Modeling Crash Risk of Highway Work Zones with Relatively Short Durations

Hong Yang, Ph.D.
(Corresponding author)
Post-doctoral Associate, Department of Civil and Urban Engineering;
Center for Urban Science and Progress (CUSP);
New York University (NYU)
One MetroTech Center, 19th Floor,
Brooklyn, NY 11201, USA
Phone: (646) 997-0548, Fax: (646) 997-0560
E-mail: hong.yang@nyu.edu

Kaan Ozbay, Ph.D.
Professor, Department of Civil and Urban Engineering;
Center for Urban Science and Progress (CUSP);
New York University (NYU)
One MetroTech Center, 19th Floor,
Brooklyn, NY 11201, USA
Phone: (646) 997-0552, Fax: (646) 997-0560
E-mail: kaan.ozbay@nyu.edu

Kun Xie,
Ph.D. Candidate, Department of Civil and Urban Engineering;
Center for Urban Science and Progress (CUSP);
New York University (NYU)
One MetroTech Center, 19th Floor,
Brooklyn, NY 11201, USA
Phone: (646) 997-0547, Fax: (646) 997-0560
E-mail: kun.xie@nyu.edu

Bekir Bartin, Ph.D.
Assistant Professor, Department of Civil Engineering
Istanbul Kemerburgaz University
Mahmutbey Dilmenler Caddesi No: 26
Bagcilar, 34217, Istanbul / Turkey
Phone: 00 1 (212) 604 01 00 Ext. 1411
E-mail: bekir.bartin@kemerburgaz.edu.tr

Word count: 5189 Texts + 3 Table + 3 Figures = 6689
Abstract: 231
Re-submission Date: November 15, 2014

Paper submitted for Publication in the
Transportation Research Record, Journal of Transportation Research Board after being presented
Transportation Research Board’s 94th Annual Meeting, Washington, D.C., 2015
Abstract
Highway work zones greatly affect operational and safety performance of traffic. Existing studies have primarily focused on exploring safety issues of long-term work zones whereas safety issues associated with a large number of relatively shorter duration work zones are seldom examined. Thus, this paper aims to present an investigation of traffic safety in these work zones with relatively short duration. Considering low frequency of crashes due to this type of short duration work zones with respect to its crash condition and non-crash condition, a “rare event” logistic regression model is developed to explore the causal relationship between a set of contributing factors and the crash risk. The proposed model accounts for the imbalance issues of events (crashes) vs. non-event (non-crash) conditions in modeling low frequency crash occurrences. In addition, the model uses actual traffic data to reduce the bias of using aggregated exposure data (i.e. averaged traffic volume over a day, month, year, etc.). The modeling results based on a case study show that the work zone length, traffic volume and lane closure are positively associated with the crash risk in those work zones with short duration. The proposed model with specific model corrections and actual input data is found to improve the depiction of the relationship between these factors and the crash risk based on the comparison of the area under the ROC (Receiver Operating Characteristics) curves of different models.
INTRODUCTION

Transportation network is frequently affected by the disruptions of the increasing number of work zones associated with construction, maintenance, and rehabilitation projects. Such issue is particularly notable for critical freeways and arterials. For example, it was estimated that about 20 percent of the highways in the nation are undergoing some kind of construction work during the peak period each year and the work zones accounted for over 20 percent of non-recurring delays and more than 87 thousand of crashes including 576 fatalities during 2010 (1, 2). This number will be more alarming when more work zones have to be deployed as a result of rapidly aging transportation networks.

Many transportation agencies continuously make great efforts to reduce the operational and safety impacts of work zones on road users as well as their own on-site workers by using various work zone deployment strategies and advanced traffic control approaches. In order to make more effective decisions, the underlying knowledge of work zone crash occurrence is essential. Therefore, a number of research projects have been dedicated to explore the potential risk factors associated with work zone crash occurrences and their consequence in recent decades (3). Despite the existence of a large body of research literature developed in recent years, work zone safety issues still need more investigation as many of the associated analysis have focused on long-term work zones. Other than the long-term projects, there are thousands of work zones with relatively short durations (i.e., from several hours to several days) that have been less frequently explored. These work zones can cause very different safety issues compared with those long-term work zones. For instance, drivers are less likely to notice work zones with short duration whereas they become familiar with the existence of long-term work zones after commuting along the same routes regularly.

Therefore, the main objective of this paper is to develop a novel statistical model for exploring the causal relationship between crash risk and a set of explanatory variables in work zones with relatively short durations. The most important contribution of this paper is that it treats the crashes for short duration work zones as rare events compared to the long-term work zone crashes for the same section of the same highway. Then a rare event regression model is adopted to specifically model this unique problem.

Unlike existing studies only exploring work zones crashes based on aggregate data and crash count models (i.e., Poisson and Negative Binomial), this study looks at the safety issue of work zones from a more comprehensive perspective namely work zones with crash as well as work zones without crash experiences. In addition, it emphasizes the importance of using data that reflect actual work zone conditions to obtain reliable modeling results.

LITERATURE REVIEW

Numerous studies have been conducted in past decades to examine the influence of work zones on roadway safety. A thorough review has been presented in (3). Despite the amount of increases varies from study to study, these studies generally found that the presence of work zones on a roadway increases the crash rate or the likelihood of crash occurrence (3). Potential contributing factors that affect work zone crash occurrence and/or consequence have also been widely explored and discussed (4-25).

Despite increasing concerns on work zone safety issues, only a limited number of research specifically modeled the causal relationship between the occurrence of crashes and contributing factors at work zones (24-34). TABLE 1 provides a summary of these studies. As seen in the table, almost all studies were focused on long-term work zones on freeways (26-29, 31, 32). Other than freeway work zones, some work zones on other types of roads were also analyzed in some studies (24, 25, 30, 33, 34). However, no study specifically examined crash risk of work zones with a relative shorter duration. Other than statistical techniques such as Poisson Regression and Linear Regression, most studies examined the underlying causes of work zone crash occurrence based on
Negative Binomial (NB) regression models. Factors such as length of a work zone, duration, and traffic counts were commonly found to be positively correlated with crash occurrence at work zones.

<table>
<thead>
<tr>
<th>Study</th>
<th>Roadway (Location)</th>
<th>WZ Duration</th>
<th>Sample Sites</th>
<th>Method</th>
<th>Exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pal and Sinha (26)</td>
<td>freeway (IN)</td>
<td>long-term</td>
<td>34</td>
<td>NB, PR, NLR</td>
<td>ADT</td>
</tr>
<tr>
<td>Khattak et al. (27)</td>
<td>freeway (CA)</td>
<td>long-term</td>
<td>36</td>
<td>NB</td>
<td>ADT</td>
</tr>
<tr>
<td>Venugopal and Tarko (28)</td>
<td>rural freeway (IN)</td>
<td>long-term</td>
<td>116</td>
<td>NB</td>
<td>ADT</td>
</tr>
<tr>
<td>Khattak et al. (29)</td>
<td>freeway (CA)</td>
<td>long-term</td>
<td>36</td>
<td>NB</td>
<td>ADT</td>
</tr>
<tr>
<td>Qiet et al. (30)</td>
<td>multiple types of road (NY)</td>
<td>long/short-term/moving</td>
<td>Unknown</td>
<td>TPR, TNB</td>
<td>AADT</td>
</tr>
<tr>
<td>Lundevaller (31)</td>
<td>freeway &amp; non-freeway (IN)</td>
<td>long-term</td>
<td>72</td>
<td>RENB</td>
<td>ADT</td>
</tr>
<tr>
<td>Srinivasan et al. (32)</td>
<td>freeway (CA, OH, NC, WA)</td>
<td>long-term</td>
<td>64</td>
<td>NB, EB</td>
<td>AADT</td>
</tr>
<tr>
<td>Yang et al. (24)</td>
<td>multiple types of road (NJ)</td>
<td>long-term</td>
<td>60</td>
<td>MENB</td>
<td>AADT</td>
</tr>
<tr>
<td>Ozturk et al. (33)</td>
<td>multiple types of road (NJ)</td>
<td>long-term</td>
<td>60</td>
<td>NB</td>
<td>AADT</td>
</tr>
<tr>
<td>Ozturk et al. (25)</td>
<td>multiple types of road (NJ)</td>
<td>long-term</td>
<td>60</td>
<td>NB</td>
<td>AADT</td>
</tr>
<tr>
<td>Chen and Tarko (34)</td>
<td>multiple types of road (IN)</td>
<td>long-term</td>
<td>72</td>
<td>RE, RP</td>
<td>ADT</td>
</tr>
</tbody>
</table>

Note: NB = Negative Binomial Regression; PR = Poisson Regression; NLR = Normal Linear Regression; TPR = Truncated Poisson Regression; TNB = Truncated NB; EB = Empirical Bayesian; RE = Random Effect; MENB = NB with Measurement Error; RP = Random Parameters; ADT = Average Daily Traffic (vehicles per day); AADT = Annual Average Daily Traffic.

The findings based on the above studies provide some useful information for work zone safety improvement. However, the findings based on the developed crash occurrence models are still questionable due to several major issues. First, these models were only built on work zones that had many crashes. Many work zones, particular those with shorter duration, that have no crash experience were not included (30). Second, data such as adjusted/estimated average daily traffic (ADT) at non-work zone periods incorporated in the models did not reflect actual traffic conditions when work zones present (27, 28, 35). The traffic condition might significantly change due to the presence of the work sites. Consequently, the deficiencies of these work zone-specific data can lead to uncertainty when quantifying work zone crash risk associated with different contributing factors (35). Finally, almost all the existing studies used the aggregated data in their models (i.e., aggregated crash counts and exposure data). These aggregated data cannot reflect the change of conditions over time in the same work zones.

To sum up, most studies in literature only provided insight into crash risk of long-term work zones. There is need to investigate crash risk of work zones/roadwork with a relative shorter duration as they are unique but frequently present on our transportation networks. In addition, data issues have to be addressed and studies based on more accurate work zone-related information are greatly needed.

**DATA DESCRIPTION**

To investigate the mechanism of crash occurrence in work zones with a relatively short duration, the work zones on a 25-mile section of the New Jersey Turnpike (NJTPK) between milepost 48 and milepost 73 were used. This section of NJTPK has 3 lanes in both northbound and southbound directions. As one of the busiest toll road corridors in the east coast, work zones present on the NJTPK can possibly have an adverse impact on traffic operations and safety. In order to minimize the impact of work zones, they are usually deployed for short durations during the off-peak periods. In 2011, 562 work zones with relatively short durations were deployed along the studied section (both directions). For instance, some of them were maintenance work zones, pavement rehabilitation, etc. Despite deploying these work zones during off-peak periods, 56 crashes occurred due to these
work zones. Therefore, it is of great interest to explore the crash risk posed by these and other similar work zones.

The studied section also has 23 remote traffic microwave sensors (RTMS) placed approximately at every one mile on the mainline, which provide real-time traffic data for these work zones. The sensors continuously measure travel speed, volume, and occupancy. The raw data were aggregated in 15-minute time intervals to capture the variation of traffic pattern. In addition, the time interval was chose considering potential errors in reported crash time. For each work zone, the traffic data from the nearest sensor was extracted for describing the actual traffic condition in the work zone. Excluding those cases with missing data, 466 work zones with 44 crashes were used in the final analysis. In total, these work zones consist of 16,285 periods (each period is 15 minutes based on the aggregation of sensor data). In other words, 44 out of the 16,285 periods had a crash, which clearly represent a very rare case of crash occurrence (0.27%). The characteristics of the data were summarized in TABLE 2. Notably, the duration of these work zones on average is 11.5 hours and the maximum one is 78.9 hours (less than 4 days).

### TABLE 2. Descriptive statistics of variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Description</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crash</td>
<td>Categorical</td>
<td>Crash occurrence =1; otherwise = 0</td>
<td>0.0027</td>
<td>0.0519</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>WZLength</td>
<td>Numerical</td>
<td>Work zone length (mile)</td>
<td>2.33</td>
<td>1.52</td>
<td>0.2</td>
<td>21.1</td>
</tr>
<tr>
<td>Volume</td>
<td>Numerical</td>
<td>Traffic volume in each time interval (veh/15min)</td>
<td>284</td>
<td>191</td>
<td>1</td>
<td>1183</td>
</tr>
<tr>
<td>Laneclosed</td>
<td>Numerical</td>
<td>Number of lanes closed in a work zone</td>
<td>1.34</td>
<td>0.49</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Duration</td>
<td>Numerical</td>
<td>Hours a work zone presents</td>
<td>11.5</td>
<td>11.7</td>
<td>0.1</td>
<td>78.9</td>
</tr>
<tr>
<td>Season</td>
<td>Categorical</td>
<td>Winter=1; otherwise=0</td>
<td>0.43</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Unlike long-term work zones that have many crashes, it should be noted that work zones with relatively short durations usually have no crash or one crash during the entire active work period. Therefore, traditional models such as those listed in TABLE 1 may be not appropriate as they mainly rely on a higher frequency of crash counts. Therefore, a more reliable and appropriate approach to describe the crash risk in these work zones is needed.

### METHODOLOGY

In this study, instead of assessing crash frequency/rate, we are interested in whether or not the presence of a short term work zone will cause a crash. Hereinafter crash risk is defined as the probability of short-term work zone having a crash. Then work zone-related factors that are hypothesized to affect the crash risk are investigated.

#### Logistic Regression

Denote $y$ as the crash occurrence at a work zone. Crash occurrence at work zones can be described as a binary outcome. Let $y = 1$ represent that a crash occurs in a work zone whereas $y = 0$ indicates no crash occurred in a work zone. A binary logistic regression model can be developed to examine the influence of factors that may affect the crash risk of a short-term work zone.

Let us define $\pi(x)$ as the probability of a work zone having a crash during its presence and $1 - \pi(x)$ as the probability of a work zone having no crash. The binary logistic regression model identifies the relationship between the log odds of the dichotomous outcome and various risk factors. It can be formulated as following equation (1):

$$\logit[\pi(x)] = \log \left[ \frac{\pi(x)}{1 - \pi(x)} \right] = \alpha + X\beta$$

(1)
Based on the above equation, the probability that a work zone having a crash can be described by the logistic distribution shown in following equation (2):

$$P(y = 1|X) = \frac{\exp(\alpha + X\beta)}{1 + \exp(\alpha + X\beta)}$$ (2)

where $\pi(x)$ is the conditional probability of the form $P(y = 1|X)$; $X$ is the vector of explanatory variables (contributing factors) that could be numerical or categorical; $\beta$ is the corresponding vector of the coefficients; and $\alpha$ is an intercept parameter.

The maximum-likelihood estimation (MLE) technique can be used to determine the regression model’s parameters. The likelihood function is constructed as equation (3). By maximizing the log likelihood expression shown in equation (4), the best estimate of parameters $\alpha$ and $\beta$ can be obtained accordingly.

$$l(\beta) = \prod_{i=1}^{n} \left[ \pi(x_i)^{y_i} \left[ 1 - \pi(x_i) \right]^{(1-y_i)} \right]$$ (3)

$$LL(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} [y_i \ln[\pi(x_i)] + (1 - y_i)\ln[1 - \pi(x_i)]]$$ (4)

where $y_i$ denotes the observed outcome at the $i^{th}$ period of a work zone, with the value of either 0 or 1 only; $i = 1, 2, ..., n$ and $n$ is total number of observed periods.

The overall goodness-of-fit of the logistic regression model is tested by likelihood ratio test. The significance of individual contributing factors within the model is examined using the Wald z statistic. Moreover, the influence of $j^{th}$ factor on work zone crash risk can be revealed by the odds ratio (OR), which is defined as equation (5):

$$OR = \exp(\beta_j)$$ (5)

where $\beta_j$ is the coefficient of the $j^{th}$ factor. OR measures the ratio of the predicted odds for a one-unit increase in a continuous variable $x_j$ or the presence of a dichotomous variable $x_j$ when other variables in the model are held constant. The 95% confidence interval of OR is $[\exp(\beta_j - 1.96s_{\beta_j}), \exp(\beta_j + 1.96s_{\beta_j})]$, where $s_{\beta_j}$ is the standard error of the coefficient $\beta_j$. An OR greater than 1 suggests that the $j^{th}$ factor increases the likelihood of crash occurrence in the work zone and vice versa.

**Rare Events Logistic Regression**

Although the binary logistic regression models have been frequently used in traffic safety studies, the models can yield biased estimates when sample data exhibits class imbalance (36). In this study, the crash occurrence at work zones with short duration can be an example of class imbalance. As noted in Section 3, there were many periods that the work zones have no crash ($y = 0$) whereas only a small number of them had crashes ($y = 1$).

In such cases, if the MLE of the binary logistic regression model is used, it can result in biased estimates of coefficients for the factors related to work zone crash occurrence (37, 38).

In order to avoid these estimation issues, (King and Zeng (37)) have developed the rare events logistic regression for the cases that include three types of corrections for the ordinary logistic regression. First, a case-control sampling design based on endogenous stratified sampling is recommended. It consists of taking all periods with a crash ($y = 1$) and a random selection of periods with no work zone crash ($y = 0$). Empirically, the proportion of periods with a crash to the periods with no crash can be set around 1:10 (39, 40). A sensitivity analysis can be performed by using different sampling proportions.
The second correction is called prior correction. The purpose of prior correction is to avoid selecting dependent variables that may introduce sampling bias on the logistic coefficients (King and Zeng, 2001). The intercept $\alpha$ is corrected using the following equation.

$$\alpha = \hat{\alpha} - \ln \left( \frac{1 - \tau \times \bar{y}}{\tau} \right)$$

(6)

where $\alpha$ is corrected intercept, $\hat{\alpha}$ is the uncorrected intercept, $\tau$ is the actual true faction of 1s (events) in the population, and $\bar{y}$ is the observed fraction of 1s (events) in the sample.

The last correction is to adjust the MLEs of the logistic regression parameters and the predictions of the estimated model. When we substitute the corrected intercept $\alpha$ in equation (6), there is an underestimation of predicated probability because of the uncertainty in the estimation of $\hat{\beta}$. Therefore, $\hat{\beta}$ has to be adjusted. King and Zeng (2001) described the procedure to obtain the "approximately unbiased" coefficients, $\hat{\beta}$. Based on the adjusted intersect and coefficients, the predicted crash occurrence probability for a given work zone is obtained as $\hat{\pi}$ based on a correction factor $C_i$ incorporated into the original estimated probability $\pi_i$. When events are rare, the predicted probability of an event using logistic regression is adjusted. In our case, it denotes the predicted work zone crash risk.

$$\hat{\pi}_i = \hat{\pi}_i + C_i$$

(7)

The correction factor $C_i$ is obtained using the following equation:

$$C_i = (0.5 - \bar{y}_i) \bar{y}_i (1 - \hat{\pi}_i) \alpha \times \alpha$$

(8)

where $\bar{y}_i$ is the event probability estimated using the bias-corrected coefficient $\alpha$; $X$ is a $1 \times (n + 1)$ vector of values for each explanatory variable; and $\alpha \times \alpha$ is the variance-covariance matrix.

To implement the rare event logistic regression model, we have processed all the traffic and work zone data, and estimated models using the open source statistical software package R. The rare event logistic regression model built in the "relogit" function of the R package "Zelig" was used to estimate the work zone crash risk (41).

RESULTS AND DISCUSSION

As mentioned in the literature review section, existing studies usually use the aggregated traffic volume (AADT) to describe the traffic condition in work zones. Since the traffic condition may greatly change during the presence of a work zone, the use of actual observed traffic data during this time period are preferred. Unlike existing studies, the present study, therefore, uses real-time sensor data as the input to the risk prediction models to accurately describe the traffic condition when a crash occurs in a work zone. FIGURE 1 illustrates the actual traffic conditions of four work zones. FIGURE 1 (a) and FIGURE 1 (b) show work zones without a crash whereas FIGURE 1 (c) and FIGURE 1 (d) show work zones that experienced crashes. It can be seen that the traffic volumes greatly change during the duration of the work zone. For example, the average traffic volume is 173 vehicles per 15 minutes for FIGURE 1 (c), but when a crash occurs then volume is observed to be 234 vehicles per 15 minutes. Larger differences are observed in FIGURE 1 (d). Therefore, the use of averaged/aggregated exposure data (i.e., ADT & AADT) obviously may yield biased results for this type of analysis. Instead of using the aggregate data, this study describes the crash risk in each 15-minute interval and uses the corresponding sensor data to capture the variation in traffic. Despite the fact that higher resolution sensor data (i.e., 30-second interval, 1-minute interval, 5-minute interval, etc.) were available, the 15-minute interval was chosen based on two major considerations: (a) the desire to have a reasonable size interval that can capture the traffic variation, and (b) the need to define a time interval that can better capture/approximate actual crashes.
Based on the traffic conditions during the time period where the crash actually occurs and work zone information, the rare event logistic regression model is estimated. The estimation results are presented in TABLE 3. The final rare event logistic regression model only includes variables that are statistically significant at the significance level of 0.1. The results suggest that all the three variables are positively associated with the crash risk in work zones with relatively short durations. More specifically, if the log-transformed work zone length increases one unit, the likelihood of having work zone crash increased, with an odds ratio of 2.415. Similarly, when the log-transformed volume increases one unit, the corresponding crash risk will be increased, with an odds ratio of 1.679. When one more lane is closed, the odds ratio is 1.662. For these work zones with relatively short duration, the seasonal impact on the crash risk was not significant. The underlying reason for this observation should be attributed to the fact that most of these work zones were only deployed during the off-peak periods under good weather condition. Since the work zone durations are short, they can be easily rescheduled to avoid severe weather impact.
### TABLE 3. Results of rare events logistic regression model (Event: Non-Event = 1:30).

| Variable        | Estimate | Std. Error | z value | Pr(>|z|) | Odds Ratio |
|-----------------|----------|------------|---------|----------|------------|
| Intercept       | -10.063  | 1.435      | -7.012  | 0.000    | ----       |
| log(WZLength)   | 0.882    | 0.317      | 2.780   | 0.005    | 2.415      |
| log(Volume)     | 0.518    | 0.219      | 2.364   | 0.018    | 1.679      |
| Lanes closed    | 0.508    | 0.304      | 1.672   | 0.095    | 1.662      |

The results presented in TABLE 3 are based on a sampling ratio of 1:30 (crash conditions vs. non-crash conditions). To test the sensitivity of the modeling results, different sampling ratios are also analyzed. FIGURE 2 shows the estimated coefficients under different sampling ratios. It can be seen that as long as the ratio of non-event conditions vs. event conditions increases (i.e. sampling ratio ≥ 30), the estimates are relatively stable. This suggests that without losing modeling accuracy, the work of data preparation can be reduced in terms of processing the real-time sensor data to reflect actual work zone conditions required by our model.

![FIGURE 2. Effect of different sampling ratios on estimates of coefficients](image)

A regular logistic regression model is also estimated for the same case study with all the observed crash periods and non-crash periods of work zones. To compare the performance of the rare events logistic regression model with the regular logistic regression model, the areas under the Receiver Operating Characteristics (ROC) curves, or simply the area under the curve (AUC) of two models are examined. This is a frequently used measure for the comparison of learning algorithms (42). According to this reference, a ROC curve is said “to dominate another one if it is always above and to the left of the other one”. If most of the curves overlap each other, the AUC is used as the measure for comparison (42). FIGURE 3 shows the ROC curves of logistic regression model vs. the rare events logistic regression model with different sample ratios. The AUCs of the rare events model...
are larger than the logistic regression model in both examples (The AUCs for rare events models are 0.6807 in FIGURE 3 (a) and 0.6804 in FIGURE 3 (b). Both are larger than the AUC of 0.6748 for the logistic regression model), which indicate that the rare events model should be preferred.

![ROC Curves](image)

FIGURE 3. ROC of rare event model vs. logistic regression model for different sampling ratios

CONCLUSIONS AND FUTURE WORK

This study examined the statistical relationships between a set of explanatory variables and the crash risk in work zones with relatively short durations. Instead of using traditional logistic regression model, i.e., Harbet et al. (19), a rare event logistic regression model was estimated to capture the characteristics of crashes occurring during the of short duration work zones on highways. The proposed rare event regression model contributes to the literature in several aspects: First, it incorporates all work zones with and without crash experience in the model. Existing studies have only focused on those work zones with crash records. Second, it accounts for the imbalance between periods with crash occurrence(s) and other large amount of periods with no crashes. In addition, instead of using aggregated crash and traffic data, this study modeled the crash risk at each time interval with the actual sensor data. The bias introduced by using aggregated data can be reduced. For instance, existing studies generally adopt Poisson and/or Negative Binomial Models to estimate work zone crash frequency. However, these models hardly differentiate the periods with crashes from others without any crash. Instead, they usually use the “duration” as a variable to aggregate these periods together based on the assumption that the crash risk of each period is the same. In contrast, by using discrete data, the work zone crash risk in each short-time interval can be reliably captured under its prevailing traffic conditions. Based on the analysis of the developed rare event logistic regression model, it is found that the work zone length, actual traffic volume, and the number of lanes closed are positively associated with the crash risk in those work zones with short duration. The sensitivity analysis shows that the modeling results are stable even when only a part of the non-event periods are sampled. The use of the rare events model also improved the prediction performance of crash occurrence compared with the traditional logistic regression model. Future work will include extending the rare event regression model to multiple sites with different characteristics.
ACKNOWLEDGEMENTS
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 any agency mentioned in the paper. The authors thank for the insightful comments and suggestions from all the anonymous reviewers that helped improve the paper.

REFERENCES


