Modeling work zone crash frequency by quantifying measurement errors in work zone length

Hong Yang, Kaan Ozbay, Ozgur Ozturk, Mehmet Yildirimoglu

A R T I C L E   I N F O

Article history:
Received 31 May 2012
Received in revised form
27 December 2012
Accepted 15 February 2013

Keywords:
Work zone
Crash frequency
Safety analysis
Measurement error
Negative binomial model
Full Bayesian

A B S T R A C T

Work zones are temporary traffic control zones that can potentially cause safety problems. Maintaining safety, while implementing necessary changes on roadways, is an important challenge for traffic engineers and researchers to confront. In this study, the risk factors in work zone safety evaluation were identified through the estimation of a crash frequency (CF) model. Measurement errors in explanatory variables of a CF model can lead to unreliable estimates of certain parameters. Among these, work zone length raises a major concern in this analysis because it may change as the construction schedule progresses generally without being properly documented. This paper proposes an improved modeling and estimation approach that involves the use of a measurement error (ME) model integrated with the traditional negative binomial (NB) model. The proposed approach was compared with the traditional NB approach. Both models were estimated using a large dataset that consists of 60 work zones in New Jersey. Results showed that the proposed improved approach outperformed the traditional approach in terms of goodness-of-fit statistics. Moreover, it is shown that the use of the traditional NB approach in this context can lead to the overestimation of the effect of work zone length on the crash occurrence.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Much of the U.S. highway infrastructure is aging and the need for maintenance, rehabilitation, and upgrading of the existing networks increases. Consequently, road users are increasingly exposed to work zone activities. Nationally, about 23,745 miles of federal-aid roadway improvement projects were underway annually from 1997 to 2001 (FHWA, 2001). On average motorists encountered an active work zone for every 100 miles driven on the national highway system (Ullman et al., 2004b). The number can be much larger considering other work zones deployed on municipal, county, and state roads.

The presence of so many work zones directly affects the safety of road users and highway workers. According to the latest safety statistics, 667 work zone fatalities occurred in the U.S. in 2009. Approximately 85 percent of those killed in work zone were drivers or passengers and the remaining 15 percent were workers. In addition to these fatalities, more than 40,000 injuries resulted from motor vehicle crashes in work zones (FHWA, 2011). As shown by many studies (Graham et al., 1978; Paulsen et al., 1978; Garber and Woo, 1990; Casteel and Ullman, 1992; Pal and Sinha, 1996; Khattak et al., 1999; Venugopal and Tarko, 2000; Khattak et al., 2002; Qi et al., 2005; Harb et al., 2008a; Ullman et al., 2008), crash rates increase in the presence of work zones compared to the normal road conditions. This rise can be attributable to the complexity of the work zone circumstance that interrupts continuing traffic flow and creates many traffic conflicts. However, precise reasons why more crashes occur at work zones may still not be clear. A thorough understanding of the risk factors associated with work zone crash occurrence is essential for the development of effective countermeasures to reduce the number of fatalities and injuries, and to enhance traffic operation and safety within work zones. However, many site- and state-specific factors need to be further investigated to better understand the reasons for work zone crashes.

Therefore, the purpose of this paper is to develop a statistical model to identify the potential relationship between crash frequency and a set of explanatory variables at work zones in New Jersey. A relative large amount of detailed work zone project files and crash data obtained from the New Jersey Department of Transportation (NJDOT) provided us with opportunities to explore extra factors that have not been addressed before. Considering the presence of measurement errors in variables, special models that
extend the use of the traditional negative binomial (NB) model are developed and parameters are estimated through the full Bayesian approach. By addressing the potential errors in variables, the developed models provide more reliable understanding of the safety impact of different contributory factors on work zone crashes.

2. Literature review

A number of studies have conducted work zone safety analysis. However, majority of them focused on the development of the descriptive statistics of work zone crash data to interpret the characteristics such as crash experience, consequences, temporal, and spatial distributions of crashes at work zones (Nemeth and Migletz, 1978; Routhail et al., 1988; Hall and Lorence, 1989; Garber and Woo, 1990; Sorock et al., 1996; Lin et al., 1997; Bryden et al., 1998; Raub et al., 2001; Chamblee et al., 2002; Garber and Zhao, 2002; Schrock et al., 2004; Bushman et al., 2005; Müngen and Gürçanlı, 2005; Salem et al., 2006; Harb et al., 2008b; Jin et al., 2008; Li and Bai, 2008; Ullman et al., 2008; Dissanayake and Akepati, 2009; Akepati and Dissanayake, 2011). Generally, the literature indicated that the presence of work zones increases the likelihood of crash occurrence (Graham et al., 1978; Paulsen et al., 1978; Garber and Woo, 1990; Casteel and Ullman, 1992; Pal and Sinha, 1996; Khattak et al., 1999; Venugopala and Tarko, 2000; Khattak et al., 2002; Qi et al., 2005; Harb et al., 2008a; Ullman et al., 2008). In addition, crashes were found to disproportionately occur across different segments of work zones. For instance, the work zone activity area was the predominant location of crashes and rear-end collisions was the predominant type of crash (Nemeth and Migletz, 1978; Hargroves, 1981; Nemeth and Rath, 1983; Pigman and Agent, 1990; Garber and Zhao, 2002; Schrock et al., 2004; Srivivasan et al., 2008; Xing et al., 2010; Zhu et al., 2010). Comparisons between daytime and nighttime work zone crashes suggest no clear evidence that crash rate significantly increased at night (Casteel and Ullman, 1992; Daniel et al., 2000; Ullman et al., 2004a; Udoka, 2005; Ullman et al., 2006, 2008).

Although CF models have been extensively used in road safety analysis, only a few studies that have specifically focused on modeling work zone crash occurrence (Pal and Sinha, 1996; Khattak et al., 1999; Venugopala and Tarko, 2000; Khattak et al., 2002; Qi et al., 2005; Srivivasan et al., 2011). Several statistical techniques have been employed to analyze CF among these existing studies. For instance, a few studies developed negative regression (NB) models to predict the expected number of crashes (Khattak et al., 1999; Venugopala and Tarko, 2000; Khattak et al., 2002; Srivivasan et al., 2011). Pal and Sinha (1996) also modeled crashes at interstate work zones in Indiana and found that a normal regression model outperformed the classical NB and Poisson models. Similarly, Qi et al. (2005) constructed the truncated NB regression model and truncated Poisson regression model to analyze the rear-end crashes at work zones in New York. The truncated NB regression model was found to have better predictive power. Other than these empirical models, Elias and Herbsman (2000) used the Monte Carlo simulation approach to develop a crash rate probability distribution function that considered the intrinsic scarcity of work zone crash data. Despite of model differences, factors most commonly found to significantly affect work zone CF included the length of the work zone, duration, and average daily traffic (ADT). Generally, the modeling results showed that work zone CF increased with increasing ADT, duration, and work zone length.

A possible reason of why few studies explored the casual factors associated with work zone CF is the deficiencies of work zone data as stated in Pal and Sinha (1996), Wang et al. (1996), Zhao and Garber (2001), and Bourne et al. (2010). For instance, work zone crash data derived from police crash reports were usually subject to a number of uncertainties (Wang et al., 1996). Explanatory variables such as ADT, work zone length, and duration were also found to be subject to measurement errors. For example, the presence of significant bias was found when using the estimated ADT instead of the actual volume during work zone conditions (Khattak et al., 1999; Venugopala and Tarko, 2000; Khattak et al., 2002). Problems in defining the length of certain work zones such as bridge works and those involving detours were identified by several studies (Wang et al., 1996; Venugopala and Tarko, 2000). Moreover, the exact starting date or ending date of a specific work zone may not be readily available to calculate the duration of that work zone project (Venugopala and Tarko, 2000). If data with deficiencies were used to develop predictive models, this could clearly lead to biased estimates of the model parameters. For instance, El-Basyouny and Sayed (2010) found that the bias could increase with the magnitude of the measurement error in traffic volume when developing a safety performance function. Therefore, to be able to develop more reliable models, a limited number of safety studies applied models that can capture measurement errors in variables and better results were obtained (Lundevaller, 2006; El-Basyouny and Sayed, 2010). Similarly, measurement errors associated with work zone data were not addressed in the literature and they should also be carefully considered. This paper thus specifically uses the measurement error models to account for the work zone crash modeling issues when the work zone length is not accurately obtained.

3. Data description

As pointed out by previous researchers (Wang et al., 1996; Garber and Zhao, 2002; Qi et al., 2005), a few agencies such as the New York State Department of Transportation (DOT), Florida DOT, Oregon DOT, and Texas DOT attempted to collect supplemental data specifically for work zone crash and/or worker accident (Bourne et al., 2010). Many others have little experience in collecting, analyzing, and using work zone crash measures to assess and track safety in their work zones (Ullman et al., 2009; Bourne et al., 2010). In New Jersey, work zone crashes as well as related information were also not perfectly archived. To investigate casual relationship between work zone characteristics and CF, data from three independent sources have been obtained and assembled in a unified database: (1) work zone data, (2) crash data, and (3) traffic data.

For work zone data, project files of 60 work zones completed between 2004 and 2010 were obtained from the NJDOT engineering document unit. Information about the proposed work zone length, mileposts of the work zone, and number of lanes operated was extracted from these individual project files. Crash data of these work zones were obtained from the crash database of the NJDOT. Information about the time, date, milepost, type of crash, road type, and posted speed was extracted for each crash record. The work zone length can be determined in two ways: (1) by using the length from the work zone project file and (2) by employing spatial-temporal diagrams of work zone crash data. Fig. 1 shows examples of the spatial–temporal distributions of the observed work zone crashes for a specific work zone site. The solid-line box indicates the proposed work zone length obtained from the project file; the dashed-line box shows the adjusted work zone length based on the reported work zone crashes. Various issues related to measuring the work zone length by these two methods were discussed in modeling Section 4.

Apart from the crash records and roadway characteristics, traffic volume data are another important exposure factor required to develop crash frequency models (Ullman et al., 2009). Notably, traffic volume should be collected during the presence of work zones. However, these actual traffic counts were usually unavailable for
the type of ex post evaluation study conducted in current study (Ullman et al., 2009; Bourne et al., 2010). Therefore, straight line diagrams from the NJDOT were used to collect traffic volume data. These diagrams provided annual average daily traffic (AADT) data for each segment of roadways. To account for the seasonal and time fluctuation under work zone conditions, AADT data were adjusted according to the seasonal and hourly correction factors from NJDOT to estimate the traffic volume of the work zones.

Based on information collected from these three sources, relationship between the variables shown in Table 1 and work zone crash frequency was examined. As property damage only (PDO) and injury crashes may have different causal factors, they were aggregated by season and analyzed separately.

4. Statistical modeling methodology

Crash occurrences are considered non-negative count data and the outcome of a set of contributing factors. To model such count data, the Poisson regression model was frequently used in crash data analysis (Miao et al., 1992; Shankar et al., 1995). However, the Poisson model cannot address potential over-dispersion issues of the data. To deal with the problem, many researchers have suggested extensions of the simple Poisson model (Hauer, 2001; Lord et al., 2005). Among the extensions, the NB model has becomes one of the extensively used alternatives to model crashes (Abdel-Aty and Radwan, 2000; Mitra and Washington, 2007). Considering the potential errors in modeled variables, we made an attempt to extend the use of the NB model by incorporating measurement errors. The following sections described the models used in this study.

4.1. The negative binomial (NB) model

\( Y_i \) denote the number of crashes at the \( i \)th work zone in a given time period. Assuming that \( Y_i \) follows the Poisson distribution with the mean \( \lambda_i \), the probability of observing \( y_i \) crashes in the work zone can be described by the basic Poisson regression model:

\[
P(Y_i = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}
\]  

Location-specific work zone explanatory variables are incorporated into the model to specify the parameter \( \lambda_i \):

\[
\lambda_i = \exp(\beta^T x_i)
\]  

where \( \beta \) is the vector of estimated coefficients of the model and \( x_i \) is the explanatory variable vector (such as duration, length, and traffic volume) for the \( i \)th work zone.

To deal with unobserved heterogeneity and to allow for unexplained randomness, a noisy measurement \( \varepsilon_i \) is introduced into Eq. (2):

\[
\lambda_i = \exp(\beta^T x_i + \varepsilon_i)
\]

where \( \varepsilon_i \) is an error term which represents a random effect of omitted explanatory variables and unmeasured heterogeneity. In the NB regression model, we assume that \( \exp(\varepsilon_i) \) is Gamma distributed:

\[
\exp(\varepsilon_i) \sim \text{Gamma}(\kappa, \kappa)
\]

where \( \kappa \) is an inverse dispersion parameter. Compared with the Poisson regression model, the NB model has the additional parameter \( \kappa \) to be estimated and the model has the following properties:

\[
P(Y_i = y_i | \lambda_i) = \frac{\Gamma(y_i + \kappa)}{y_i! \Gamma(\kappa)} \left( \frac{\lambda_i}{\kappa + \lambda_i} \right)^{\kappa} \left( \frac{\lambda_i}{\kappa + \lambda_i} \right)^{y_i}
\]  

\[
E(Y_i) = \lambda_i; \quad \text{var}(Y_i) = \lambda_i + \frac{\lambda_i^2}{\kappa}
\]

The variance \( \text{var}(Y_i) \) shown in Eq. (6) is always larger than the mean \( E(Y_i) \). Thus, the NB model is more appropriate for modeling the over-dispersed crash data.

Based on the variables listed in Table 1, the expected number of crashes in a given period can be specified as follows:

\[
\ln(\lambda_i) = \alpha_0 + \alpha_1 \ln(L_i) + \alpha_2 \ln(Q_i) + \sum_{j=1}^{m} \beta_j X_{ij}, \quad i = 1, \ldots, n
\]

where \( \lambda_i \) is the expected number of crashes in given time period; \( L_i \) is the work zone length; \( Q_i \) is the traffic volume during the period of study; \( X_{ij} \) represents the \( j \)th explanatory variable at the \( i \)th work zone; and \( \alpha_0, \alpha_1, \alpha_2 \) and \( \beta_j \) are the model parameters.
Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crash</td>
<td>Numerical</td>
<td>Crash counts at work zone every three months</td>
</tr>
<tr>
<td>Length</td>
<td>Numerical</td>
<td>Length of the work zone</td>
</tr>
<tr>
<td>Traffic volume</td>
<td>Numerical</td>
<td>Adjusted volume/lane (by direction, season, hour)</td>
</tr>
<tr>
<td>Road system</td>
<td>Indicator</td>
<td>Interstate = 0, state = 1</td>
</tr>
<tr>
<td>Light condition</td>
<td>Indicator</td>
<td>Daytime = 0, nighttime = 1</td>
</tr>
<tr>
<td>Number of intersection</td>
<td>Numerical</td>
<td>Number of intersections within a work zone</td>
</tr>
<tr>
<td>Operated lanes</td>
<td>Indicator</td>
<td>Number of operating lanes in work zone</td>
</tr>
<tr>
<td>Work zone speed limit</td>
<td>Numerical</td>
<td>Work zone posted speed limit (mph)</td>
</tr>
<tr>
<td>Speed reduction</td>
<td>Numerical</td>
<td>Reduction in posted speed limit (mph)</td>
</tr>
<tr>
<td>Dropped lanes</td>
<td>Numerical</td>
<td>Number of lanes closed in a work zone</td>
</tr>
<tr>
<td>Number of ramps</td>
<td>Numerical</td>
<td>Number of ramps within a work zone</td>
</tr>
</tbody>
</table>

\[ \text{Crash counts at work zone every three months} \]
\[ \text{Length of the work zone} \]
\[ \text{Adjusted volume/lane (by direction, season, hour)} \]
\[ \text{Interstate = 0, state = 1} \]
\[ \text{Daytime = 0, nighttime = 1} \]
\[ \text{Number of intersections within a work zone} \]
\[ \text{Number of operating lanes in work zone} \]
\[ \text{Work zone posted speed limit (mph)} \]
\[ \text{Reduction in posted speed limit (mph)} \]
\[ \text{Number of lanes closed in a work zone} \]
\[ \text{Number of ramps within a work zone} \]

4.2. Modeling measurement errors in work zone length

Using the aforementioned methods, the length measurement of each work zone used has important accuracy issues as none of the methods described in Section 3 to obtain the work zone length was perfect.

First, the length extracted from the project plans provided by the NJDOT engineering document unit is likely to be inaccurate. These project plans may not reflect the final work zone layouts, which may be slightly changed in the field. For instance, the length of one of the studied projects 1368 m. This length only indicates that many tasks such as widening, grading, and paving have to be done for that section. It does not represent the actual work zone setting which may include other components such as buffer area, termination area, and so on. Moreover, work zone projects generally have multiple stages of tasks and the length of a work zone can vary according to the progress of the project. The length information for each stage was not clearly described in most of the project plans.

Second, the length of a work zone identified using the spatial–temporal diagrams of work zone crashes might have estimation errors. As shown in Fig. 1, the length of the work zone is estimated as the difference between the lowest and highest mileposts in which work zone crashes are observed. The estimation relies on the locations of the observed crashes. Obviously, this brings a major bias into work zone length measurement, mainly because of the randomness of crashes locations. If the observed crashes are not uniformly distributed along the entire work zone, the estimated work zone length is biased. In addition, the reported milepost of an observed crash might not be accurate.

Therefore, these imperfect data sources imply that either the length based on spatial–temporal diagrams or the length obtained from project plans are likely to have measurement errors.

To consider the measurement errors in the work zone length, a classical measurement-error model has been proposed. Specifically, we assume that the length on the log-scale is measured as the true value on the log-scale plus an additive error. The model structure can be expressed as:

\[
\ln(L_t) = \ln(F_t) + \tau_{it}, \quad i = 1, \ldots, n; t = 1, \ldots, T_i
\]

where \(L_t\) denotes the measured work zone length in each time period (in terms of season), \(F_t\) denotes the true length, and \(\tau_{it}\) is the measurement error term. \(T_i\) is the total number of periods for the work zone \(i\). Assume that error term \(\tau_{it}\) and \(F_t\) are independent and \(\tau_{it}\) follows a normal distribution \(N(0, \sigma^2)\), the measured work zone length follows the log-normal distribution shows here:

\[
\ln(L_t) \sim N(\ln(F_t), \sigma^2), \quad i = 1, \ldots, n; t = 1, \ldots, T_i
\]

Eq. (9) is a classical measurement error model that captures the relationship between measured work zone length \(L_t\) in a given season \(t\) and the (unknown) actual length \(F_t\). As \(F_t\) is unknown, it can be assumed to be a latent variable that follows a log-normal distribution:

\[
\ln(F_t) \sim N(\mu_{\mu}, \sigma_{\mu}^2), \quad i = 1, \ldots, n
\]

Eq. (10) represents the seasonal variation model describing the distribution of the unknown work zone length over the work zone duration. Because the dataset is not homogeneous, each work zone has its own expected length \(\mu_{\mu}\). A common seasonal variation parameter, \(\sigma_{\mu}^2\), is assumed. Instead of using the measured work zone length, the true (unknown) length obtained from Eq. (10) is incorporated into Eq. (7) and the mean function can be rewritten as follows:

\[
\ln(\lambda_i) = \alpha_0 + \alpha_1 \ln(F_{t}) + \alpha_2 \ln(Q_{it}) + \sum_{j=1}^{m} \beta_j X_{ij}, \quad i = 1, \ldots, n
\]

Eq. (8)–(11) represents the fundamental of NB model with a measurement error (MENB) in work zone length. The adjust work length based on spatial–temporal diagrams of work zone crashes was assumed to be the measured length \(L_t\). The work zone length determined from the individual work zone project files was assumed to be the expected length \(\mu_{\mu}\).

5. Model estimation

5.1. Full Bayesian estimation

All model parameters are estimated using the Full Bayesian method, which uses Monte Carlo Markov Chain (MCMC) sampling. The WinBUGS statistical software package was used. This estimation methodology has been widely used in road safety studies (El-Basyouny and Sayed, 2009; Lan et al., 2008; Yanmaz-Tuzel and Ozbay, 2010); MCMC approach repeatedly samples from the posterior distribution and generates chains of random points. Once the distribution of the simulated chains is observed to converge to the target posterior distribution, full Bayesian estimates of the model parameters are obtained from the remaining iterations. Trace plots of the chains and the Brooks–Gelman–Rubin (BGR) statistic are used to monitor the convergence. The BGR diagnostic first proposed by Brooks and Gelman (1998) examines average widths of 80 percent intervals of the pooled sample as well as each individual sample. When the average width of the individual samples approach the average width of the pooled sample, BGR statistic approaches 1. In other words, convergence is assumed to occur when the ratio is close to 1 (less than 1.2 is often sufficient to indicate convergence). The iterations up to the convergence point are excluded as burn-in samples and the remaining iterations are then used for parameter estimation. An inference is considered to be reliable when the Monte Carlo error relative to the standard deviation of the estimated parameter is less than 0.05 (Burnham and Anderson, 2002).
5.2. Prior distributions

Full Bayesian estimates of the parameters can be obtained by specifying prior distributions when they are available. Prior distribution is used to capture some known information about the distribution of each parameter. In the absence of strong prior information, uninformative priors can be used to obtain the Full Bayesian estimates. The normal distribution with zero mean and a relative large variance is a frequently used diffuse prior distribution for the regression parameters (El-Basyouny and Sayed, 2010). In this study, since we did not have known prior distributions, we have also employed normal distribution as the prior distribution for the parameters in Eqs. (7) and (11). Gamma (0.001, 0.001) was used as the uninformative priors for parameters $k$, $\sigma_e^{-2}$, and $\sigma_r^{-2}$ mentioned in Section 4.

5.3. Model selection

The Deviance information criterion (DIC) is used to compare alternative models as well as to determine which model outperforms the others. DIC is calculated as follows:

$$D = \hat{D} + p_D$$

where $D = E[D(\beta)] = -2 \log [L(Y|\beta)]$ and $p_D = \hat{D} - D(\hat{\beta})$. $\hat{D}$ is the posterior mean; $D(\hat{\beta})$ is the point estimate which measures how well the model fits the data via a log-likelihood function and unknown parameters of the model $\beta$. $Y$ is the data; $L(Y|\beta)$ is the likelihood function; $p_D$ is a measure of model complexity, which defines the effective number of parameters; and $\hat{D}$ is the posterior mean of $\beta$.

In the literature, the model with the lowest DIC value is assumed to be the best estimated model. A difference in DIC that is larger than 10 suggests that the model with a smaller DIC should be favored. On the other hand, if the difference in DIC is less than 5 then it is recommended that it might be misleading to report the model with the lower DIC as the best model (Gelman et al., 1996; Spiegelhalter et al., 2002).

To compare the goodness-of-fit of different models, the procedure developed by Gelfand (1996) was used. The procedure involves a leave-one-out cross-validation process: a data set of $n$ observations is repeatedly split into a training set of size $n-1$ and a test set of size of 1 and a prediction of the conditional predictive ordinate (CPO) of the test set is made based on the remaining training set. Assuming ith observation namely, $Y_{i,obs}$, is left out and the corresponding CPO, based on model $M_b$ is denoted as Eq. (13):

$$CPO_i = f(Y_{i,obs}|Y_{-i,obs}, M_b)$$

where $CPO_i$ is equal to the predictive density $f(Y_{i,obs}|Y_{-i,obs}, M_b)$ and $Y_{-i,obs}$ represents the all but ith observation used to estimate the model $M_b$. For comparison, the CPO$\beta$ ratio is calculated as follows:

$$CR_i = \frac{f(Y_{i,obs}|Y_{-i,obs}, M_b)}{f(Y_{i,obs}|Y_{-i,obs}, M_b)}$$

$CR_i > 1$ indicates ith observation $Y_{i,obs}$ supports model $M_b$ whereas $CR_i < 1$ suggests that ith observation $Y_{i,obs}$ supports model $M_b$. The model supported by more observations is then preferred. See Gelfand (1996) for extensive discussion of this approach from a Bayesian perspective. It should be noted that when $n$ is relative large the leave-one-out approach could be computationally intensive. Therefore, rather than running the approach for all $n$ observations, this cross-validation approach is applied to a number of randomly selected observations in this study.

6. Simulation study

To investigate the bias that might be arising from using the NB model, a Monte Carlo simulation through WinBUGS was conducted before analyzing the actual dataset. In simulation, two variables including the work zone length with measurement error and traffic volume are assumed to affect work zone crashes. It is assumed that the traffic volume $Q_i$ follows the normal distribution $\ln(Q_i) \sim N(\ln(5000), \ln(5000)^2)$ and the true work zone length $F_i$ follows the lognormal distribution so that $\ln(F_i) \sim N(\ln(\mu_f), \sigma_f^2)$, where $\mu_f = 4$, $\sigma_f = 0.4$. In addition, the measured work zone length $L_i$ is assumed to follow the lognormal distribution so that $\ln(L_i) \sim N(\ln(F_i), \sigma_e^2)$. To explore the potential bias caused by the measured work zone length with different level of errors, five scenarios including $\sigma_e = 0.1$, $0.2$, $0.3$, $0.4$, and $0.5$ are tested. Smaller $\sigma_e$ represents the smaller variation in the work zone length. The magnitude of the measurement error relative to the variation in the work zone length is assessed by using reliability ratio $\sigma_e^2/(\sigma_f^2 + \sigma_e^2)$. Hence the reliability ratios of the five scenarios are calculated as 0.941, 0.800, 0.640, 0.500, and 0.390. A total of $n = 150$ simulated crash counts $y_{i,obs}(i = 1, 2, ..., n)$ are generated from a binomial model link function with $\ln(\lambda_i) = \alpha_0 + \alpha_1 L_i + \alpha_2 \ln(Q_i)$ and $\kappa = 2$, where $\alpha_0 = -3.0$, $\alpha_1 = 1.0$, $\alpha_2 = 0.25$. Once the dataset is created, Eq. (7) is used to re-estimate the parameters of the NB model. Then Eq. (8) through Eq. (11) is used to re-estimate the MENB model in WinBUGS. Two chains of 10,000 iteration each are used. The first 5000 iterations are used as burn-in runs.

Table 2 presents the modeling results for the simulated dataset using both the NB and the MENB models. This results shows that the MENB model was able to reproduce the “true” parameter values. All coefficients were statistically significant based on their 95-percent Bayesian credible interval (CI). However, when the $\sigma_e$ increased the reliability ratio decreased and the bias of the estimated results increased. The bias of the estimated parameter $a_1$ ranges from 8.9 percent to 54.1 percent. For instance, the scenario 1 shows that the estimated parameter $a_1$ of work zone length is 0.911. It is close to the true parameter $a_1 = 1.0$ but still has a bias of 8.9 percent. All the estimated results of $a_1$ in scenarios 3, 4 and 5 are far from the true value. Moreover, their 95-percent CI values also do not cover the true value. These findings are consistent with the previous simulation study conducted by El-Basyouny and Sayed (2010) that showed that the presence of measurement errors would cause bias in parameter estimations. In the meantime, estimations of other parameters are also affected (no specific direction) due to the bias caused by the variable with measurement errors.

7. Results and discussion

Considering the performance of the MENB model in the simulation study, the actual work zone dataset is investigated. The posterior estimates of model parameters are obtained via WinBUGS using two independent Markov chains. The convergence of each model’s parameters was monitored using the BGR statistic (below 1.2) and also using visual approaches such as observing trace plots. In addition, the ratio of the Monte Carlo error relative to the standard deviation of each parameter is about 0.012–0.084.

Tables 3 and 4 present the parameter estimation and the 95 percent Bayesian CI for modeling PDO crashes and injury crashes, respectively.

Results of a conventional NB model and the alternative improved model proposed in this study (MENB) are summarized in Tables 3 and 4. Note that these are the final results that only keep the significant variables after implementing the Gibbs variable selection in WinBUGS. Based on the results, longer work zone length and higher volume per lane are positively associated with
### Table 2
Results of simulated study based on different models.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>NB</th>
<th>MENB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Scenario 1 ((\sigma_\gamma = 0.4, \sigma_\alpha = 0.1))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a_0 = -3.0)</td>
<td>-2.705</td>
<td>0.416</td>
</tr>
<tr>
<td>(a_1 = 1.0)</td>
<td>0.911</td>
<td>0.200</td>
</tr>
<tr>
<td>(a_2 = 0.25)</td>
<td>0.234</td>
<td>0.018</td>
</tr>
<tr>
<td>(k = 2)</td>
<td>1.923</td>
<td>0.397</td>
</tr>
<tr>
<td>Scenario 2 ((\sigma_\gamma = 0.4, \sigma_\alpha = 0.2))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a_0 = -3.0)</td>
<td>-2.520</td>
<td>0.359</td>
</tr>
<tr>
<td>(a_1 = 1.0)</td>
<td>0.755</td>
<td>0.176</td>
</tr>
<tr>
<td>(a_2 = 0.25)</td>
<td>0.237</td>
<td>0.016</td>
</tr>
<tr>
<td>(k = 2)</td>
<td>1.787</td>
<td>0.356</td>
</tr>
<tr>
<td>Scenario 3 ((\sigma_\gamma = 0.4, \sigma_\alpha = 0.3))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a_0 = -3.0)</td>
<td>-2.194</td>
<td>0.330</td>
</tr>
<tr>
<td>(a_1 = 1.0)</td>
<td>0.571</td>
<td>0.161</td>
</tr>
<tr>
<td>(a_2 = 0.25)</td>
<td>0.230</td>
<td>0.016</td>
</tr>
<tr>
<td>(k = 2)</td>
<td>1.745</td>
<td>0.346</td>
</tr>
<tr>
<td>Scenario 4 ((\sigma_\gamma = 0.4, \sigma_\alpha = 0.4))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a_0 = -3.0)</td>
<td>-2.190</td>
<td>0.357</td>
</tr>
<tr>
<td>(a_1 = 1.0)</td>
<td>0.566</td>
<td>0.162</td>
</tr>
<tr>
<td>(a_2 = 0.25)</td>
<td>0.234</td>
<td>0.017</td>
</tr>
<tr>
<td>(k = 2)</td>
<td>1.716</td>
<td>0.337</td>
</tr>
<tr>
<td>Scenario 5 ((\sigma_\gamma = 0.4, \sigma_\alpha = 0.5))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a_0 = -3.0)</td>
<td>-2.011</td>
<td>0.361</td>
</tr>
<tr>
<td>(a_1 = 1.0)</td>
<td>0.459</td>
<td>0.160</td>
</tr>
<tr>
<td>(a_2 = 0.25)</td>
<td>0.228</td>
<td>0.017</td>
</tr>
<tr>
<td>(k = 2)</td>
<td>1.710</td>
<td>0.346</td>
</tr>
</tbody>
</table>

### Table 3
Modeling results for work zone PDO crashes.

<table>
<thead>
<tr>
<th>Variable</th>
<th>NB CF model</th>
<th>MENB CF model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>2.5%</td>
</tr>
<tr>
<td>Constant</td>
<td>-5.757</td>
<td>0.588</td>
</tr>
<tr>
<td>Light condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln (length)</td>
<td>0.781</td>
<td>0.060</td>
</tr>
<tr>
<td>ln (traffic)</td>
<td>0.674</td>
<td>0.054</td>
</tr>
<tr>
<td>Work zone speed limit</td>
<td>-0.021</td>
<td>0.007</td>
</tr>
<tr>
<td>Road system</td>
<td>0.342</td>
<td>0.113</td>
</tr>
<tr>
<td>Operated lanes</td>
<td>0.536</td>
<td>0.047</td>
</tr>
<tr>
<td>Dropped lanes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>1.861</td>
<td>0.143</td>
</tr>
<tr>
<td>(\sigma_\gamma)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\sigma_\alpha)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td>4326.19</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4
Modeling results for work zone injury crashes.

<table>
<thead>
<tr>
<th>Variable</th>
<th>NB CF model</th>
<th>MENB CF model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>2.5%</td>
</tr>
<tr>
<td>Constant</td>
<td>-4.927</td>
<td>0.821</td>
</tr>
<tr>
<td>Light condition</td>
<td>-0.382</td>
<td>0.177</td>
</tr>
<tr>
<td>ln (traffic)</td>
<td>0.257</td>
<td>0.102</td>
</tr>
<tr>
<td>Road system</td>
<td>1.047</td>
<td>0.139</td>
</tr>
<tr>
<td>Operated lanes</td>
<td>0.652</td>
<td>0.133</td>
</tr>
<tr>
<td>Dropped lanes</td>
<td>0.432</td>
<td>0.108</td>
</tr>
<tr>
<td>(\sigma_\gamma)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\sigma_\alpha)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td>2738.45</td>
<td></td>
</tr>
</tbody>
</table>
no smooth transition, many vehicles are forced to merge, and some of late merge attempts might be a possible reason for the higher crash frequency.

Comparing the NB and MENB models, the inverse dispersion parameter \( k \) of the NB model is less than that of the MENB model. This significance of \( k \) validates the assumption that negative binomial model is a better alternative for Poisson model to represent over-dispersed crash counts (Abdel-Aty and Radwan, 2000). In addition, the larger value of \( 1/k \) in the NB model implies that apparent over-dispersion can be partially affected by measurement error in the zone work length. The reliability ratios for PDO and injury crash models are 0.776 and 0.781, respectively, which indicate the presence of a relatively high magnitude of measurement errors in the observed work zone length. Without addressing the measurement error, the coefficient of work zone length in the NB model can also lead to a bias on other parameters in the model.

The comparisons of the model diagnostic criteria estimated from each model are also shown in Tables 3 and 4. For POD crash modeling, the DIC for the NB model is 4326.19 and 4313.93 for the MENB model. For injury crash model, the DICs are 2738.45 and 2735.22 for NB and MENB models, respectively. For both PDO and injury crash modeling, allowing accounting for the measurement error in the work zone length produces smaller DIC values. Thus model selection based on DIC suggests that the MENB is preferred, particularly for the PDO crash model.

In addition, the leave-one-out cross validation approach is used to compare the two estimated models. It is well known that the computational requirements of the Full Bayesian (FB) estimation are very high even for a single iteration given the fact that FB requires thousands of iterations to converge. We also have a relatively larger set of observations compared to some of previous similar studies. Thus, in order to be able to complete this specific validation in a reasonable amount of time, 50 randomly selected samples were used to conduct the leave-one out tests. It took around two days to complete the computations with this randomly generated set. It is safe to assume that the same exercise with the complete data set would have taken around 40–45 days to complete. Thus, we used this randomly selected data set approach as an alternative way to demonstrate the performance of our modeling approach through leave-one-out validation procedure. However, as a future task, it is also possible to try other more computationally efficient validation approaches such as k-fold cross validation that can be more practical when Bayesian approach is used as the estimation method in conjunction with a relatively large data set.

According to Eqs. (13) and (14), the CPO and CR for each sample are estimated. Fig. 2 shows the results of CR for 50 tested samples. The results suggest that the leave-one-out method perfectly predicts the PDO MENB model in 36 (diamond dots in Fig. 2) out of the 50 cases based on the comparison of CPOs between the MENB and NB models. Similarly, the injury MENB model is also preferred in 33 (diamond dots in Fig. 2) out of the 50 cases. Therefore, based on DIC and the cross validation tests, the MENB model performs better in terms of modeling work zone crashes in the presence of the measurement errors in work zone lengths.

To better understand the marginal effects of the variables, elasticity of each variable is also examined. In general, elasticity is calculated as:

\[
e^{\lambda_j} = \frac{\partial \lambda_j}{\partial x_j} \frac{x_j}{\lambda_j}
\]  

(15)

where \( E \) represents the elasticity; \( \lambda_j \) is the expected crash frequency of the work zone \( i \); and \( x_j \) is the \( j \)th explanatory variable associated with work zone \( i \). Eq. (15) is applied to Eq. (3):

\[
e^{\lambda_j} = \beta_j x_j
\]  

(16)

where \( \beta_j \) is the coefficient of the \( j \)th corresponding variable.

### Table 5: Elasticity estimates for explanatory variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>PDO CF model</th>
<th>Injury CF model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light condition</td>
<td>NB</td>
<td>MENB</td>
</tr>
<tr>
<td>In (length)</td>
<td>0.690</td>
<td>0.748</td>
</tr>
<tr>
<td>In (traffic)</td>
<td>6.181</td>
<td>4.929</td>
</tr>
<tr>
<td>Work zone speed limit</td>
<td>-0.884</td>
<td>-1.452</td>
</tr>
<tr>
<td>Road system</td>
<td>0.929</td>
<td>-</td>
</tr>
<tr>
<td>Operated lanes</td>
<td>0.896</td>
<td>0.741</td>
</tr>
<tr>
<td>Dropped lanes</td>
<td>-</td>
<td>0.180</td>
</tr>
</tbody>
</table>

Inclusion of a logarithmic transformation of variables in the NB model gives the opportunity to test the proportionalities between the variables and crash counts. The estimated log-transformed model parameters directly indicate the elasticity of corresponding variables.

Please note that Eq. (16) is valid only for the continuous explanatory variables, not for dummy variables. Pseudo-elasticity can be used to approximate the elasticity of these dummy variables (Lee and Manmering, 2002). In case where the covariates are dummy variables, pseudo-elasticity indicates the incremental change in the CF produced by change in the corresponding dummy variables. Pseudo-elasticity can be calculated as follows:

\[
e^{\lambda_j} = \exp (\beta_j) - 1
\]  

(17)

The elasticity for each explanatory variable is shown in Table 5. The elasticity of the MENB model indicates that a one percent increase in the work zone length on the log-scale resulted in about 0.75 percent increase of PDO crashes and a 0.87 percent increase in injury crashes. Similarly, a one percent increase in traffic volume on the log-scale resulted in about a 4.93 percent and 1.84 percent increase in PDO crashes and injury crashes, respectively. When the work zone speed limit increased by one percent, the PDO crashes were reduced by about 1.45 percent. This finding about work zone speed supports the practical conclusion that if the presence of a work zone does not seriously affect traffic conditions (in terms of both operations and safety), a relatively higher speed limit or normal speed limit generally should be posted in the work zone. In addition, other variables including operated lanes and number of dropped lanes within the work zone collectively are observed to affect both types of work zone crash occurrences.

For indicator variables such as light condition and road system type, the interpretation of elasticity is different. The estimated elasticity indicates the change in CF given the existence of a condition (indicator = 1). For example, during nighttime, PDO and injury crash occurrences in MENB models will be reduced by about 48 percent and 80 percent, respectively. Similarly, if the road system is a state highway or other lower level road, the expected number of injury crash counts will increase by about 61 percent as per the MENB model, and the PDO crash occurrence rate will not significant change.

As shown in Table 5, the elasticity of the work zone length in the NB model is less than that in the MENB model. The impact of work zone length on work zone crash occurrences is somehow under emphasized. Such biased understanding of the impact of work zone length on safety can lead to biased selection of the optimal work zone length (Chien and Schonfeld, 2001), which in turn may result in unreliable decisions for estimation of work zone user (crash) costs, design of work zone project contracting strategies, and impact of other safety related management strategies.
8. Conclusions

In this study, statistical relationships between a set of explanatory variables and work zone crash frequencies were examined using extensive and detailed work zone data collected in New Jersey. We extended the traditional NB model by incorporating the effects that arise from measurement errors related to the work zone length. A new model to estimate the work zone CF – namely, the MENB – is proposed and estimated.

The modeling results suggest that both work zone length and traffic volume are positively associated with crash occurrence in work zones. This finding confirmed outcomes of previous studies by Pal and Sinha (1996), Venugopal and Tarko (2000), Khattak et al. (2002), and others. The crash frequency during nighttime was less than that of daytime. The elasticity of the coefficient in the MENB model suggested crash occurrence rate was 48 percent and 80 percent less at nighttime for PDO crash and injury crash, respectively. The detailed information obtained from individual work zone project files provided us with opportunities to examine additional factors than the ones considered in previous studies. First, two parameters related to work zone speed were specially considered in the models and the results suggest that more variations in the speed in work zones can result in more crashes. In addition, parameters that represent the type of roadway and the complexity of the work zone (in terms of the number of operated lanes, closed lanes, number of intersections, and ramps within work zones) were investigated. The results imply that work zones on state highways tend to have higher crash occurrence rate than that of the interstate highways. Increased complexity of work zones in terms of above factors was also attributable to more work zone crashes.

Because work zone length or configuration may change as the construction schedule progresses, using a fixed-length measurement throughout the work zone duration in the NB model was found to lead to bias on the estimation of the impact of work zone length. It also affects the estimation of other explanatory variables such as traffic volume and number of lane closure. By considering the measurement errors, the MENB model provides a better fit to the data than the traditional NB model, as indicated by DIC and cross validation tests. Such results confirmed the findings of the simulation study. Despite better performance, the proposed MENB model is not meant to justify the collection of low-quality data. If the work zones were not subject to such errors in length measurements, both models can provide comparable findings.

The CF model proposed in this study can be enhanced in several ways. The proposed CF model can be improved to account for measurement errors related to other explanatory variables. For instance, traffic volume used in the model might not be precise enough. The best way is to collect traffic data during the actual time period when the work zone exists. Comparisons with other approaches such as random effect model would be helpful to further examine the performance of the MENB model. Moreover, statistical distributions other than normal distribution can be tested to test the existence of better representations of the measurement errors in variables. In addition, other factors such as work zone setup (traffic control plans) are known to influence crashes and congestion. Therefore, in the future they should be carefully addressed given the availability of such valuable information.

Acknowledgements

The authors would like to thank the New Jersey Department of Transportation for providing the data necessary in this paper. The authors also appreciate the valuable comments from the anonymous reviewers to improve this paper. The contents of this paper reflect views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents of the paper do not necessarily reflect the official views or policies of the agencies.

References

Transportation Research Record. Journal of the Transportation Research Board 1650, 90–98.


