DEVELOPING A CLUSTERING-BASED EMPIRICAL BAYES ANALYSIS METHOD FOR HOTSPOT IDENTIFICATION

FINAL PROJECT REPORT

by

Yinhai Wang
Yajie Zou
John Ash
Ziqiang Zeng
University of Washington

Sponsorship
US Department of Transportation

for
Pacific Northwest Transportation Consortium (PacTrans)
USDOT University Transportation Center for Federal Region 10
University of Washington
More Hall 112, Box 352700
Seattle, WA 98195-2700

In cooperation with US Department of Transportation-Research and Innovative Technology Administration (RITA)

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Pacific Northwest Transportation Consortium, the U.S. Government and matching sponsor assume no liability for the contents or use thereof.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
Developing a Clustering-based Empirical Bayes Analysis Method for Hotspot Identification	04/30/2016
	6. Performing Organization Code
7. Author(s)	**8. Performing Organization Report No.**
Yinhai Wang, Yajie Zou, John Ash, and Ziqiang Zeng	
9. Performing Organization Name and Address	**10. Work Unit No. (TRAIS)**
PacTrans Pacific Northwest Transportation Consortium University Transportation Center for Region 10 University of Washington More Hall 112 Seattle, WA 98195-2700	**11. Contract or Grant No.** DTRT13-G-UTC40
12. Sponsoring Organization Name and Address	**13. Type of Report and Period Covered**
United States of America Department of Transportation Research and Innovative Technology Administration	Research
	14. Sponsoring Agency Code

15. Supplementary Notes

Report uploaded at www.pacTrans.org

16. Abstract

Hotspot identification (HSID) is a critical part of network-wide safety evaluation. Put simply, HSID involves ranking sites (e.g., roadway segments or intersections) on the basis of observed and/or estimated safety so they may be prioritized for treatment. Typical methods for ranking sites are often rooted in use of the Empirical Bayes (EB) method to estimate safety from both observed crash history and crash frequency predictions based on similar sites. Such procedures are an improvement over naïve methods that consider only observed crash frequencies/rates as they can account for regression-to-the-mean bias and are less subject to random variation in the crash data. That said, the performance of the EB method is highly related to the selection of a reference group of sites similar to the target site from which the safety performance function (SPF) used to predict crash frequency in the EB method 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 were developed. The performances of the clustering-based EB methods were then compared by using real crash data. Here, HSID results, when computed with Texas undivided rural highway cash data, suggested that all three clustering-based EB analysis methods are preferable over conventional statistical methods. Therefore, HSID accuracy may be further improved by properly classifying roadway segments on the basis of the heterogeneity in the data.

17. Key Words	18. Distribution Statement
Clustering, Empirical Bayes Method, Safety Performance Function, Hotspot Identification, K-means	No restrictions.

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified.	Unclassified.	46	NA

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Table of Contents

List of Tables

List of Abbreviations

ADT: Average Daily Traffic

AF: Accident Frequency

AR: Accident Rate

BIC: Bayesian Information Criterion

CD: Curve Density

EB: Empirical Bayes

GFMNB-g: Generalized Finite Mixture of Negative Binomial Regressions with g Components

HSID: Hotspot Identification

LW: Lane Width

MCT: Method Consistency Test

NB: Negative Binomial

NCHRP: National Cooperative Highway Research Program

PacTrans: Pacific Northwest Transportation Consortium

PDF: probability density function

SCT: Site Consistency Test

SW: Shoulder Width

TRDT: Total Rank Differences Test

Acknowledgments

The PI and Co-PI would like to give our great appreciation to the UW team researchers and Ph.D. students including John Ash and Dr. Ziqiang Zeng for their effort and contributions on this research.

Executive Summary

Hotspot identification (HSID) is a critical part of network-wide safety evaluation. Put simply, HSID involves ranking sites (e.g., roadway segments or intersections) on the basis of observed and/or estimated safety so that they may be prioritized for treatment. Typical methods for ranking sites are often rooted in use of the Empirical Bayes (EB) method to estimate safety from both observed crash history and crash frequency predictions based on similar sites. Such procedures are an improvement over naïve methods that consider only observed crash frequencies/rates, as they can account for regression-to-the-mean bias and are less subject to random variation in the crash data. That said, the performance of the EB method is highly related to the selection of a reference group of sites, and their similarity to the target site, from which the safety performance function (SPF) used to predict crash frequency in the EB method 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 were developed. The performances of the clustering-based EB methods were then compared by using real crash data. Here, HSID results, when computed with Texas undivided rural highway cash data, suggested that all three clustering-based EB analysis methods are preferable over conventional statistical methods. Thus, HSID accuracy may be further improved by properly classifying roadway segments on the basis of the heterogeneity in the data.

Chapter 1 Introduction

1.1 Problem Statement

Network screening to identify sites with a potential for safety treatments is an important task in road safety management (Persaud et al., 2010). The identification of sites with promise, also known as crash hotspots or hazardous locations, is the first step in the overall safety management process (Montella, 2010). 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 (2010) for roadway safety management. This method is relatively insensitive to random fluctuations in accident frequency by combining clues from two sources, the historical crash record for the site and the expected number of crashes derived from a safety performance function (SPF) for similar sites (or a reference group). The EB method can correct for regression-to-the mean bias and refine the predicted mean of an entity. Furthermore, it is relatively simple to implement in comparison to the fully Bayesian approach. Although the EB method has several advantages, there are a few issues associated with the methodology that may limit its widespread application. First, the accuracy of the EB method depends largely on the selection of the reference population or grouping of similar sites, and the definition of "similar" is a somewhat open question. When the safety performance function is estimated, the crash data are often collected from different geographic locations to ensure the adequacy of sample size for valid statistical estimation. As a result, the aggregated crash data often contain heterogeneity. When an EB analysis is conducted, the reference group must be similar to the target group in terms of geometric design, traffic volumes, etc. However, manually identifying such a reference group is rather time consuming 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. Therefore, 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 for convenience, which may produce questionable results.

1.2 Project Goals

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

Chapter 2 Background and Methodology

2.1. Hotspot Identification Methods

One common HSID method is the accident frequency (AF) method. As the name suggests, accident frequencies are computed for each site (e.g., intersection or road segment) of interest over a given time period. Then, sites are ranked on the basis of AF, and hotspots are defined as sites whose accident frequency exceeds some threshold value (Deacon et al., 1975). Alternately, some percentage of sites with the highest crash frequencies (e.g., the top 10 percent) can be taken as hotspots. When sites are ranked, it is desirable to select sites from similar locations (e.g., along a given stretch of highway with similar geometric characteristics) to help avoid biases that may result from characteristics attributed to the roadway functional classification or driver behavior. AF methods, while simple to implement, are certainly not without their flaws. First and foremost, they do not consider any measure of exposure. Therefore, sites with higher traffic volumes will typically be overrepresented (Hauer, 1996). Additionally, they have trouble distinguishing between actual hotspots and sites with high accident frequencies due to random fluctuations in crash counts (Deacon et al., 1975; Cheng and Washington, 2008).

The problem of accounting for exposure in an HSID method can be accommodated by considering accident rate (AR) instead of accident frequency. For this reason, some analysts have normalized accident frequency by traffic count to get an AR in units of accidents per 100 million vehicle miles traveled. AR can be calculated as shown in equation 2.1 (Golembiewski and Chandler, 2011). Once accident rates have been computed, sites can be ranked, and those above some threshold value or within some percentage of all sites can be selected as hotspots, just as is the case for the AF method.

While AR methods may seem to solve a major issue with the use of AF methods, they are also not without their flaws. First of all, they make an implicit assumption that accident count and traffic volume are linearly related, an assumption that is often untrue in practice. Furthermore, normalizing by volume can give low-volume sites high ARs and cause such sites to be overrepresented in the data (Hauer, 1996; Persaud et al., 1999).

$$AR = \frac{C*100{,}000{,}000}{V*365*N*L} \tag{2.1}$$

where,

C = number of crashes observed over the analysis period;

V = traffic volume (e.g., AADT);

N = number of years in the analysis period;

L = segment length in miles; and

All other variables defined as aforementioned.

While AF and AR methods are easily to implement, they have difficulty accounting for randomness in crash data. Therefore, another popular HSID method was developed, that being the Empirical Bayes (EB) method presented by Abbess et al. (1981). Since its introduction decades ago, the EB method has been used numerous times in many safety studies (Cao et al., 2012; Mountain et al., 1996; Zou et al., 2015). One of the key advantages of using the EB method is that it accounts for regression-to-the-mean (RTM) bias. Put simply, RTM is a statistical phenomenon that describes how there is an increased likelihood that a time period with a relatively high (low) crash frequency will precede a time period with relatively low (high) crash frequency. If not accounted for properly in analysis procedures, RTM can cause the reduction in crash frequency to be overestimated (Hauer 1996, AASHTO, 2010). The EB method

can also help improve precision when limited amounts of historical accident data are available for analysis for 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 expected accident count at similar location as determined from a regression model (Hauer et al., 2002). The regression model is generally referred to as a safety performance function (SPF) and typically takes into account roadway characteristics (e.g., lane width, shoulder width, etc.) and traffic characteristics (e.g., average daily traffic etc.) at similar sites. To date, the most popular choice for the SPF has been a negative binomial (NB) regression model (Mannering and Bhat, 2014). Similar to a Poisson model, the NB model can also be used to model count data; however, it removes the restrictive assumption that the crash mean must equal the variance as is the case for the Poisson distribution. Before the EB method is used, it is first important to fully understand the NB regression model. The following describes the derivation following Cameron and Trivedi (1998) and Zou et al. (2017).

For the NB model, one assumes that the number of crashes, y_i, is conditionally Poisson (Cameron and Trivedi, 1998).

$$p(y_i|\lambda_i) = \frac{\lambda_i^{y_i} * \exp(-\lambda_i)}{y_i!} \tag{2.2}$$

where,

y_i = crash count at site i; and

λ_i = mean crash rate at site i.

It is then assumed that the Poisson parameter, λ_i, which is itself modeled via regression, follows a two-parameter gamma distribution with shape parameter φ and rate parameter φ/μ. That is to say,

$$y_i|\lambda_i \sim Pois(\lambda_i) \tag{2.3}$$

$$\lambda_i|\phi, \boldsymbol{\beta} \sim Ga(\phi, \phi/\mu_i) \tag{2.4}$$

where,

β = vector of regression coefficients; and

$\mu_i = \exp(\beta' x_i)$ = mean crash count at site i.

For the gamma distribution, $(\lambda_i) = \mu_i$ and $Var(\lambda_i) = \frac{\mu_i^2}{\phi}$.

Then, the joint distribution of y and λ has a probability density function (PDF) as follows (subscript i omitted without loss of generality). From the following, one can also see that a mixing distribution (gamma) is being combined with a Poisson distribution, hence the alternate naming of an NB model as the Poisson-gamma model.

$$p\,(y, \lambda) = p\,(y|\lambda)(\lambda) \tag{2.5}$$

$$= \frac{\lambda^y * \exp(-\lambda)}{y!} * \frac{\left(\frac{\phi}{\mu}\right)^{\phi}}{\Gamma(\phi)} * \lambda^{\phi-1} * \exp\left(-\lambda\left(\frac{\phi}{\mu}\right)\right) \tag{2.6}$$

$$= \frac{\left(1+\frac{\phi}{\mu}\right)^{y+\phi}}{\Gamma(y+\phi)} \int_0^{\infty} \lambda^{y+\phi-1} * \exp\left(-\lambda * \left(1 + \left(\frac{\phi}{\mu}\right)\right)\right) \tag{2.7}$$

One can then obtain the marginal distribution for the number of crashes by integrating with respect to the mixing distribution, p(λ), as follows:

$$p(y) = \int_0^\infty p(y|\lambda)p(\lambda)_{d\lambda} \tag{2.8}$$

$$= \frac{\left(1+\frac{\phi}{\mu}\right)^{y+\phi}}{\Gamma(y+\phi)} \int_0^\infty \lambda^{y+\phi-1} * \exp\left(-\lambda * \left(1+\left(\frac{\phi}{\mu}\right)\right)\right) \tag{2.9}$$

By noting that the integrand, $\lambda^{y+\phi-1} * \exp\left(-\lambda * \left(1+\left(\frac{\phi}{\mu}\right)\right)\right)$, is a kernel of $Gamma\ (y+\ ^{\phi,\,1+\frac{\phi}{\mu}})$ for λ, one can obtain the following:

$$\frac{\left(1+\frac{\phi}{\mu}\right)^{y+\phi}}{\Gamma(y+\phi)} \int_0^\infty \lambda^{y+\phi-1} * \exp\left(-\lambda * \left(1+\left(\frac{\phi}{\mu}\right)\right)\right) = 1 \tag{2.10}$$

$$\int_0^\infty \lambda^{y+\phi-1} * \exp\left(-\lambda * \left(1+\left(\frac{\phi}{\mu}\right)\right)\right) = \frac{\Gamma(y+\phi)}{\left(1+\frac{\phi}{\mu}\right)^{y+\phi}} \tag{2.11}$$

So,

$$p(y) = \frac{(\phi/\mu)^{\phi}}{\Gamma(y+1)\Gamma(\phi)} * \frac{\Gamma(y+\phi)}{\left(1+\frac{\phi}{\mu}\right)^{y+\phi}} \tag{2.12}$$

$$= \frac{\Gamma(y+\phi)}{\Gamma(y+1)*\Gamma(\phi)} * \left(\frac{\phi}{\mu+\phi}\right)^{\phi} * \left(\frac{\mu}{\mu+\phi}\right)^{y} \tag{2.13}$$

Alternately, letting α=1/φ, the following is obtained:

$$p(y) = \frac{\Gamma\left(y+\alpha^{-1}\right)}{\Gamma(y+1)*\Gamma(\alpha^{-1})} * \left(\frac{\alpha^{-1}}{\mu+\alpha^{-1}}\right)^{\alpha^{-1}} * \left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y} \tag{2.14}$$

Regardless of the form considered, p(y) is the PDF for the negative binomial distribution (i.e., the distribution assumed for the crash count under the NB regression model). The marginal mean and variance of the crash count, y, are as follows:

$$(y) = E\{E(y|\lambda)\} = E(\lambda) = \mu \tag{2.15}$$

$$Var(y) = E\{Var(y|\lambda)\} + Var\{E(y|\lambda)\} = \mu + \frac{\mu^2}{\phi} = \mu + \alpha\mu^2 \tag{2.16}$$

From the preceding, it can be seen that unlike the Poisson distribution, the NB distribution allows for the mean to exceed the variance, a property known as over-dispersion that is quite common in crash data.

After derivation of the NB model, the EB estimate can be derived as follows according to Zou et al. (2017). First, consider the posterior distribution of λ given y:

$$p(\lambda|y) = \frac{p(y,\lambda)}{p(y)} \propto p(y|\lambda)p(\lambda) \tag{2.17}$$

$$\propto \lambda^{y+\phi-1} * \exp(-\lambda * (1 + \phi/\mu)) \tag{2.18}$$

Noting that $\lambda^{y+\phi-1} * \exp(-\lambda(1 + \phi/\mu))$ is a kernel of Gamma(y+φ,1+φ/μ) for λ, the following can be obtained:

$$\lambda|y \sim Gamma(y + \phi, 1 + \phi/\mu) \tag{2.19}$$

Then,

$$E(\lambda|y) = \frac{y+\phi}{1+\phi/\mu} \tag{2.20}$$

$$= \left(\frac{\mu}{\mu+\phi}\right) * y + \left(\frac{\phi}{\mu+\phi}\right) * \mu \tag{2.21}$$

And,

$$Var(\lambda|y) = \left(\frac{\mu}{\mu+\phi}\right) E(\lambda|y) \tag{2.22}$$

The EB estimate of crash count for a given site is then defined as $(\lambda|y)$ and $\frac{\phi}{\mu+\phi}$ is considered to be a weighting factor. Hence, the EB estimate for site i is alternately written as:

$$\hat{\hat{\mu}}_i = w_i * \hat{\mu}_i + (1 - w_i) * y_i \tag{2.23}$$

where,

$\hat{\mu}_i$ = EB estimate for site i;

$\hat{\mu}_i$ = crash count for site i estimated from SPF; and

All other variables as defined previously.

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 threshold are then considered to be hotspots. Besides the EB method, another relatively common HSID method is rooted in so-called "accident reduction potential" (ARP). At first, 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 came from a regression model developed from data at sites similar to the target. The idea was later refined to use the EB estimate instead of the observed accident count because of the EB estimate's robustness to RTM bias and other randomness often present in crash data. On the basis of this refinement, the ARP is computed as follows:

$$ARP_i = w_i * \hat{\mu}_i + (1 - w_i) * y_i - \hat{\mu}_i \quad (2.24)$$

where,

ARP_i = accident reduction for site i; and

All other variables are as defined previously.

Overall, the larger the ARP value, the higher the estimated chances of reducing accidents at a given site, the logic being that the given site has an EB estimate of accidents that is much greater than the accident count estimated for similar sites. Conversely, small values of ARP imply that such sites may not be ideal candidates for treatment as they are estimated to experience accident counts that are quite similar to what is expected at similar sites. Furthermore, research on ARP has raised the question about what set of available predictors should be used in the SPFs that are part of the ARP calculation. Persaud et al. (1999) noted that one may consider using a full set of predictors for the SPF in the EB estimate, while using only a subset of all available predictors in the model for the expected accident count. This reasoning is based on the notion that geometric features specific to the site that cannot be corrected should be considered as a base scenario whose impact can in some sense be subtracted from the true conditions at the site. Stated alternately, the SPF in the EB estimate should consider factors that can be changed to help decrease the expected crash count at a given site.

2.2. Clustering for Selection of Similar Sites

In the following section, we introduce three methods that can be used to group data into different clusters. As previously mentioned, 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 we hoped would align to some degree with the underlying sub-populations from which the crash data were generated. Then separate NB regression models (i.e., SPFs) can be

developed on the basis of each cluster, and EB estimates can then be computed by using an SPF that considers sites that truly are "similar" to the site in question.

2.2.1 Generalized Finite Mixture of NB Regression Models

As is the case for the conventional NB model, the generalized finite mixture of NB regression models with g components (GFMNB-g) assumes that the crash count y follows a Poisson distribution with mean λ. It is important to note that in this derivation, the subscript i denoting the site index is omitted without loss of generality. Unlike the typical NB case, however, the GFMNB-g assumes that λ is from a g-component finite mixture of gamma distributions. That is to say (Zou et al., 2017):

$$y|\lambda \sim Poisson(\lambda) \tag{2.25}$$

$$p(y) = \sum_{j=1}^{g} w_j p_j(\lambda) \tag{2.26}$$

where,

$p_j(\lambda) \sim Gamma(\phi_j, \phi_j/\mu_j)$; and

$w_j > 0$ is a weight value for component j (note that $\sum w_j = 1$).

Then, following Zou et al. (2017) and Gharib (1995), the marginal distribution for y (the number of crashes) is derived according to the following steps:

$$p(y) = \int_0^\infty p(y|\lambda)p(\lambda)_{d\lambda} \tag{2.27}$$

$$= \int_0^\infty \frac{\lambda^y * e^{-\lambda}}{y!} * p(\lambda)d\lambda \tag{2.28}$$

$$= \sum_{j=1}^{g} \left\{ w_j \frac{(\phi_j/\mu_j)^{\phi_j}}{y!\Gamma(\phi_j)} \int_0^\infty \lambda^{y+\phi_j-1} e^{-\lambda(1+\phi_j/\mu_j)} d\lambda \right\} \tag{2.29}$$

$$= \sum_{j=1}^{g} w_j \left\{ \frac{\Gamma(y+\phi_j)}{\Gamma(y+1)\Gamma(\phi_j)} \left(\frac{\phi_j}{\mu_j+\phi_j} \right)^{\phi_j} \left(\frac{\mu_j}{\mu_j+\phi_j} \right)^{y} \right\} = \sum_{j=1}^{g} w_j NB(\mu_j, \phi_j) \tag{2.30}$$

For clarity, we present the marginal distribution of y for site i as follows and note that such a distribution takes the form of the GFMNB-g model presented in Park and Lord (2009).

$$p(y_i|\boldsymbol{x_i}, \boldsymbol{\Theta}) = \sum_{j=1}^{g} w_j \left\{ \frac{\Gamma(y_i+\phi_j)}{\Gamma(y_i+1)\Gamma(\phi_j)} \left(\frac{\phi_j}{\mu_{ij}+\phi_j} \right)^{\phi_j} \left(\frac{\mu_{ij}}{\mu_{ij}+\phi_j} \right)^{y_i} \right\} = \sum_{j=1}^{g} w_j NB(\mu_{ij}, \phi_j) \tag{2.31}$$

We then obtain the marginal mean and variance of y as shown in the following (Zou et al., 2017).

$$E(y_i|\boldsymbol{x_i}, \boldsymbol{\Theta}) = \sum_{j=1}^{g} w_j \mu_{ij} \tag{2.32}$$

$$Var(y_i|\boldsymbol{x_i}, \boldsymbol{\Theta}) = E(y_i|\boldsymbol{x_i}, \boldsymbol{\Theta}) + \left(\sum_{j=1}^{g} w_j \mu_{ij}^2 \left(1 + \frac{1}{\phi_j} \right) - E(y_i|\boldsymbol{x_i}, \boldsymbol{\Theta})^2 \right) \tag{2.33}$$

where,

w_j = the weight of component j (weight parameter), with $w_j > 0$, and $\sum_{j=1}^{g} w_j = 1$;

g = the number of components;

$\mu_{ij} = \exp(\mathbf{x}_i \boldsymbol{\beta}_j)$, the mean rate of component j ;

$\mathbf{x}_i$ = a vector of covariates;

$\boldsymbol{\beta}_j$ = a vector of the regression coefficients for component j ;

$\boldsymbol{\Theta}=\{(\phi_1,...,\phi_g),(\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_g),\mathbf{w}\} = \{(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_g),\mathbf{w}\}$ for $i = 1,2,...,n$; and,

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

Conventional (i.e., non-generalized) FMNB-g models apply a fixed weight factor w_j. Generalized models (i.e., GFMNB-g models), however, use a weight parameter that is not fixed but rather a function of the covariates used in development of the NB regression models themselves. The equation for developing the weight parameter is shown as equation 2.34. 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 depend on the site-specific values of the covariates. Zou et al. (2014) demonstrated how this additional flexibility can lead to better classification results than those obtained from the conventional FMNB-g model.

$$\frac{w_{ij}}{w_{ig}} = e^{\gamma_{0j}} e^{\boldsymbol{\gamma}_j \mathbf{x}_i} \tag{2.34}$$

where,

w_{ij} = the estimated weight of component j at segment i ;

$\boldsymbol{\gamma}_j = (\gamma_{0j}, \gamma_{1j}, \gamma_{2j},...,\gamma_{mj})'$ are the estimated coefficients for component j ,

m is the number of coefficients; and,

$\mathbf{x}_i$ = a vector of covariates.

2.2.2 K-Means Clustering

The K-Means clustering algorithm is often attributed to Lloyd (1982), 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, average daily traffic (ADT), etc.). 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), 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 data set 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 equation 2.35. Then, the second step recalculates the center for each cluster. Pseudo-code for the algorithm is shown in the following.

$$d(x_i, x_{i'}) = \sum_{j=1}^{m} \left(x_{ij} - x_{i'j}\right)^2 = \|x_i - x_{i'}\|^2 \tag{2.35}$$

where,

d(.) = Euclidean distance between two points;

i = data point index, ranging from 1:n;

j = variable index, ranging from 1:m for m variables; and

||.|| = the two norm of two data points.

K-Means Algorithm

While the cluster assignments are still changing…

Cluster-Assignment Step

$$C(i) = argmin_{1 \le k \le K} \|x_i - m_k\| \tag{2.36}$$

where,

C(i) = cluster assignment for data point x_i;

m_k = center of cluster k; and

All other variables defined as previous.

Center-Update Step

$$m_k = \frac{1}{|C(k)|}\sum_{i \epsilon C(k)} x_i \tag{2.37}$$

where,

|C(k)| = cardinality (number of data points) in cluster C(k); and

All other variables 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 data set), and at each step, the two clusters with the smallest intergroup dissimilarity measure are joined to form a new cluster. Hence, the

number of clusters is reduced by one at each successive step. For complete linkage clustering the intergroup dissimilarity is defined as follows (Hastie et al., 2008):

$$d(A,B) = \max_{i \epsilon A, i' \epsilon B} d_{ii'} \tag{2.38}$$

where,

A, B = two arbitrary clusters; and

$$d_{ii'} = \|x_i - x_{i'}\| \tag{2.39}$$

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 data likely have some degree of heterogeneity, as if they are generated from multiple distinct sub-populations. For this reason, 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. Therefore, three types of clustering algorithms (GFMNB-g model based, k-means clustering, and hierarchical clustering with complete linkage) were proposed to cluster the data into distinct subgroups that would 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 so that all points assigned to/belonging to a given cluster are closer or more similar to the points in that cluster than to those of any other cluster (Hastie et al., 2008).

Clustering methods present an ideal means to represent and describe heterogeneity within crash data. Therefore, we applied clustering-based EB methods in this study as a new

means of hotspot identification. For these methods, the three types of clustering previously mentioned were considered, and the classification method for HSID purposes had four main steps. First, the full set of input crash data was 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 was 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 the number of clusters and number of components in the GFMNB-g model is up to the analyst. The second step of the algorithm involved splitting the data into g groups on the basis of the results of the applied clustering algorithm. The third step of the algorithm involved estimating 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, then two SPFs would be estimated, and the data in each of the two groups would have EB estimates calculated through application of the corresponding SPF. Fourth, the EB estimates for all sites across all g subgroups were aggregated and ranked, after which, hotspots were identified on the basis of threshold values or other methods. From this point forward, the classification HSID methods previously discussed will be referred to as the GFMNB- based EB method, the K-means-based EB method, and the hierarchical-based EB method.

A summary of the classification-based EB method for HSID is shown in table 2.1.

Table 2.1 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 on the basis of 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.

Chapter 3 Hotspot Identification Method Evaluation Criteria

In order to evaluate the performance of HSID methods, some kind of standardized test procedures are needed. Ultimately, analysts are concerned with an HSID method's capability to find high-risk sites and to 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. Therefore, we considered three commonly used tests attributed to Cheng and Washington (2008).

3.1 Site Consistency Test

The first hotspot identification evaluation procedure considered for this project was the site consistency test (SCT) (Cheng and Washington, 2008). As the name implies, the goal of the test is to try and uncover consistent performance in a method over time. The underlying idea here is that sites with high accident risk will typically exhibit high accident frequencies over time, assuming that no safety treatments have been applied and that no other major changes have occurred in site-specific conditions. Therefore, a well-performing HSID method should be able to detect high-risk hotspots at multiple points in time as a result of their high crash counts and corresponding high crash risk. In the following description, subscript i denotes a time period, while subscript j denotes and HSID method. The SCT states that the best HSID method will identify the greatest number of accidents at high-risk sites in a future time period i+1. High-risk sites are selected by taking c*n sites, where c is a number less than or equal to 1.0, and n is the total number of sites, and using this subset as the input for each method in the SCT test statistic.

Given a total of n=1,000 sites and a scenario in which we consider the top 5 percent in terms of crash counts as hotspots (i.e., c=0.05), then c*n=0.05*1000=50 sites will be selected as high risk. Such sites will then be ranked in descending order on the basis of their EB estimates (although other criteria, such as AF, AR, etc., could be used in general) from 1, 2., …, c*n. Then, the following test statistic (equation 3.1) is computed for each method, and the best method is denoted as the one that yields the highest value of T_x (Cheng and Washington, 2008). Hence, larger values from the SCT for a given method, all else being equal, indicate better performance of the method.

$$_{(j)} = \sum^{n}_{k=n-cn} C_{k,method=j(i),i+1} \quad (3.1)$$

where,

$T_{SC(j)}$ = site consistency test (SCT) test statistic for method j

C = number of crashes;

i = time period index;

j = HSID method index; and

k = site index.

3.2 Method Consistency Test

Another test for HSID method performance testing developed by Cheng and Washington (2008) is the Method Consistency Test (MCT). The test is similar to the SCT in that it is rooted in the idea that sites that are truly high risk will display poor safety performance across multiple time periods if no safety treatments have been applied and no other large changes in site-specific conditions have occurred. Unlike the SCT, however, the MCT considers whether the same sites are considered as high risk across multiple time periods, as opposed to looking solely at crash counts. If we consider c*n hotspots as defined by HSID method j, the MCT says that the best

performing method is the one that finds the highest number of consistent hotspots over time periods i and i+1. The equation for the MCT test statistic is equation 3.2. All else being equal, methods with higher values of TMC will be determined to be better-performing HSID methods.

$$T_{MC(j)} = \{k_{n-cn}, k_{n-cn+1}, \dots, k_n\}_{j,i} \cap \{k_{n-cn}, k_{n-cn+1}, \dots, k_n\}_{j,i+1} \quad (3.2)$$

where,

$T_{MC(j)}$ = method consistency test (MCT) test statistic for method j;

i = time period index;

j = HSID method index; and

k = site index.

3.3 Total Rank Difference Test

The final test considered for HSID-method performance evaluation in this study was the Total Rank Difference Test (TRDT) (Cheng and Washington, 2008). As with the two previously discussed tests, it is also a consistency test of an HSID method across different time periods, suggesting that high-risk sites will remain high risk (i.e., experience high crash counts) if no safety treatments have been applied and no major changes in site-specific conditions have occurred. The test works by calculating the sum of differences between the ranks of the top c*n high-risk sites for time period i and the ranks of the same sites in time period i+1. It is important to note that the sites considered for time period i+1 have to be the same as those selected as hotspots for time period i; therefore, it is possible that some of these sites are no longer considered in the top c*n hotspots in time period i+1. The TRDT test statistic is calculated as shown in equation 3.3 (Cheng and Washington, 2008).

$$T_{TRDT(j)} = \sum_{k=n-cn}^{n} \left(\Re(k_{j,i}) - \Re(k_{j,i+1}) \right) \quad (3.3)$$

where,

$T_{TRDT(j)}$ = total rank difference test (TRDT) test statistic for method j;

$\Re(k_{j,i})$ = rank of site k from method j for time period i;

i = time period index;

j = HSID method index; and

k = site index.

Chapter 4 Data and Analysis

4.1 Data Description

In order to examine the effectiveness of the methodology presented herein, the research team chose to work with a data set used in many previous safety studies, that being the Texas rural undivided highway data set. 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 (Lord et al., 2008). Since the previously discussed tests for evaluating the hotspot identification methods require data from different time periods for comparison purposes, the data set comprising 1,499 observations was broken down into two temporal subsets.

The first subset, called “Time Period 1”, contained the data from the original data set recorded for 1997 and 1998. The second subset, called “Time Period 2”, contained the data from the original data set recorded for 1999, 2000, and 2001. Thus, the union of these two subsets was the original data set with 1,499 points. Variables collected to describe the segments and to be considered as independent variables in the analysis included average daily traffic over 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 was the number of crashes observed on each segment over the analysis period, and another variable, segment length (L, in miles) was considered to be an offset in the regression. Summary statistics on the data set are presented in table 4.1

Table 4.1 Summary statistics for road segments in the Texas rural undivided highways data set

Variable	Time Period 1 (1997 and 1998)			Time 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)

4.2 Modeling Results

The modeling results from the NB and GFMNB-g models developed with the Texas rural undivided highway data set are presented in this section. When the NB model was developed, a log link function was used; hence, the mean response in terms of estimated number of crashes for segment i can be presented in the following form (equation 4.1), found by exponentiation of both sides of the estimated NB regression equation that uses segment length as an offset (equation 4.2).

$$\mu_i = \beta_0 L_i F_i^{\beta_1} e^{\beta_2 * LW_i + \beta_3 * SW_i + \beta_4 * CD_i} \tag{4.1}$$

$$\ln\left(\frac{\mu_i}{L_i}\right) = \beta_0 + \beta_1 * \ln(F_i) + \beta_2 * LW_i + \beta_3 * SW_i + \beta_4 * CD_i \tag{4.2}$$

where,

μ_i = estimated number of crashes at site i during the study time period;

L_i = length of site (segment) i in miles;

F_i = traffic flow (average daily traffic during the study period) at site i;

LW_i = lane width in feet for segment i;

SW_i = total shoulder width in feet for segment i;

CD_i = curve density (curves per mile) for segment i; and

β_0, β_1, β_2, β_3, β_4 = estimated regression coefficients.

In this study, two different NB regression models were developed with the Texas data, one for each of the two time periods. The model results (e.g., estimated regression coefficients, dispersion parameter (α), standard error (SE), etc.) are shown in table 4.2. In terms of their signs, the estimated coefficients seemed reasonable. For instance, the coefficients for average daily traffic and curve density were positive, indicating that increasing values in their corresponding covariates would lead to more crashes as expected. The estimated coefficients for lane width and shoulder width were negative, indicating that wider lanes and shoulders would lead to a decrease in crashes. Ultimately, the NB models developed here were only used in the conventional EB HSID method.

Table 4.2 NB model results for time periods 1 and 2

Estimates	Time Period 1 (1997 and 1998)		Time Period 2 (1999, 2000 and 2001)	
	Value	**SE**	**Value**	**SE**
Intercept ln(β_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	

After the typical NB model was estimated, GFMNB-g models were also estimated from the Texas data. The mean response, in terms of number of crashes, for each component, at each site is expressed in equation 4.3

$$\mu_{j,i} = \beta_{j,0} L_i F_i^{\beta_{j,1}} e^{\beta_{j,2} * LW_i + \beta_{j,3} * SW_i + \beta_{j,4} * CD_i} \quad (4.3)$$

where,

$\mu_{j,i}$ = estimated number of crashes at site i during the study time period for component j;

$\beta_{j,0}$, $\beta_{j,1}$, $\beta_{j,2}$, $\beta_{j,3}$, $\beta_{j,4}$ = estimated regression coefficients; and

All other variables are as described previously.

The weight parameter applied in the GFMNB-g model is defined as a function of all possible explanatory variables as follows (equation 4.4).

$$\frac{w_{ij}}{w_{ig}} = e^{\gamma_{0,j}} e^{\gamma_{1,j} * L_i + \gamma_{2,j} * F_i + \gamma_{3,j} * LW_i + \gamma_{4,j} * SW_i + \gamma_{5,j} * CD_i} \quad (4.4)$$

where,

w_{ij} = estimated weight coefficient for component j of the GFMNB-g model at segment i;

$\boldsymbol{\gamma_j} = (\gamma_{0,j}, \gamma_{1,j}, \gamma_{2,j}, \ldots, \gamma_{m,j})'$ are the estimated coefficients used to determine the weight coefficient for component j; and

m = number of coefficients (predictors).

To further study classification-based EB methods, GFMNB-g models were developed from the crash data for time periods 1 and 2. Data in each time period were used to estimate the finite mixture models with g components, i.e., g separate NB models were estimated for each type of mixture model and then 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 the GFMNB-g models are estimated, perhaps the biggest issue 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. (2010) 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 a goodness of fit metric such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) that balances the number of components and overall model fit (measured via log-likelihood). Eluru et al. (2012) noted that the BIC is more stringent than the AIC in terms of applying a penalty based on the number of components, and therefore, it

may be more robust in terms of preventing over-fitting. Therefore, the BIC (equation 4.5) was selected as the means of choosing the number of components for the finite mixture models in each of the two time periods.

$$BIC_j = -2 * loglikelihood_j + g_j * \log(n) \quad (4.5)$$

where,

$loglikelihood_j$ is the loglikelihood of model j;

g = number of components in finite mixture model j; and

n = sample size (i.e., number of sites in the data set).

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. Table 4.3 indicates that use of g=2 (i.e., finite mixture models with two components) led to the best goodness-of-fit, as indicated by the lowest value of BIC. Hence, the optimal number of components was selected as g=2 and the GFMNB-g models could then be indicated as GFMNB-2 models. It is important to note that in general, g=2 may not be the optimal number of components, and the choice will depend on the data. That said, the BIC was a reasonable method to use to select g.

By examining tables 4.2 and 4.3, one can see that the BIC values reported for the GFMNB-2 models were smaller than those for the regular NB model in the corresponding time period, suggesting that the mixture models had better goodness-of-fit. Furthermore, the choice of g=2 based on the BIC seemed to suggest the existence of two distinct subpopulations within the crash data corresponding to each time period instead of a lone data population.

Table 4.3 BIC values for GFMNB-g models with g = 2, 3, and 4 components

Time Period 1			
	Number of Components (g)		
Model	**2**	**3**	**4**
GFMNB-g	5833.49	5868.99	5940.98
Time Period 2			
	Number of Components (g)		
Model	**2**	**3**	**4**
GFMNB-g	6755.59	6810.57	6873.91

The results of the GFMNB-2 modeling procedures in terms of coefficients, standard errors, coefficients used to determine component weighting, and dispersion parameters are shown in table 4.4. In some cases, certain covariates were found to be insignificant at the 95 percent confidence level; however, they were still included in the model. The signs of all coefficients are intuitive and consistent with those from the conventional NB model shown in table 4.2.

Table 4.4 Parameter estimates for the GFMNB-2 models

Method	Component	Statistic	ln(β_0)	β_1 (ln(F))	β_2 (LW)	β_3 (SW)	β_4 (CD)	α
Time Period 1								
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
	Estimate		δ_0	δ_1	δ_2	δ_3	δ_4	δ_5
			64.211	-199.09	0.019	-6.908	1.499	-18.618
Time Period 2								
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
	Estimate		δ_0	δ_1	δ_2	δ_3	δ_4	δ_5
			2.282	4.8630	-0.0003	-0.091	-0.066	0.187

* Not significant at 5% significance level; † SE = Standard Error.

4.3 Grouping Results

On the basis of the results of the GFMNB-g fitting procedure, the authors determined that a GFMNB model with 2 components fit the data best. Therefore, 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 previously described 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.

Since the GFMNB-g model essentially provided a soft clustering of the data (i.e., data points are assigned to each group with some probability level), road segments were classified into one group (i.e., a hard clustering) by assigning them to the component with the higher posterior probability. From Zou et al. (2014) and Rigby and Stasinopolous (2009), the posterior probability that data corresponding to observation y_i are from component j of the GFMNB-g is given in equation 4.6.

$$\hat{\varepsilon}_{ij}^{(r+1)} = p\left(\delta_{ij} = 1 \middle| y_i, \widehat{\boldsymbol{\Theta}}^{(r)}, \boldsymbol{x}_i\right) = \frac{\hat{w}_{ij}^{(r)} f_j(y_i|\widehat{\boldsymbol{\theta}}_j^{(r)}, \boldsymbol{x}_i)}{\sum_{k=1}^{g} \hat{w}_{ik}^{(r)} f_k(y_i|\widehat{\boldsymbol{\theta}}_k^{(r)}, \boldsymbol{x}_i)} \tag{4.6}$$

where,

δ_{ij} = indicator variable denoting group/component membership;

$\hat{w}_{ij}^{(r)} = p\left(\delta_{ij} = 1|\widehat{\boldsymbol{\Theta}}^{(r)}\right)$ = prior probability that y_i is from component j, given $\hat{}^{(r)}$, which is estimated from the r^{th} iteration of the expectation-maximization algorithm (which is used to fit the GFMNB-g); and

All other variables as defined previously.

Table 4.5 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 data set (as described previously) are presented. From the table, it can be seen that in general, mean values for the lane width, shoulder width, and segment length did 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 differed dramatically between components. Additionally, there was 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 have come from underlying subpopulations in which traffic volume was a defining characteristic for subpopulation membership and thus a good descriptor of the heterogeneity in the data.

Table 4.5 Summary statistics of each component for time periods 1 and 2

Method	Component (Sample)	Statistic	Crashes	F	LW	SW	L
Time Period 1 (years 1997 and 1998)							
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 (years 1999 to 2001)							
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

With crash data clustered for each time period according to the three aforementioned clustering methods, EB estimates were then obtained after an NB regression model had been estimated for each of the two components corresponding to a given clustering method for a given time period. When results are interpreted, one should consider the sample sizes used to estimate the NB models. For example, "Component 1" (i.e., one grouping) for Time Period 2, as defined by hierarchical clustering with complete linkage, had only 66 data points. Therefore, 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.

4.4 Test Results

Evaluations of six different HSID method— (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—were conducted by using the three main tests from Cheng and Washington (2008). As all test procedures involved comparison across two different time periods, we used the time periods as defined in table 4.1. Furthermore, we considered three different scenarios in terms of the number of high-risk sites selected for consideration under each HSID method. These scenarios corresponded to considering 1 percent, 5 percent, and 10 percent of all sites as high risk (i.e., c={0.01, 0.05, 0.10}). For example, in this study, when c=0.10, a total of approximately 150 sites (i.e., ~10 percent of the 1,499 total sites) were considered as high risk, and their data were used in calculation of the test statistics for the various HSID methods.

Table 4.6 shows the results of the six HSID methods considered under the SCT. As previously discussed, the goal of the SCT is to measure the 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 is. From the table, it can be seen that the worst performing method across all cut-off levels for high-risk site identification (i.e., all c values) was the AR method. When 1.0 percent of sites were considered as high risk, the conventional EB method, K-means-based EB method, and the Hierarchical-based EB method all perform equally well. For the cases in which 5 percent and 10 percent of sites were considered as high risk, the K-means-based EB method was 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 gave a value quite close to those obtained by the K-means-based method, indicating that it performed nearly as well in HSID.

Table 4.6 Results of the site consistency test for various methods

Method	**c = 0.01**	**c = 0.05**	**c = 0.10**
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

* Number in bold indicates the best result under each cut-off level.

The results of the evaluation of the six HSID methods in terms of the MCT are shown in table 4.7. The MCT was designed to assess consistent identification of the same high-risk sites across different time periods. Therefore, 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 4.7, one can see that across all three cut-off levels for proportions of sites to consider as high risk, the GFMNB-based EB method

performed the best. That said, for the case in which 10 percent of sites were considered as high risk, the K-means-based method performed just as well. Additionally, for all cut-off levels, the clustering-based EB methods (e.g., GFMNB-, K-means-, and hierarchical-based) exhibited quite similar performance. As was the case for the SCT, the AR method consistently performed the worst across all three cut-off levels for proportions of sites to consider as high risk.

Table 4.7 Results of method consistency test for various methods

Method	**c = 0.01**	**c = 0.05**	**c = 0.10**
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

* Number in bold indicates the best result under each cut-off level.

Table 4.8 presents the results of the TRDT evaluation of the HSID procedure. Again, this test is based on consistent of 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 yielded the best HSID performance across all three cut-off levels of proportions of sites to consider as high risk. Under this test, the other clustering-based EB methods (e.g., K-means-based and hierarchical-based) outperformed the naïve AF and AR methods across all cut-off values and also outperformed the EB method for the 5 percent and 10 percent cut-offs. As was the case for the preceding two HSID performance tests, the AR method of HSID consistently performed the worst across all three cut-off levels of proportions of sites to consider as high risk.

Table 4.8 Results of total rank difference test for various methods

Method	c = 0.01	c = 0.05	c = 0.10
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

* Number in bold indicates the best result under each cut-off level.

Overall, the preceding tests indicated that the GFMNB-based EB method exhibited the strongest HSID performance in all three tests and across the different cut-off levels of proportion of sites to consider as high risk. That said, the results obtained from the other clustering-based EB methods (e.g., K-means-based and hierarchical-based) were usually close and tended to outperform the AF, AR, and standard EB methods. From all tests, it appeared that the AR method performed the worst. One possible explanation for this behavior may be that because the test sites were rural road segments, many may have exhibited low ADT values and thus, as previously discussed, low-volume sites may be over-represented as high risk since the AR calculation normalizes by traffic count. Ultimately, HSID methods that themselves made use of the EB method in computing safety estimates prior to site ranking appeared to perform better than the naïve AF and AR methods. This finding was consistent with many previous studies, including Cheng and Washington (2008), Cheng and Washington (2005), Montella (2010), and Wu et al. (2014).

4.5 Discussion

The preceding analysis showed that the GFMNB-based EB procedure for HSID performed the best when evaluated with the three discussed test procedures from Cheng and Washington (2008) with the Texas rural undivided highway data set. That said, all EB-based

methods typically outperformed the naïve methods, especially the AR HSID method. One possible reason that the EB-based HSID methods performed better may be their use of both the observed historical crash data and predicted crash counts from similar sites with the SPF. Furthermore, the EB methods were able to adjust for RTM bias. That said, the conventional EB method is not without its limitations for HSID. The main limitation, perhaps, occurs when there is a substantial degree of heterogeneity in the crash data, so that the crash data seem to arise from different subpopulations. Such heterogeneity could occur when large amounts of crash data are collected from areas that differ dramatically both geographically and with respect to a variety of other site-specific conditions. Often, 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 project. The idea behind these methods was to group the overall set of crash data (i.e., the full list of study sites) into smaller subsets so that the sites in each subset were more similar to sites within their groups than to sites in other groups (i.e., minimize within-group variance and maximize between-group variance) in terms of features such as traffic volume, lane width, and other predictors. Furthermore, 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 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 the 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 be further compounded when analysts interpret the biased results and make erroneous inferences/conclusions based on them. Therefore, it is important that one be cognizant of the sample sizes of the clusters and the impacts they may have on model estimates and resulting inference (Lord, 2006). Ultimately, as always, analysts are encouraged to interpret all results, especially those corresponding to regression models developed from small samples (e.g., 100 or fewer sites) with caution.

Chapter 5 Conclusions

This study proposed 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. In general, the clustering-based EB method for HSID has the following steps:

(1) Use the GFMNB-g model, K-means algorithm, or hierarchical clustering algorithm to cluster the full set of crash data into g groups.

(2) Separate the data into g groups on the basis of the results of the clustering.

(3) Estimate g NB regression models, one for each of the g subgroups, and use the corresponding SPF to get the EB estimates for each site.

(4) Aggregate the EB estimates for all sites, rank the sites, and identify hotspots.

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 methods, by using three methods for comparing performance in HSID across different time periods developed by Cheng and Washington (2008). The HSID results based on applying the methodology to Texas undivided rural highway crash data suggested that all three clustering-based EB analysis methods are preferable to the conventional statistical methods. Additionally, the HSID accuracy can be possibly improved by properly classifying roadway segments on the basis of heterogeneity in the crash data (i.e., clustering the data before developing SPFs for use in the EB estimates). That said, caution should always be taken when roadway segments are classified into clusters, as inappropriate classification of roadway segments can result in erroneous results (e.g, biased coefficient estimates from SPFs developed from small sample sizes). Future work could investigate the 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).

References

Abbess, C.R., Jarrett, D.F., Wright, C.C., (1981*).”Accidents at blackspots: Estimating the effectiveness of remedial treatment with special reference to the “regression to mean” effect”*. Traffic Engineering and Control, vol. 22,10, 535-542.

American Association of State Highway and Transportation Officials (AASHTO). (2010). *Highway Safety Manual*, AASHTO, Washington, D.C.

Cameron, A. C., and Trivedi, P. K. (1998). *Regression analysis of count data*, Cambridge University Press, Cambridge, UK.

Cao, X., Xu, Z., and Huang, A. Y. (2012). “Safety Benefits of Converting HOV Lanes to HOT Lanes: Case Study of the I-394 MnPASS.” *Institute of Transportation Engineers Journal*, 82, 3237.

Cheng, W., and Washington, S. P. (2005). “Experimental Evaluation of Hotspot Identification Methods.” *Accident Analysis and Prevention*, 37, 870-881.

Cheng, W., and Washington, S. (2008). “New Criteria for Evaluating Methods of Identifying Hot Spots.” *Transportation Research Record*, 2083, 76-85.

Deacon, J. A., Zegeer, C. V., and Deen, R. C. (1975). “Identification of Hazardous Rural Highway Sections.” *Transportation Research Record*, 543, 16-33.

Eluru, N., Bagheri, M., Miranda-Moreno, L. F., Fu, L. 2012. “A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings.” *Accident Analysis and Prevention*, 47, 119-127.

Gharib, M. (1995). “Two characteristics of a gamma mixture distribution.” *Bulletin of the Australian Mathematical Society*, 52, 353-358.

Golembiewski, G. A., and Chandler, B. (2011). *Roadway Departure Safety: A Manual for Local Rural Road Owners*, FHWA, Washington, D.C.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Stanford, CA.

Hauer, Ezra. (1996). Identification of Sites with Promise. Transportation Research Record. 1542. 54-60. 10.3141/1542-09.

Hauer, E., Harwood, D. W., Council, F. M., and Griffith, M. S. (2002). “Estimating Safety by the Empirical Bayes Method: A Tutorial.” *Transportation Research Record*, 1784, 126-131.

Highway Safety Manual (2010)

Lloyd, S. P. (1982). "Least squares quantization in PCM." *IEEE Transactions on Information Theory*, 28, 129-137.

Lord, D. (2006). "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 and Prevention*, 38(4), 751-766.

Lord, D., et al. (2008). *NCHRP Web-Only Document 126: Methodology to Predict the Safety Performance of Rural Multilane Highways – Contractor's Final Report for NCHRP Project 1729*, Transportation Research Board, Washington, D. C.

Mannering, FL, Bhat, C.R., Analytic Methods in Accident Research: Methodological frontier and future directions. (2014). 1-22

Montella, A. (2010). "A Comparative Analysis of Hotspot Identification Methods." *Accident Analysis and Prevention*, 42, 571-581.

Mountain, L., Fawaz, B., and Jarrett, D. (1996). "Accident Prediction Models for Roads with Minor Junctions." *Accident Analysis and Prevention*, 28, 695-707.

Park, B.-J., and Lord, D. (2009)."Application of finite mixture models for vehicle crash data analysis." *Accident Analysis and Prevention,* 41, 683-691.

Park, B.-J., Lord, D., and Hart, J. D. (2010). "Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis." *Accident Analysis and Prevention,* 42**,** 741-749.

Persaud, B. N., Lyon, C., and Nguyen, T. (1999). "Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement." *Transportation Research Record*, 1665, 7-12.

Persaud, B.N., Hadayeghi, A., Shalaby, A. (2010) "*Development of Planning Level Transportation Safety Modeuls usting Full Bayesian Semiparametric Additive Techniques"*, Journal of Transporatation Safety & Security, 2:1, 45-68.

Rigby, B., and Stasinopolous, M. (2009). *A Flexible Regression Approach using Gamlss in R*, University of Lancaster, Lancashire, England.

Wu, L., Zou, Y., and Lord, D. (2014). "Comparison of Sichel and Negative Binomial Models in Hot Spot Identification." *Transportation Research Record,* 2460**,** 107-116.

Zou, Y., Wu, L., and Lord, D. (2013). "Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis." *Accident Analysis and Prevention,* 50, 1042-1051.

Zou, Y., Zhang, Y., and Lord, D. (2014). "Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models." *Accident Analysis and Prevention,* 1, 39-52.

Zou, Y., Ash, J.E., Park, B.J., Lord, D., and Wu, L., 2017. Application of Finite Mixture of Negative Binomial Regression Models in Estimating Empirical Bayes Estimates. Paper submitted for publication.