ARTICLE IN PRESS

Accident Analysis and Prevention xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Predicting crash frequency for multi-vehicle collision types using multivariate Poisson-lognormal spatial model: A comparative analysis

Mehdi Hosseinpour[a,*], Sina Sahebi[b], Zamira Hasanah Zamzuri[c], Ahmad Shukri Yahaya[d], Noriszura Ismail[c]

[a] Department of Civil Engineering, Central Tehran Branch, Islamic Azad University (IAUCTB), Tehran, Iran
[b] Department of Civil Engineering, Sharif University of Technology, Tehran, Iran
[c] School of Mathematical Sciences, Universiti Kebangsaan Malaysia, 43600, Bangi, Malaysia
[d] School of Civil Engineering, Universiti Sains Malaysia, 14300, Nibong Tebal, Malaysia

ARTICLE INFO

Keywords:
Collision type
Multivariate Poisson lognormal model
Spatial correlation
Two-stage model

ABSTRACT

According to crash configuration and pre-crash conditions, traffic crashes are classified into different collision types. Based on the literature, multi-vehicle crashes, such as head-on, rear-end, and angle crashes, are more frequent than single-vehicle crashes, and most often result in serious consequences. From a methodological point of view, the majority of prior studies focused on multivehicle collisions have employed univariate count models to estimate crash counts separately by collision type. However, univariate models fail to account for correlations which may exist between different collision types. Among others, multivariate Poisson lognormal (MVPLN) model with spatial correlation is a promising multivariate specification because it not only allows for unobserved heterogeneity (extra-Poisson variation) and dependencies between collision types, but also spatial correlation between adjacent sites. However, the MVPLN spatial model has rarely been applied in previous research for simultaneously modelling crash counts by collision type. Therefore, this study aims at utilizing a MVPLN spatial model to estimate crash counts for four different multi-vehicle collision types, including head-on, rear-end, angle, and sideswipe collisions. To investigate the performance of the MVPLN spatial model, a two-stage model and a univariate Poisson lognormal model (UNPLN) spatial model were also developed in this study. Detailed information on roadway characteristics, traffic volume, and crash history were collected on 407 homogeneous segments from Malaysian federal roads. The results indicate that the MVPLN spatial model outperforms the other comparing models in terms of goodness-of-fit measures. The results also show that the inclusion of spatial heterogeneity in the multivariate model significantly improves the model fit, as indicated by the Deviance Information Criterion (DIC). The correlation between crash types is high and positive, implying that the occurrence of a specific collision type is highly associated with the occurrence of other crash types on the same road segment. These results support the utilization of the MVPLN spatial model when predicting crash counts by collision manner. In terms of contributing factors, the results show that distinct crash types are attributed to different subsets of explanatory variables.

1. Introduction

Road traffic crashes in Malaysia have become a major socio-economic problem for road safety agencies and imposed a huge cost on the economy. For the year 2016, over 7000 people died in road accidents nationwide, a 6.7% increase compared to 2015 (IRTAD, 2017). Based on the Royal Malaysian Police (RMP) crash database, multi-vehicle crashes, including head-on, rear-end, angle, and sideswipe crashes, account for about 60% and 75% of total and fatal crashes, respectively. These figures indicate that multi-vehicle crashes are more frequent than single-vehicle crashes, and result in more serious consequences. Thus, it is highly justifiable to put more efforts into reducing the frequency and severity of multi-vehicle crashes. Safety researchers routinely apply crash prediction models (CPMs) to establish the relationship between crash occurrence and a set of explanatory variables (Xie et al., 2007; Geedipally and Lord, 2008; Li et al., 2008; Abdel-Aty and Haleem, 2011). In the context of crash frequency modelling, numerous studies have been conducted to relate traffic crashes to various risk indicating factors. From a methodological viewpoint, a number of published studies have modelled traffic crashes by collision type or injury severity.

* Corresponding author.
E-mail addresses: h_mehdi61@yahoo.com, meh.hosseinpour@iauctb.ac.ir (M. Hosseinpour).

https://doi.org/10.1016/j.aap.2018.05.003
Received 23 January 2018; Received in revised form 22 March 2018; Accepted 3 May 2018

As documented by previous studies (Kim et al., 2006; Geedipally et al., 2010; Christoforou et al., 2011; Aguero-Valverde et al., 2016), modelling crash frequencies by collision type (i.e., crash-type model) can gain a better understanding of the effect of various factors on each crash type, leading to develop more effective countermeasures. In this regard, various modelling specifications have been developed in the literature. These models include univariate count models (e.g., the standard negative binomial (NB) model, Poisson-lognormal model, zero-altered model), two-stage approach, and multivariate count models (e.g., multivariate Poisson (MVP), multivariate Poisson-lognormal (MVPLN) model, and multivariate spatial models). However, much of the literature has applied univariate count models to separately model crash frequencies by collision manner, while two-stage models and multivariate modelling techniques have recently been applied. A limitation associated with univariate count models is that they cannot account for correlations, which may exist between different levels of classification (e.g., severity levels or collision types). The two stage method integrates the outputs of a crash-count model (e.g., NB model) and a discrete choice model (e.g., multinomial logit model), in which the former is adopted to estimate the total predicted crashes while the latter is employed to predict the probabilities of crash occurrence for each collision type and then to allocate the predicted crash counts to each collision type. The two-stage model is more appropriate for modelling the counts of crashes by collision type in cases where the crash data are characterized by small sample size and low sample mean (Geedipally et al., 2010; Wang et al., 2011; Aguero-Valverde et al., 2016). Regarding multivariate modelling specifications, an advantage of these models over their univariate counterparts is that they can accommodate possible correlations (i.e., presence of interdependencies) between different crash types, as they jointly model crash counts by collision manner. Among the multivariate specifications, the MVPLN model with spatial effects has been found to be more promising, since it not only accounts for unobserved heterogeneity (extra-Poisson variation) and possible interdependencies between crash types, but also spatial correlation between adjacent sites. The idea behind the inclusion of spatial correlation in the multivariate models is that neighbouring locations have similar geographical and environmental characteristics, and hence tend to be correlated instead of independent. Therefore, the presence of spatial correlations in crash data should be accounted for. Otherwise, neglecting to consider spatial dependence leads to underestimation of variability, which results in biased parameter estimates (Lee et al., 2015). Several researchers have recently proposed the MVPLN spatial model in order to address the issue of spatial correlation between neighbouring locations (Aguero-Valverde, 2013; Wang and Kockelman, 2013; Barua et al., 2014; Imprialou et al., 2016). However, despite advantages of the MVPLN spatial models, they were rarely applied in the safety literature to estimate the crash frequencies for different vehicle-to-vehicle collision types, which are considered as commonly-occurring crash types (Aguero-Valverde et al., 2016). Therefore, this study aims to develop a MVPLN model with spatial effects for simultaneously estimating crash frequencies for four different multi-vehicle collision types, including head-on, rear-end, angle, and sideswipe collisions. The intent is to investigate the effects of various explanatory variables on the frequency of crashes across these four collision types while controlling for possible spatial correlations between adjacent roadways.

It is important to note that in the vast majority of the prior studies, the appropriateness of the multivariate modelling approaches were compared only to their univariate counterparts, UNPLN, or to other crash-type models, such as the MVP or conventional NB models. Nevertheless, it is also interesting here to find out how accurately the MVPLN spatial model can estimate the counts of different collision types when compared with other advanced models. To do this, a two-stage model is also developed using the same data set. The reason for developing a two-stage model in this paper is that this model, despite its advantages, has rarely been applied in the literature (notable exceptions are works by Wang et al. (2011) and Geedipally et al. (2010) as discussed later). Therefore, this study intends to extend research by conducting a comparative analysis between the MVPLN spatial and the two-stage model to determine which model provides better statistical fit. Moreover, to evaluate the appropriateness of the MVPLN spatial model, it is also compared with separate univariate spatial Poisson-lognormal (PLN) models and the MVPLN without spatial correlation. To accomplish the objectives of this study, 4-year (2007-2010) crash data were collected on 407 homogenous segments from Federal Roads in Malaysia. The organisation of this paper is as follows: Section 2 presents a review of previous research that has been focused on crash frequency modelling. Section 3 describes the methodology for estimating and comparing the proposed modelling techniques that are employed in this paper to model crash counts by type of collision. Section 4 presents the characteristics of the data used in this research. In Section 5, the results of model calibration and comparison are discussed followed by the discussion of the estimated parameters. The last section presents the key findings and conclusions of the study and offers specific recommendations and ideas for future research.

2. Literature review

There are a large body of studies in the literature that has estimated crash frequency either by collision type or injury severity. In this context, much of the earlier studies used conventional univariate count models, such as Poisson-gamma (i.e., NB), Poisson-lognormal, and zero-inflated models, to separately relate crash counts to a set of explanatory variables by collision type (Shankar et al., 1995; Greibe, 2003; Abdel-Aty et al., 2005; Qin et al., 2005; Kim et al., 2006; Ulfarsson et al., 2006; Jonsson et al., 2007, 2009; Venkataraman et al., 2013). However, a main concern when developing separate crash-type models is that they fail to correctly predict crash frequencies in cases where crash data are characterised by a small sample size and low sample mean (Lord and Mannering, 2010; Wang et al., 2011). By disaggregating total crashes into different collision types (e.g., head-on, rear-end, and rollover), the resulting subset will exhibit smaller sample sizes and lower sample means. This issue may cause less stable results and biased parameter estimates (Xie et al., 2007; Li et al., 2008; El-Basyouny and Sayed, 2009a; Bham et al., 2011; Aguero-Valverde et al., 2016). As documented by Lord (2006); Lord and Miranda-Moreno (2008), and Park and Lord (2009), this issue can negatively influence the estimation of the dispersion parameter of a Poisson-gamma model (i.e., negative binomial). They suggested that under such conditions, Poisson-lognormal models are preferred over Poisson-gamma models. More recently, other researchers introduced a more sophisticated methodology in which the total number of crashes is initially fitted by an aggregate count model. Then, the crash counts for each collision type are allocated using a discrete choice model. This approach is termed as the two-stage model (Wang et al., 2011). The literature has shown that the use of discrete choice models is more appropriate for modelling crash types when crash data for collision types exhibit the small sample size and low sample mean problem (Geedipally et al., 2010; Bham et al., 2011; Wang et al., 2011). The two-stage approach has been rarely applied in the safety literature. Notable exceptions are Geedipally et al. (2010) and Wang et al. (2011).

However, a critical drawback to the use of both univariate crash-type models and two-stage models is that they fail to accommodate the possible correlations that may exist across collision types. These correlations may be caused by shared latent information arising from common omitted factors or unobserved error terms. Ignoring such simultaneity may reduce the precision of parameter estimates and model efficiency (Aguero-Valverde and Jovanis, 2009; El-Basyouny and Sayed, 2009a; Lee et al., 2015). To cope with this problem, multivariate modelling approaches, which simultaneously model crash counts either by injury severity or collision type, have recently been proposed in the safety literature. The most popular and notable multivariate