Collection

Road crash data collection and handling in many developing countries is very poor relative to the level of accuracy required for crash analysis. For example, the existing road traffic crash data collection system by police in Ethiopia needs many improvements and a standardised form for describing road crashes including vehicles and persons involved. Misclassification of crash severity, non-georeferenced crash spots, below standard crash data recording form, and underreporting of crashes are the major problems the road crash data collection system in Ethiopia is facing. In crash prediction modelling, lack of such information regarding the crash spot negatively influences the prediction rate and the reliability and interpretation of the crash prediction models (Bhavsar et al., 2020).

Theoretically, to develop efficiently representative crash prediction models, a random sample data should be taken from similar road types or intersections. In this case, several authors recommended the minimum sample sizes in different ways. The AASHTO Highway Safety Manual (2014) recommends examining a sample of 30–50 locations with a total of at least 100 crashes per year. Other authors recommended minimal sample

sizes such as at least 200 crashes (Jonsson, 2005) and 300 crashes (Srinivasan and Bauer, 2013). Many others, on the other hand, were criticising the one-size-fits-all approach. One of them was Lord (2006) who recommended a sample size dependent on the sample mean, such as 200 segments for an average of 5 crashes per segment and 1000 segments for an average of 1 crash per segment. Although road crash data sets are mostly characterised by small sample size and low sample mean due to too many zero-crash records and high skewness (Mekonnen and Sipos, 2022).

Therefore, care should be taken when selecting the suitable modelling technique that can handle these statistical problems associated with road crash data.
Traffic volume is the other and one of the most important factors in predicting road crashes, most identified as a significant explanatory variable in many CPM in previous research works. However, traffic volume is prone to errors since the typical measure of traffic volume measure (AADT) is a composite of several vehicle types. Another decision-making required in crash prediction modelling is the time period for the crash and AADT data. Mostly, longer time periods may present problems due to changes in conditions over time, such as significant increases in traffic volumes or layout changes, which may not reflect the current traffic situation anymore. In research by Ambros et al. (2016), 4 years period was found to be sufficient for developing a CPM after employing multiple consistency checks. Typically, a compromise should be reached between the necessity for early investigation of new treatments and the need to accumulate enough crashes to permit analysis (Elvik, 2010).

A robust and reliable road safety data system is crucial for accurately understanding the scale and nature of road traffic fatalities. The latest report by the WHO Global Status Report on Road Safety revealed that in many developing countries, including Ethiopia, the ratio between WHO-estimated and reported national road traffic fatalities greater than 5, indicating a significant level of underreporting (WHO,2023). Underreporting often stems from limited data collection systems, lack of institutional capacity, and inadequate resources to monitor road crashes effectively. This results in discrepancies between reported national data and WHO estimates, masking the true picture of the problem.

High-quality road safety data is essential for evidence-based decision-making, as it provides policymakers with a clear picture of the problem and trends. Accurate and reliable data allows for better allocation of budgets toward critical road safety interventions, such as infrastructure improvements, enforcement of traffic laws, and public awareness campaigns. Additionally, it enables targeted interventions where they are needed most, improving their effectiveness. Without reliable data, governments risk underestimating the road safety crisis, leading to insufficient funding, poorly designed interventions, and missed opportunities to save lives. Strengthening road safety data management systems is, therefore, a priority to drive informed policies and achieve measurable progress in reducing road traffic fatalities

2.2 Road Network Segmentation

Crash analysis on large road networks requires segmenting the network into smaller units. Different ways of road segmentation have been studied in road safety studies in which the most common methods include constant-length segments, homogeneous properties, and segments with interconnected crash locations (Nair and Bhavathrathan, 2022). Some researchers use constant-length segmentation where relatively short road segments were identified as undesirable in crash rate estimations (Cenek et al., 1997; Cafiso et al., 2010, Bhavsar et al., 2021). For CPMs to be developed in the length of the road section, the segment is taken as an important factor in several research works. Both extreme long and short roadway segments can have an impact on the results of a safety analysis (Lu et al., 2013, Green, 2018). (P. Resende and Benekohal, 1997; Bhavsar et al., 2020) advised that the segment length should be at least 0.8 km long to develop a reliable crash prediction model and argue that shorter sections would result in incorrect statistics and correlations between roadway and traffic conditions. The main problem related to short road segments is the possibility that a road feature in one segment triggered a crash officially located in another segment (Bhavsar et al., 2020).

The AASHTO Highway Safety Manual (2014) suggests the use of homogeneous segments in terms of AADT, number of lanes, curvature, presence of a ramp at interchange, lane width, outside and inside shoulder widths, median width, and clear zone width. Although there is no prescribed minimum segment length for prediction models to work, it is recommended that the segments be at least 0.16 km long. Cafiso et al. (2018) concluded that if a longer segment can give good results in terms of goodness of fit, and they are still of engineering interest for safety analysis, the conclusion is that the use of a longer segment can be the better solution for segmentation of a road network.

2.3 Selection of Explanatory Variables

Theoretically, the research literature should be used to guide the selection of explanatory variables based on previously reported crash and injury risk factor findings. Although in practice data availability is simply the deciding factor (Ambros et al., 2018). Previous researchers identified many risk factors that affect the occurrence of road crashes, including driver behavior, vehicle factors, roadway characteristics, traffic volumes, environmental factors and time factors (Berhanu, 2004, Ackaah and Salifu, 2011, Abdulhafedh, 2016,

Abdulhafedh, 2017). Therefore, the variables included should be variables that:

- Have been focused in previous studies to exert a major influence on the number of crashes;
- Can be measured in a valid and reliable way;
- Are not endogenous, that is dependent on other explanatory variables included or on the dependent variable in the model.

Based on the number of explanatory variables, crash prediction models can be divided in to simple models involving risk exposure only (traffic volume and segment length) as independent variable and multivariate models that use further variables, usually geometric characteristics and consistency measures (Persaud, 2001). Different authors warned against overfitting models, which include too many insignificant factors, after they discovered that adding more predictors isn't as helpful as expected (Peltola et al., 1994; Sawalha and Sayed, 2006, Wood et al., 2013, Saha et al., 2015). On the contrary, models that contained geometry and design consistency variables were found to be more reliable than others (Hauer, 1997; Persaud, 2001, Ambros et al., 2016; Cafiso et al. (2018). Many researchers agreed that it is best to aim towards concise models, which include as few explanatory variables as possible as these models result in simple interpretation and understanding, as well as easy updating (Eenink et al., 2008; Ambros et al., 2016).

2.4 Data and Methodological Issues

2.4.1. Over-dispersion in Road Crash Data

Overdispersion occurs when the variance is greater than the mean of the road crash data, which is caused by heterogeneity of the subjects (Mekonnen and Sipos, 2022). In many circumstances, the variance of road crash data exceeds the mean value, meaning that the data are overdispersed (Lord and Mannering, 2010, Mohammed et al., 2018). The Poisson regression model, which is considered fundamental in road traffic crash data, usually face regression difficulties due to characteristics of the crash data such as overdispersion (Basu and Saha, 2017). Poisson regression models cannot handle over and under dispersion, and they can be affected by low sample-means and can generate biased results in small samples. In count data like road crash data, over-dispersion can be affected by different factors such as the clustering of data, unaccounted temporal correlation, and model misspecification (Madanat, 1995). Adopting the common Poisson regression model for an overdispersed data set can result in biased and inconsistent parameter estimates and incorrect inferences about the explanatory variables (Cameron and Trivedi, 1998; Park and Lord, 2007, Lord and Mannering, 2010).

2.4.2. Under-dispersion in Road Crash Data

Underdispersion is the other characteristics of road crash data, which means the variance is less than the mean of the road crash data, even though it is not as common as overdispersion in road crash data set (Lord and Mannering, 2010). Many traditional models developed based on count data like road crash data have been found to produce incorrect parameter estimates due to an under-dispersed data set (Oh et al., 2006). This under-dispersion problem cannot be handled by ordinary Poisson regression models; hence different modelling techniques are proposed as discussed in the next sections.

2.4.3. Low Sample Mean and Small Sample Size

Road crash data are mostly characterised by small number of crash records due to the high cost of the crash data collection process, or there may be predominantly many numbers of zero crash records on roadway segments over a long period of time. A road crash data set with low sample mean and small sample size can create an estimation problem. If the data are characterised by small sample mean and/or a preponderance of zero records, the distribution of crash counts will be skewed excessively toward zero, which can result in incorrectly estimated parameters and erroneous inferences (Lord and Mannering, 2010). According to Lord and Miranda-Moreno (2008) the dispersion parameter of the negative binomial model may not be correctly estimated when road crash data characterised by a small sample size and low sample mean values are used.

3. Road Crash Modelling Techniques

There are many factors that contribute to road crashes, including the characteristics of the roadway environment, vehicles, and the behaviour of road users. Over the past years, CPMs have become the fundamental scientific tools of road safety management. Several research works have been carried out on the development of CPMs to estimate crash frequencies on a particular road segment or intersection.

Typically, the Poisson, Negative binomial, Poisson lognormal, and zero inflated models are usually considered appropriate because of their flexibility and ease in estimating the parameters (Basu and Saha, 2017). Over the years, several regression techniques have been experimented while modelling the crash data in order to investigate their compatibility in terms of goodness of fit, and the most common techniques include negative binomial and Poisson lognormal. The details of road crash modelling techniques, from the basic modelling techniques to mostly practised and state-of-the-art ones, are discussed below followed by a table summary of their strengths and weaknesses along with their suitability for over-dispersed data.

3.1. Multiple Linear Regression Models

Multiple linear regression modelling techniques have been used to model the relationship between an outcome variable, in this case, crash rate or crash frequency, and two or more explanatory variables by fitting a linear equation assuming normally distributed errors (Kutner et al., 2005). Today it is not widely used because accident data frequently violates linear regression analysis assumptions such as normal error structure and constant error variance (Al-Qadi et al., 2008; Mekonnen and Sipos, 2022). The general form of multiple regression models can be expressed as:

$$y_{it} = \beta_0 + \sum_k \beta_k x_{itk} + \varepsilon_{it}$$

where:
y_{it}: number of observed crashes in the segment i at time t
x_{it}: explanatory variables or factors contributing to crashes
β_0: intercept (constant term)
β_k: regression coefficients
ε_{it}: the model's error term (also known as the residuals

Generally multiple linear regression modelling techniques have limitations to adequately describe the random, non-negative, discrete, and typically infrequent events, which are all characteristics of road crashes (Abdel-Aty and Radwan, 2000, Ackaah and Salifu, 2011).

3.2. Poisson Regression Model

Many scholars claim that the Poisson regression model is more suited than multiple linear regression models, since crashes are unavoidable, discrete, and more probably random events (Abdulhafedh, 2017, Mekonnen and Sipos, 2022). However, Poisson models have drawbacks such as the mean must equal the variance of road crashes. However, the variance of road crash data usually exceeds the mean value, indicating that the data are overdispersed (Lord and Mannering, 2010, Mohammed et al., 2018). To overcome the overdispersion problem researchers suggest modifying the Poisson regression model by using a correction that accounts for overdispersion in to “quasi-Poisson” (QP) model or use the negative binomial distribution (Maher and Summersgill, 1996, Moksony and Hegedűs, 2015).

3.3 Negative Binomial (Poisson-Gamma) Model

Crash data have special properties, such as overdispersion, which occur when the variance is greater than the mean of the observations. Overdispersion causes underestimation or deflation of standard errors of estimations (Shirazi et al., 2016). Due to the over-dispersion problem when using Poisson regression, several authors use the negative binomial as an alternative modelling method (Agresti, 2002, Berhanu, 2004, Cafiso et al., 2010).
Cafiso et al. (2010) tried to establish CPMs for two-lane rural road sections using a combination of geometry, exposure, context, and consistency variables related to road safety performance with a five-year crash analysis period assuming a negative binomial distribution error structure. Finally, they found three significant models based on practical considerations, statistical significance, and goodness-of-fit indices. An extensive crash study conducted in Ethiopia by Berhanu (2004) tried to fit both quasi-Poisson and Negative-Binomial models and concluded that the Negative-Binomial model was generally preferable in handling overdispersion problems. However, the negative binomial model has limitations in handling cases of under-dispersion of the data count, which occurs when mean is greater than the variance (Amoros et al., 2003; Oh et al., 2006).

3.4. Poisson Log-Normal Model

The Poisson-lognormal model, in which the error component is Poisson-lognormal rather than gamma distributed, was created to solve the shortcomings of the NB models to better handle small sample sizes and low sample mean values (Lord and Miranda-Moreno, 2008; Daniels et al., 2010). The Poisson-log-normal model is similar to the Poisson-gamma model, except that the Poisson rate is represented by a log-normal distribution. Model estimation in Poisson-lognormal modelling is more complex because the Poisson-lognormal distribution does not have a closed form and can still be adversely affected by small sample sizes and low sample mean values (Miaou and Lord, 2003). When crash data are characterised by low sample mean values and a small sample size, Poisson lognormal models offer a better alternative than Poisson gamma models (Lord and Miranda-Moreno, 2008).

3.5. Zero-Inflated Poisson and Negative-Binomial Model

It is common to come across crash data with a lot of zero observations which are called zero-inflated data because the number of zeros is more than expected when performing negative binomial or Poisson regression models (Jang et al., 2010). Zero-inflated models were constructed to handle data with a considerable number of zeros or more zeros than a typical Poisson or negative binomial/Poisson-gamma model would expect (Lord and Mannering, 2010).
Zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) modelling approach have been developed, with a dual-state crash system, with one state being the zero-crash state, which can be considered essentially safe for the observation period, and the other being the non-zero crash state. This approach has recently been used by a number of researchers (Jiang et al., 2013, Xu et al., 2017, Xu et al., 2019).

3.6. Conway-Maxwell Poisson Regression Model

The Conway-Maxwell Poisson regression is a generalisation of the Poisson distribution to handle both under-dispersed and over-dispersed crash data or a combination of both using a variable called dispersion parameter. This is especially important to handle the underdispersion in crash data that cannot be modelled by the Poisson model or the Negative Binomial model (Abdulhafedh, 2017). The Conway-Maxwell Poisson regression could be negatively affected by the low sample mean and small sample size bias. Therefore, it has been limited in the application of road safety analysis (Mor et al., 2019). Lord and Guikema (2012) and Boatwright et al. (2006) adopted the Conway-Maxwell Poisson regression technique to develop crash prediction models.

3.7. Random-Effects Model

The random effect model assumes that road crash data are hierarchical in nature in which the lowest level of the hierarchy represents the crashes themselves, while the type of location on the road network at which the crash occurred represents the higher-level hierarchy. Therefore, the main assumption of this model is that there may be correlations between crashes occurring at the same location, so these crashes may share unobserved or unrecorded characteristics related to the location. These unobserved characteristics could include low pavement friction, poor pavement condition, poor reflectivity of road signs, and other similar factors (Mohammad N. Al-Maraf and Somasundaraswaran, 2018). Due to serial correlation in road crash data, nonhierarchical models seem to be inappropriate since crash data variables are likely to have location specific effects. In such cases, random-effects model is more suitable which account for correlation within clusters by introducing random effects in the population-based models (Mor et al., 2019).

3.8. Multiple Logistic Modeling

The multiple logistic regression modelling technique is adopted when the value of the dependent variable ranges between 0 and 1. It is used to build models that provide a measure of the probability of injury or noninjury crash binary outcomes (Mor et al., 2019). The multiple logistic regression modelling technique is suitable to study the effect of one variable while controlling for other variables (Mohammad N. Al-Maraf and Somasundaraswaran, 2018).

3.9. Artificial Neural Networks (ANNs)

When a linear function fails to adequately explain the relationship between the dependent and explanatory variables in crash modelling, nonlinear approximators like fuzzy logic and neural networks have been investigated (Abdulhafedh, 2017). Artificial neural networks (ANNs) are a class of computational intelligence tools that can be used to solve problems involving prediction and categorisation. ANNs can simulate very complex nonlinear functions with great reliability employing a learning process that is analogous to the cognitive system's learning method in the human brain (Mohammed et al., 2018). The input layers, hidden layers, and output layers make up the network body (Abdulhafedh, 2017).

In a supervised learning process, these models can be trained to estimate any nonlinear function to a specified degree of accuracy using a learning technique (such as back propagation) that produces the desired result (Mohammed et al., 2018). ANNs have an advantage over statistical models in that the ANNs do not require the establishment of a predefined relationship between dependent and independent variables and can be easily applied in the analysis (Ma et al., 2008). In a study conducted to compare the prediction performance between the NB model, the Poisson model and the ANN, the overall prediction performance of the ANN was found to be much better than the NB and Poisson models for the testing data (Abdulhafedh, 2016).

3.10 Summary of Crash Modeling techniques

Several modelling techniques have been used to develop crash prediction models for the estimation of predicted crashes or crash rates on road segments or intersections. The advantage and disadvantage and suitability of the models for the over dispersed data are summarised in Table 1.

Table 1. Summary of crash modeling techniques

Model Type	Advantages	Disadvantages	AOD
MLR	Easy to estimate predicted crash frequencies	Unable to describe adequately the random, non-negative, discrete, and typically sporadic events	NA
P	Handle the discrete and more likely random events Most fundamental model for count data and easy to estimate	Cannot handle over- and under-dispersion; Highly influenced by the low sample mean and small sample size bias	NA
NB	Easy to estimate; Can handle the over-dispersion problem	unable to handle under-dispersion; Negatively affected by the low sample mean and small sample size bias	A
PN	More flexible than the Poisson gamma to handle over dispersion	Unable to handle under-dispersion; can be adversely affected by the low sample mean and small sample size bias (less than the Poisson-gamma)	NA
ZiP and NB	Can handles datasets with large number of zero-crash observations	Zero- inflated negative binomial can be adversely influenced by the low sample mean and small sample size bias	A

CMP	Can handle under- and over-dispersion or combination of both using a dispersion parameter	Could be negatively influenced by the low sample mean and small sample size bias;	A
REM	Can handle temporal correlation (location specific effects)	The results from this technique may not be transferable to other data sets because the results are observation specific	NA
ML	Suitable to study the effect of one variable while controlling for other variables	Used to analyze only crash binary outcomes	NA
ANN	Non parametric approach which does not require an assumption about data distribution; Usually provides better statistical fit than traditional parametric models	Complex estimation process It cannot extrapolate the results	A

Note: AOD: Applicability for over-dispersed data, NA: Not Applicable, A: Applicable, MLR: multiple linear regression, P: Poisson, NB: negative binomial, PN: Poisson-log normal, ZiP: Zero-inflated Poisson, NB: negative binomial, CMP: Conway Maxwell-Poisson, REM: Random Effect Model, ML: multiple logistic regression, ANN: Artificial neural networks

4. Application of Crash Prediction Models in Black Spot Identification

The black spot can be defined as any location that has a higher predicted number of road crashes than other similar locations as a result of local risk factors existing in the location which is considered as the first step in the road crash reduction process (Elvik, 2008). It is any location where there is a concentration of road crashes even though its precise definition varies from country to country. Norwegian, for example, define black spot as any location with a maximum length of 100m at which at least 4 injury crashes have occurred within 5 years. In Portugal, a black spot is any road section with a maximum length of 200m with at least 5 crashes in the year of analysis. Miranda-Moreno et al. (2005) investigated the relative performance and decision implications of three models, the Poisson lognormal, heterogeneous negative binomial, and the traditional negative binomial model, for improving ranking locations for road safety. They concluded that the selection of modelling techniques and ranking criteria can result in considerably different lists of hazardous locations.

Based on the type of crash data used in the black spot identification process, black spot analysis can be done using numerical approach or model-based approaches (Mohammad N. Al-Maraf and Somasundaraswaran, 2018). The numerical approach uses historical crash data to define black spots, in which locations with crash records higher than the average crash frequency are identified as black spots. The model-based approach, on the other hand, uses statistical crash prediction models to identify black spot locations where the observed number of crashes significantly exceeds the number of expected crashes on roads with similar geometric and traffic characteristics the road segment is identified as black spot (AASHTO, 2014).

Road crash record data and crash prediction models do not address a statistical phenomenon known as Regression to Mean Effect (RME), whereby the number of road crashes at a particular site fluctuates up or down around a long-term average, which can lead to an overestimation of effectiveness of an intervention. For example, road crashes at a given site may be higher in a given period and low for the next period without any improvement in road safety. To solve the RME problem, the empirical Bayes (EB) approach has been introduced. The EB approach is used to identify black spot locations based on their potential for Safety Improvement (PSI), which is the difference between predicted and expected crashes at the particular road site (Mohammad N. Al-Maraf and Somasundaraswaran, 2018).

Therefore, the best approach to determine black spot locations is using the expected crash frequency, not the recorded crashes, where the combination of the recorded crashes and the model estimates for a given site is used to estimate the expected crash frequency (Elvik, 2007). The suitable method to do this is applying the empirical Bayes method of analysis where the expected crash frequencies are estimated using the following formula.

$$E(\lambda/r) = \alpha.\lambda + (1-\alpha).r$$

Where:
r is the recorded number of crashes on the given site
λ is predicted number of crashes as estimated by crash prediction models
α is a parameter that determines the weight given to the estimated number of crashes using CPMs for similar sites when combining it with the recorded number of crashes to estimate the expected (adjusted) number of crashes for a particular site.

After 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 segment or intersection. In general, CPMs can be used to identify hazardous locations as a list of spots that are being prioritised for further researches of engineering with distinguished road crash patterns, contributing factors, and potential resolution. Furthermore, cost-effective projects are often selected to get the best results from limited resources (Mohammad N. Al-Maraf and Somasundaraswaran, 2018).

5. Conclusions and Future Directions

Crash prediction models are very important mathematical tools in road safety programmes that can be used by different governmental institutions that oversee road safety, vehicles, and driver education in making decisions about improving road safety. From previous studies it is concluded that multiple linear regression models are not suitable for road crash modelling from a methodological point of view due to their limitations to describe adequately the random, nonnegative, discrete, and typically infrequent events like road crashes. Traditional crash prediction methods such as Poisson and negative binomial regression techniques have been extensively used by many researchers in road safety for developing CPMs due to their ability to analyse crash data while preventing the possibility of negative expected values. However, crash frequency data is characterized by formidable problems such as overdispersion, under dispersion, low sample means and small sample size, underreporting of crashes, omitted variables bias, and issues related to functional form and fixed parameters which prevent the adoption of traditional crash modelling techniques.

To solve these problems, innovative methodological approaches and advanced modelling techniques such as Conway-Maxwell Poisson regression models and artificial neural network techniques have been introduced to improve the statistical validity of findings. Since road crash record data and crash prediction models do not address the regression to mean effect phenomenon, the Empirical Bayes (EB) approach should be introduced to determine the expected crash frequency by combining the crash record data and the predicted crash frequency. Then the Potential for Safety Improvement (PSI) should be calculated as the difference between the predicted and expected crashes at the particular site. These values are then used to decide whether a particular roadway segment or intersection is a black spot location or not.

Future research should prioritize in improving the quality and reporting of crash data by using emerging technologies such as connected vehicle systems, GPS, and sensor-based data collection to address issues like underreporting and data rawness, especially in low- and middle-income nations. Researchers should also explore hybrid CPMs that combine traditional statistical techniques with advanced machine learning approaches to advantageously handle complex road crash data characteristics. The outputs of CPMs, such as the Potential for Safety Improvement (PSI), should be integrated into transport policies and infrastructure planning to guide resource allocation toward high-risk locations effectively. Furthermore, efforts should focus on developing CPMs that are scalable and transferable across diverse regions and contexts, ensuring their applicability in low-resource settings. Finally, future studies should examine alternative explanatory variables, including driver behavior patterns, vehicle technology advancements, and land use changes, to enhance the predictive accuracy and utility of CPMs in road safety management.

References

[1] AASHTO (2014). "Highway safety manual user guide." National Cooperative Highway Research Program: 17-50.

[2] Abdel-Aty, M. A. and A. E. Radwan (2000). "Modeling traffic accident occurrence and involvement." Accident Analysis & Prevention 32(5): 633-642.

[3] Abdulhafedh, A. (2016). "Crash frequency analysis." Journal of transportation technologies 6(04): 169.

[4] Abdulhafedh, A. (2017). "Road crash prediction models: different statistical modeling approaches." Journal of transportation technologies 7(02): 190.

[5] Ackaah, W. and M. Salifu (2011). "Crash prediction model for two-lane rural highways in the Ashanti region of Ghana." IATSS research 35(1): 34-40.

[6] Agresti, A. (2002). Categorical data analysis. Hoboken, NJ: wiley.

[7] Al-Qadi, I. L., et al. (2008). Efficient Transportation and Pavement Systems: Characterization, Mechanisms, Simulation, and Modeling, CRC Press.

[8] Ambros, J., et al. (2018). "An international review of challenges and opportunities in development and use of crash prediction models." European transport research review 10(2): 1-10.

[9] Ambros, J., et al. (2016). "Developing updatable crash prediction model for network screening: case study of Czech two-lane rural road segments." Transportation research record 2583(1): 1-7.

[10] Amoros, E., et al. (2003). "Comparison of road crashes incidence and severity between some French counties." Accident Analysis & Prevention 35(4): 537-547.

[11] Basu, S. and P. Saha (2017). "Regression models of highway traffic crashes: a review of recent research and future research needs." Procedia engineering 187: 59-66.

[12] Berhanu, G. (2004). "Models relating traffic safety with road environment and traffic flows on arterial roads in Addis Ababa." Accident Analysis & Prevention 36(5): 697-704.

[13] Bhavsar, R., et al. (2020). "Development of Model for Road Crashes and Identification of Accident Spots." International journal of intelligent transportation systems research.

[14] Bhavsar, R., et al. (2021). "Development of model for road crashes and identification of accident spots." International journal of intelligent transportation systems research 19(1): 99-111.

[15] Boatwright, P., et al. (2006). Conjugate analysis of the Conway-Maxwell-Poisson distribution 1(2): 363–374.

[16] Cafiso, S., et al. (2018). "Investigating the influence of segmentation in estimating safety performance functions for

roadway sections." Journal of traffic and transportation engineering (English edition) 5(2): 129-136.

[17] Cafiso, S., et al. (2010). "Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables." Accident Analysis & Prevention 42(4): 1072-1079.

[18] Cameron, A. C. and P. K. Trivedi (1998). Regression Analysis of Count Data. Cambridge, Cambridge University Press.

[19] Cenek, P., et al. (1997). "Road environment and traffic crashes." Transfund New Zealand Research Report (79).

[20] Daniels, S., et al. (2010). "Explaining variation in safety performance of roundabouts." Accident Analysis & Prevention 42(2): 393-402.

[21] Eenink, R., et al. (2008). "Accident prediction models and road safety impact assessment: recommendations for using these tools." Institute for Road Safety Research, Leidschendam.

[22] Elvik, R. (2007). State-of-the-art approaches to road accident black spot management and safety analysis of road networks. Norway, Institute of Transport Economics: 1-126.

[23] Elvik, R. (2008). "A survey of operational definitions of hazardous road locations in some European countries." Accident Analysis and Prevention: 1830–1835.

[24] Elvik, R. (2010). Assessment and applicability of road safety management evaluation tools: Current practice and state-of-the-art in Europe.

[25] Ganguly, R., et al. (2014). "Traffic Volume and Accident Studies on Nh-22 Between Solan and Shimla, India." European Scientific Journal.

[26] Green, E. R. (2018). Segmentation strategies for road safety analysis, University of Kentucky. PhD.

[27] Hauer, E. (1997). Observational before/after studies in road safety. estimating the effect of highway and traffic engineering measures on road safety.

[28] Jang, H., et al. (2010). "Bayesian analysis for zero-inflated regression models with the power prior: Applications to road safety countermeasures." Accident Analysis & Prevention 42(2): 540-547.

[29] Jiang, X., et al. (2013). "Investigating the influence of curbs on single-vehicle crash injury severity utilizing zero-inflated ordered probit models." Accident Analysis & Prevention 57: 55-66.

[30] Jonsson, T. (2005). "Predictive models for accidents on urban links: A focus on vulnerable road users."

[31] Kutner, M. H., et al. (2005). Applied linear statistical models, McGraw-Hill New York.

[32] Lord, D. and S. D. Guikema (2012). "The Conway-Maxwell-Poisson model for analyzing crash data." Applied Stochastic Models in Business and Industry 28(2): 122–127.

[33] Lord, D. and F. Mannering (2010). "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives." Transportation research part A: policy and practice 44(5): 291-305.

[34] Lord, D. and L. F. Miranda-Moreno (2008). "Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective. Safety science 46(5): 751-770.

[35] Lu, J., et al. (2013). Clustering-based roadway segment division for the identification of high-crash locations. Journal of Transportation Safety & Security 5(3): 224-239.

[36] Ma, J., et al. (2008). "A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods." Accident Analysis & Prevention 40(3): 964-975.

[37] Madanat, W. H. W. I. (1995). "Poisson regression models of infrastructure transition probabilities." Transport Engineering 121: 1-09.

[38] Maher, M. J. and I. Summersgill (1996). "A comprehensive methodology for the fitting of predictive accident models." Accident Analysis & Prevention 28(3): 281-296.

[39] Mekonnen, A. A. and T. Sipos (2022). "Crash Prediction Models and Methodological Issues." Periodica Polytechnica Transportation Engineering.

[40] Miaou, S.-P. and D. Lord (2003). "Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods." Transportation research record 1840(1): 31-40.

[41] Miranda-Moreno, et al. (2005). "Alternative Risk Models for Ranking Locations for Safety Improvement." Journal of the Transportation: 1–8.

[42] Mohammad N. Al-Maraf and K. Somasundaraswaran (2018). "Review of Crash Prediction Models and their Applicability in Black Spot Identifcation to Improve Road Safety." Indian Journal of Science and Technology 11(5).

[43] Mohammed, A. A., et al. (2018). "Classification of traffic accident prediction models: a review paper." International Journal of Advances in Science Engineering and Technology 6(2): 35-38.

[44] Moksony, F. and R. Hegedűs (2015). "The use of Poisson regression in the sociological study of suicide." Corvinus Journal of Sociology and Social Policy 5(2).

[45] Mor, N., et al. (2019). An Overview of Challenges and Opportunities in Development and Use of Accident Prediction Models.

[46] Nair, S. R. and B. Bhavathrathan (2022). "Hybrid segmentation approach to identify crash susceptible locations in large road networks." Safety science 145: 105515.

[47] Oh, J., et al. (2006). Accident prediction model for railway-highway interfaces. Accident Analysis & Prevention 38(2): 346-356.

[48] P. Resende and R. Benekohal (1997). Effects of Roadway Section Length on Accident Model. Proceedings of TC&TF in the 21st century: challenges, innovations, and opportunities, Chicago.

[49] Park, E.-S. and D. Lord (2007). Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity Transportation research record: 1-6.

[50] Peltola, H., et al. (1994). Why use a complicated accident prediction model when a simple one is just as good. 22nd European Transport Forum (The PTRC Summer Annual Meeting).

[51] Persaud, B. N. (2001). Statistical methods in highway safety analysis.

[52] Saha, D., et al. (2015). Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees. Accident Analysis & Prevention 79: 133-144.

[53] Sawalha, Z. and T. Sayed (2006). Traffic accident modeling: some statistical issues. Canadian Journal of Civil Engineering 33(9): 1115-1124.

[54] Shirazi, M., et al. (2016). A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis & Prevention 91: 10-18.

[55] Sørensen, M. and R. Elvik (2008). Black Spot Management and Safety Analysis of Road Networks - Best Practice Guidelines and Implementation Steps.

[56] Srinivasan, R. and K. M. Bauer (2013). Safety performance function development guide: Developing jurisdiction-specific SPFs, United States. Federal Highway Administration. Office of Safety.

[57] Wegman, F. (2014). Analyzing road design risk factors for run-off-road crashes in the Netherlands with crash prediction models. Journal of safety research 49: 121. e121-127.

[58] WHO (2015). Global Status Report on Road Safety. Geneva, World Health Organization. 15.

[59] WHO (2023). Global Status Report on Road Safety. Geneva, World Health Organization.

[60] Wood, A., et al. (2013). Updating outdated predictive accident models. Accident Analysis & Prevention 55: 54-66.

[61] Xu, C., et al. (2019). Investigating the factors affecting secondary crash frequency caused by one primary crash using zero-inflated ordered probit regression. Physica A: Statistical Mechanics and its Applications 524: 121-129.

[62] Xu, P., et al. (2017). Revisiting crash spatial heterogeneity: a Bayesian spatially varying coefficients approach Accident Analysis & Prevention 98: 330-337.

[63] Yannis, G., et al. (2017). Road traffic accident prediction modelling: a literature review. Proceedings of the institution of civil engineers-transport, Thomas Telford Ltd.