A Comparative Analysis of Intersection Hotspot Identification: Fixed vs. Varying Dispersion Parameters in Negative Binomial Models

Yi Meng, Lingtao Wu*, Chaolun Ma, Xiaoyu Guo and Xiubin (Bruce) Wang

*Corresponding author
Email: l-wu@tti.tamu.edu

aTexas A&M Transportation Institute, Bryan, USA;
bZachry Department of Civil Engineering, Texas A&M University, College Station, USA
A Comparative Analysis of Intersection Hotspot Identification: Fixed vs. Varying Dispersion Parameters in Negative Binomial Models

Network screening for crash hotspot is the first step in roadway safety management. The Empirical Bayes (EB) method has been widely used for ranking sites. The negative binomial (NB) distribution is the most frequently used model for developing safety performance functions (SPFs). The dispersion parameter of NB models plays a critical role in the EB process. There are primarily two forms of dispersion parameters: fixed and varying. Previous studies illustrated that SPFs with varying dispersion parameters showed better performance in modeling crash data, and the Highway Safety Manual has adopted the varying form for segment SPFs. However, it is unclear whether intersection models should utilize the varying form. The primary objective of this paper is to examine the performance of intersection SPFs with varying dispersion parameters in hotspot identification. Five models with varying forms of dispersion parameters (VDP) were developed at 1,943 unsignalized intersections in Texas, and the results were compared with the traditional EB approach with a fixed dispersion parameter (EB-FDP). It is found that EB approach with two proposed VDP were superior to EB-FDP in hotspot identification. Safety analysts and practitioners are encouraged to consider varying forms of dispersion parameter in future work when analyzing intersection crashes.

Keywords: varying dispersion parameter, hotspot identification, safety performance function, Empirical Bayes
1. Introduction

Intersections take a small portion of roadway space but belong to the site types where fatal crashes happen most frequently. According to the Federal Highway Administration (FHWA), traffic fatalities at intersections increased consistently in recent years. The conflict points of roadway user’s trajectory movements generally happen at intersections, making them the concentration of traffic crashes. Intersection safety is one of the seven emphasis areas in the Strategic Highway Safety Plan (https://www.texasshsp.com/). Statistical modeling techniques have been widely used to quantitatively estimate the safety of roadway entities as well as to rank sites. Among these techniques, the empirical Bayes (EB) method is an established, state-of-art method in crash analysis since it solves the regression-to-mean (RTM) problem. Safety researchers have extensively applied the EB method in analyzing intersection safety, identifying higher risk intersections, and estimating the effectiveness of countermeasures, such as left-turn phasing (Srinivasan et al., 2012), flashing yellow arrows used in protected permitted phasing (Appiah et al., 2018) and adaptive traffic-signal control (Ma et al., 2016). Many of them incorporated the EB method with other methods or considerations. For example, Gross et al. (2013) examined the crash modification factor (CMF) estimation for converting signalized intersections to roundabouts using the EB method together with a cross-sectional analysis for the sake of enriching sample size and unveiling the compatibility between the results of cross-sectional and before-after studies. Naznin et al. (2016) explored crash analysis from the perspective of transit-roadway interaction by applying EB on crash estimation of streetcars at road section and intersections. Yang et al. (2016) conducted pedestrian safety at two-way stop-controlled intersections with EB incorporated with the probability density function for pedestrians considering motorist yielding behavior and pedestrian arrival time.
The EB estimate of safety is a combination of observed crashes and predicted number of crashes. In the commonly used Negative Binomial (NB) distribution structure, the dispersion parameter plays a critical role in the EB estimation. All the aforementioned studies assumed a constant or fixed dispersion parameter (FDP) for all entities. Even though the fixed dispersion parameter reflects the data variation to some extent with the prior Gamma distribution in the NB distribution structure, it may not precisely capture the heterogeneity of real-world data. Therefore, data-driven structures are needed for capturing these extra variations. The idea of varying dispersion parameter (VDP) for NB distribution was first proposed by Hauer (2001), which noted that the dispersion parameter was proportional to roadway segment length and thus affected the weight that used in the EB method. Since the EB estimate is the weighted average of NB model prediction and the actual crash number, the change of dispersion parameter would substantially affect the EB estimate of crashes. This finding has been supported by several studies (Cafiso, et al., 2010; Geedipally et al., 2009). Although the first version of the Highway Safety Manual (HSM) applied a functional form for the dispersion parameter on the safety performance function (SPF) of roadway segments, a fixed dispersion parameter was used for the SPFs of the intersections. As features of the intersections differ from those of the roadway segments, the formula form used for estimating varying dispersion parameter of roadway segments cannot be applied to the intersection data without appropriate modifications.

More importantly, EB method is among the most popular choices of network screening for identifying hazardous sites. Numerous studies on hotspot identification (HSID) aimed at finding good indicators such as crash frequency, crash rate, property damage only, accident reduction potential, etc. (Deacon and Zegeer, 1974; Hauer and Persaud, 1984; Lord and Park, 2008b; Stokes and Mutabazi, 1996). The EB estimate of total crash frequency is shown to be more consistent and
reliable than other methods (Montella, 2010; Qu and Meng, 2014; Yu et al., 2014) for HSID. It was demonstrated by several studies that a varying dispersion parameter is a remedy for improving the performance of EB estimation on the before–after evaluation of the effects of implemented countermeasures (Geedipally and Lord, 2008; Lord and Park, 2008a; Miaou and Lord, 2003; Mitra and Washington, 2007). However, the performance of intersection SPF{s with varying dispersion parameter has not been fully examined, especially in the application of network screening. Consequently, the objective of this paper is two-fold. First, this paper evaluates the effect of the dispersion parameter in terms of its performance in HSID and model statistical fits. Second, this paper provides proper functional forms of the dispersion parameter that can improved the performance of EB in HSID.

The rest of this paper includes the following content. Section 2 reviews literature pertaining to the varying form of dispersion parameters. Section 3 describes the proposed intersection safety models with varying dispersion parameters. A summary of the dataset for this paper is provided in section 4. In section 5, details of our findings and modeling results are presented and concluded. This paper ends with discussions and potentials for future research in section 6.

2. Background

Since the main assumption of the Poisson model is violated due to the over-dispersed nature of crash data, NB distribution is introduced and widely used in SPF development. However, the heterogeneity of the real-world dataset makes the classic NB model suffer from underestimating the true data dispersion level (Zou et al., 2015). Recently, substantial statistical models were built to address this issue such as negative binomial-Dirichlet, Sichel (the Poisson-generalized inverse
Gaussian distribution), negative binomial-Lindley with random parameters, and Poisson inverse Gaussian regression model (Shaon et al., 2018; Shiraziet al., 2016; Zha et al., 2016; Zou et al., 2015).

The dispersion parameter (inverse of the gamma shape parameter) captures the actual dispersion level of the data and thus attracts the attention of safety researchers. According to Lord (2008), the over-dispersion of datasets is caused by some unmeasured uncertainties related to the unobservable variables. In addition to new statistical models, two other strategies were applied to deal with the over-dispersion issue. One is by replacing fixed dispersion parameters to varying dispersion parameters, and the other is by building multivariate models for the crash expectation.

Existing literature associated with the varying dispersion parameter can be categorized into two categories: varying dispersion parameter on crash estimation for roadway segments (Cafiso et al., 2010; Geedipally et al., 2009) and intersections (Geedipally and Lord, 2008; Lord and Park, 2008b; Miaou and Lord, 2003; Mitra and Washington, 2007). In the study of Lord and Park, (2008b), the dispersion parameter was estimated each year with different functional forms, and the time-varying dispersion parameter showed better performance on the generalized NB model than the fixed multi-year model on rural three-leg intersections. Geedipally and Lord (2008) examined the effect of varying dispersion parameter with respect to the confidence interval on three-leg rural unsignalized intersections in California. It was found that the involvement of varying dispersion parameters would shrink the confidence interval length, which made the prediction more precise. In the meantime, models with varying dispersion parameters provided better statistical fits, especially for flow-only models. However, the parameterization of the dispersion parameter is limited. Geedipally (2009) provided a broader scope of the parameterization by evaluating 10 structures of relationships among the dispersion parameter, roadway segment length and AADT
in three datasets. Results showed that any functional form (flow-only, length-only or the combination of length and flow) could not be generalized for all datasets and the linear relationship of the dispersion parameter and segment length proposed by Hauer (2001) did not always outperform other functional forms.

Mitra and Washington (2007) reported that the effect of varying dispersion parameter became insignificant with a well-defined mean function that incorporates multiple covariates. It was pointed out that there was a trade-off between the number of covariates involved and the use of variance function in the NB model. Based on this study, Cafiso et al. (2010) made a comparison between the AADT only model and multivariable models, with or without varying dispersion parameter on rural two-lane roads selected from Italy and Ohio datasets. The results confirmed the finding of Mitra and Washington (2007), indicating that the application of varying dispersion parameter on multivariable models would not improve the model fit. For the AADT only model, varying dispersion parameter provides a better fit with smaller Akaike information criterion (AIC) value. Their study assumed that the dispersion parameter was supposed to be only related to segment length, which is questionable because the functional form for the dispersion parameter is site-type specified and factors like AADT may truly reflect specific dataset features (Geedipally et al., 2009).

Additionally, few studies have investigated the effect of varying dispersion parameter on the crash hotspot identification. Miranda et al. (2005) used the conditional mean of marginal and posterior distributions in the rank of hazardous sites to compare the performance of the NB model and Poisson-lognormal model with the variability in the dispersion parameter allowed. With the number of selected hotspots as the only standard of model evaluation, the varying dispersion parameter was shown to be beneficial to the accident estimates. Lord and Park (2008b) visually
explored the effect varying dispersion parameter in hazardous sites identification by plotting ranks from fixed model versus ranks from model with varying dispersion parameter.

To sum up, model statistical fits was used as the only standard to evaluate the effect of the varying dispersion parameter in existing studies. The application of dispersion functions in the EB method would improve model statistical fits in the development SPF of roadway segments. In a similar fashion, several studies have examined the SPF of intersections with varying forms of dispersion parameters and it was demonstrated that the varying dispersion parameter would slightly improve model statistical fits only in some cases. Additionally, studies exploring the effect of the varying dispersion parameter on hotspots ranking are rather rare. To the best of authors’ knowledge, no systematic quantitative method has been applied in the evaluation of VDPMs (varying dispersion parameter models) for roadway intersections.

3. Methodology

The commonly used NB distribution (also known as Poisson-gamma distribution) is a mixture of Poisson distributions with the Poisson rate $\lambda_{i,t}$ following gamma distribution, or:

$$y_{i,t} | \lambda_{i,t} \sim \text{Poisson}(\lambda_{i,t}), \forall i \in \{1,2,3 \ldots N\}, t \in T$$

$$\lambda_{i,t} \sim \text{Gamma}(\tau, \theta), \forall i \in \{1,2,3 \ldots N\}, t \in T$$

Where $y$ denotes the crash frequency and $N$ denotes the total number of entities studied. The subscripts $i$, $t$ represent entities index and study period, respectively.

The probability distribution of $\lambda_{i,t}$ with a shape parameter $\tau$ and a scale parameter $\theta$ is given by (with subscripts eliminated for simplicity):

Where $y$ denotes the crash frequency and $N$ denotes the total number of entities studied. The subscripts $i$, $t$ represent entities index and study period, respectively.
\[ f(\lambda|\tau, \theta) = \frac{1}{\theta^{\tau}} \times \frac{\lambda^{\tau-1} \times \exp(-\lambda/\theta)}{\Gamma(\tau)} \]  

(1)

Summing out \(\lambda\) in Equation 1, the probability mass function (PMF) of the NB distribution can be obtained as:

\[ p(y|\tau, \theta) = \int_0^{\infty} \frac{\lambda^y \times \exp(-\lambda)}{y!} \times \frac{\lambda^{\tau-1} \times \exp(-\lambda/\theta)}{\Gamma(\tau)} d\lambda \]  

(2)

Using gamma function and the PMF can be rewritten as (Hilbe, 2011):

\[ p(y|\mu, \alpha) = \frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1) \times \Gamma(\frac{1}{\alpha})} \times \left(\frac{\alpha \times \mu}{1+\alpha \times \mu}\right)^y \times \left(\frac{1}{1+\alpha \times \mu}\right)^{1/\alpha} \]  

(3)

Where, \(y\) is the response variable, \(\mu\) indicates the mean response of the observation, and \(\alpha\), which can also be interpreted as \(\frac{1}{\tau}\), is the dispersion parameter.

### 3.1 Model evaluation by statistical fits

In the HSM, the dispersion parameter is fixed for intersection SPFs, while the varying dispersion parameter (VDP) is only applied for roadway segment SPFs. In this paper, we explored several functional forms for the dispersion parameter. It is worth mentioning that this paper considers the flow-only model. Compare to multivariate models, flow-only models have three advantages that make them easier to calibrate when applied to different jurisdictions (Persaud et
al., 2002). First, the intersection SPFs with flow-only parameters require less data input than SPFs that contain multiple variables (such as intersection skew angle, number of left/right turning lanes, intersection lightening, etc). In many cases, those variables are not fully recorded or in poor quality for the model calibration. Second, even if the essential variables are available to researchers, it could be challenging to obtain reliable crash modification factors (CMFs) for these variables. Taking the variable number of left-turn lanes as an example, the CMF reported in the CMF Clearinghouse varies from 0.49 to 1.02. Unreliable CMFs will result in inaccurate calibration factors. Finally, with more variables involved, the specification of a proper functional form for the varying dispersion parameter would be much harder to determine.

VDP Model 1 to 5 below specify five different forms of the dispersion function (variance function) considered in this paper:

VDP Model 1(M1): $\alpha = \exp \left( \frac{\gamma}{\log(AADT_{major})+\log(AADT_{minor})} \right)$ \hspace{1cm} (4)

VDP Model 2(M2): $\alpha = \exp \left( \frac{\gamma}{\log(AADT_{major}+AADT_{minor})} \right)$ \hspace{1cm} (5)

VDP Model 3(M3): $\alpha = \exp \left( \frac{\gamma}{\log(AADT_{major})+\log(AADT_{minor})} \right)$ \hspace{1cm} (6)

VDP Model 4(M4): $\alpha = \exp \left( \frac{\log(AADT_{major})+\log(AADT_{minor})}{\gamma} \right)$ \hspace{1cm} (7)

VDP Model 5(M5): $\alpha = \exp \left( \frac{\log(AADT_{major}+AADT_{minor})}{\gamma} \right)$ \hspace{1cm} (8)
With dispersion functions given above, crash frequency estimations deriving from the EB procedures are conducted respectively for each model. First, the mean estimate takes the form of:

\[
\hat{Y}_{i,t,m} = e^{\beta_0} \ast AADT_{major}^{\beta_1} \ast AADT_{minor}^{\beta_2}
\]  

(9)

Where \(\hat{Y}_{i,t,m}\) is mean estimation and the subscript, \(m \in M = \{m|0,1,2..5\}\) specifies models used. \(m=0\) stands for the EB method with fixed dispersion parameter whereas \(m=1...5\) refers to the EB method with varying dispersion parameters estimated via model \(m\). All \(\beta\)'s in function (9) and \(\gamma\)'s in model 1 to model 5 are estimated simultaneously with method of maximum likelihood coded in R. Since there is no R package that is ready for the analysis of this study, the authors developed log-likelihood function specifically for the varying dispersion parameter crash models, and used package “bbmle” (https://cran.r-project.org/web/packages/bbmle/index.html) to obtain MLE estimates.

Further, with the weight obtained with \(w_i = \frac{1}{1+\alpha \ast \hat{Y}_{i,t,m}}\), the model prediction of crash for each study period is written as:

\[
Y_{i,t,m} = w_i \ast (\hat{Y}_{i,t,m}) + (1 - w_i) \ast y_{i,t,m}
\]  

(10)

Lastly, in this research, AIC is chosen as a measure of effectiveness (MOEs) for model fitting. Let \(p_i\) be the number of parameters estimated and \(\hat{L}_i\) be the maximum likelihood function for each model, the AICs is calculated as follows:

\[
AIC_i = 2 \ast p_i - 2 \ast \log (\hat{L}_i)
\]  

(11)
3.2 Model evaluation by hotspot identification tests

This part documents the numerical tests used for investigating the performance of varying dispersion parameter in crash hotspot identification. The evaluation method mainly follows the methods proposed by Cheng and Washington, (2008) and Guo et al. (2019), but necessary changes are made to fulfill the research requirements. The crash hotspots in this paper are identified based on crash estimations from each model discussed in the previous section. All entities are ranked in descending order based on the corresponding model estimates. A threshold \( \epsilon \) is defined for selecting intersections with top \( \epsilon \) concerns, where \( \epsilon \) is a percentage. For example, \( \epsilon = 10\% \) means top 10\% of all \( N \) intersections with high crash estimations are selected for model comparison and evaluation.

Three tests are applied to evaluate performances of these models, including estimation consistency test, site consistency test and a modified rank difference test. A two-period scheme is required for these tests, denoted by \( t \in \{0, 1\} \). The estimation consistency test (T1) will reveal the consistency of risk level identification for one site in both study periods. The assumption behind this test is that if one spot is screened as high risk in the first period, it tends to be high risk in the period afterwards given no countermeasure occurs. Mathematically, if \( \epsilon N \) high-risk intersections are selected for model evaluation, the final score of T1 is represented by:

\[
T_{1m} = \sum_{i=t}^{N} \bar{Y}_{i,t,m}, \forall i \in \{N - \epsilon N, N - \epsilon N + 1 \ldots N\}, t = 1, m \in M
\]  

(12)
Where N is total number of sites, t refers to study periods and \( m \in M \) represents different methods for estimating \( Y_{i,t,m} \), the mean estimation of crash frequency.

In contrast to T1 which examines the overall stability of cumulative crash frequencies of top \( \epsilon \) intersections, site consistency test (T2) focuses on the consistency of the number of high crash rate sites over both study periods. Let \( \Theta_{t,m} \) denotes the set contains hazardous intersections, where \( |\Theta_{t,m}| = \epsilon N \). The robustness of one model requires an intersection \( I_i \) to be very likely within the top \( \epsilon \) in the first period, (\( I_i \in \Theta_{0,m} \)), if it is recognized as a high crash rate site in period afterwards (\( I_i \in \Theta_{1,m} \)). Note that the homogeneity assumption needs to hold for the validation of T2, that is, geometrical and operational features of the intersection stay unchanged. Intersections with countermeasures are excluded from this test since the countermeasures are supposed to affect entities’ safety performance significantly. The site consistency test is given by:

\[
T^2_m = \left| \bigcap_{i=1}^{\epsilon N} \Theta_{t,m} \right|, \forall i \in \{N - \epsilon N, N - \epsilon N + 1 ... N\}, m \in M
\]  

(13)

Differences between ranks in the prior period and latter period reflect the consistency in model prediction. Apparently, a more consistent model provides smaller changes among analysis periods than an inconsistent one. Following this logic, the rank difference test (T3) is developed as a supplement for T2. On a general basis, T2 concerns the number of intersections identified as high risk in both periods of the EB procedure. T3 aims at capturing the detailed change of the ranks \( R_{t,m} \) in the form of summation, in other words, the cumulative absolute difference between estimations in different study periods using the aforementioned models. Analytically, the test score is formulated as follows:
\[
T3_m = \sum_{i=0}^{N-\epsilon N} |R_{t=0,m}(I_i) - R_{t=1,m}(I_i)|, \forall i \in \{N - \epsilon N, N - \epsilon N + 1 \ldots N\}, m \in M
\] (14)

One method (i.e., \(m\)) outperforms others (i.e., \(m'\)) if and only if \(T1_m > T1_{m \in M}\), \(T2_m > T2_{m \in M}\), \(T3_m < T3_{m \in M}\).

### 4. Data Description

To examine the performance of intersection SPFs with varying form dispersion parameters, the authors collected 1,943 rural 3-leg unsignalized stop-controlled intersections in Texas. Two years (2017 and 2018) of crash data on these intersections were extracted from the Crash Records Information System (CRIS), managed by the Texas Department of Transportation (TxDOT), and the volume on major and minor roadways were gathered from the Texas Roadway Inventory Database. Since this paper considers the flow-only model, variables such as lane, median and shoulder width, functional class of crossing roadways, have been used as filters for similar intersections. The summary statistics of the data are shown in Table 1. 1,000 of the intersections were randomly selected to develop the SPFs, and the remaining 943 were used for model performance evaluation.

From the perspective of data distribution, the data shows higher variances than means for all features concerned and clearly displays the over-dispersion nature of crash data. Given the information in Table 1, it is assumed that the number of crashes at each intersection is Poisson distributed with its mean following the Gamma distribution. The study includes a two-year scope: period 1 refers to the year 2017 and period 2 refers to the year 2018. No countermeasures are applied, which guarantees the consistency of factors not included in the model.
5. Results

Model statistical fits results, as well as test scores for hazard intersection identification, are presented in the following subsections.

5.1 Results of Statistical Fits

Table 2 records statistical fits of the five proposed varying dispersion parameter models (VDPMs) and the fixed dispersion parameter model (FDPM), with standard deviation in parentheses. Overall, all proposed models and the FDPM provide similar estimations on the mean of parameters. The parameters of AADT on major road ($\beta_1$) and AADT on minor road ($\beta_2$) are quite similar across all five models and the FDPM. Judging from the standard deviations of $\beta_1$ and $\beta_2$, the sampling distributions of these two parameters are very stable no matter which model is applied. This indicates the reliability of the application of varying dispersion parameters on the NB model which conforms to earlier studies. Compared to M1, M2 and M3, M4 and M5 have relatively smaller estimated absolute values and larger standard deviations with intercepts ($\beta_0$). All the parameters discussed above ($\beta_0, \beta_1$ and $\beta_2$) are statistically significant under a 0.05 significant level.

The FDPM provides a fixed dispersion parameter ($\alpha$) of value 1.83. For proposed VDPMs, the estimated value of $\gamma$ characterizes the dispersion function. The standard deviation of $\gamma$ in M1, M2 and M3 are close to their mean estimates and are not statistically significant at 0.05 significant. This suggests that these models are not good choices in the application of varying
dispersion parameters, as their functional form cannot reflect the real behavior of the dispersion parameter. However, M4 and M5 provide standard deviations of $\gamma$ and p-values that are close to 0. Compared to the study of Geedipally and Lord (2008), single parameter dispersion function gives much reliable parameter estimations.

The models in Table 2 are listed by ascending AIC order values in the last column. The AIC value of the FDPM is slightly smaller than M1 to M5, where varying dispersion parameters are applied. However, the difference between those AIC values are trivial. The max discrepancy in the AIC values lies in between the FDPM and M5, where 997.24-992.28=4.96. Since this value is less than 5, it is concluded that the proposed models neither significantly differ from the FDPM nor differ from each other (Shahdah, Saccomanno, & Persaud, 2015).

Figure 1 displays the scatter plot of the crash estimated with the FDPM (horizontal axis) vs. with five proposed models (vertical axis) in period 1. For each subplot, the level of data point’s concentration along the diagonal indicates the level of similarity between associated model predictions of crash frequency. Starting from top left to the bottom right, data points in those subplots clearly show a tendency of scattering out of the diagonal. This conforms to the results of statistical fits where AIC increases monotonically from M1 to M5. In detail, M1 and M2 are more similar to the FDPM with most of points lies on the diagonal, M3 presents mild difference while M4 and M5 show the greatest difference from the FDPM. The scatter plot of period 2 is shown in Figure 2, which displays a similar scattering tendency as Figure 1 does. Comparing crash frequency predicted in two study periods for each model, it appears that the number of high crash frequency sites increases for all models, which might attribute to the effect of outliers.

5.2 Results of Hotspot Identification Tests
Followed the numerical tests documented in the previous subsection, the performance of EB method with fixed and proposed varying dispersion parameters in HSID were analyzed in this part. The EB method with fixed dispersion parameter (EB-FDP) does not outperform any EB method with varying dispersion parameter (EB-VDP) in the hotspot identification tests as it did in the model fitting. Test scores under different thresholds are presented in the following tables with color-coded heat maps where dark color indicates a better model performance. EB-M1 to EB-M5 refers to EB with VDP Model 1 to EB with VDP Model 5. Overall, EB-M4 and EB-M5 outperform other models under all tests with threshold 2.5%, 5%, 7.5% and 10% as shown in Table 3, Table 4 and Table 5.

Table 3 displays the scores of the estimation consistency test (T1). A higher test score indicates better consistency in model estimations. In identifying top 2.5% and 5% (about 25 and 50 sites) hazardous sites, the results show that EB-M4 and EB-M5 perform equivalently best by providing the highest cumulative number of accidents, following closely by EB-M3. EB-FDP, EB-M1 and EB-M2 report the lowest cumulative crash frequency. In examining the top 7.5% (about 75 sites) hazardous sites, EB-M3 scores 86, which is the lowest among all models whereas the highest scores are obtained from EB-M4 and EB-M5. When it comes to the top 10% crash hotspots (approximately 100 sites), EB-M3 has the highest score of 111, following by EB-M4 and EB-M5. It turns out that, EB-M4 and EB-M5 are the most stable models that provide consistent outputs of crash estimations in period 2, especially in cases for identifying top few percentages.

The model consistency test (T2) reflects the number of sites that are classified as hazardous sites in both periods. Similarly, a higher test score indicates better model performance in rank consistency. Judging from the scores displayed in Table 4, the outperformance of EB-M4 and EB-
M5 are obvious. Particularly, EB-M4 and EB-M5 are able to screen 3, 3, 6 and 10 more sites consistently across the two study periods than EB-FDP and EB-M1, which perform poorly in T2. EB-M2 and EB-M3 show slightly better performance than EB-FDP and EB-M1, as can be seen from the different shades of colors in the heat map. It is worth mentioning that even though T2 and T1 provide similar results that EB-M4 and EB-M5 are the best choices in hotspot identification, T2 is much clearer than T1 as the former provides much bigger inter-model discrepancies in the test scores.

Likewise, the rank difference test (T3) endorses that EB-M4 and EB-M5 are remarkable. It should be noted that a lower score in T3 shows smaller rank differences between crash estimations in two periods, and thus indicates a better model performance. For example, when we consider 48 intersections (ε = 5%) with the highest crash estimations in period 1, the cumulative rank differences of these intersections between two periods is 3,357 with EB-FDP and 2,745 with EB-M4 (see Table 5). As we are looking for identifiers with small estimation deviation between study periods, EB-M4 beats the EB-FDP with a much smaller score. The heat map of T3 displays strong polarization where EB-M4 and EB-M5 are colored with dark orange but other alternatives are very light, indicating large inter-model discrepancies of T3 scores compared to T1 and T2. This is because T3 involves more details (both rank and crash estimation) than T1 (period 2 crash estimation only) and T2 (rank differences between two study periods only), it is able to provide model selection in a much smaller scale and the results are consistent within a certain range of sites required (ε ∈ [2.5%, 10%] in this case).

Combining the results of model statistical fits and model performance in HSID, it is concluded that if a model performs better in statistical fits, it does not necessarily performs better in HSID. For certain functional forms, better model statistical fits may lead to better performance
in HSID. For instance, EB-M1 has an AIC values of 994.57, which is higher than the model with fixed dispersion parameter, which has an AIC value of 992.28. The first two rows in Table 3 to 5 shows that EB with fixed dispersion parameter provides better performance in T1, T2 and T3 than EB with M1. However, some functional forms that provide better model statistical fits does not guarantee better performance in HSID (e.g., M4 and M5). The AIC value of M4 and M5 are 997.22 and 997.24, which is not as good as the fixed dispersion parameter model (FDPM). Yet their performance in HSID is superior to FDPM, which can be seen from the last two rows of Tables 3 to 5.

6. Conclusion and Discussion

The heterogeneous nature of traffic crash data defies the assumption that the dispersion parameter in the NB model is constant among all sites over study periods (Hauer, 2001). Several studies have confirmed this hypothesis and suggested that the varying dispersion parameter should be applied in the SPFs. The HSM has adopted the varying form of dispersion parameters in segment SPFs. In light of the recent findings, researchers explored different dispersion functions for different types of intersections such as urban 4-leg signalized, rural 4-leg signalized, rural 3-leg unsignalized etc., (Geedipally and Lord, 2008; Lord and Park, 2008b; Miaou and Lord, 2003; Mitra and Washington, 2007), yet their results diverged.

Stimulated by this issue, intersection SPFs with various structures of the dispersion functions are tested. This paper has tested many flow-only functional forms and selected 5 models with stable and statistically significant parameters. As statistical fits and performance tests in HSID
are both commonly used in evaluating SPFs of the intersections, this paper evaluate the proposed functional form of the dispersion parameter in both aspects. It is found that model statistical fits are sensitive to the functional form of the dispersion parameter. Only certain forms can improve model statistical fits such as those proposed in Miaou and Lord (2003) and Geedipally et al., (2008). The values of AICs in this paper have differences less than 5 for all models, indicating that the proposed models neither significantly differ from the FDPM nor differ from each other.

Since EB crash estimation is one of the most commonly used methods by highway agencies in practice for the identification of high crash risk sites (Montella, 2010), its performance with the varying form of the dispersion parameter worth to be explored. To yield a quantitative understanding of the varying dispersion parameter hypothesis, three numerical tests are conducted based on the rank of model estimates of crash mean.

Consistently, all tests have shown that EB-VDP with dispersion parameter positively related to major and minor AADT (M4 and M5) are more stable, consistent and robust than the EB-FDP in identifying the top 2.5%, 5%, 7.5% and 10% hazardous intersections. This indicates that the dispersion functions M4 and M5 are able to capture the extra variations of the intersection crash data. This observation suggests that models with proper forms of dispersion functions are more stable in model estimates due to the emphasis on the considered covariates and the exclusion of other explanatory variables that may related to each other, which better justifies the basic assumptions of the EB method. Moreover, When the number of explanatory variables are limited, functional forms with dispersion parameters positively related to major and minor AADT such as M4 and M5 are highly recommended for practitioners. When there are multiple variables, a varying form of the dispersion parameter is still recommended, but the proper form need to be cautiously selected.
Consequently, three conclusions are reached for this paper:

- Both model statistical fits and model performance in HSID are sensitive to the functional form of the varying dispersion parameter.

- Model statistical fits and model performance in HSID should be used as different model evaluation and selection standard in crash analysis since better model statistical fits does not guarantee better performance in HSID.

- The application of proper functional form of the dispersion parameter is highly recommended for HSID as it will significantly improve prediction stability and consistency.

Eventually, this paper shows a potential for future exploration of performances of dispersion functions with multiple parameters in hotspot identification and on other intersection types.


Table 1. Data Summary

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major AADT</td>
<td>3,312</td>
<td>53</td>
<td>27,774</td>
<td>3,430.72</td>
</tr>
<tr>
<td>Minor AADT</td>
<td>206.5</td>
<td>10</td>
<td>11,409</td>
<td>449.59</td>
</tr>
<tr>
<td>Annual Crash Count</td>
<td>0.23</td>
<td>0</td>
<td>8</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 2. Summary of Model Fit Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Intercept/(\beta_0)</th>
<th>Parameter of (\log(\text{MajorADT})/\beta_1)</th>
<th>Parameter of (\log(\text{MinorADT})/\beta_2)</th>
<th>Parameter of dispersion function/(\gamma)</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDPM</td>
<td>-9.2004 (0.776)</td>
<td>0.7748 (0.090)</td>
<td>0.2779 (0.080)</td>
<td>1.8263 (&lt;0.001)&lt;sup&gt;1&lt;/sup&gt;</td>
<td>992.28</td>
</tr>
<tr>
<td>M1</td>
<td>-9.2165 (0.786)</td>
<td>0.7762 (0.090)</td>
<td>0.2786 (0.081)</td>
<td>8.1488 (6.506)&lt;sup&gt;2&lt;/sup&gt;</td>
<td>994.57</td>
</tr>
<tr>
<td>M2</td>
<td>-9.2048 (0.783)</td>
<td>0.7745 (0.090)</td>
<td>0.2791 (0.081)</td>
<td>5.2856 (4.059)&lt;sup&gt;2&lt;/sup&gt;</td>
<td>994.35</td>
</tr>
<tr>
<td>M3</td>
<td>-9.2067 (0.794)</td>
<td>0.7761 (0.090)</td>
<td>0.2769 (0.083)</td>
<td>22.2026 (20.565)&lt;sup&gt;2&lt;/sup&gt;</td>
<td>995.03</td>
</tr>
<tr>
<td>M4</td>
<td>-9.1899 (0.827)</td>
<td>0.7745 (0.094)</td>
<td>0.2762 (0.087)</td>
<td>-216.4190 (&lt;0.001)</td>
<td>997.22</td>
</tr>
<tr>
<td>M5</td>
<td>-9.1900 (0.827)</td>
<td>0.7745 (0.094)</td>
<td>0.2761 (0.087)</td>
<td>-347.3837 (&lt;0.001)</td>
<td>997.24</td>
</tr>
</tbody>
</table>

Note: Number in parentheses indicates standard deviation. <sup>1</sup>The fixed dispersion parameter of the traditional EB method. <sup>2</sup>Not significant at the 95% level.

Table 3. Accumulated Results of Estimation Consistency Test (T1) *

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Top (\epsilon) Hazardous Sites of Total 994 sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\epsilon = 2.5%)</td>
</tr>
<tr>
<td>EB-FDP</td>
<td>49</td>
</tr>
<tr>
<td>EB-M1</td>
<td>49</td>
</tr>
<tr>
<td>EB-M2</td>
<td>49</td>
</tr>
<tr>
<td>EB-M3</td>
<td>49</td>
</tr>
<tr>
<td>EB-M4</td>
<td>54</td>
</tr>
<tr>
<td>EB-M5</td>
<td>54</td>
</tr>
</tbody>
</table>

Note: *Darker color indicates a better performance.
Table 4. Accumulated Results of Model Consistency Test (T2) *

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Top $\epsilon$ Hazardous Sites of Total 994 sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon = 2.5%$</td>
</tr>
<tr>
<td>EB-FDP</td>
<td>9</td>
</tr>
<tr>
<td>EB-M1</td>
<td>9</td>
</tr>
<tr>
<td>EB-M2</td>
<td>10</td>
</tr>
<tr>
<td>EB-M3</td>
<td>9</td>
</tr>
<tr>
<td>EB-M4</td>
<td>12</td>
</tr>
<tr>
<td>EB-M5</td>
<td>12</td>
</tr>
</tbody>
</table>

Note: *Darker color indicates a better performance.

Table 5. Accumulated Results of Rank Difference Test (T3) *

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Top $\epsilon$ Hazardous Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon = 2.5%$</td>
</tr>
<tr>
<td>EB-FDP</td>
<td>1,257</td>
</tr>
<tr>
<td>EB-M1</td>
<td>1,251</td>
</tr>
<tr>
<td>EB-M2</td>
<td>1,257</td>
</tr>
<tr>
<td>EB-M3</td>
<td>1,230</td>
</tr>
<tr>
<td>EB-M4</td>
<td>804</td>
</tr>
<tr>
<td>EB-M5</td>
<td>803</td>
</tr>
</tbody>
</table>

Note: *Darker color indicates a better performance.
Figure 1. EB crash estimation with FDPM (tradition) vs. with proposed VDPMs-period 1
Figure 2. EB crash estimation with FDPM (tradition) vs. with proposed VDPMs-period 2