Hindawi
Journal of Advanced Transportation
Volume 2017, Article ID 5230248, 9 pages
https://doi.org/10.1155/2017/5230248

Research Article

Developing a Clustering-Based Empirical Bayes Analysis Method for Hotspot Identification

Yajie Zou,[1] Xinzhi Zhong,[1] John Ash,[2] Ziqiang Zeng,[3,4] Yinhai Wang,[1] Yanxi Hao,[5] and Yichuan Peng[1]

[1] *The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, China*
[2] *University of Washington, P.O. Box 352700, Seattle, WA 98195-2700, USA*
[3] *Uncertainty Decision-Making Laboratory, Sichuan University, Chengdu 610064, China*
[4] *Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195, USA*
[5] *College of Urban Railway Transportation, Shanghai University of Engineering Science, 333 Longteng Road, Shanghai 201620, China*

Correspondence should be addressed to Yanxi Hao; haoyx0316@163.com and Yichuan Peng; yichuanpeng1982@hotmail.com

Received 15 June 2017; Revised 10 October 2017; Accepted 15 October 2017; Published 22 November 2017

Academic Editor: Chunjiao Dong

Hotspot identification (HSID) is a critical part of network-wide safety evaluations. Typical methods for ranking sites are often rooted in using the Empirical Bayes (EB) method to estimate safety from both observed crash records and predicted crash frequency based on similar sites. The performance of the EB method is highly related to the selection of a reference group of sites (i.e., roadway segments or intersections) similar to the target site from which safety performance functions (SPF) used to predict crash frequency will be developed. As crash data often contain underlying heterogeneity that, in essence, can make them appear to be generated from distinct subpopulations, methods are needed to select similar sites in a principled manner. To overcome this possible heterogeneity problem, EB-based HSID methods that use common clustering methodologies (e.g., mixture models, K-means, and hierarchical clustering) to select "similar" sites for building SPFs are developed. Performance of the clustering-based EB methods is then compared using real crash data. Here, HSID results, when computed on Texas undivided rural highway cash data, suggest that all three clustering-based EB analysis methods are preferred over the conventional statistical methods. Thus, properly classifying the road segments for heterogeneous crash data can further improve HSID accuracy.

1. Introduction

Network screening to identify sites (i.e., roadway segments or intersections) with promise for safety treatments is an important task in road safety management [1–7]. The identification of sites with promise, also known as crash hotspots or hazardous locations, is the first task in the overall safety management process [8]. One widely applied approach to this task is the popular Empirical Bayes (EB) method. The EB method is described and recommended in the *Highway Safety Manual* [9] for roadway safety management. This method is relatively insensitive to random fluctuations in the frequency of accidents with two clues combined, the observed crash frequency of the site and the expected number of crashes calculated from a safety performance function (SPF) for homogeneous sites (or the reference group) [10, 11]. The EB method can correct for regression-to-the-mean bias and refine the predicted mean of an entity [12]. Further, it is relatively simple to implement compared to the fully Bayesian approach.

Although the EB method has several advantages, there are a few issues associated with the methodology which may limit its widespread application. First, the selection of the reference population (i.e., similar sites) influences the accuracy of the EB method. When estimating the safety performance function, the crash data are usually obtained from distinct geographic sites to ensure a sufficient sample size for valid statistical estimation [10]. As a result, the aggregated crash data often contain heterogeneity. When conducting an EB analysis, the reference group must be similar to the target

group in terms of geometric design, traffic volume, and so on. Manually identifying such a reference group is a rather time consuming task for transportation safety analysts whose time could be better spent elsewhere. Second, the EB procedure is relatively complicated and requires a transportation safety analyst with considerable training and experience to implement it for a safety evaluation. Thus, the training investment required to prepare analysts to undertake EB evaluations can be a barrier. As a result, some quick and dirty conventional evaluation methods may be applied as a compromise of convenience, which may produce questionable results.

Given that the specification of correct reference groups is critical for the accuracy of the EB methodology, the primary objective of this research will examine different clustering algorithms (e.g., centroid-based clustering, connectivity-based clustering, and distribution-based clustering) and develop a procedure to identify appropriate reference groups for the EB analysis.

2. Hotspot Identification Methods Used in Comparison

2.1. Conventional Hotspot Identification Methods. One common HSID method is the accident frequency (AF) method. Sites are ranked based on AF and hotspots are defined as sites whose accident frequency exceeds some threshold value [13]. The problem of accounting for exposure in the AF method can be accommodated by considering accident rate (AR) instead of accident frequency. As such, AR methods have been used by some analysts and normalize accident frequency by traffic count. While AF and AR methods are easy to implement, they have difficulty accounting for randomness in crash data. As such, another popular HSID method was developed, that being the Empirical Bayes method presented by Abbess et al. [14]. Since its introduction decades ago, the EB method has been used numerous times in many safety studies [15–23]. One of the key advantages of using the EB method is that it accounts for the regression-to-the-mean (RTM) bias. The EB method can also help improve precision in cases where limited amounts of historical accident data are available for analysis at a given site. At its core, the EB method forecasts the expected crash count at a particular site as a weighted combination of (1) the accident count at the site based on historical data and (2) the estimated number of accidents at similar locations as determined from a regression model [24]. The regression model is generally referred to as a SPF and typically takes into account roadway and traffic characteristics (e.g., average daily traffic) at similar sites. To date, the most popular choice for the SPF has been a Negative Binomial regression model [25–27]. In terms of HSID via the EB method, EB estimates are computed for each site and then sites are ranked according to such estimates. Sites exceeding some thresholds are then considered as hotspots. Besides the EB method, another relatively common HSID method is rooted in so-called "accident reduction potential" (ARP). The ARP metric used for ranking sites was computed by subtracting the estimated accident count from the observed accident count at a given site, where the estimated accident count comes from a regression model developed from data at similar sites to the target. Among different HSID methods, the EB method is probably the most widely applied approach for screening sites with potential for safety treatment.

2.2. Clustering for Selection of Similar Sites. In the following section, we present three methods that can be used to group data into different clusters. As aforementioned, crash data often exhibit heterogeneity that can affect model estimates if not properly accounted for. The idea here is to cluster crash data into different groups that hopefully align to some degree with the underlying subpopulations from which the crash data are generated. Then, separate Negative Binomial (NB) regression models (i.e., SPFs) can be developed based on each cluster and EB estimates can then be computed using an SPF that hopefully considers sites that truly are "similar" to the site in question.

2.2.1. Generalized Finite Mixture of NB Regression Models. The generalized finite mixture of NB regression models with g components (GFMNB-g) assumes that y_i follows a mixture of NB distributions, as shown as follows [28]:

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

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

$$\mathrm{Var}\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right) = E\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right) + \left(\sum_{j=1}^{g} w_j \mu_{ij}^2 \left(1 + \frac{1}{\phi_j}\right) - E\left(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta}\right)^2\right), \tag{3}$$

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 $\boldsymbol{\theta}_j$ is vectors of parameters for the component j.

For GFMNB-g models, the equation for developing the weight parameter is shown in (4). By using a function of the covariates, the GFMNB-g model makes it possible for each site to have different weights for each component that depends on the site-specific values of the covariates. Zou et al.

[28] demonstrated how this additional flexibility can lead to better classification results

$$\frac{w_{ij}}{w_{ig}} = e^{\gamma_{0j}} e^{\boldsymbol{\gamma}_j \mathbf{x}_i}, \tag{4}$$

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

2.2.2. K-Means Clustering. The K-means clustering algorithm is often attributed to Lloyd [29] and Anderson [30], and it is one of the most popular clustering algorithms in use today. Inputs to the algorithm are the data points; here, each data point can be viewed as one of the road segments in the crash data set and its corresponding descriptive variables (e.g., lane width and average daily traffic (ADT)). With the data in hand, K cluster centers are initialized. Cluster centers can be chosen as random points in the feature space (i.e., points that do not exist in the data set could be selected) and random data points in the feature space (i.e., only points in the dataset can be selected) or through a variety of other methods. For this project, the initialization using K random data points in the dataset was used. The algorithm then proceeds in an iterative process until it converges, where convergence is defined as the point at which the cluster assignments do not change. The first step in the iteration assigns each data point to the cluster such that the distance between that cluster center and the data point itself is smallest; the distance metric used for this work is Euclidean distance defined as shown in (5). Then, the second step recalculates the center for each cluster. Pseudocode for the algorithm is shown in the following:

$$d(x_i, x_{i'}) = \sum_{j=1}^{m} \left(x_{ij} - x_{i'j}\right)^2 = \left\|x_i - x_{i'}\right\|^2, \tag{5}$$

where $d(\cdot)$ is Euclidean distance between two points; i is data point index, ranging from $1:n$; j is variable index, ranging from $1:m$ for m variables; and $\|\cdot\|$ is the two norms of two data points.

K-Means Algorithm

Cluster-Assignment Step

$$C(i) = \underset{1\le k\le K}{\arg\min} \left\|x_i - m_k\right\|, \tag{6}$$

where $C(i)$ is cluster assignment for data point x_i; m_k is center of cluster k; and all other variables are defined as previous.

Center-Update Step

$$m_k = \frac{1}{|C(k)|} \sum_{i\in C(k)} x_i, \tag{7}$$

where $|C(k)|$ is cardinality (number of data points) in cluster $C(k)$ and all other variables are defined as previous.

2.2.3. Hierarchical Clustering. Hierarchical clustering methods differ from K-means clustering in that the results do not depend on the number of clusters used (i.e., the results will always be the same for a given number of clusters) nor an initialization. Rather, they are rooted in the use of a dissimilarity measure defined between clusters that is defined in terms of all possible pairwise combinations of data points within two given clusters. In this research, agglomerative (i.e., bottom-up) hierarchical clustering in the form of complete linkage clustering was considered. Agglomerative clustering methods (e.g., complete linkage, single linkage, and average linkage) take the data points (i.e., road segments and their corresponding descriptors) as inputs and begin with each data point as its own cluster; a lone data point forming its own cluster is also known as a singleton. For complete linkage clustering, the algorithm proceeds in a total of $n - 1$ steps (i.e., one step less than the total number of data points in the dataset) and at each step, the two clusters with the smallest intergroup dissimilarity measure are joined to form a new cluster. Thus, in each successive step, the number of clusters is reduced by one. For complete linkage clustering the intergroup dissimilarity is defined as follows [31]:

$$d(A, B) = \max_{i\in A, i'\in B} d_{ii'}, \tag{8}$$

where A, B are two arbitrary clusters and

$$d_{ii'} = \left\|x_i - x_{i'}\right\|. \tag{9}$$

Thus, for each step of the complete linkage clustering algorithm, the two clusters with the smallest value of the maximum between-cluster distance are joined.

2.3. Classification-Based EB Methods. At this point it is important to clarify the main contribution of this work. It is well-known that aggregated crash likely has some degree of heterogeneity, as if they are generated from multiple distinct subpopulations. As such, if one were able to try to capture this heterogeneity and group the data into different units, ideally based upon the subpopulations from which they were generated, better estimates of safety and HSID rankings could likely be obtained. Thus, three types of clustering algorithms (GFMNB-g model based, k-means clustering, and hierarchical clustering with complete linkage) are proposed to cluster the data into distinct subgroups that hopefully correspond to the subpopulations from which the data were generated. The main idea/application of clustering is to define groups (i.e., clusters) of data points such that all points assigned to/belonging to a given cluster are closer/more similar to the points in that cluster than any other cluster [31].

Clustering methods present an ideal means to represent/describe heterogeneity within crash data. As such, we apply clustering-based EB methods in this study as a new means of hotspot identification. For these methods, three types of clustering as aforementioned are considered, and the classification method for HSID purposes has four main steps as follows. First, the full set of input crash data is clustered into g clusters via the GFMNB-g model, k-means clustering algorithm, or hierarchical clustering algorithm. In this study,

the number of clusters considered is set equal to the number of components selected for the GFMNB-g model, which was itself selected on the basis of the Bayesian information criterion (BIC). Ultimately, however, the choice for selection of both number of clusters and number of components in the GFMNB-g model is up to the analyst. The second step of the algorithm involves splitting the data into g groups based on the results of the applied clustering algorithm. The third step of the algorithm calls for estimation of an NB regression model (i.e., SPF) for each of the g subgroups/clusters from the data and using these SPFs in further generation of EB estimates for each site. For example, if $g = 2$, two SPFs will be estimated and the data in each of the two groups will have EB estimates calculated through application of the corresponding SPF. Fourth and finally, the EB estimates for all sites across all g subgroups are aggregated and ranked, after which hotspots identification is based on threshold values or other methods. From this point forward, the classification-based HSID methods aforementioned will be referred to as follows: GFMNB-based EB method, the K-means-based EB method, and hierarchical-based EB method, respectively. A summary of the classification-based EB method for HSID is shown in Table 1.

2.4. Evaluation Criteria for Hotspot Identification Methods. For the purpose of evaluating the performance of HSID methods, some kind of standardized test procedures are needed. Ultimately, analysts will be concerned with an HSID method's capability to find accident prone sites and properly rank sites according to risk. These concerns are directly related to the overarching objective of prioritizing safety treatments at hotspots in a limited-funding environment. While a multitude of tests are available and determining which test is optimal may not be clear, one might argue that "good" performance (to be described in the forthcoming test descriptions) across multiple tests could be a reasonable indicator of a method's overall performance in HSID. As such, we consider three commonly used tests attributed to Cheng and Washington [32]: the Site Consistency Test (SCT), the Method Consistency Test (MCT), and the Total Rank Difference Test (TRDT). For more information about these three tests, interested readers are referred to Cheng and Washington [32].

3. Data and Analysis

3.1. Data Description. In order to examine the effectiveness of the methodology presented herein, the research team chose to work with a dataset used in many previous safety studies, which being the Texas rural undivided highway crash dataset. The dataset contains crash counts collected over 1,499 rural undivided highway segments over a span of five years, 1997–2001, for the National Cooperative Highway Research Program (NCHRP) 17–29 project [33]. Since the aforementioned tests for evaluation of the hotspot identification methods require data from different time periods for comparison purposes, the dataset comprised of 1,499 observations was broken down into two temporal subsets. The first subset, called "Time Period 1" henceforth, contains the data from the original dataset recorded for 1997 and 1998. The second subset, called "Time Period 2" henceforth, contains the data from the original dataset recorded for 1999, 2000, and 2001. Thus, the union of these two subsets is the original dataset with 1,499 points. Variables collected to describe the segments and be considered as independent variables in the analysis include average daily traffic during the analysis period (F), lane width (LW, in feet), total shoulder width (i.e., the sum of shoulder width on both sides of the roadway in feet, SW in feet), and curve density (i.e., the number of curves per mile, CD). The dependent variable in the analysis is the number of crashes observed on each segment over the analysis period, and another variable, segment length (L, in miles), was considered as an offset in the regression. Summary statistics on the dataset are presented in Table 2.

3.2. Modeling Results. To study classification-based EB methods, GFMNB-g models were developed from the crash records in Time Periods 1 and 2, respectively. Data in each time period was used to estimate the finite mixture models with g components; that is, g separate NB models were estimated for each type of mixture model that are combined together to form a weighted estimate. Then, the GMFNB-g model was used as the basis for the clustering-based EB methods. That is to say, the number of components used in the model was selected as the basis for the number of clusters to use for grouping the crash data under each of the aforementioned clustering methods.

When estimating the GFMNB-g models, perhaps the main problem is to determine how many components should be used in the model (i.e., to select g). In order to select the number of components for each model in each time period, the method presented in Park et al. [34] was applied in this study. Under this approach, the analyst builds finite mixture models with increasing numbers of components (from two upwards) and selects the final model (and number of components) through goodness-of-fit metrics, such as the Akaike Information Criterion (AIC) or the previously mentioned BIC, which balance the number of components and overall model fit (measured via log-likelihood). Eluru et al. [35] noted that BIC is more stringent than AIC in terms of applying a penalty based on number of components, and thus it may be more robust in terms of preventing overfitting. As such, BIC was selected as the means of choosing the number of components for the finite mixture models in each of the two time periods:

$$\text{BIC}_j = -2 * \text{loglikelihood}_j + g_j * \log(n), \quad (10)$$

where loglikelihood_j is the log-likelihood of model j; g is the number of components in finite mixture model j; and n is the sample size (i.e., number of sites in dataset).

In this study, GFMNB-g models were developed from the crash data in Time Periods 1 and 2 with increasing numbers of components $g = 2$, 3, or 4. Figure 1 indicates that use of $g = 2$ (i.e., finite mixture models with two components) leads to the best goodness-of-fit as indicated by the lowest value of BIC. Hence, the number of components is selected as $g = 2$ and

Table 1: Classification-based EB method for HSID.

Step	Description
(1)	Use the GFMNB-g model, K-means algorithm, or hierarchical clustering algorithm to cluster the data into g groups
(2)	Separate the data into g groups based on the results of clustering
(3)	Estimate g NB regression models, one for each of the g subgroups, and use the corresponding SPF to get EB estimates for each site
(4)	Aggregate the EB estimates for all sites, rank the sites, and identify hotspots

Table 2: Summary statistics for road segments in Texas rural undivided highways dataset.

Variable	Time Period 1 (1997 and 1998)			Time Period 2 (1999–2001)		
	Minimum	Maximum	Mean (SD†)	Minimum	Maximum	Mean (SD†)
Crash number	0	59	2.93 (4.81)	0	78	4.58 (7.81)
F	40	24000	6391 (3835.01)	43.33	25333.3	6761.8 (4149.84)
LW (ft)	9.75	16.5	12.57 (1.59)	9.75	16.5	12.57 (1.59)
SW (ft)	0	40	9.96 (8.02)	0	40	9.96 (8.02)
CD	0	18.07	1.43 (2.35)	0	18.07	1.43 (2.35)
L (miles)	0.1	6.28	0.55 (0.68)	0.1	6.28	0.55 (0.68)

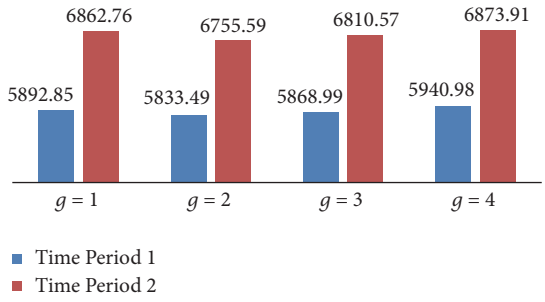

Figure 1: BIC values for GFMNB-g models.

the GFMNB-g models can now be indicated as GFMNB-2 models. It is important to note that, in general, $g = 2$ may not always be the optimal number of components and the choice will depend on the data. That said, BIC is a reasonable method to use to select g.

By examining Figure 1, one can see that the BIC values reported for the GFMNB-2 models are not as large as those for the regular NB model ($g = 1$) in the corresponding time period suggesting that the mixture models have better goodness-of-fit. Further, the choice of $g = 2$ based on BIC seems to suggest the existence of two distinct subpopulations within the crash data corresponding to each time period instead of a lone data population.

3.3. Grouping Results. According to the results of the GFMNB-g fitting procedure, it was determined that a GFMNB model with two components fit the data best. Thus, for each of the clustering-based-EB procedures for HSID, the full set of crash data was split into two groups for each time period (i.e., four groups total) from which NB models were estimated and corresponding EB estimates were calculated. That is to say, given the crash data for Time Periods 1 and 2, the three aforementioned clustering algorithms (i.e., k-means, hierarchical with complete linkage, and estimation of a GFMNB-g model) were applied to group the data from each time period into two clusters, for which EB estimates were computed.

Table 3 shows grouping results for each component, under each clustering method for both time periods considered in the study. For each component, the sample size along with the mean and standard deviation (SD) for each variable in the dataset (as described previously) is presented. From the table, it can be seen that, in general, mean values for the lane width, shoulder width, and segment length do not differ much between components. That said, in some cases, particularly for the groupings based on hierarchical clustering for Time Period 2, the mean number of crashes differs dramatically between components. Additionally, there is a substantial difference in the mean values of average daily traffic (F) between components for all clustering methods considered in both time periods. Such a trend suggests that the data considered here may come from underlying subpopulations where traffic volume is a defining characteristic for subpopulation membership and thus a good descriptor of the heterogeneity in the data.

With crash data clustered for each time period according to the three aforementioned clustering methods, EB estimates were then obtained after estimating an NB regression model for each of the two components corresponding to a given clustering method for a given time period. When interpreting results henceforth, one should consider the sample sizes used to estimate the NB models. For example, the sample size of "Component 1" (i.e., one grouping) for Time Period 2 as defined via hierarchical clustering with complete linkage has only 66 data points. Thus, modeling results associated with this group (namely, the results of the SPF and corresponding EB estimates) and the overall EB estimates for Time Period 2 as determined via hierarchical clustering (i.e., the aggregation of the EB estimates for components 1 and 2) should be interpreted with caution.

Table 3: Characteristics of each component.

Method	Component (sample)	Statistic	Crashes	F	LW	SW	L
			Time Period 1				
K-means	Component 1 (527)	Mean	5	10547.19	12.88	10.08	0.54
		SD	6.58	3013.67	1.76	8.45	0.6
	Component 2 (972)	Mean	1.796	4138.07	12.4	9.89	0.55
		SD	2.92	1820.303	1.46	7.77	0.69
Hierarchical	Component 1 (473)	Mean	5.063	10895.39	12.9577	10.24	0.53
		SD	6.58	2988.96	1.803	8.51	0.52
	Component 2 (1026)	Mean	1.939	4314.87	12.39	9.83	0.565
		SD	3.27	1924.28	1.44	7.778	0.72
GFMNB-2	Component 1 (738)	Mean	3	8191	12.58	11.22	0.29
		SD	4.45	3867.52	1.62	8.31	0.17
	Component 2 (761)	Mean	2.85	4646	12.57	8.74	0.81
		SD	5.13	2878.96	1.56	7.53	0.85
			Time Period 2				
K-means	Component 1 (972)	Mean	2.68	4364.14	12.4	9.89	0.55
		SD	4.71	2035.808	1.46	7.77	0.69
	Component 2 (527)	Mean	8.07	11184.04	12.88	10.08	0.54
		SD	10.6	3343.19	1.76	8.459	0.609
Hierarchical	Component 1 (66)	Mean	16.69	18144.65	13.59	12.39	0.64
		SD	17.71	3095.972	2.03	8.46	0.633
	Component 2 (1433)	Mean	4.02	6237.53	12.52	9.84	0.5496
		SD	6.52	3366.45	1.548	7.98	0.66
GFMNB-2	Component 1 (452)	Mean	6.27	9145.69	12.95	12.98	0.26
		SD	8.6	4457.79	1.74	8.12	0.15
	Component 2 (1047)	Mean	3.85	5732.65	12.41	8.66	0.68
		SD	7.33	3546.66	1.49	7.61	0.76

3.4. Test Results. In this section, evaluations of six different HSID methods, (1) AF, (2) AR, (3) EB (here, all data are considered as being from one population), (4) GFMNB-based EB method, (5)K-means-based EB method, and (6) hierarchical-based-EB method, are conducted using the three main tests from Cheng and Washington [32]. As all test procedures involve comparison across two different time periods, we use the time periods as defined in Table 2. Further, we consider three different scenarios in terms of the number of high-risk sites selected for consideration under each HSID method. These scenarios correspond to considering 1%, 5%, and 10% of all sites as high-risk (i.e., $c = 0.01$, 0.05, and 0.10). For example, in this study, when $c = 0.10$, a total of approximately 150 sites (i.e., ~10% of the 1,499 total sites) will be considered as high-risk, and their data will be used in calculation of the test statistics for the various HSID methods.

Table 4 shows the results of the six HSID methods considered under the Site Consistency Test. As aforementioned, the goal of the SCT is to measure consistency of a method in identifying sites as high-risk over time. The underlying principle is that high-risk sites should show consistently high crash counts over time, and thus the higher the value for the SCT statistic, the better performing the HSID method. From the table, it can be seen that the worst performing method across all cutoff levels for high-risk site identification (i.e., all c values) is the AR method. When one percent of sites are considered as high-risk, the conventional EB method, K-means-based EB method, and hierarchical-based EB method all perform equally well. For the cases in which 5% and 10% of sites are considered as high-risk, the K-means-based EB method is identified as the best performing HSID method according to the SCT. That said, in both of these cases, the value of the SCT test statistic for the hierarchical-based-EB method gives a value quite close to those obtained by the K-means-based method, indicating that it also seems to perform nearly as well in HSID.

The results of the six HSID methods being evaluated in terms of the Method Consistency Test are shown in Table 4. The MCT is designed to assess consistent identification of the same high-risk sites across different time periods. As such, the higher the value of the MCT test statistic, the better the performance of the HSID method (i.e., higher values imply that more sites were identified as high-risk in both time periods considered). From Table 4, one can see that, across all three cutoff levels for proportions of sites to consider as high-risk, the GFMNB-based EB method performs the best. That said, for the case in which 10% of sites are considered as high-risk, the K-means-based method performs just as well. Additionally, for all cutoff levels, the results of all clustering-based EB methods (e.g., GFMNB-, K-means-, and

Table 4: Results for various methods using three different tests.

Method	$c = 0.01$	$c = 0.05$	$c = 0.10$
Site Consistency Test (SCT)			
AF	269	1109	1911
AR	110	570	1051
EB	**361**	1376	2182
GFMNB-based EB method	329	1352	2115
K-means-based EB method	**361**	**1396**	**2186**
Hierarchical-based EB method	**361**	1395	2171
Method Consistency Test (MCT)			
AF	7	43	88
AR	2	27	63
EB	7	47	100
GFMNB-based EB method	**8**	**51**	**103**
K-means-based EB method	7	49	**103**
Hierarchical-based EB method	7	47	99
Total Rank Difference Test (TRDT)			
AF	365	7599	20721
AR	5944	24259	49548
EB	217	3543	14132
GFMNB-based EB method	**162**	**3226**	**10195**
K-means-based EB method	220	3273	12391
Hierarchical-based EB method	220	3420	14068

Note. Bold number indicates the best result.

hierarchical-based) exhibit quite similar performance. As was the case for the SCT, the AR method consistently performs the worst across all three cutoff levels for proportions of sites to be considered as high-risk.

Table 4 presents the results of the HSID-procedure evaluation under the Total Rank Difference Test. Again, this test is based on consistent identification of high-risk sites across time periods, but here, the rankings of sites identified as high-risk in one time period are compared to the rankings of the same sites in another time period. Hence, the smaller the value of the TRDT test statistic, the better the performance of the method in HSID. From the table, it can be seen that the GFMNB-based EB method yields the best HSID performance across all three cutoff levels of proportions of sites to be considered as high-risk. Under this test, the other clustering-based EB methods (e.g., K-means-based and hierarchical-based) outperform the naïve AF and AR methods across all cutoff values and also outperform the EB method for the 5% and 10% cutoffs. As was the case for the preceding two HSID performance tests, the AR method of HSID consistently performs the worst across all three cutoff levels of proportions of sites to be considered as high-risk.

Overall, the preceding tests indicate that the GFMNB-based EB method appears to exhibit the strongest HSID performance in all three tests and across the different cutoff levels of proportion of sites to be considered as high-risk. That said, the results obtained from the other clustering-based EB methods (e.g., K-means-based and hierarchical-based) are usually close and tend to outperform the AF, AR, and standard EB methods. From all tests, it appears that the AR method performs the worst. One possible explanation for this behavior may be that since the test sites are rural road segments, many may exhibit low ADT values and thus, as aforementioned, low-volume sites may be overrepresented as high-risk since the AR calculation normalizes by traffic count. Ultimately, HSID methods that themselves make use of the EB method when computing safety estimates prior to site ranking appear to perform better than the naïve AF and AR methods. This finding is consistent with many previous studies including [17, 36, 37].

3.5. Discussion. From the preceding analysis, it appears that the GFMNB-based EB procedure for HSID performs the best when evaluated with the three aforementioned test procedures [32] on the Texas rural undivided highway crash dataset. That said, it seems that all EB-based methods typically outperform the naïve methods, especially the AR HSID method. One possible reason the EB-based HSID methods may perform better is due their use of both the observed historical accident data and predicted accident count from similar sites with the SPF. Further, the EB methods are able to adjust for RTM bias. That said, the conventional EB method is not without its limitations for HSID. The main limitation, perhaps, arises when there is a substantial degree of heterogeneity in the crash data making it such that the crash data seem to arise from different subpopulations. Such heterogeneity could arise when large amounts of crash data collected from areas that differ dramatically geographically and with respect to a variety of other site-specific conditions. Oftentimes, crash data are aggregated in an effort to ensure

that sufficient sample sizes are available for model estimates (i.e., in an effort to reduce the standard error value of regression coefficients). In order to remedy this issue of not accounting for heterogeneity in the data, three clustering-based EB methods were proposed in this report. The idea behind these methods was to group the overall set of crash data (i.e., full list of study sites) into smaller subsets such that the site in each subset was more similar to sites within their groups than sites in other groups according to features, such as traffic volume, lane width, and other predictors. Further, it was hoped that such clustering could potentially help uncover the underlying groups/subpopulations from which the data could have been generated. Indeed, it appears that the clustering-based EB methods that applied *k*-means-, hierarchical-, and GFMNB-based clustering were able to analyze heterogeneous data and outperform more conventional methods in terms of HSID.

While the clustering-based EB methods for HSID have several benefits, they are not without their limitations. Perhaps the largest limitation of clustering-based EB methods is that, in some cases, they can cluster data into groups with relatively small sample sizes. Then, regression models (i.e., SPFs) developed from these small samples are more likely to exhibit their own issues such as biases in their coefficient estimates. This issue can further be compounded when analysts interpret the biased results and have the potential to make erroneous inferences/conclusions. As such, it is important that one be cognizant of the sample sizes of the clusters and what impacts they may have on model estimates and resulting inference [38]. Ultimately, as always, analysts are encouraged to interpret all results, especially those corresponding to regression models developed from small samples (e.g., 100 or less sites) with caution.

4. Conclusions

This study introduced three clustering-based EB methods for hotspot identification purposes. The clustering methods considered were the GFMNB-g model, *K*-means clustering, and hierarchical clustering with complete linkage. The newly developed clustering-based EB methods for HSID were compared in terms of performance to conventional HSID methods, including the EB method, as well as the naïve AF and AR method, with three methods comparing performance in HSID across different time periods as developed by Cheng and Washington [32]. When studying the HSID results based on applying the methodology to Texas undivided rural highway crash data, the results suggest that all three clustering-based EB analysis methods are preferred over the conventional statistical methods. Additionally, it seems that the accuracy of HSID can be enhanced by appropriately classifying roadway segments according to the heterogeneity of the crash data (i.e., clustering the data before developing SPFs for use in EB estimates). That said, one should always be cautious when classifying roadway segments into clusters as inappropriate classification of roadway segments can lead to erroneous results (e.g., biased coefficient estimates from SPFs developed from small sample sizes). Although the proposed clustering-based EB method is not yet ready for practical application, transportation safety analysts may use the clustering-based EB method to calculate the EB estimates and avoid manually identifying similar groups within the heterogeneous crash dataset, a task that may be difficult as the underlying subpopulations in the data are usually unknown. Future work could evaluate development of a performance measure to evaluate the overall HSID performance of the three clustering-based EB methods (i.e., to determine which clustering method is best and when it is best to use each).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Dr. Dominique Lord from Texas A&M University for graciously providing us with the Texas crash data. This research is sponsored jointly by the Pacific Northwest Transportation Consortium (Contract no. DTRT13-G-UTC40), the National Natural Science Foundation of China (Grant no. 51608386 and Grant no. 71601143), and Shanghai Sailing Program (Grant no. 16YF1411900).

References

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

[2] C. Dong, D. B. Clarke, X. Yan, A. Khattak, and B. Huang, "Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections," *Accident Analysis & Prevention*, vol. 70, pp. 320–329, 2014a.

[3] X. Ye, V. M. Garikapati, D. You, and R. M. Pendyala, "A practical method to test the validity of the standard Gumbel distribution in logit-based multinomial choice models of travel behavior," *Transportation Research Part B: Methodological*, 2017.

[4] C. C. Xu, P. Liu, W. Wang, and Z. B. Li, "Evaluation of the impacts of traffic states on crash risks on freeways," *Accident Analysis & Prevention*, vol. 47, pp. 162–171, 2012.

[5] Y. J. Zou, J. J. Tang, L. T. Wu, K. Henrickson, and Y. H. Wang, "Quantile analysis of factors influencing the time taken to clear road traffic incidents," in *Proceedings of the Institution of Civil Engineers-Transport*, vol. 170, pp. 296–304, 2017a.

[6] J. Tang, F. Liu, W. Zhang, R. Ke, and Y. Zou, "Lane-changes prediction based on adaptive fuzzy neural network," *Expert Systems with Applications*, vol. 91, pp. 452–463, 2018.

[7] L. Fawcett, N. Thorpe, J. Matthews, and K. Kremer, "A novel Bayesian hierarchical model for road safety hotspot prediction," *Accident Analysis & Prevention*, vol. 99, pp. 262–271, 2017.

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

[9] H. S. Manual, *American Association of State Highway and Transportation Officials (AASHTO)*, DC, Washington, Wash, USA, 2010.

[10] Y. Zou, K. Henrickson, L. Wu, Y. Wang, and Z. Zhang, "Application of the empirical Bayes method with the finite mixture model for identifying accident-prone spots," *Mathematical*

Problems in Engineering, vol. 2015, Article ID 958206, 10 pages, 2015.

[11] Y. J. Zou, H. Yang, Y. L. Zhang, J. J. Tang, and W. B. Zhang, "Mixture modeling of freeway speed and headway data using multivariate skew-t distributions," *Transportmetrica a-Transport Science*, vol. 13, pp. 657–678, 2017b.

[12] Y. Zou, D. Lord, Y. Zhang, and Y. Peng, "Comparison of sichel and negative binomial models in estimating empirical bayes estimates," *Transportation Research Record*, no. 2392, pp. 11–21, 2013.

[13] J. A. Deacon, C. V. Zegeer, and R. C. Deen, Identification of hazardous rural highway locations, 1974.

[14] 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.

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

[16] 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.

[17] L. Wu, Y. Zou, and D. Lord, "Comparison of sichel and negative binomial models in hot spot identification," *Transportation Research Record*, vol. 2460, no. 1, pp. 107–116, 2014.

[18] M. A. Mohammadi, V. A. Samaranayake, and G. H. Bham, "Safety effect of missouri's strategic highway safety plan: missouri's blueprint for safer roadways," *Transportation Research Record*, vol. 2465, pp. 33–39, 2014.

[19] W. Cheng, W. H. Lin, X. Jia, X. Wu, and J. Zhou, "Ranking cities for safety investigation by potential for safety improvement," *Journal of Transportation Safety & Security*, pp. 1–22, 2017.

[20] C. Xu, W. Wang, and P. Liu, "A genetic programming model for real-time crash prediction on freeways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 574–586, 2013.

[21] C. Dong, S. H. Richards, B. Huang, and X. Jiang, "Identifying the factors contributing to the severity of truck-involved crashes," *International Journal of Injury Control and Safety Promotion*, vol. 22, no. 2, pp. 116–126, 2015.

[22] J. J. Tang, F. Liu, Y. J. Zou, W. B. Zhang, and Y. H. Wang, "An improved fuzzy neural network for traffic speed prediction considering periodic characteristic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, pp. 2340–2350, 2017.

[23] S. Zhang, J. J. Tang, H. X. Wang, Y. H. Wang, and S. An, "Revealing intra-urban travel patterns and service ranges from taxi trajectories," *Journal of Transport Geography*, vol. 61, pp. 72–86, 2017.

[24] 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.

[25] F. L. Mannering, V. Shankar, and C. R. Bhat, "Unobserved heterogeneity and the statistical analysis of highway accident data," *Analytic Methods in Accident Research*, vol. 11, pp. 1–16, 2016.

[26] 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.

[27] C. Dong, D. B. Clarke, S. H. Richards, and B. Huang, "Differences in passenger car and large truck involved crash frequencies at urban signalized intersections: an exploratory analysis," *Accident Analysis & Prevention*, vol. 62, pp. 87–94, 2014b.

[28] 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.

[29] S. P. Lloyd, "Least squares quantization in PCM," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[30] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359–364, 2009.

[31] J. Z. Lu, "The elements of statistical learning: data mining, inference, and prediction," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 173, no. 3, pp. 693-694, 2010.

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

[33] D. Lord, S. Geedipally, B. Persaud et al., "Methodology to predict the safety performance of rural multilane highways [Final Report for NCHRP Project 17-29], 2008".

[34] 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.

[35] 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.

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

[37] L. Wu, D. Lord, and Y. Zou, "Validation of crash modification factors derived from cross-sectional studies with regression models," *Transportation Research Record*, vol. 2514, pp. 88–96, 2015.

[38] 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.