Analytic Methods in Accident Research 17 (2018) 32–46

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

Multivariate random parameters zero-inflated negative binomial regression for analyzing urban midblock crashes

Chenhui Liu [a,b,*], Mo Zhao [c], Wei Li [d], Anuj Sharma [a]

[a] Department of Civil, Construction, and Environmental Engineering, Iowa State University, InTrans, 2711 South Loop Drive, Suite 4700, Ames, IA 50010-8664, United States
[b] Department of Statistics, Iowa State University, United States
[c] Virginia Department of Transportation, 530 Edgemont Rd, Charlottesville, VA 22903, United States
[d] Department of Information Systems, Statistics, and Management Science, The University of Alabama, 370 Alston Hall, 361 Stadium Drive, Tuscaloosa, AL 35487-0226, United States

A R T I C L E I N F O

Article history:
Received 12 August 2017
Received in revised form 7 March 2018
Accepted 8 March 2018
Available online 14 March 2018

Keywords:
Unobserved heterogeneity
Multivariate random parameters zero-inflated negative binomial model
Crash frequency
Urban midblock segments
Bayesian

A B S T R A C T

Urban midblock crashes are influenced mainly by traffic operation and roadway geometric features. In this paper, 10-year crash data from 1,506 directional urban midblock segments in Nebraska were analyzed using the multivariate random parameters zero-inflated negative binomial model to account for unobserved heterogeneity produced by correlations across segments, correlations across crash collision types, excessive zero crashes, and over dispersion. The multivariate random parameters zero-inflated negative binomial model was superior to many common crash frequency models in terms of both goodness of fit and prediction accuracy. Compared with the multivariate fixed parameters zero-inflated negative binomial model, the multivariate random parameters zero-inflated negative binomial model identified fewer key influencing factors and revealed segment-specific effects of these factors on different crash types. It showed that the number of lanes, annual average daily traffic per lane, and segment length might have non-positive effects on crash frequencies. Segments with a speed limit of 45 mph had fewer crashes than did those with lower speed limits, and there were fewer crashes on the segments in Omaha than on those in Lincoln. It was also found that neither the presence of a shoulder, on-street parking, or one-way traffic, nor lane width had significant influences on crash frequencies. These findings are informative for transportation agencies to take correct and efficient measures to accommodate diverse transportation demands without reducing traffic safety. By contrast, the fixed parameters model produced results consistent with intuition, but the results were insufficient to provide actionable recommendations.

1. Introduction

Traffic crashes can be divided into junction crashes and non-junction crashes based on where they occur (National Center for Statistics and Analysis, 2017). Non-junction crashes, also referred to as midblock crashes, are crashes that occur on roadway segments. In 2015, they accounted for 41.7% of the total number of crashes and 63.3% of fatal crashes in the United

* Corresponding author at: Department of Civil, Construction, and Environmental Engineering, Iowa State University, InTrans, 2711 South Loop Drive, Suite 4700, Ames, IA 50010-8664, United States.

E-mail addresses: cliu9@iastate.edu (C. Liu), mo.zhao@vdot.virginia.gov (M. Zhao), anujs@iastate.edu (A. Sharma).

https://doi.org/10.1016/j.amar.2018.03.001
2213-6657/

States (National Center for Statistics and Analysis, 2017). Thus, reducing midblock crashes is critical for improving traffic safety. Although midblock crashes are usually not directly influenced by junctions, they are greatly influenced by traffic operation and roadway geometric factors, which are much more complex on urban roadways than on rural roadways. On one hand, urban roadway segments usually have large traffic volumes and face diverse traffic demands, which might increase crash opportunities; for example, an increase in the number of crosswalks might increase the frequency of pedestrian crashes. On the other hand, urban development might limit or even reduce available roadway space, which might also increase crash risk; for example, vehicle lanes may be narrowed to make room for biking lanes and on-street parking. This predicament requires transportation agencies to determine what traffic operation and roadway geometric factors really influence the frequency of urban midblock crashes so that they can take effective measures to accommodate traffic demands without reducing traffic safety.

Previous studies have shown that important traffic operation and roadway geometric factors influencing midblock crashes include traffic volume (Bonneson and Mccoy, 1997; Greibe, 2003; Dumbaugh, 2006; Zhang et al., 2012; Manuel et al., 2014; Ferreira and Couto, 2015), speed limit (Greibe, 2003; Dumbaugh, 2006; Pande et al., 2010), on-street parking (Bonneson and Mccoy, 1997; Greibe, 2003), lane width (Greibe, 2003; Manuel et al., 2014), median type (Bonneson and Mccoy, 1997; Sawalha and Sayed, 2001), median width (Dumbaugh, 2006), number of lanes (Sawalha and Sayed, 2001; Greibe, 2003; Dumbaugh, 2006), land use (Bonneson and Mccoy, 1997; Sawalha and Sayed, 2001; Greibe, 2003), pavement condition (Usman et al., 2010; Xiong et al., 2014; Zeng and Huang, 2014), access points (Lee et al., 2011; Zeng and Huang, 2014), and so on. However, studies' findings have often been inconsistent, that is, some factors might have had different effects in different studies. For example, speed limit was found to be not significant for midblock crash frequencies on a 27-mile urban arterial in Florida Department of Transportation District 5 (Dumbaugh, 2006), whereas it was the most important variable for midblock crash frequencies on a 19.659-mile corridor of U.S. Route 19 in Pasco County, Florida (Pande et al., 2010). This inconsistency implies that, in practice, the effects of some factors on crashes might be location specific. Ignoring this unobserved heterogeneity might produce biased and inefficient estimated parameters, leading to erroneous inferences and predictions (Mannering et al., 2016).

One solution to account for unobserved heterogeneity across observations in crash frequency analysis is to adopt random parameters count data models (Lord and Mannering, 2010; Chen and Tarko, 2014; Venkataraman et al., 2014; Barua et al., 2015, 2016; Coruh et al., 2015; Alarifi et al., 2017; Bhat et al., 2017; Chen et al., 2017; Rista et al., 2017). Compared to fixed parameters models assuming the same effects of factors on all observations, random parameters models can capture the observation-specific effects of factors on crash frequency and have also been widely applied in crash injury severity analyses (Russo et al., 2014; Zhao and Khattak, 2015, 2017, Behnood and Mannering, 2016, 2017a, 2017b; Naik et al., 2016; Anderson and Hernandez, 2017; Fountas and Anastasopoulos, 2017; Seraneeprakarn et al., 2017) and crash rate analyses (Anastasopoulos, 2016). Especially, for the data where one entity has multiple observations, such as panel data, group-specific random parameters models may be adopted to account for heterogeneity among groups (Wu et al., 2013; Sarwar et al., 2017). More details about random parameters formulations can be seen in the study by Mannering et al. (2016).

Crash data usually can be divided into multiple types based on different criteria. For example, midblock crashes can be divided based on the type of collision: rear-end crashes, right-angle crashes, side-swipe (same direction) crashes, single-vehicle crashes, overturn crashes, and so on. A single factor might be expected to have different effects on different collision types, causing different outcomes. Thus, identifying the specific significant factors for each collision type is important for transportation agencies so they can take accurate countermeasures to reduce specific types of collision. When these crashes are jointly analyzed, multivariate count data models are necessary, as univariate models may produce biased and inefficient results because the unobserved heterogeneity often present across crash types is ignored (Huang et al., 2008; Dong et al., 2014a; Mannering et al., 2016). Most multivariate count data models in literature were derived from the multivariate Poisson log-normal (MVPLN) model (Ma et al., 2008; El-Basyouny and Sayed, 2009; Aguero-Valverde and Jovanis, 2010; Barua et al., 2014; Zhan et al., 2015; Serhiyenko et al., 2016; Huang et al., 2017; Osama and Sayed, 2017; Zhao et al., 2017; Wang et al., 2018), which is flexible enough to accommodate various correlations among crash types, but it does not work well for crash data with excess zeros (Dong et al., 2014a). In addition to the multivariate Poisson log-normal model, the natural extensions of the Poisson and negative binomial (NB) models to multivariate data, i.e., the multivariate Poisson (MVP) model (Johnson et al., 1997; Ma and Kockelman, 2006) and the multivariate negative binomial (MVNB) model (Anastasopoulos et al., 2012; Chen et al., 2017), also have been used in some studies. The multivariate Poisson/negative binomial models assume positive correlations across crash types, but they cannot deal with crash data with excess zeros either, as the marginal distribution per crash type is still a Poisson/negative binomial model.

The zero-inflated models are often adopted for univariate count data with excess zeros (Lambert, 1992; Lord et al., 2005). The excess zeros in crash frequency data can be explained in two ways for zero-inflated models. One explanation is that there is a two-state crash-generating process: (i) a normal count state and (ii) an accident-free state, which can be thought of as a nearly safe state, with accidents occurring extremely rarely (Malyshkina and Mannering, 2010). The other explanation is that there is a two-state crash-reporting process: (i) one in which accidents did occur, but they were not reported for some reason, such as for minor crashes, which were not necessary to report, or hit-and-run crashes, i.e., a crash-underreporting state, and (ii) one in which all accidents that occurred were reported, i.e., a normal crash reporting state. This explanation applies to many scenarios, as crash underreporting has been found to be common in practice (Hauer and Hakkert, 1988; Elvik and Mysen, 1999; Yamamoto et al., 2008; Lord and Mannering, 2010; Yannis et al., 2014). Both explanations may justify the application of zero-inflated models in our case, although it is difficult to determine what the truth is by observing the