COMPARING CRASH ESTIMATION TECHNIQUES FOR RANKING OF SITES IN A NETWORK SCREENING PROCESS

Lalita Thakali
iTSS Lab, University of Waterloo, Canada

Liping Fu
iTSS Lab, University of Waterloo, Canada
Southwest Jiaotong University Chengdu, P. R. China

Tao Chen
Department of Economics, University of Waterloo, Canada
Big Data Research Center, Jianghan University Wuhan, P. R. China

ABSTRACT

Network screening, a process for an effective and efficient management of road safety programs, relies on crash prediction techniques to quantify the relative risks of given sites. The two most commonly used statistical approaches are- cross-sectional model-based approach and Empirical Bayesian (EB) approach as they are known for reducing regression-to-mean bias problem of a simple crash history-based method. Meanwhile, relatively the EB approach is known to be a robust technique as it accounts for a site-specific risk level while still incorporating the risk estimates obtained from a cross-sectional model. Common to both the approaches is they are relatively convenient to apply and easy to interpret due to a defined mathematical equation used to relate crashes and the potential explanatory variables. However, pre-specification of such relations are challenging as the true cause-effects are not known. One approach is to use trial and error process to select for the final relation. Nonetheless, potential misspecification may still remain which consequently could result in an inaccurate list of crash hotspots in a network screening process. As an alternative to model-based approach, this study applies kernel regression (KR), which is a data-driven nonparametric method. In addition, the KR method is extended in a similar framework of EB approach to account for site-specific risk levels. All these techniques when applied in a case study using crash data from Highway 401 of Ontario, Canada showed that some deviations exist between the methods, particularly when applied in the ranking of sites in a network screening process.

Keywords: Network screening, Ranking, Empirical Bayesian method, Kernel regression, Negative binomial model

1. INTRODUCTION

Network screening is an essential component of any road safety improvement programs. By prioritizing the sites of interest for a given safety improvement program, network screening can ensure that the limited resources allocated for the safety program be utilized in a most cost effective way. Network screening often starts with estimation of risk levels at all sites of interest followed by their ranking based on some risk measures. The sites that are ranked at the top most are then identified as crash hotspots for further consideration of safety treatments. A variety of risk measures such as average crash rate (crash per vehicle-kilometers), average crash frequency (crash per km-year), weighted crash frequency based on crash severities can be used as the ranking criteria for hotspot identification, depending the interest and priority set by the stakeholders or road agencies (Laughland et al., 1975; Deacon et al., 1975; Mcguigan, 1981; Mcguigan, 1982; Stokes and Mutabazi, 1996; HSM, 2010). While the choice of these ranking criteria lies in the interest and priority factors set by road agencies, selecting a robust technique to estimate risk measures is critical to make sure that genuine crash hotspots are identified during the process of network screening.
Several statistical approaches can be employed to predict crash risk at individual sites. Among them, the most widely adopted one in the context of both research and state-of-practice is the parametric approach. This includes a number of count models that are framed either in generalized linear, Empirical Bayesian or full Bayesian modeling approaches (Greibe, 2003; Saccamanno et al., 2004; Miranda-Moreno, 2005; Geedipally and Lord, 2010; Higle & Witkowski 1988; Hauer, 1996; Montella, 2010; Persaud et al., 1999; HSM, 2010; Miaou and Song, 2005; Miranda-Moreno et al., 2005; Miranda-Moreno et al., 2007; Huang et al., 2009; Miranda-Moreno et al., 2013; Wang et al., 2014). The fundamental concept to these model-based techniques is that the risk measures are estimated from a given crash data through a process of defining a relation of crashes and the factors imposing the crash risk. However, establishing such relations prior to determining the magnitude of model parameters is a challenging task. Failure to capture such true relation may result in a biased estimate of ranking criteria, and consequently, these affect the results of network screening.

Another alternative approach to determining ranking criteria could be using a data-driven nonparametric technique which has rarely been explored for the purpose of network screening. The fact that it is specification free may provide a significant advantage for improving the accuracy of crash frequency (Persaud et al., 1999). One of the examples could be using kernel regression which provides an estimate of conditional average crash frequency without the need for prior specification of a relation of crashes and its influencing factors. As the implementation of nonparametric techniques is relatively new for this particular road safety application, it may be important to examine how these two approaches differ, particularly in the ranking of sites.

The objective of this study is to investigate the practical implications of alternative crash prediction techniques in terms of network screening. Two parallel methods are compared- 1) NB model with KR method and 2) traditional NB based EB method with KR-based EB method. The latter method is introduced in this study to account for site-specific crash history. Crash data from Highway 401 of Ontario, Canada is used as a case study. The paper hereafter is arranged as follows. Section 2 discusses some of the common estimation techniques used in network screening. Section 3 presents estimation techniques employed for this study and followed by a case study in Section 4. Finally, the conclusions are made in Section 5.

2. LITERATURE REVIEW

In the past, when the practice of statistical techniques was not well matured, road agencies used simply observed crash frequency (or rate) as ranking criteria in network screening process. However, these conventional approaches do not account for uncertainty in crash occurrence and thus suffers from the regression-to-mean effect. In a network screening study by Cheng and Washington (2005), they evaluated the performance of conventional approach of using simple crash count with the EB approach. For the conventional approach, two ranking criteria were used. First was based on the observed crash frequencies where a set of sites were ranked in descending order and the top most sites were selected as hotspots, and the second criterion was by establishing a threshold value and comparing it with the observed crash counts. In the latter, the threshold value was calculated as a summation of average observed crashes and the confidence interval. When the observed crash exceeded threshold value then the sites were classified as hotspots. Similarly, for the EB approach, ranking criterion was the expected crash frequency obtained from the combination of model estimates (NB model) and observed crashes. The study showed that the EB approach significantly outperformed the conventional method. The study further concluded that the importance of an EB approach is especially critical when there is high heterogeneity in crash counts. Similar conclusions were made in a study by Elvik (2008).

Recently, the cross-sectional model-based and EB approach have been the most extensively used approach for network screening. The commonly used ranking criteria i.e., expected crash frequency (or rate) can be obtained using cross-sectional crash models such as Poisson and NB model in a generalized linear modeling approach (Saccamanno et al., 2001; Greibe, 2003; Saccamanno et al., 2004; Miranda-Moreno, 2005; Geedipally and Lord, 2010). Meanwhile, these criteria can also be obtained from the EB approach where NB model is extended in a framework of Bayesian approach (Higle & Witkowski 1988; Hauer, 1996; Montella, 2010; Persaud et al., 1999; HSM, 2010). Mathematically, the EB estimates are the combination of model estimates and observed crashes. While implementing such criterion, one has to be cautious, especially when the site-specific historical crash data are not available (HSM, 2010). Another advanced form of parametric models are using Full Bayesian approach (Miaou and Song, 2005; Miranda-Moreno et al., 2005; Miranda-Moreno et al., 2007; Huang et al., 2009; Miranda-Moreno et al., 2013; Wang et al., 2014). When the dataset and sample mean is large, the use of full Bayesian approach do not significantly contribute to improving the estimation
results compared to generalized linear modeling approach (Lord & Miranda-Moreno, 2008; Miranda-Moreno et al., 2013).

Some studies have compared network screening across different estimation techniques that are used to determine the ranking criteria. For example, Saccamanno et al. (2001) applied Poisson model (cross-sectional model-based approach) and EB method for the identification of hotspots in a two-lane highway in Italy using crash frequency as the ranking criteria. The result concluded that the numbers of hotspots identified by the EB approach were less than that from the Poisson model. Furthermore, the authors mentioned that the results from the Poisson model may have biased estimation due to its inability to account over-dispersion nature of crash data. Comparatively, the advantages of EB method is- first, it uses a NB model which takes into account of over-dispersion in the data structure unlike in a Poisson model; second, the precision of prediction is improved by considering site-specific crash history under a Bayesian framework. Similarly, Saccamanno et al (2004) applied Poisson and NB models for ranking of highway-rail grade crossing and concluded that comparatively the NB model performs better. The result could be attributed due NB to account dispersion in crash data. In another study by Miranda-Moreno et al (2005), a significant difference was observed between EB and generalized linear models in ranking of highway-railway grade crossings, thus emphasizing on the importance of method selection. Similarly, Huang et al (2009) compared EB and full Bayesian approach using NB and Poisson-lognormal model structure to identify crash hotspots of signalized intersection. First, the sites were ranked based on mean crash frequency estimated from the individual methods using three years (2004-2009) of crash data. Among them, a certain number of sites, e.g., 5%, 10% of total sites were considered as hotspots. These hotspots were then compared with the “true” hotspots obtained from crash counts using 10 years (1997-2006) repressing a long-term average. The study concluded that the full Bayesian approach showed better performance in identifying the crash hotspots.

An alternative approach for improving the modeling part, whether when used independently in a cross-sectional model-based approach or in an EB framework, could be using nonparametric approach. While the application of this approach is relatively new in road safety, past few studies have demonstrated evidence of some promising results by comparing with the NB model. The methods used were- regression tree (Karlaftis and Golias, 2002), artificial neural network (Xie et al., 2007; Chang, 2005), multivariate adaptive regression splines (MARS) (Abdel-Aty & Haleem, 2011), support vector machine (Li et al., 2008), kernel regression (Thakali et al., 2013) and others. However, these studies are limited to only exploring the effects of causal factors on crashes and their comparisons to the parametric counterpart models.

3. PROPOSED CRASH ESTIMATION TECHNIQUES

This section provides a theoretical background on various crash prediction techniques used in network screening.

3.1 Cross-sectional Modeling Approach

Negative binomial (NB) model: Negative binomial model is one of the most popular count models used in network screening. It is also known as Poisson-gamma model as it is derived from the Poisson model with inclusion of a gamma-distributed error term thereby releasing the restriction of equal mean and variance condition of Poisson model (Maher and Summersgill, 1996; Lord and Manering, 2010; Cameron and Trivedi 2013). Let \( Y \) be a number of crashes occurring at a certain site for a specified time period (here year). Assuming \( Y \) follows a Poisson distribution, i.e., \( Y|\theta \sim \text{Poisson}(\theta) \), where \( \theta = \mu^{NB} \cdot e^\epsilon \) with \( e^\epsilon \) – a random term being included for modelling additional variation in the count data. If \( e^\epsilon \) is assumed to follow a Gamma distribution with both of its parameters equal and greater than zero (i.e., \( (\alpha, \alpha) \), where \( \alpha \) is known as a dispersion parameter), the resulting \( Y \) would follow a NB distribution. In road safety analysis, it is customary that the conditional expected crash frequency \( (\mu^{NB}) \) is assumed to be a function of site attributes in a form similar to the one shown in Eq. 1. The dispersion parameter could also be modelled as a function of site attributes, as shown in Eq. 2. Note that the model coefficients are estimated using crash data in a framework of generalized linear modeling approach with an objective to maximize the likelihood of crash occurrence.

\[
\begin{align*}
\mu^{NB} &= (exposure)^{\beta_0 e^{(\beta_0 + \sum \beta_i x_i)}} \\
\alpha &= e^{\theta_0 + \sum \gamma_i x_i}
\end{align*}
\]
where, in Eq. 1 - $\mu^{NB}$ is expected crash frequency, *exposure* is defined as the product of annual traffic volume and the road section length, $\beta_1$ is exponent of the exposure, $\beta_0$ is intercept, $\beta_i$ is coefficient of explanatory variable $x_i$ in the NB model and $x_i$ is $i$th explanatory variable. Similarly, in Eq. 2 - $\gamma_0$ is intercept, $\gamma_1$ is coefficient of explanatory variables in the dispersion-parameter model.

**Kernel regression (KR):** Kernel regression, a data-driven nonparametric technique, relaxes the exponential restriction of the conditional mean by employing the kernel density estimation process (Pagan and Ullah, 1999; Silverman, 1984). Following the same notations as the NB model, we represent KR method using Eq. 3 and Eq. 4:

$$Y := m(X) + \varepsilon$$

where, the identification is $E[\varepsilon|X] = 0$. It is important to note that there is no functional assumption imposed on $m(\cdot)$, which is interpreted as the conditional mean of $Y$ given $X$. According to Pagan and Ullah (1999), for an $x$ in the support of $X$ and a random sample $(X_i, Y_i)$ for $i = 1, ..., n$,

$$\mu^{KR} = \hat{m}(x) := \sum_{j=1}^{n} \omega_j(x, X_1, ..., X_n)Y_j$$

where,

- $\hat{m}(x)$ (or $\mu^{KR}$) is called the kernel nonparametric estimator of $m(x)$,
- $\omega_j(x, X_1, ..., X_n) = K \left( \frac{X_j - x}{b} \right) / \sum_{j=1}^{n} K \left( \frac{X_j - x}{b} \right)$,
- $K(U) := \Pi_{d=1}^{D} k_d(U^{(d)})$ and each $k_d$ called a kernel or kernel function, satisfying: $k_d(u) > 0$, $k_d(u) = k_d(-u)$, $\int k_d(u) du = 1$, $\int u^2 k_d(u) du < \infty$. $K$ is called the product kernel, $D$ is the number of input variables.
- $b$ is called the bandwidth, which is a $D \times 1$ vector with each coordinate be a sequence of positive numbers such that $b^d \downarrow 0$ (goes down to 0 monotonically) and $nb^d \to \infty$ for all $d = 1, ..., D$.

It is easy to show that $\omega_j > 0$ and $\sum_{j=1}^{n} \omega_j = 1$. Therefore, the estimated number of crashes given $X = x$ could be viewed as a specific weighted average of $Y_j$’s. Such weights are determined jointly by the kernel and bandwidth. The most common choice of kernel is the density function of a standard normal random variable. In theory, the choice of the kernel makes little difference when the sample large and the conditions listed above are satisfied. For fixed $b_n$, larger weights are given to the observations closer to $x$. Therefore, the kernel regression is a local fitting technique as opposed to a parametric approach that chooses one curve of a certain shape to fit all the data points. By down-weighing the observations that are further apart, kernel nonparametric estimator uses more relevant information for estimation, hence, it could capture variations that are overlooked by parametric models. Apart from this Kernel regression technique, there are other nonparametric methods available, such as Spline and Orthogonal polynomial. However, within this family of nonparametric models, KR is one of the well-established and widely applied techniques (Livanis et al., 2009). It has also been argued that all these methods are in an asymptotic sense essentially equivalent to kernel regression (Hardle and Mammen 1993; Silverman, 1984).

The critical component of kernel nonparametric estimator is the bandwidth, which determines the “distance” between $X_j$’s and $x$. Though the kernel nonparametric regression estimator is free from misspecification, i.e., it converges to the truth when sample size approaches infinity, it is biased for a finite sample. A smaller bandwidth reduces the bias but inflates the variance, while a bigger bandwidth reduces variance at the cost of bigger bias. This natural trade-off between the bias and variance helps to pin down the desirable bandwidth that minimizes the mean squared error of the estimator. A detailed discussion of possible choices of bandwidth could be found in Pagan and Ullah (1999). In this paper, we adopt a variation of the Silverman’s rule of thumb. Similar rule of thumb is also applied in past studies (Lavergne and Vuong, 2000; Gu et al., 2007; Dudek, 2012). The bandwidth for the $j^{th}$ variable is calculated as:

$$b_d = \left( \frac{4}{D+1} \right)^{\frac{1}{4D}} \ast n^{-\frac{1}{4+D}} \ast \sigma_d$$

where $\sigma_d$ is the standard deviation of the corresponding $d^{th}$ variable, D is the total number of variables.
3.2 Empirical Bayesian (EB) Approach

Empirical Bayesian (EB) is another extensively used approach in a network screening. Intuitively, including a site-specific crash history with expected crashes from similar sites can provide a better representation of expected risk levels at the selected sites. An effort to combine this information is popular in an EB framework using Eq. 6, where weights are calculated using Eq. 7 and Eq. 8. Here, we followed same notations as used by Hauer (1997).

\[ E(k/K) = wE(k) + (1 - w)K \]  
\[ w = \frac{1}{1 + \frac{\text{VAR}(k)}{E(k)}} \]  
\[ \text{Var}(k) = SD - E(k) \]

where, \( E(k/K) \) = Expected crashes on a given site (e.g., a road segment, or an intersection) given that K crashes occurred, \( E(k) \) = Expected crashes (or mean) referenced from similar sites (e.g., estimated from a crash model), \( K = \) Observed crashes on that given site, \( w = \) weight; \( \text{Var}(k) \) is variance of expected crashes from similar sites, \( SD \) is defined as square difference between the observed crash counts and crash estimates from a model.

As given by Eq. 6 to Eq. 8, an EB framework requires estimates of \( E(k) \) and the weight. For this, Hauer (1997) proposed two methods- 1) method of sample moment and 2) method of multivariate regression. In the first method, a simple sample mean is used to estimate the value of \( E(k) \) (also referred here as mean) and \( \text{Var}(k) \) (also referred here as variance) which together are used to determine the weight (Eq. 7). While this method is simple to apply with very few assumptions made, however, it does not, account for possible effects of safety related factors that could explain variation in crashes between sites. This issue is addressed by using the second method of regression approach. This approach takes into account of effects due to site-specific factors by developing their relation with crash frequency. In the past, use of multivariate regressions from a family of parametric models has been the common tradition. As in this study, we used two different approaches to estimate the crash risk, we name respective EB method after their name, either NB-based EB method or KR-based EB method. Note that the use term EB in association to KR may be argued due to the fact that the derivation does not come purely from the Bayesian analysis. Rather it is explained from an approach of combining two random variables and Hauer (1997) showed that both the approaches result in the same form.

**NB-based EB method:** The EB method can be categorized depending on the types of count models used in its formulation. Some of the examples include- EB based on NB model (Hauer., 1997; Miranda-Moreno et al., 2005; Montella, 2010, HSM, 2010), EB based on Sichel model (Zou et al., 2013), EB based on Poisson-lognormal model (Miranda-Moreno et al., 2005). Among these, the first one, based on the NB model, is the most commonly used method. In context of NB-based EB method, we make following adjustments:

- **E(k):** Replace the estimate as- \( E(k) = \mu^{NB} \)
- **Weights (w):** Hauer & Persaud (1987), in a study using NB based EB method, observed a systematic relation of mean (i.e., \( E(k) \)) and variance (i.e., \( \text{Var}(k) \)). A quadratic function, given by Eq. 9, was used to establish the relation. After substituting the values of mean and variance in Eq. 7, the expression obtained for weight is given by Eq. 10. Since then, in NB based EB method, it has been a standard procedure to apply proposed relation of mean-variance to the weights (Persaud et al., 1999; HSM, 2010).

\[ \text{Var}(k) = [E(k)]^2 \times \alpha; \text{where, } \alpha \text{ is a dispersion parameter} \]

\[ w = \frac{1}{1 + \frac{[E(k)]^2 \times \alpha}{E(k)}} = \frac{1}{1 + E(k) \times \alpha} = \frac{1}{1 + \mu^{NB} \times \alpha} \]

Finally, Eq. 6 is used to obtain the final NB based EB estimates by substituting the values of \( E(k) \) and \( w \).
**KR-based EB method**: The commonly used methods to obtain an estimate of E(k) in an EB framework are based on parametric approach. In context of KR-based EB method, the following method is proposed:

- **E(k)**: Replace the estimates as- \(E(k) = \mu^KR\) (ref to Eq. 4)
- **Weights (w)**: The steps involved in determining the weights (w) are not as straight forward as in previously discussed NB-based EB method. For this particular method, we trace back to its initial form in Eq. 7, where it is represented as a function of variance and mean of crash estimates from similar sites. Following three steps are needed to determine the weights:
  1. Estimate the variance associated with each site using Eq. 8.
  2. Use kernel regression to establish a relation of mean and variance. The detailed process are described in the section Cross-sectional modeling approach- KR method. Note that, establishing a mean-variance relation here is a univariate case.
  3. Use Eq. 7 to calculate weights associated with each site.

Finally, Eq. 6 is used to obtain the final KR-based EB estimates by substituting the values of E(k) and w.

**4. APPLICATION OF CRASH ESTIMATION TECHNIQUES IN NETWORK SCREENING – A CASE STUDY**

This paper presents a case study using crash data from one of the busiest highways in North America - Highway 401, located in the provinces of Ontario, Canada. The entire highway stretches for 817 km of which a total of 800 km were covered in this study. Three different sources of data used are - 1) historical crash records from 2000 to 2008 extracted from MTO’s Accident Information System (AIS); 2) historical AADT data for the same years from MTO’s Traffic Volume Inventory System (TVIS); and 3) road geometric features from MTO’s Highway Inventory Management System (HIMS) database. A Geographical Information System (GIS) tool was used to process and integrate all the datasets on an annual basis over individual homogenous sections (HSs) which represented a segment with similar road features and traffic levels, resulting in 418 unique HS sections. Table 1 presents a variable list including summary statistics of processed datasets from nine years (2000-2008). Further details about the data processing are described in our previous paper (Thakali et al., 2016).

<table>
<thead>
<tr>
<th></th>
<th>Total Crash (per year)</th>
<th>AADT (veh/hr)</th>
<th>Exposure (MVK)</th>
<th>Commercial AADT (veh/hr)</th>
<th>Median width (m)</th>
<th>Shoulder width- right (m)</th>
<th>Curve deflection (1/m, in 1000)</th>
<th>Shoulder width- left (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>23.81</td>
<td>76633</td>
<td>41.79</td>
<td>13993</td>
<td>11.11</td>
<td>3.14</td>
<td>0.19</td>
<td>1.60</td>
</tr>
<tr>
<td>St.dev.</td>
<td>50.02</td>
<td>91476</td>
<td>54.05</td>
<td>6719</td>
<td>6.14</td>
<td>0.28</td>
<td>0.35</td>
<td>1.19</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>12000</td>
<td>1.66</td>
<td>0</td>
<td>0.60</td>
<td>2.60</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Max</td>
<td>468</td>
<td>442900</td>
<td>611.41</td>
<td>42076</td>
<td>30.50</td>
<td>4.00</td>
<td>1.86</td>
<td>5.19</td>
</tr>
</tbody>
</table>

*Sample size: 3762*

**Network screening process**- It is a systematic process of ranking sites that suffer from unacceptably high risk of crashes. Four steps are involved: 1) selection of ranking criteria; 2) selection of a method to estimate expected crash frequencies; 3) ranking of sites, and 4) selection of high-risk sites (or crash hotspots). Two ranking criteria are used for this study, including crash frequency/km and crash rate (crash frequency/exposure) as determined by normalizing the estimated crash frequency by length and exposure, respectively.

**Estimation methods and their relative performance**: The expected crash frequencies required for ranking criteria are obtained using the methods discussed previously. Prior to their applications for network screening, we present a summary of the modeling results including the performance measures. For this, the dataset mentioned in Section 4 was divided into a training set (2000-2006) and validation set (2007-2008). Table 2 provides the model results using the training set. Note that the variables included in a NB model are significance at a level of significance of 5%. While, in KR method, a similar statistical test is not possible due to non-parameterization. However, we used an
indicator called Variable Importance (VI) measure which was determined based on effect of each variable on improving goodness-of-fit values of the KR model. The result showed that all the variables has high VI values suggesting for their inclusion. The performance of each method was determined using goodness-of-fit measures- sum of absolute deviation (SAD) and sum of square deviation (SSD) computed on the validation set. Note that the lower these values, the higher the performance level of given crash estimation technique. The results in Table 3 indicate that the KR method (in both approaches) performed relatively better in comparison to the NB model. All analyses were performed using the statistical software platform- “R”.

Table 2: Summary of model results

<table>
<thead>
<tr>
<th>Variables</th>
<th>NB model</th>
<th>CR</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficients</td>
<td>Std. error</td>
<td>t value</td>
<td>P value</td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.16</td>
<td>0.15</td>
<td>-7.528</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ln(Exposure) (MVK)</td>
<td>0.84</td>
<td>0.02</td>
<td>50.75</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>AADT (Commercial)</td>
<td>5E-05</td>
<td>0.00</td>
<td>18.69</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Median width (m)</td>
<td>-0.01</td>
<td>0.00</td>
<td>-5.23</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Median shoulder (m)</td>
<td>-0.10</td>
<td>0.01</td>
<td>-9.14</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Shoulder width (m)</td>
<td>0.16</td>
<td>0.04</td>
<td>3.51</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Deflection (1/m)</td>
<td>0.001</td>
<td>-0.09</td>
<td>0.05</td>
<td>-1.69</td>
</tr>
</tbody>
</table>

Dispersion parameter (\(\alpha\))

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Std. error</th>
<th>t value</th>
<th>P value</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.51</td>
<td>0.03</td>
<td>-15.10</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Length (km)</td>
<td>-0.83</td>
<td>0.04</td>
<td>-20.11</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>19303</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is noted that a mean-variance relation was established for the KR-based EB method as shown in Figure 1. This relation was directly used to compute the weights associated with each site to obtain final expected crash frequencies in an EB framework. However, the weights for NB-based EB method were directly determined using dispersion parameter model presented in Table 1.

Figure 1: mean and variance relation from KR method
Table 3: Summary of performance measures

<table>
<thead>
<tr>
<th>Methods</th>
<th>SAD</th>
<th>SSD</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cross-sectional modeling approach</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KR</td>
<td>6822</td>
<td>425475</td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>9291</td>
<td>1026236</td>
<td></td>
</tr>
<tr>
<td><strong>Empirical Bayesian (EB) approach</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KR based EB</td>
<td>4504</td>
<td>190392</td>
<td></td>
</tr>
<tr>
<td>NB based EB</td>
<td>4793</td>
<td>261577</td>
<td></td>
</tr>
</tbody>
</table>

For EB approach

1. Obtain EB estimates for the last two years of model set i.e., 2005-2006.
2. Adjust estimated values in Step 1 by using a factor $= \frac{\mu_{2007-2008}}{\mu_{2005-2006}}$ to obtain estimates for validation.
3. Obtain SAD and SSD by using estimated values in Step 2 and observed crashes in the validation set (2007-2008).

Note: \( SAD = \sum_{i=1}^{n} |\hat{y}_i - y_i| \) and \( SSD = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \) where, \( y_i \) is \( i^{th} \) observed crashes; \( \hat{y}_i \) is estimated crashes of the \( i^{th} \) observation and \( n \) is total number of observations. A method with smaller SAD and higher SSD implies that the quality of \( \hat{y}_i \) is quite uneven.

**Ranking of sites:** All the methods previously described were used to estimate the crash frequency for each road section (418 unique HS sections) using the validation set. Then, for crash frequency criterion, the estimated crash values were normalized by section length, whereas for rate criterion they were normalized by exposure level measured in millions vehicle kilometer. Finally, road sections were ranked in descending order of each criterion. Figure 2 and 3 present scatter plots of one to one ranking of road sections obtained from the selected methods for each criterion. We can observe some deviations in ranking as some of the sites are found significantly off the diagonal line. Spearman’s correlation (SC) coefficients are used to determine the correlation between the two methods, where large SC values represent high correlation and vice versa for the low values. As summarized in Table 4, KR and NB model in a cross-sectional modeling approach showed relatively lower correlation compared to those when they were used in an EB approach. One of the reasons of their high correlation when applied in the EB framework could be due to the involvement of site-specific crash history in both the methods. Meanwhile, relatively, ranking correlations are higher with frequency criterion when compared to the rate criterion. The result from example suggests that the choice of criteria for ranking may have a significant impact on the relative performance of each method.

Figure 2: Ranking comparison based on cross-sectional modeling approach- (a) crash frequency/km and (b) crash frequency/million vehicle km
5. CONCLUSIONS

The main purpose of this study was to introduce kernel regression (KR), a data-driven approach, for network screening and examine its performance in comparison to negative binomial (NB) model, one of the most extensively used parametric model in road safety literature. These two methods were applied in two different types of crash prediction methods - cross-sectional modeling and Empirical Bayesian (EB) approach. One of the appealing parts of the KR method is, unlike in the parametric models, there is no need of prior specification for defining a relation of crashes and its influencing risk factors. Therefore, such method is expected to capture any underlying complex relations for improved crash estimations. An empirical study of Highway 401 of Ontario was considered to examine the differences in implications of proposed estimation techniques in ranking of sites in a network screening process. Summary of findings are as follows:

- There exist differences among the methods in their ability to predict crashes where KR method outperformed the NB model in both the approaches.
- Deviations in estimation methods were visually observed in one to one ranking of sites (road sections). This was further confirmed by computing Spearman’s correlation coefficients (SC) for each pair. Meanwhile, the SC’s varied depending on types of ranking criteria. The findings from this study showed that the crash rate criterion results higher deviation among the methods than the crash frequency criterion.

While this study has demonstrated a potential use of kernel regression as an alternative to parametric models in network screening, more case studies may be required to confirm their relative performance. Moreover, emphasis should be given to inclusion of more samples as the performance of kernel regression is expected to increase significantly as demonstrated by Thakali et al. (2016). Meanwhile, the use of the EB approach must be considered with care, especially in a case of major changes in the road features and missing of site-specific crash data (HSM, 2010).
ACKNOWLEDGEMENTS

This research was jointly supported by Ontario Ministry of Transportation (MTO) and National Science and Engineering Research Council of Canada (NSERC).

REFERENCES


