Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 958206, 10 pages
http://dx.doi.org/10.1155/2015/958206
Publication Year 2015

Research Article

Application of the Empirical Bayes Method with the Finite Mixture Model for Identifying Accident-Prone Spots

Yajie Zou,[1] Kristian Henrickson,[2] Lingtao Wu,[3] Yinhai Wang,[2] and Zhaoru Zhang[4]

[1]*Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, Shanghai 201804, China*
[2]*Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195-2700, USA*
[3]*Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, USA*
[4]*Institute of Oceanology, Shanghai Jiao Tong University, Shanghai 200240, China*

Correspondence should be addressed to Lingtao Wu; wulingtao@gmail.com and Yinhai Wang; yinhai@uw.edu

Received 18 January 2015; Accepted 2 April 2015

Academic Editor: Paolo Maria Mariano

Hotspot identification (HSID) is an important component of the highway safety management process. A number of methods have been proposed to identify hotspots. Among these methods, previous studies have indicated that the empirical Bayes (EB) method can outperform other methods for identifying hotspots, since the EB method combines the historical crash records of the site and expected number of crashes obtained from a safety performance function (SPF) for similar sites. However, the SPFs are usually developed based on a large number of sites, which may contain heterogeneity in traffic characteristic. As a result, the hotspot identification accuracy of EB methods can possibly be affected by SPFs, when heterogeneity is present in crash data. Thus, it is necessary to consider the heterogeneity and homogeneity of roadway segments when using EB methods. To address this problem, this paper proposed three different classification-based EB methods to identify hotspots. Rural highway crash data collected in Texas were analyzed and classified into different groups using the proposed methods. Based on the modeling results for Texas crash dataset, it is found that one proposed classification-based EB method performs better than the standard EB method as well as other HSID methods.

1. Introduction

For the purpose of prioritizing safety improvements on roadway network, identifying sites with consistently elevated accident risk, often referred to as hotspots or black spots, is of critical importance. To address this need, a number of analytical methods for hotspot identification (HSID) have been developed over the last several decades, with the overarching objective of optimizing the allocation of limited funding. An inaccurate HSID method will result in inefficient allocation of safety treatment resources, with potentially serious costs in terms of overall safety performance of the network. The need for accurate methods to identify and prioritize accident-prone locations is underscored by the U.S. 2012 Federal Moving Ahead for Progress in the 21st Century Act (MAP-21), which emphasizes data-driven crash risk analysis and safety treatment prioritization. Further, the performance reporting requirements outlined in MAP-21 provide an additional layer of incentive for public agencies to maximize the impact of safety spending by selecting and treating sites with high improvement potential.

A number of papers in past years have focused on the accident frequency (AF) or accident rate (AR) based HSID methods, which rely on observed accident counts as the primary measure of accident risk. Because sites are ranked and identified based on observed accident data only, there is no mechanism for identifying sites with elevated risk (due to some combination of geometric and traffic characteristics) but few accidents. Further, these methods cannot distinguish between actual high risk locations and those with higher occurrence of accidents due to random fluctuations. The empirical Bayes (EB) HSID method addresses these issues by combining two clues, the historical crash record of the entity and the expected number of crashes obtained from a safety performance function (SPF) for similar entities. This approach is less sensitive to random fluctuations in

accident frequency and in theory can identify truly high risk locations with greater accuracy. Building on the EB approach, additional methods have been developed based on estimated accident reduction potential (ARP). Such methods attempt to quantify the difference between the actual accident count at the location of interest (as estimated using EB method) and the expected accident count for similar locations, under the supposition that this difference represents the potential for improvement.

Unfortunately, the definition of "similar" sites is a somewhat open question, and the accuracy of the EB method depends largely on the selection of the reference population. When estimating the SPF, the studied crash data are often collected from different geographic locations to ensure the adequacy of sample size for valid statistical estimation. Since crash data observed from different geographic locations may exhibit different characteristics, the aggregation of these crash data may result in heterogeneity. For the aggregated crash data, it is reasonable to assume that the sites with different combinations of characteristics (i.e., geometric design features) can constitute distinct subpopulations [1]. Under this assumption, applying the EB method to the entire crash data may become inappropriate because the second clue of the EB method requires the homogeneity of the crash data. Therefore, as proposed by some previous studies [2–8], it is reasonable to assume that the crashes on highway entities (i.e., road segments or intersections) are generated from a certain number of hidden subpopulations. Since entities are heterogeneous across but homogeneous within the subpopulations, the EB method can be applied to the crash data in each subpopulation.

The primary objectives of this research are (1) to propose three different classification-based EB methods to identify accident-prone sites and (2) to compare the proposed methods with the commonly used HSID approaches (i.e., AF, AR, EB, and ARP methods). To accomplish the objective of this study, the finite mixture of NB regression models with fixed or varying weight parameter is considered to classify the crash data into different subgroups. The effectiveness of the proposed HSID approaches is examined using the crash data collected at 4-lane undivided rural highways in Texas and performance is evaluated using criteria proposed by Cheng and Washington [9].

2. Background

2.1. Hotspot Identification Methods. AF based HSID methods have been in use for many years. Such approaches typically rank locations or segments along a highway by observed accident count over a specified time interval and define hotspots as those exceeding some critical value [10]. Road segments or intersections are ranked by accident count among similar locations (such as along a relatively homogeneous section of highway), to insure that the identified hotspots represent specific opportunities for remediation instead of some inherent characteristic of a particular roadway class or driver population. One criticism often raised with regard to the AF method is that this approach lacks the ability to differentiate between actual hotspots and locations with increased accident frequency attributable to the randomness of traffic accidents [9, 10].

It is readily apparent that, all else being equal, a segment with higher traffic volume can be expected to have a higher accident count, and so hotspots identified using AF methods tend to overrepresent high volume locations that may or may not be amenable to remediation efforts [11]. In response, AR methods have been developed which rely on accident count per unit traffic volume for HSID, typically in units of accidents per million vehicle miles traveled. Similar to the AF methodology, sites are ranked by accident rate and those exceeding a critical value are identified as hotspots. Implicit in this approach is the assumption that accident count and exposure are linearly related, which is often not the case. In addition, by normalizing accident count by entering traffic volume, locations with very low traffic volume are sometimes over represented [11, 12].

The EB method for traffic accident HSID was introduced by Abbess et al. [13] to address issues with existing methodologies, most notably regression-to-the-mean (RTM) bias and low precision due to limited accident history. It has since been refined and widely used in a range of safety performance modeling applications [14–16]. In the EB crash modeling procedure, the expected number of crashes at a location is estimated by combining two pieces of information: (1) the accident count at the location of interest and (2) the expected accident count at locations determined to be similar based on traffic and roadway characteristics [17]. It is assumed that the actual accident count for the location of interest is available, and the expected accident count for similar locations is generally estimated from the SPF. The SPF describing accident counts as a function of traffic volume, lane width, and so forth is typically fitted using the negative binomial (NB) regression model. The observed accident count for a given roadway segment is combined with the expected value estimate as shown in

$$\widehat{\widehat{\mu}}_i = w_i \widehat{\mu}_i + (1 - w_i)\, y_i, \tag{1}$$

where $\widehat{\widehat{\mu}}_i$ is the EB estimate of the expected number of crashes per year for site i; $\widehat{\mu}_i$ is the estimated number of crashes per year by the SPF for given site i (estimated using a NB model); $w_i = 1/(1+\alpha\widehat{\mu}_i)$ is the weight factor estimated as a function of $\widehat{\mu}_i$ and dispersion parameter α; and y_i is the observed number of crashes per year at site i.

Another measure often used in HSID is ARP. It was originally suggested that ARP be estimated as the difference between the observed accident count at the site of interest and the expected count as estimated from a set of reference sites. More recently, it has been proposed that the observed accident count at the site of interest be replaced by the EB-estimated accident count. This approach can account for random fluctuations in accident frequency and so gives a better estimate of the true safety of the location of interest. Using the EB-estimated accident count, the ARP is calculated as shown in

$$\mathrm{ARP}_i = w_i \widehat{\mu}_i + (1 - w_i)\, y_i - \widehat{\mu}_i, \tag{2}$$

where ARP_i = ARP for site i.

Persaud et al. [12] suggested that a better estimate of the true ARP can be derived using a full predictor set in the EB-estimated accident count and a subset of available regressors (i.e., those not describing a correctable site-specific geometric feature) in the expected accident count model. This way, the estimated ARP is a measure of the difference between the EB-estimated "true" safety and the expected safety of what could be considered a base scenario.

2.2. Negative Binomial Model. In highway safety, the dispersion parameter of NB models refines the estimates of the predicted mean when the EB method is used. So far, the NB distribution is the most frequently used model by transportation safety analysts to generate SPFs [18]. The NB model has the following structure: the number of crashes y_i during some time period is assumed to be Poisson distributed, which is defined by

$$p(y_i \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y_i!}, \tag{3}$$

where λ is the mean response of the observation.

The NB distribution can be viewed as a mixture of Poisson distributions where the Poisson rate is gamma distributed. For the complete derivation of the NB model, the reader is referred to Hilbe [19]. The probability density function (PDF) of the NB is defined as follows:

$$f(y_i \mid \mu, \alpha) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(1/\alpha)\,\Gamma(y_i + 1)} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y_i} \left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}, \tag{4}$$

where μ is the mean response of the observation and α is the dispersion parameter.

Compared to the Poisson distribution, the NB distribution can allow for overdispersion.

2.3. Finite Mixture of NB Regression Models. This study adopts a g-component finite mixture of NB regression models (termed as the FMNB-g model) to classify heterogeneous crash data. For the FMNB-g model, it is assumed that the marginal distribution of y_i follows a mixture of NB distributions, as shown in the following:

$$f_Y(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{j=1}^{g} w_j \mathrm{NB}(\mu_{ij}, \phi_j) = \sum_{j=1}^{g} w_j \left[\frac{\Gamma(y_i + \phi_j)}{\Gamma(y_i+1)\,\Gamma(\phi_j)} \left(\frac{\mu_{ij}}{\mu_{ij}+\phi_j}\right)^{y_i} \cdot \left(\frac{\phi_j}{\mu_{ij}+\phi_j}\right)^{\phi_j}\right] \tag{5}$$

$$E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{j=1}^{g} \mu_{ij} w_j \tag{6}$$

$$\mathrm{Var}(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) = E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}) + \left(\sum_{j=1}^{g} w_j \mu_{ij}^2 \left(1 + \frac{1}{\phi_j}\right) - E(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta})^2\right), \tag{7}$$

where w_j is the weight of component j (weight parameter), with $w_j > 0$, and $\sum_{j=1}^{g} w_j = 1$; g is the number of components; $\mu_{ij} = \exp(\mathbf{x}_i \boldsymbol{\beta}_j)$, the mean rate of component j; $\mathbf{x}_i$ is a vector of covariates; $\boldsymbol{\beta}_j$ is a vector of the regression coefficients for component j; $\boldsymbol{\Theta} = \{(\phi_1, \ldots, \phi_g), (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_g), \mathbf{w}\} = \{(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g), \mathbf{w}\}$ for $i = 1, 2, \ldots, n$; and $\boldsymbol{\theta}_j$ are the vectors of parameters for the component j.

The term w_j is assumed to be fixed for the FMNB-g models. However, instead of estimating a fixed weight parameter, a generalized FMNB-g model (GFMNB-g model) can be derived by modeling the varying weight w_{ij} as a function of covariates, shown in (8). The GFMNB-g model allows each entity to have a different weight that is dependent on the sites' attributes (i.e., covariates). Previous studies [1] have shown that the GFMNB-g model can provide more reasonable classification results than the FMNB-g model. Note that the GFMNB-g model has the same PDF shown in (5), but the weight factors are calculated using

$$\frac{w_{ij}}{w_{ig}} = e^{\gamma_{0j}} e^{\boldsymbol{\gamma}_j \mathbf{x}_i}, \tag{8}$$

where w_{ij} is the estimated weight of component j at segment i; $\boldsymbol{\gamma}_j = (\gamma_{0j}, \gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{mj})'$ are the estimated coefficients for component j, m being the number of coefficients; and $\mathbf{x}_i$ is a vector of covariates.

2.4. Classification-Based EB Methods. Since the aggregated crash data may contain heterogeneity, the FMNB-g and GFMNB-g models are proposed to classify the crash data into different subpopulations. Besides the mixture model, a simple mean-based grouping method is also considered and compared with the FMNB-g and GFMNB-g models. After separating the aggregated crash data into different subgroups, the EB estimates are calculated based on the crash data in each individual subpopulation. The three grouping methods and the procedure for prioritizing the hotspots are described as follows.

The first classification method assumes the crash data are generated from two subpopulations (i.e., one subpopulation contains accident-prone sites and the other consists of low-risk sites). To separate the sites into two groups, the mean of the number of crashes across the entire dataset is calculated. The sites with the observed number of crashes greater than the mean are labeled as the accident-prone group, and the sites with the observed number of crashes smaller than the mean are labeled as low-risk group. The mean-based classification method for hotspot identification consists of three steps. First, separate the entire crash data into two subgroups using the crash mean as the threshold value. Second, the NB regression model is estimated using the

crash data in accident-prone group and the corresponding EB estimates are calculated; likewise, the NB regression model is estimated using the crash data in the low-risk group and the EB estimates are obtained. Third, the two sets of EB estimates obtained from step two are aggregated and ranked; then the hotspots are identified based on the aggregated EB estimates. Hereinafter, we denote this HSID approach as the mean-based EB method.

The second and third classification methods assume that the aggregated crash data contain heterogeneity. Heterogeneity implies that crash data are generated from different subpopulations (i.e., crash data in the same subpopulation share common characteristics, while crash data across the subpopulations may exhibit different characteristics). Given the advantages of finite mixture models in describing the heterogeneity in crash data, FMNB-g and GFMNB-g models are used to classify the crash data into g components based on the site characteristics. The FMNB-based or GFMNB-based classification method for hotspot identification consists of four steps. First, fit the FMNB-g or GFMNB-g model to entire crash data, and select the number of components using the Bayesian information criterion (BIC). Second, after determining the number of components, separate the entire crash data into g subgroups using the FMNB-g or GFMNB-g model. Third, the NB regression model is estimated using the crash data in each subgroup and the corresponding EB estimates are obtained. Fourth, the g sets of EB estimates obtained from step three are aggregated and ranked; then the hotspots are identified based on the aggregated EB estimates. Hereinafter, the second and third HSID approaches are denoted as the FMNB-based EB method and the GFMNB-based EB method, respectively.

3. Hotspot Identification Method Evaluation Criteria

With the objective of prioritizing safety treatments, the performance of a HSID method must be described in terms of both its ability to identify truly high risk sites and its ability to accurately rank sites by accident risk. In combination, the three tests proposed by Cheng and Washington [9] address both of these performance considerations.

3.1. Site Consistency Test. Cheng and Washington [9] introduced the Site Consistency Test (SCT) as a measure of a HSID method's consistency over subsequent time periods. It is based on the idea that actual high risk sites will experience consistently elevated accident frequencies, which means that a desirable HSID method should identify as hotspots sites which can be expected to have poor safety performance in subsequent time periods assuming no safety treatments or other changes have been applied. The SCT considers the sites identified as hotspots by method j during time period i and compares the methods based on the sum of accidents at high risk sites during future time period $i + 1$. The optimal HSID method as determined by the SCT is the one with the greatest number of accidents occurring on the sites identified as high risk by that method during time period $i + 1$. Out of n sites, a threshold is defined for each method such that $c \times n$ are designated as high risk by each method. For example, with $n = 200$ total sites and $c = 0.05$, 10 sites will be selected by each method under comparison. With sites ranked $1, 2, \ldots, c \times n$ in order of increasing accident risk, the test statistic for each method is defined by Cheng and Washington [9] as shown below in (9). The best method according to the SCT, then, is that which produces the highest T_{sc}:

$$T_{sc(j)} = \sum_{k=n-cn}^{n} C_{k,\text{method}=j(i),i+1}, \tag{9}$$

where n is the total number of sites under analysis; C is the accident count for site k; c is the threshold for high risk sites, defined as the fraction of all sites n that are designated high risk; j is the HSID method which is being compared; and i is the time period of observation.

3.2. Method Consistency Test. Similar to the SCT, the Method Consistency Test (MCT) is based on the notion that actual high risk sites will consistently experience poor safety performance, assuming that operating conditions are similar and that no safety treatments have been applied. However, the MCT estimates the performance of a HSID method by the extent to which the same sites are identified as hotspots over two consecutive time periods. Given $c \times n$ hotspots identified by each method j, the best performing method is that which identifies the greatest number of hotspots that are consistent between time period i and time period $i + 1$. From Cheng and Washington [9], the performance of method j is computed as shown in

$$T_{MC(j)} = \{k_{n-cn}, k_{n-cn+1}, \ldots, k_n\}_{j,i} \cap \{k_{n-cn}, k_{n-cn+1}, \ldots, k_n\}_{j,i+1}. \tag{10}$$

3.3. Total Rank Differences Test. Similar to the MCT, the Total rank Differences Test (TRDT) is a measure of a HSID method's consistency across multiple time periods, again assuming that none of the sites have undergone safety treatments. However, the TRDT explicitly takes into account the safety performance ranking assigned by each method and estimates performance based on the ranking consistency between subsequent time periods. While the MCT would assign a high score to a method which consistently identifies the same sites in two time periods as being in the top $c \times 100\%$ in terms of accident risk, the TRDT considers the relative rankings of all sites identified in time period i. The performance of each method is calculated as the sum of differences between the rank assigned to all $c \times n$ high risk sites in time period i and the rank assigned to the same $c \times n$ sites in time period $i + 1$. Note that the sites identified in time period i are compared by rank in the two time periods whether or not they are identified as high risk in time period $i + 1$. The performance measure for the TRDT is computed as described by Cheng and Washington 2008 [9], shown in

$$T_{TRDT(j)} = \sum_{k=n-\alpha n}^{n} \left(\Re\left(k_{j,i}\right) - \Re\left(k_{j,i+1}\right)\right), \tag{11}$$

Table 1: Summary statistics of characteristics for individual road segments in the Texas data for Periods 1 and 2.

Variable	Period 1 (1997 and 1998)			Period 2 (1999–2001)		
	Min.	Max.	Mean (SD)	Min.	Max.	Mean (SD)
Number of crashes	0	59	2.93 (4.81)	0	78	4.58 (7.81)
Average daily traffic over the study period (F)	40	24000	6391 (3835.01)	43.33	25333.3	6761.8 (4149.84)
Lane width (LW) (ft)	9.75	16.5	12.57 (1.59)	9.75	16.5	12.57 (1.59)
Total shoulder width (SW) (ft)	0	40	9.96 (8.02)	0	40	9.96 (8.02)
Curve density (CD)	0	18.07	1.43 (2.35)	0	18.07	1.43 (2.35)
Segment length (L) (miles)	0.1	6.28	0.55 (0.68)	0.1	6.28	0.55 (0.68)

SD: standard deviation.

Table 2: Modeling results of NB models for Periods 1 and 2.

Estimates	Period 1 (years 1997 and 1998)		Period 2 (years 1999, 2000, and 2001)	
	Value	SE	Value	SE
Intercept $\ln(\beta_0)$	−7.836	0.497	−8.116	0.453
Ln(average daily traffic) β_1	1.093	0.054	1.137	0.048
Lane width β_2	−0.044	0.020	−0.055	0.018
Total shoulder width β_3	−0.013	0.004	−0.012	0.004
Curve density β_4	0.026	0.014	0.024	0.013
α	0.825	0.062	0.792	0.049
Log-likelihood	2924.490		3409.444	
AIC	5860.980		6830.880	
BIC	5892.850		6862.763	

SE: standard error.

where $\Re(k_{j,i})$ is the rank of site k from method j for time period i.

4. Data Description

The crash dataset used for comparing different HSID methods was collected at 4-lane undivided rural segments in Texas as a part of NCHRP 17–29 research project. This dataset records crash data collected on 1,499 undivided rural segments over a five-year period from 1997 to 2001. In order to compare different HSID methods using the three criteria, the five-year crash data were divided into two time periods, Period 1 (years 1997 and 1998) and Period 2 (years 1999, 2000, and 2001). Table 1 provides the summary statistics for individual road segments in the Texas data for time periods 1 and 2.

5. Results

5.1. Modeling Results. This section describes the modeling results for the NB, FMNB-g, and GFMNB-g models. For the NB model, a mean function of the form shown in the following is adopted:

$$\mu_i = \beta_0 L_i F_i^{\beta_1} e^{\beta_2 * \mathrm{LW}_i + \beta_3 * \mathrm{SW}_i + \beta_4 * \mathrm{CD}_i}, \quad (12)$$

where μ_i is the estimated number of crashes at segment i over the study period; L_i is the segment length in miles for segment i; F_i is the traffic flow (average daily traffic over the study period) traveling on segment i; LW_i is the lane width in feet for segment i; SW_i is the total shoulder width in feet for segment i; CD_i is curve density (curves per mile) for segment i; and β_0, β_1, β_2, β_3, and β_4 are estimated coefficients.

The NB model is applied to the Texas data, and the parameters were estimated separately for Periods 1 and 2. The results of the fitted NB models are shown in Table 2, along with the NB dispersion parameter (α) estimated for each time period. The estimated coefficients are reasonable and consistent between the two time periods. This fitted NB model is used in the EB and ARP HSID methods only.

For the FMNB-g and GFMNB-g models, the mean functional form for each component is adopted as follows:

$$\mu_{j,i} = \beta_{j,0} L_i F_i^{\beta_{j,1}} e^{\beta_{j,2} * \mathrm{LW}_i + \beta_{j,3} * \mathrm{SW}_i + \beta_{j,4} * \mathrm{CD}_i}, \quad (13)$$

where $\mu_{j,i}$ are the estimated numbers of crashes at segment i for component j and $\beta_{j,0}$, $\beta_{j,1}$, $\beta_{j,2}$, $\beta_{j,3}$, and $\beta_{j,4}$ are the estimated coefficients for component j.

For the GFMNB-g model, the weight parameter is modeled using all available explainable variables:

$$\frac{w_{ij}}{w_{ig}} = e^{\gamma_{0j}} e^{\boldsymbol{\gamma}_{1j} * L_i + \boldsymbol{\gamma}_{2j} * F_i + \boldsymbol{\gamma}_{3j} * \mathrm{LW}_i + \boldsymbol{\gamma}_{4j} * \mathrm{SW}_i + \boldsymbol{\gamma}_{5j} * \mathrm{CD}_i}, \quad (14)$$

where w_{ij} is the estimated weight of component j at segment i and $\boldsymbol{\gamma}_j = (\gamma_{0j}, \gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{mj})'$ are the estimated coefficients for component j, m being the number of coefficients.

Table 3: BIC values for FMNB-g and GFMNB-g models with number of components $g = 2$, 3, and 4.

Model	Number of components		
	2	3	4
	Period 1		
FMNB	5904.54	5932.72	5969.38
GFMNB	5833.49	5868.99	5940.98
	Period 2		
FMNB	6820.89	6857.04	6870.91
GFMNB	6755.59	6810.57	6873.91

To evaluate the performance of classification-based EB methods, the crash data in Periods 1 and 2 are separated into different subgroups using the mean-based classification method and FMNB-g and GFMNB-g models. For each period, the NB regression model is estimated using the crash data in each subgroup and the corresponding EB estimates are calculated. For each period, the hotspots are identified based on the EB estimates obtained from each subgroup.

To determine the number of components for FMNB-g and GFMNB-g models, an approach suggested by Park et al. [20] is applied to the crash data in Periods 1 and 2. Specifically, this approach fits a series of models with an increasing number of components and selects the most plausible model using model choice criteria. As discussed by Eluru et al. [21], compared to the Akaike information criterion (AIC), the BIC imposes a higher penalty on overfitting with excess parameters. Thus, the BIC is preferred for selecting the optimal number of components. For the crash data in Periods 1 and 2, we applied the FMNB-g and GFMNB-g models with an increasing number of components $g = 2$, 3, 4. As shown in Table 3, $g = 2$ is preferred for both models, which provides the smallest BIC value. Overall, based on reported BIC values, we selected the optimal number of components $g = 2$ and use the FMNB-2 and GFMNB-2 models to group the crash data. Compared to the standard NB model, the GFMNB-2 model has a significantly smaller BIC value. The goodness-of-fit statistics suggest that the crash data may contain two subpopulations with different characteristics, rather than a single population.

The parameter estimation results for Periods 1 and 2 are provided in Table 4. Table 4 shows that some models include some insignificant covariates. Note that, for the FMNB-2 model in Period 1, the estimated coefficient for variable curve density in component 1 is counterintuitive and it may indicate that the FMNB-2 model provides an unreasonable group classification. Due to the inappropriate grouping, the estimated coefficient is counterintuitive.

5.2. Grouping Results. The FMNB-2 and GFMNB-2 models were used to classify the crash data observed in Period 1 (years 1997 and 1998). Based on the modeling results of FMNB-2 and GFMNB-2 models, the 1,499 road segments were classified into two groups by assigning each site to the component with the highest posterior probability. The posterior probability is used to calculate the probability that observation y_i is from component j [1]. In the EM algorithm, at iteration $r + 1$, the posterior probability $\hat{\varepsilon}_{ij}^{(r+1)}$ that observation y_i is from component j, given y_i and $\widehat{\boldsymbol{\Theta}}^{(r)}$, is defined as [22]

$$\begin{aligned}\hat{\varepsilon}_{ij}^{(r+1)} &= p\left(\delta_{ij} = 1 \mid y_i, \widehat{\boldsymbol{\Theta}}^{(r)}, \mathbf{x}_i\right) \\ &= \frac{\widehat{w}_{ij}^{(r)} f_j\left(y_i \mid \widehat{\boldsymbol{\theta}}_j^{(r)}, \mathbf{x}_i\right)}{\sum_{k=1}^{g} \widehat{w}_{ik}^{(r)} f_k\left(y_i \mid \widehat{\boldsymbol{\theta}}_k^{(r)}, \mathbf{x}_i\right)},\end{aligned} \tag{15}$$

where δ_{ij} is the indicator variable and $\widehat{w}_{ij}^{(r)} = p(\delta_{ij} = 1 \mid \widehat{\boldsymbol{\Theta}}^{(r)})$ is the prior probability that observation y_i is from component j, given $\widehat{\boldsymbol{\Theta}}^{(r)}$, which is estimated from iteration r.

The grouping results from mean-based classification method and FMNB-2 and GFMNB-2 models are provided in Table 5 for Periods 1 and 2. Note that the means and standard deviations of the variables are calculated for each group. For Periods 1 and 2, the two components generated from mean-based classification method show a significant difference in the mean value of number of crashes; on the other hand, the difference in the mean values of other variables is not very noticeable. For Periods 1 and 2, the two components in GFMNB-2 model demonstrate a remarkable difference in the mean values of segment length and curve density. Note that the difference in the mean values of variables is not significant in FMNB-2 model. For each period (i.e., Period 1 or 2), the NB regression model is estimated using the crash data in each subgroup and the corresponding EB estimates are obtained. For mean-based EB method, FMNB-based EB, and GFMNB-based EB methods, the hotspots are identified based on the EB estimates and the three criteria are calculated.

5.3. Testing Results. This section describes the HSID method performance evaluation and comparison results, which was conducted using the tests described in Section 3. For each test, seven HSID methods are compared over the two time periods identified in Table 1. The following high risk cutoff points are used to compare the methods: $c = \{0.99, 0.90, 0.95\}$. The c value in this case describes the fraction of the total number of points that will be identified as high risk. For example, the selected dataset contains 1,499 highway segments, which will result in approximately 150 high risk sites being identified as hotspots given $c = 0.90$.

As described previously, the SCT is a measure of a HSID method's ability to identify truly high risk sites, which it does by comparing the safety performance of sites identified as hotspots during an out of sample test observation period. Table 6 shows the accident count during time period 2, for hotspots identified by each method during time period 1. It is clear from these results that the EB with GFMNB-based classification method performs best for all high risk cutoff points c. The worst performing subpopulation EB method is FMNB-based classification in all cases.

The MCT is a measure of a HSID method's consistency over two subsequent time periods. That is, the best performing method will identify the largest number of sites

Table 4: Parameter estimates for the FMNB-2 and GFMNB-2 models.

Method	Component	Statistic	$\text{Ln}(\beta_0)$	β_1	β_2	β_3	β_4	α
Period 1								
FMNB-2	1	Estimate	−6.275	0.991	−0.056	−0.013	−0.188	0.626
		SE	0.628	0.067	0.025	0.005	0.024	0.091
	2	Estimate	−10.026	1.248	−0.035*	−0.014	0.216	0.450
		SE	0.660	0.072	0.025	0.005	0.016	0.115
GFMNB-2	1	Estimate	−6.045	0.830	−0.044*	−0.011	0.079	0.482
		SE	0.628	0.066	0.026	0.005	0.016	0.107
	2	Estimate	−3.906	0.669	−0.027*	−0.024	0.080	0.894
		SE	0.899	0.103	0.027	0.005	0.028	0.087
			γ_0	γ_1	γ_2	γ_3	γ_4	γ_5
	Estimate		64.211	−199.09	0.019	−6.908	1.499	−18.618
Period 2								
FMNB-2	1	Estimate	−7.424	1.113	−0.049*	−0.015	0.027*	0.833
		SE	0.685	0.071	0.028	0.006	0.020	0.075
	2	Estimate	−8.085	1.086	−0.068	−0.009	0.029	0.247
		SE	0.473	0.051	0.018	0.004	0.014	0.125
GFMNB-2	1	Estimate	−3.138	0.715	−0.089	−0.036	0.067	0.708
		SE	0.731	0.077	0.026	0.005	0.021	0.088
	2	Estimate	−7.111	1.004	−0.082	−0.013	0.041	0.344
		SE	0.487	0.051	0.020	0.004	0.014	0.091
			γ_0	γ_1	γ_2	γ_3	γ_4	γ_5
	Estimate		2.282	4.8630	−0.0003	−0.091	−0.066	0.187

*Not significant at 5% significance level; SE: standard error.

Table 5: Summary statistics of each component for Periods 1 and 2.

Method	Component (sample)	Statistic	Crashes	F	LW	SW	CD	L
Period 1 (years 1997 and 1998)								
Mean-based classification method	Component 1 (1007)	Mean	0.68	5461.00	12.63	10.62	1.45	0.40
		SD	0.77	3376.75	1.60	7.95	2.53	0.40
	Component 2 (492)	Mean	7.51	8296.00	12.46	8.61	1.39	0.87
		SD	6.15	4012.35	1.55	7.98	1.92	0.93
FMNB-2	Component 1 (545)	Mean	3.91	6274	12.51	9.99	1.69	0.42
		SD	4.76	3573.48	1.52	7.99	2.85	0.45
	Component 2 (954)	Mean	2.37	6458	12.61	9.95	1.28	0.63
		SD	4.75	3977	1.63	8.03	2	0.75
GFMNB-2	Component 1 (738)	Mean	3	8191	12.58	11.22	0.7	0.29
		SD	4.45	3867.52	1.62	8.31	1.54	0.17
	Component 2 (761)	Mean	2.85	4646	12.57	8.74	2.14	0.81
		SD	5.13	2878.96	1.56	7.53	2.75	0.85
Period 2 (years 1999 to 2001)								
Mean-based classification method	Component 1 (421)	Mean	12.92	9420.99	12.46	8.51	1.39	0.96
		SD	10.79	4374.44	1.59	7.97	1.82	0.98
	Component 2 (1078)	Mean	1.33	5723.28	12.62	10.53	1.44	0.39
		SD	1.31	3556.24	1.58	7.97	2.53	0.39
FMNB-2	Component 1 (289)	Mean	10.29	7601.30	12.68	9.89	1.44	0.38
		SD	10.72	4495.77	1.60	7.97	2.46	0.34
	Component 2 (1210)	Mean	3.22	6561.29	12.55	9.98	1.43	0.60
		SD	6.21	4039.09	1.58	8.03	2.32	0.72
GFMNB-2	Component 1 (452)	Mean	6.27	9145.69	12.95	12.98	0.85	0.26
		SD	8.60	4457.79	1.74	8.12	1.90	0.15
	Component 2 (1047)	Mean	3.85	5732.65	12.41	8.66	1.68	0.68
		SD	7.33	3546.66	1.49	7.61	2.48	0.76

Table 6: Accumulated results of Site Consistency Test of various methods.

Method	$c = 0.99$	$c = 0.95$	$c = 0.90$
AF	620	1967	3079
AR	341	1482	2342
EB	636	1999	3068
ARP	536	1574	2298
Mean-based EB method	618	1980	3085
FMNB-based EB method	575	1937	3025
GFMNB-based EB method	**637**	**2042**	**3096**

Number in bold indicates the best result under each cutoff level.

Table 7: Accumulated results of Method Consistency Test of various methods.

Method	$c = 0.99$	$c = 0.95$	$c = 0.90$
AF	7	46	107
AR	6	43	85
EB	8	49	**109**
ARP	7	46	86
Mean-based EB method	7	48	108
FMNB-based EB method	6	46	105
GFMNB-based EB method	**9**	**50**	106

Number in bold indicates the best result under each cutoff level.

Table 8: Accumulated results of Total Rank Differences Test of various methods.

Method	$c = 0.99$	$c = 0.95$	$c = 0.90$
AF	131	3244	10138
AR	232	8804	24745
EB	110	2722	9032
ARP	1494	12985	37715
Mean-based EB method	111	2757	9015
FMNB-based EB method	246	3565	12059
GFMNB-based EB method	**92**	**2533**	**8517**

Number in bold indicates the best result under each cutoff level.

that are consistent between two observation periods. The results in Table 7 show that the EB with GFMNB-based classification method performs best for all cutoff values (except for 0.90, when the EB method is slightly better). The least consistent subpopulation EB method is again FMNB-based classification.

The TRDT compares the accident risk rankings over two subsequent observation periods, with the best method producing the smallest sum of rank differences between time periods 1 and 2. Table 8 shows the results of the TRDT in which, unlike the SCT and MCT, a lower value indicates better performance. As with the SCT and MCT, the GFMNB-based classification method performs best overall with the lowest sum rank differences for all values of c.

To summarize the results of the three HSID performance tests, the EB with GFMNB-based classification method performed the best overall in all three tests. The most substantial relative performance advantage offered by the GFMNB-based method is in the TRDT, more so at higher values of c. The regular EB and mean-based classification EB methods performed somewhat comparably in all tests, and both outperformed the AF, AR, ARP, and FMNB-based classification methods. Generally, EB methods (i.e., EB, mean-based EB, FMNB-based EB, and GFMNB-based EB methods) perform better than other methods (i.e., AF, AR, and ARP), which is consistent with previous studies [9, 23–25].

6. Discussion

The HSID results in Section 5.3 suggest that the EB with GFMNB-based classification method provides the most accurate ranking in identifying crash-prone locations for the Texas dataset. The possible explanation for this is that the EB method is based on two clues, the historical crash record of the entity and the expected number of crashes obtained from a safety performance function for similar entities. If the aggregated crash data contain heterogeneity (i.e., the crash data are collected from different regions with different characteristics), the requirement for the second clue is not satisfied. As a result, although the EB method increases the precision of estimation and corrects for the RTM bias, the advantage of EB method may be restricted to some extent when analyzing the heterogeneous crash data. Therefore, the GFMNB-g model is adopted to separate the aggregated crash data into a certain number of homogeneous subgroups and the EB method is applied to the sites in each subpopulation. Since the entities in the same subgroup share common characteristics, the proposed EB with GFMNB-based classification method is capable of analyzing the heterogeneous crash data.

There are several other important points worth mentioning. First, the results in this study support the findings in some previous studies [1, 26] that the modeling of the weight parameter is necessary when using the finite mixture of NB regression models to analyze the crash data. For example, as shown in Section 5.3, the EB with FMNB-based classification method even provides worse HSID results than the simple EB with mean-based classification method. Second, since some crash datasets may contain a small number of observations, one important issue associated with EB with GFMNB-based classification method is that this approach may suffer from the small sample bias problem. As discussed by Lord [27], data characterized by small sample size can result in biased estimated coefficients for the NB regression models. If only a small number of observations (e.g., 100 sites or less) are assigned to one subgroup, the estimated coefficients for this subgroup may be unreliable and erroneous inferences may be drawn from the EB method. Thus, when applying the EB with GFMNB-based classification method for HSID, transportation safety analysts are suggested to examine the size of the data sample for each subpopulation to ensure reliable parameter estimation and sound statistical inferences.

7. Summary and Conclusions

This study proposed three different classification model based EB methods to identify hotspots, that is, mean-based

classification, FMNB, and GFMNB methods. The new classification-based EB methods were evaluated against four conventional HSID methods (i.e., AF, AR, EB, and ARP) using the new criteria proposed by Cheng and Washington [9]. The important findings can be summarized as follows: first, for the considered Texas crash dataset, the EB with GFMNB-based classification method yields better results in identifying hotspots than the standard EB and other methods. This implies that the HSID accuracy can be possibly improved by properly classifying roadway segments based on the heterogeneity in crash data. Second, caution should be taken when classifying roadway segments. Inappropriate classification of roadway segments can result in worse results. And finally, the EB methods generally perform better than other methods, which is consistent with previous studies. For future works, accident datasets collected at other locations should be used to further examine the performances of the GFMNB-based EB method.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to thank Dr. Dominique Lord from Texas A&M University for graciously providing them with the Texas data.

References

[1] Y. Zou, Y. Zhang, and D. Lord, "Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models," *Analytic Methods in Accident Research*, vol. 1, pp. 39–52, 2014.

[2] B.-J. Park and D. Lord, "Application of finite mixture models for vehicle crash data analysis," *Accident Analysis& Prevention*, vol. 41, no. 4, pp. 683–691, 2009.

[3] I. Chang and S. W. Kim, "Modelling for identifying accident-prone spots: bayesian approach with a poisson mixture model," *KSCE Journal of Civil Engineering*, vol. 16, no. 3, pp. 441–449, 2012.

[4] B.-J. Park, D. Lord, and C. Lee, "Finite mixture modeling for vehicle crash data with application to hotspot identification," *Accident Analysis & Prevention*, vol. 71, pp. 319–326, 2014.

[5] M. A. Mohammadi, V. A. Samaranayake, and G. H. Bham, "Crash frequency modeling using negative binomial models: an application of generalized estimating equation to longitudinal data," *Analytic Methods in Accident Research*, vol. 2, pp. 52–69, 2014.

[6] J. Lu, A. Gan, K. Haleem, and W. Wu, "Clustering-based roadway segment division for the identification of high-crash locations," *Journal of Transportation Safety & Security*, vol. 5, no. 3, pp. 224–239, 2013.

[7] R. Bandyopadhyaya and S. Mitra, "Fuzzy cluster based method of hotspot detection with limited information," *Journal of Transportation Safety Security*, vol. 7, no. 4, pp. 307–323, 2015.

[8] J. de Oña, G. López, R. Mujalli, and F. J. Calvo, "Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks," *Accident Analysis & Prevention*, vol. 51, pp. 1–10, 2013.

[9] W. Cheng and S. Washington, "New criteria for evaluating methods of identifying hot spots," *Transportation Research Record*, vol. 2083, pp. 76–85, 2008.

[10] J. A. Deacon, C. V. Zegeer, and R. C. Deen, " Identification of hazardous rural highway locations," *Transportation Research Record*, vol. 543, pp. 16–33, 1975.

[11] E. Hauer, "Identification of sites with promise," *Transportation Research Record*, vol. 1542, pp. 54–60, 1996.

[12] B. Persaud, C. Lyon, and T. Nguyen, "Empirical bayes procedure for ranking sites for safety investigation by potential for safety improvement," *Transportation Research Record*, vol. 1665, no. 1, pp. 7–12, 1999.

[13] C. Abbess, D. Jarrett, and C. C. Wright, "Accidents at blackspots: estimating the effectiveness of remedial treatment, with special reference to the 'regression-to-mean' effect," *Traffic Engineering & Control*, vol. 22, no. 10, pp. 535–542, 1981.

[14] X. Cao, Z. Xu, and A. Y. Huang, "Safety benefits of converting HOV lanes to HOT lanes: case study of the I-394 MnPASS," *ITE Journal*, vol. 82, no. 2, pp. 32–37, 2012.

[15] L. Mountain, B. Fawaz, and D. Jarrett, "Accident prediction models for roads with minor junctions," *Accident Analysis & Prevention*, vol. 28, no. 6, pp. 695–707, 1996.

[16] Y. Zou, L. Wu, and D. Lord, "Modeling over-dispersed crash data with a long tail: examining the accuracy of the dispersion parameter in negative binomial models," *Analytic Methods in Accident Research*, vol. 5-6, pp. 1–16, 2015.

[17] E. Hauer, D. W. Harwood, F. M. Council, and M. S. Griffith, "Estimating safety by the empirical bayes method: a tutorial," *Transportation Research Record*, no. 1784, pp. 126–131, 2002.

[18] F. L. Mannering and C. R. Bhat, "Analytic methods in accident research: methodological frontier and future directions," *Analytic Methods in Accident Research*, vol. 1, pp. 1–22, 2014.

[19] J. M. Hilbe, *Negative Binomial Regression*, Cambridge University Press, Cambridge, UK, 2nd edition, 2011.

[20] B.-J. Park, D. Lord, and J. D. Hart, "Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis," *Accident Analysis & Prevention*, vol. 42, no. 2, pp. 741–749, 2010.

[21] N. Eluru, M. Bagheri, L. F. Miranda-Moreno, and L. Fu, "A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings," *Accident Analysis & Prevention*, vol. 47, pp. 119–127, 2012.

[22] B. Rigby and M. Stasinopoulos, *A Flexible Regression Approach Using Gamlss in R*, 2013, http://www.gamlss.org/wp-content/uploads/2013/01/Lancaster-booklet.pdf.

[23] W. Cheng and S. P. Washington, "Experimental evaluation of hotspot identification methods," *Accident Analysis & Prevention*, vol. 37, no. 5, pp. 870–881, 2005.

[24] A. Montella, "A comparative analysis of hotspot identification methods," *Accident Analysis and Prevention*, vol. 42, no. 2, pp. 571–581, 2010.

[25] L. Wu, Y. Zou, and D. Lord, "Comparison of sichel and negative binomial models in hot spot identification," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2460, pp. 107–116, 2014.

[26] Y. Zou, Y. Zhang, and D. Lord, "Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis," *Accident Analysis and Prevention*, vol. 50, pp. 1042–1051, 2013.

[27] D. Lord, "Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter," *Accident Analysis & Prevention*, vol. 38, no. 4, pp. 751–766, 2006.