Accident Analysis and Prevention 71 (2014) 319–326

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Finite mixture modeling for vehicle crash data with application to hotspot identification

Byung-Jung Park[a,1], Dominique Lord[b,*], Chungwon Lee[c,2]

[a] Department of Transportation Engineering, Myongji University, South Korea
[b] Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, United States
[c] Department of Civil and Environmental Engineering, Seoul National University, South Korea

ARTICLE INFO

Article history:
Received 22 April 2014
Received in revised form 26 May 2014
Accepted 27 May 2014

Keywords:
Finite mixture model
Negative binomial
Overdispersion
Vehicle crash data
Hotspot identification
False positive and negative

ABSTRACT

The application of finite mixture regression models has recently gained an interest from highway safety researchers because of its considerable potential for addressing unobserved heterogeneity. Finite mixture models assume that the observations of a sample arise from two or more unobserved components with unknown proportions. Both fixed and varying weight parameter models have been shown to be useful for explaining the heterogeneity and the nature of the dispersion in crash data. Given the superior performance of the finite mixture model, this study, using observed and simulated data, investigated the relative performance of the finite mixture model and the traditional negative binomial (NB) model in terms of hotspot identification. For the observed data, rural multilane segment crash data for divided highways in California and Texas were used. The results showed that the difference measured by the percentage deviation in ranking orders was relatively small for this dataset. Nevertheless, the ranking results from the finite mixture model were considered more reliable than the NB model because of the better model specification. This finding was also supported by the simulation study which produced a high number of false positives and negatives when a mis-specified model was used for hotspot identification. Regarding an optimal threshold value for identifying hotspots, another simulation analysis indicated that there is a discrepancy between false discovery (increasing) and false negative rates (decreasing). Since the costs associated with false positives and false negatives are different, it is suggested that the selected optimal threshold value should be decided by considering the trade-offs between these two costs so that unnecessary expenses are minimized.

1. Introduction

The identification of accurate crash hotspots has been an important research topic in highway safety because it directly affects the efficient use of resources for safety improvements. Identifying a safe site as hazardous or identifying an unsafe site as safe could result in inefficient investments and additional loss of lives. While a hotspot, also referred to as a blackspot (Maher and Mountain, 1988; Elvik, 2007), site with promise (Hauer, 1996; Hauer et al., 2002), or site with high potential for safety improvement (Persaud, 1999), can be generally defined as a location (roadway segment, intersection or interchange) with high crash risk, it has been defined in many different ways depending on how to measure the crash risk at a particular location. For example, Hakkert and Mahalel (1978) proposed that a hotspot be defined as a site that has a crash frequency which is significantly higher than expected at some prescribed level of significance. McGuigan (1981) proposed the use of potential for accident reduction, as the difference between the observed and expected number of crashes at a site given exposure. More recently, Elvik (2008) proposed a theoretical definition of a hotspot as being any location that has a higher expected number of accidents than other similar locations as a result of local risk factors.

A naïve approach to identifying hotspots is to rank locations based on their observed accident frequencies. However, because of the rare and random nature of accident occurrences, this approach tends to be very sensitive to random variations. Miaou and Song (2005) illustrated the limitation associated with the naïve or raw crash-risk approach in ranking using simple simulation procedures. To better address the random fluctuation, researchers have used statistical modeling-based approaches that apply random effect or

* Corresponding author. Tel.: +1 979/458 3949; fax: +1 979/845 6481.
E-mail addresses: bjpark@mju.ac.kr (B.-J. Park), d-lord@tamu.edu (D. Lord).
[1] Tel.: +82 31 330 6499; fax: +82 31 330 2885.
[2] Tel.: +82 2 880 7368; fax: +82 2 873 2684.

http://dx.doi.org/10.1016/j.aap.2014.05.030
0001-4575/

Bayesian methods and compared their relative performances in identifying hotspots (Miranda-Moreno et al., 2005; El-Basyouny and Sayed, 2006). Miranda-Moreno et al. (2005), for example, pointed out that various models and ranking criteria can lead to different lists of hazardous locations. In order to evaluate the performance of different hotspot identification methods, Cheng and Washington (2008) developed four new evaluation tests and applied them to select the most appropriate hotspot identification method among the crash count ranking, the crash rate ranking, the crash reduction potential, and the empirical Bayes (EB) method. In their subsequent papers (Cheng et al., 2010a,b), they also applied those tests to answer the question regarding which kind of criteria (crash counts vs. crash rates or crash counts vs. crash reduction potential) should be employed to identify hotspots. While Cheng and Washington (2008) and Montella (2010) showed that the EB approach is the most consistent and reliable method for identifying hotspots based on innovative robust evaluation criteria, Huang et al. (2009) explained an essential theoretical advantage of the full Bayesian (FB) approach over the EB approach. Using Singapore intersection crash data, they found that the FB hierarchical model significantly outperformed the EB approach in correctly identifying hotspots. In the light of their work, we also adopted the FB approach for this study.

Although many alternative statistical models and ranking criteria are available in the literature for identifying hotspots in highway safety analyses, the main difficulty arises from the inability to differentiate between sites that are truly high risk from those that happen to have experienced random fluctuations during a period of observation (Cheng and Washington, 2005). In this respect, some researchers have recently adopted epidemiological criteria, such as "sensitivity" or "specificity"[3] to compare different statistical models or ranking criteria for identifying hotspots (Cheng and Washington, 2005; Miranda-Moreno, 2006; Elvik, 2008). These criteria can provide information about "false positives" (identifying a safe site as a hotspot) and "false negatives" (identifying a hotspot as a safe site). These criteria along with others will also be used in this paper to compare the relative performance of alternative models for identifying hotspots.

During the past few years, many methodological innovations in the development of statistical models have been made for analyzing vehicle crash data to overcome the overdispersion problem. Various types of crash prediction models used by highway safety analysts are well summarized in Lord and Mannering (2010) and more recently in Mannering and Bhat (2014). Among them, applications of a finite mixture regression model have gained an interest from highway safety researchers because of its considerable potential for addressing unobserved heterogeneity (Park and Lord, 2009; El-Basyouny and Sayed, 2010; Zou et al., 2013, 2014). Finite mixture models assume that the observations of a sample arise from two or more unobserved components with unknown proportions, which allows a great modeling flexibility over traditional single aggregate models. For example, using urban 4-legged signalized intersection crash data in Toronto, Park and Lord (2009) showed the possible existence of two distinct sub-populations in the data, each having different regression coefficients and degrees of over-dispersion, and recommended that transportation safety analysts use finite mixture models over a traditional single aggregate model, especially when the data are suspected to belong to different groups. Zou et al. (2013) demonstrated the advantages of the finite mixture model with varying weight parameters over a fixed weight parameter model using two datasets, the same data described in Park and Lord (2009) and 4-lane undivided rural segments in Texas. In short, both fixed and varying weight parameter finite mixture models have been shown to be useful for explaining the heterogeneity and the nature of the dispersion in crash data.

Given the superior performance of finite mixture models for vehicle crash data analysis, there is a need to investigate whether or not this type of model would result in important differences in various highway safety analyses as compared to the commonly used models, such as the negative binomial (NB) regression model. Therefore, the objectives of this study are, first, to investigate the relative performance of two alternative models (i.e., the two-component finite mixture of NB regression model (FMNB-2) and the NB regression model) in terms of hotspot identification, and second, to demonstrate what the consequences will be if a misspecified model is used for hotspot identification. Both empirical and simulation data were used to achieve these objectives.

2. Finite mixture of NB regression model

This section describes the model structure and parameter estimation method of K-component finite mixture of negative binomial (NB) regression models (referred to as FMNB-K). More details and general structure of finite mixture models can be found in McLachlan and Peel (2000) and Früwirth-Schnatter (2006).

2.1. Model structure of FMNB-K

The underlying assumption of finite mixture of regression models is that there are a finite number (K) of unobservable categories of observations and the heterogeneity arises from different values of regression coefficients caused by missing variables. The probability density function, mean and variance of the FMNB-K are expressed as follows:

$$p(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} w_k NB(\mu_{i,k}, \phi_k) = \sum_{k=1}^{K} w_k \left[\frac{\Gamma(y_i + \phi_k)}{\Gamma(y_i + 1)\Gamma(\phi_k)} \left(\frac{\mu_{i,k}}{\mu_{i,k} + \phi_k} \right)^{y_i} \left(\frac{\phi_k}{\mu_{i,k} + \phi_k} \right)^{\phi_k} \right] \quad (1)$$

$$E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \mu_{i,k} w_k \quad (2)$$

$$Var(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) + \left(\sum_{k=1}^{K} W_k \mu_{i,k}^2 \left(1 + \frac{1}{\phi_k} \right) - E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta})^2 \right) \quad (3)$$

where,

y_i is a random variable of ith observation ($i = 1,2,\ldots n$);
w_k is the weight of component k which Sum to 1 ($\sum_{k=1}^{K} w_k = 1$);
$\mu_{i,k} = \exp(\mathbf{x_i}\boldsymbol{\beta}_k)$ is the mean of component k;
x_i is a vector of covariates;
$\boldsymbol{\beta}_k$ and ϕ_k are the regression coefficients and the dispersion parameter of the NB;
distribution for component k; and
$\boldsymbol{\Theta} = \{(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k), (\phi_1, \ldots, \phi_k), (w_1, \ldots, w_K)\}$ is a vector of all unknown parameters.

It can be seen that when $\phi_k = 0$ in each component the FMNB-K model reduces to the finite mixture of Poisson regression models (FMP-K). The FMNB models, therefore, allow for additional heterogeneity within components not captured by the covariates. If additional heterogeneity is present within the components, the

[3] $Sensitivity = \frac{number\ \ of\ \ detected\ hotspots}{number\ of\ true\ hotspots}$; $Specificity = \frac{number\ of\ detected\ non-hotspots}{number\ of\ true\ non-hotspots}$.