Accident Analysis and Prevention 90 (2016) 73–81

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Accident prediction model for public highway-rail grade crossings

Pan Lu[a,*], Denver Tolliver[b]

[a] Assistant Professor of Transportation, Upper Great Plains Transportation Institute, Dept. 2880, North Dakota State University, Fargo, ND 58108-6050, USA
[b] Director and Professor of Transportation, Upper Great Plains Transportation Institute, Dept. 2880, North Dakota State University, Fargo, ND 58108-6050, USA

ARTICLE INFO

Article history:
Received 25 June 2015
Received in revised form 15 January 2016
Accepted 18 February 2016
Available online 27 February 2016

Keywords:
Accident prediction
Railroad grade crossing
Poisson
Quasi-Poisson
Gamma
Binomial
Bernoulli
Under-dispersion

ABSTRACT

Considerable research has focused on roadway accident frequency analysis, but relatively little research has examined safety evaluation at highway-rail grade crossings. Highway-rail grade crossings are critical spatial locations of utmost importance for transportation safety because traffic crashes at highway-rail grade crossings are often catastrophic with serious consequences. The Poisson regression model has been employed to analyze vehicle accident frequency as a good starting point for many years. The most commonly applied variations of Poisson including negative binomial, and zero-inflated Poisson. These models are used to deal with common crash data issues such as over-dispersion (sample variance is larger than the sample mean) and preponderance of zeros (low sample mean and small sample size). On rare occasions traffic crash data have been shown to be under-dispersed (sample variance is smaller than the sample mean) and traditional distributions such as Poisson or negative binomial cannot handle under-dispersion well. The objective of this study is to investigate and compare various alternate highway-rail grade crossing accident frequency models that can handle the under-dispersion issue. The contributions of the paper are two-fold: (1) application of probability models to deal with under-dispersion issues and (2) obtain insights regarding to vehicle crashes at public highway-rail grade crossings.

1. Introduction

More than 97% of highway-rail crossings are railroad grade crossings (RGCs) where the highway and railroad tracks are at the same elevation and allow traffic to alternate between railroads and highways. RGCs are critical spatial locations of utmost importance for transportation safety because traffic crashes at RGCs are often catastrophic with serious consequences including fatalities, injuries, extensive property damage and delays in railway and highway traffic (Raub, 2009). The consequences can be further exacerbated if collisions involve freight trains carrying hazardous materials which could spill resulting environmental disaster or increased danger to those nearby. From 1996 to 2014, 26% of RGC accidents in North Dakota involved hazardous material. The need to improve traffic safety has been a major social concern in the United States for decades. Transportation agencies and other stakeholders must identify the factors that contribute to the likelihood of RGC collision to be able to better predict probability of crashes and provide direction for RGC designs and policies that will reduce the number of crashes.

A large amount of literature exists in regard to evaluation of vehicle accidents on roadway intersections or roadway segment crashes (Young and Liesman, 2007; Campbell, 1991; Chen and Chen, 2011; Schoor et al., 2001; Chen et al., 2011; Zhang et al., 2013, 2012; Wang and Abdel-Aty, 2008; Xie et al., 2014; Kwon et al., 2015; Jung et al., 2014; Chen and Xie, 2014; Kashani et al., 2014). Although several studies have been carried out to investigate accidents at RGCs (Belle et al., 1975; Gitelman and Hakkert, 1997; Yan et al., 2010; Austin and Carson, 2002; Raub, 2009; Oh et al., 2006), our literature review found relatively little research effort has focused on RGC accidents compared to roadway accidents (Oh et al., 2006). Salmon et al. (2013) pointed that because of the limited research effort, various aspects of RGC performance remain poorly understood. Therefore, safety evaluation (i.e., accident frequency prediction) of RGCs is needed to re-examine both prediction methods and contribution factors at RGCs (Austin and Carson, 2002).

This paper seeks to investigate RGC crashes and contributing factors through the application of statistical models to crash data. Following a review of earlier highway-rail crossing accident

* Corresponding author. Fax: +1 701 231 1945.
E-mail addresses: pan.lu@ndsu.edu (P. Lu), denver.tolliver@ndsu.edu (D. Tolliver).

http://dx.doi.org/10.1016/j.aap.2016.02.012

prediction methods, the authors suggest improved accident prediction model alternatives that allows for handling the under-dispersion issue and provide the greatest insight into public RGC related crashes, especially in North Dakota.

This paper is divided into six sections. Section 2 presents a literature review on crash models and prediction methods to deal with under-dispersion issues regarding to highway-rail grade crossings. Section 3 describes the details of statistical methodologies used for estimating the models. Section 4 provides details about the data used in the paper. Section 5 covers the application of models to the dataset and described the results of the analysis and comparisons. Section 6 summarizes the key results of the study and provides recommendations for further work.

2. Background and literature review

Because crash accident data are random, discrete, and non-negative in nature, hence generalized linear models (GLMs) (known as the Frequentist approach in the statistical literature) have been frequently used to investigate the relationship between crash frequency and contributing factors. Poisson regression has been commonly used to model crash frequency because of the discrete and non-negative nature of crash data. GLMs have several advantages because they are believed to be better suited for discrete and nonnegative crash frequency data (Zhang et al., 2012). However, Lord and Mannering (2010) pointed out that although the GLMs possess desirable elements for describing accidents, these models face various data challenges which have been shown to be a potential source of error in terms of incorrectly specifying statistical models that can result in incorrect prediction and explanatory factors. The most common crash data issue is over- or under-dispersion. Over-dispersion is where sample variance is greater than sample mean. In many collision databases, the variance in accident frequency is higher than the mean. Over-dispersion arises from the unmeasured uncertainties associated with the observed or unobserved variables (Lord and Park, 2008). And on rare occasions, crash data can display under-dispersion where sample variance is less than the sample mean (Oh et al., 2006). These issues are problematic (Lord and Mannering, 2010) because the most common count-data modeling approach requires that the variance be equal to the mean. Over- and under-dispersed data would lead to inconsistent standard errors for parameter estimates when using the traditional Poisson distribution (Cameron and Trivedi, 1998). Because of this, Poisson regression is usually a good modeling starting point (Oh et al., 2006). When data shows over-dispersion, some modifications to the standard Poisson model are available to account for over-dispersion such as the Negative Binomial (NB) model (Lord and Mannering, 2010). When under-dispersion arise, less common models such as Gamma probability count model is believed to be capable of handling under-dispersion issues (Oh et al., 2006). Poisson and NB models have been shown to have great limitations when applied to under-dispersed crash data (Oh et al., 2006). This research will explore the potential GLM model options to handle under-dispersed RGC crash data by (1) demonstrating the general forms of various models and (2) Investigating and comparing models that may handle under-dispersed data with ND public RGC crash data.

It is worthwhile to note that Bayesian method gain its popularity over the last few years in highway safety research, mainly due to its ability to handle so called regression-to-mean (RTM) problem in highway safety which is often associated with traditional frequentist approaches such as maximum likelihood estimation to estimate the parameters in GLMs. RTM refers to the tendency of high or low crashes in one time period to regress or to return to the mean in subsequent time periods. The distinction between frequentist approach and Bayesian approach lies in the assumption about the parameters. Frequentist method assumed the response variable was random and the parameters and variance were fixed and unknown while Bayesian approach assumes that both the response variable and parameters are random. Thus, many researchers developed empirical or full Bayes methods in highway safety analysis (Lord and Park, 2008; Aguero-Valverde, 2013; Ezra Hauer, 2001) and found promising results in better predictions, estimations and less error. However, Bayes method focuses on the randomness of parameters but does not assume under- or over-dispersion distribution thus Bayes approach is not considered in this study.

3. Statistical methodology

3.1. Poisson model

Non-negative integer count data are often approximated well by the Poisson regression model. In Poisson regression the dependent variable is an observed count that follows the Poisson distribution, probability of RGC i having y_i crash (where $y_i y_i$ is the expected number of 0, 1, 2,. . .) is given by:

$$P(y_i) = \frac{e^{(-\lambda_i)}\left(\lambda_i{}^{y_i}\right)}{y_i!} = \frac{e^{-\mu}\left(\mu^{y_i}\right)}{y_i!} \quad (1)$$

Where λ_i is the predicted count or Poisson parameter for RGC i, which is equal to RGC i's expected number of crashes per year, $E[y_i]$ or μ. In other words, if we say the probabilities of number of crashes follow a Poisson distribution with parameter λ, the mean of the crashes,y_i, is its parameter lambda. Poisson parameter Lambda is the total number of events divided by the number of units in the data. Poisson Model is specifying the Poisson parameter λ_i as a function of explanatory variables, more specifically, Poisson model let the logarithm of the mean depend on a vector of explanatory variables xs. That is, for a given set of predictors,x_i, the categorical outcome follows a Poisson distribution with rateλ_i. The most commonly selected functional form (or function link) is in log-linear form:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} \quad (2)$$

Where βsare the estimated regression coefficients and xs are all the explanatory variables. In this model, the regression coefficient β_j represents the expected change in the log of the mean per unit change in the predictorx_j. The log link is the canonical link function for the Poisson distribution. Techniques to find an estimates of the regression coefficients, $\hat{\beta}$, to achieve maximized log likelihood can be applied and the expected value of the response is modeled. One important property of the Poisson distribution model is that it restricts equal mean and variance of the distribution:

$$\mathrm{Var}[Y] = \mathrm{E}[Y] = \mu \quad (3)$$

If the mean is not equal to the variance of the crash counts, then the data are said to be either over- or under-dispersed and the resulting parameter estimated will be biased (Cameron and Trivedi, 1998). Crash data has been found to often exhibit over-dispersion due to unmeasured variances associated with the observed or unobserved variables (Lord and Park 2008).

3.2. Negative binomial model

The negative binomial as a special case of Poisson–gamma mixture model is a variant of the Poisson model designed to deal with over-dispersed data. The negative binomial model relaxes the