A Joint Framework for Static and Real-time Crash Risk Analysis

Shamsunnahar Yasmin
Postdoctoral Associate
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 407-823-4815, Fax: 407-823-3315
Email: shamsunnahar.yasmin@ucf.edu

Naveen Eluru*
Associate Professor
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 407-823-4815, Fax: 407-823-3315
Email: naveen.eluru@ucf.edu

Ling Wang
Postdoctoral Associate
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 407-823-0300, Fax: 407-823-3315
Email: lingwang@knights.ucf.edu

Mohamed A. Abdel-Aty
Professor
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 407-823-4535, Fax: 407-823-3315
Email: m.aty@ucf.edu

*Corresponding author

Abstract

The current research effort bridges the gap between traditional crash risk and real-time crash risk models by developing a joint model that accommodates for both dimensions in developing crash risk analysis models. Specifically, we develop a joint reactive and proactive crash modeling framework by coupling the monthly crash risk and real-time crash risk in a unified econometric framework for a microscopic analysis unit. In the joint modeling approach, we propose and estimate an alternative to the case-control binary logit based real-time crash risk analysis by proposing a multinomial logit based approach where time periods serve as alternatives and the chosen alternative is the time period in which crash occurs. The joint model also allows us to accommodate for the common unobserved factors that increase the likelihood of a crash in microscopic unit to affect the real-time crash risk propensity. We demonstrate the application of the proposed approach by using data on roadway segments from three expressways in Central Florida (State Roads 408, 417, and 528) for 29 months. The monthly crash risk component is examined by using binary logit model employing different static roadway attributes (roadway geometry and operational attributes). The real-time crash risk component is examined by using a multinomial logit model employing different real-time traffic attributes (volume, speed, lane occupancy and environmental conditions). The outcome of the proposed approach allows us to predict both the monthly and real-time crash risk components simultaneously in a single econometric framework.

Keywords: real-time crash risk; case-control binary logit, multinomial logit models; joint econometric frameworks; expressways

1. INTRODUCTION

Traditionally, statistical analysis in road safety research has evolved based on police reported crash databases along two major streams: crash frequency analysis and crash severity analysis. The first stream of research is focused on identifying attributes that result in traffic crashes and propose means to reduce the occurrence of traffic crashes (see Lord and Mannering, 2010; Yasmin and Eluru, 2016 for a review of these studies). The second stream of work examines crash events and identifies factors that impact the crash outcome and suggests countermeasures to reduce crash related consequences (injuries and fatalities) (see Savolainen et al., 2011; Yasmin and Eluru, 2013). However, analysis based on police reported crash databases is "after the fact" i.e. after the occurrence of the crash and is predominantly *reactive* in nature. On the other hand, with the availability of emerging data collection techniques, researchers have the opportunity to develop *proactive* real time crash analysis models based on real-time traffic data (Mannering and Bhat, 2014). For example, advanced transportation information system (ATIS) based databases have become efficient and readily accessible for safety professionals. The access to such data has resulted in growing interest among safety researchers to evaluate real-time crash risk. Findings from these studies can assist in the development and deployment of proactive traffic management strategies to ameliorate hazardous traffic conditions (see Roshandel et al., 2015 for a review of these studies). Research in reactive and proactive safety areas has progressed independently. The current research effort bridges the gap between these two streams by developing a joint model that accommodates for both dimensions in developing crash risk analysis models. Specifically, we formulate and estimate a joint econometric framework to analyze the monthly crash risk and real-time crash risk components simultaneously, while also modeling the real-time crash risk component based on a "sampling of alternatives" approach within a multinomial logit based structure.

2. EARLIER RESEARCH

In existing safety literature, studies that evaluated microscopic crash risk of highway/freeway segments can be classified along two broad categories: (1) reactive crash risk assessment models and (2) proactive crash risk assessment models.

The *first group of studies* in the transportation safety area are focused on identifying critical factors contributing to crash occurrences for a particular roadway entity – traditionally known as crash frequency or crash prediction models (also often referred to as reactive, aggregate or static crash risk models). The crash count events examined in traditional crash frequency models are aggregated at a spatial unit for a given period of time (such as month or year). These models are developed by using relatively low-resolution historical data related to roadway geometry, roadway condition, traffic characteristics, operational attributes and/or environmental data (Chen et al., 2016; Ding and Gou, 2016; Dinu and Veeraragavan, 2011). In examining crash occurrences in the long-term at a micro-level, statistical modeling approaches considered include ordinary least square regression, logistic regression, bayesian regression and a wide variety of count regression approaches (see Lord and Mannering, 2010; Yasmin and Eluru, 2016 for a detailed review). Outcome of these studies are predominantly used to identify effective countermeasures to improve roadway design and/or operational attributes, to identify black spots and to evaluate roadway safety policies and interventions (Chang and Kim, 2012; Geurts and Wets, 2003; Saccomanno et al., 2001). These models are integral to the development of safety performance function (SPF) and

crash modification factors (Hariharan et al., 2016; Park et al., 2016). However, these models are not applicable for identifying short-term hazardous traffic conditions that may lead to crash occurrences and thus are poorly suited for developing crash avoidance systems using advanced traffic management interventions (Abdel-Aty and Pande, 2007).

The *second group of studies* – referred to as real-time crash risk models (also often referred to as proactive, disaggregate or dynamic crash risk models), link real-time crash likelihood with hazardous microscopic traffic flow conditions prior to crash occurrences (see Roshandel et al., 2015; Xu et al., 2015 for detailed review). These studies are developed due to the appealing feature of predicting crash risk proactively. The underlying assumption of these studies is that real-time traffic, geometric and weather conditions can identify 'crash prone' traffic conditions. These models are focused on identifying crash precursors that are likely to lead to crash occurrence in dynamic traffic environment using high-resolution traffic data (such as traffic monitoring data for 5-10 minute intervals), weather characteristics and road geometry (Abdel-Aty et al., 2008; Lee et al., 2002; Oh et al., 2001; Wang et al., 2015; Xu et al., 2013). In developing real-time crash risk models, researchers have predominantly resorted to case-control sampling design (Abdel-Aty and Pande, 2007; Sun and Sun, 2015; Xu et al., 2016). Case-control studies are usually retrospective studies where cases - outcomes of interest - are matched with a control group. In real-time crash prediction models, a crash is considered as 'case' and non-crash events for similar situational conditions are considered as 'control' within case-control study design. It is an efficient and cost effective method in designing studies for rare events. The case-control approach can take two forms: unmatched or matched. The major difference between these two study designs lies in selecting the random sample of controls from the crash-free observations. In the matched design approach, control confounding factors (such as segment characteristics or time) are used to identify controls. On the other hand, these factors are considered directly in the model estimation for unmatched case-control design (Bruce et al., 2008). It is important to note that the differences in efficiency between model estimates from these two sampling designs is marginal (Thompson et al., 1982). The econometric structures employed in developing real-time crash risk models include log-linear model, binary logistic regression model, multilevel binary logistic regression model, Bayesian logistic regression model, dynamic Bayesian network model and several data mining techniques (Abdel-Aty et al., 2007; Basso et al., 2018; Lee et al., 2002; Qu et al., 2012; Zheng et al., 2010; Wang et al., 2017a; Wang et al., 2017b; Wu et al., 2018; Xie et al., 2017; You et al., 2017). A majority of the studies to date used the matched case–control design and binary logit model to examine possible relationships between crash precursors and crash risk for dynamic traffic environment (Roshandel et al., 2015).

3. CURRENT STUDY IN CONTEXT

3.1 Addressing Binary Logistic Regression Model Limitations

From the literature review it is evident that the predominantly used model to study real-time crash occurrence is the binary logit model to differentiate crash occurrence (yes or no) in a time period. The rationale behind this approach is to generate an estimated crash risk for the chosen time period (5-minute interval or similar). There are two major limitations for the case control approaches employed in literature. *First*, by design, crash occurrence is a rare event; there is significant evidence from econometrics and statistics literature to indicate that traditional binary logit models (or logistic regression approaches) with extremely small shares for one of the alternatives are likely

to yield biased/incorrect model estimates (see for example Calabrese and Osmetti, 2013; King and Zeng, 2001).

Second, identifying the appropriate number of control cases for a case-control study is far from "well-defined". The model estimates from the case-control study are likely to be unbiased only if the case-control ratio considered in the study design is representative of the ratio from the overall population. However, given that crash is a rare-event and collection of all control events against such low incidence events are expensive and to some extent infeasible, researchers have considered various values of this ratio. Most of the existing real-time crash risk studies used case-control study design with a ratio ranging from 1:1 to 1:10[1]. Thus the databases considered are predominantly biased towards an underrepresentation of non-crash cases. Earlier studies implicitly assume that the influence of the altered sample is captured by the constant in the binary logit model and does not affect other model parameters. While theoretically under certain conditions the assumption is possibly valid – empirical studies have shown that model estimates of other parameters are also likely to be biased (see Wooldridge, 2010; Yasmin and Eluru, 2013). Most recently, Theofilatos et al. (2018) considered crash as a rare event and examined real-time crash risk by using the Firth methods (a bias correction). Furthermore, even if the parameters are unbiased model estimates from case-control studies cannot be used to calculate relative risk directly without employing corrections for the constant (see Zhang and Kai, 1998 for detailed discussion). The case control model outputs can only be used to calculate the odds ratio (Mann, 2003). It is surprising to note that very few studies in safety literature have explicitly discussed this issue with the case-control study design.

In this context, the current research effort contributes to real-time crash risk analysis methodologically by proposing an alternate approach to the case-control binary logit based approach. Specifically, we propose to consider a "sampling of alternatives" based approach with a multinomial logit based formulation to examine real-time crash risk. In our approach, time periods (5 minute interval) serve as alternatives and the outcome to be studied is the time period in which crash occurs. This is a significantly different approach to the traditional approach in literature. In the binary approach, the number of alternatives are limited to 2. However, the challenge is in determining the appropriate amount of crash and non-crash records for analysis. In our proposed approach, the number of outcome contexts (or observations) is based on the number of crash events on a spatial unit (such as a freeway segment) in prescribed time period (such as a month). Given the reformulation, the dependent variable is the outcome of time period within the month for a segment. Any time interval in the month could be a potential alternative for crash occurrence. Thus, we have translated the issue of "data size" into an issue of "outcome set size". In the context of multinomial logit model (and other categorical modeling approaches), there are relatively straight forward approaches to handle such large outcome set sizes. The most commonly employed approach is to consider a random sample of time periods in addition to the time period that the crash has occurred and estimate the model with this outcome set. These approaches have been employed in a host of transportation scenarios including residential/work location choice (Waddell et al., 2007), destination choice (Scott and He, 2012), and bicycle sharing system station choice (Faghih-Imani and Eluru, 2015).

[1] In existing real-time crash risk studies, case-control ratio of 1:4 has been widely accepted with an argument that the improvement in term of statistical power is negligible beyond a case-control ratio of 1:4 (Ahmed and Abdel-Aty, 2012).

3.2 Bringing Together Static and Real-Time Crash Analysis

Based on the review of earlier research, it also evident that static and real-time models in examining crash risks have evolved independently. The static crash risk models are developed based on historical aggregate level information. On the other hand, current framework for real-time crash risk models choose crash occurrence events and generate control cases for similar situational conditions. Thus only a sample of segments are considered in the eventual model development. A potential result of this approach is the absence of any consideration of a reasonably large number of segments because crash occurrence is a rare event. In that case, the analysis is inherently biased toward crash occurring locations (and their controls). To provide a real-time crash risk assessment this approach might not result in a truly universal sample.

In our study, to alleviate this potential mis-selection of segments, we propose an econometric framework with two components. The first component is monthly crash risk component for all segments and the dependent variable is characterized as a binary outcome: no crash or crash. Thus, all segments part of the study region are considered in our first component model. The unit of analysis is a month across different roadway segments for the study period. We employ the binary logit (BL) model in the current study context[2]. The reader would recognize that the monthly crash risk dependent variable is not as rare as the crash risk at the time period level.

For the second component, we replace the traditional case control binary approach with a "sampling of alternatives" based multinomial logit based formulation to examine real-time crash risk. The alternatives in the real-time crash risk model are time intervals of 5 minute duration in the month for every segment. Naturally, the universal outcome set will be very large and would result in computational tractability challenges (Ben-Akiva and Lerman, 1985; Train, 2009). To resolve this, we adopt a sampling approach to generate 30 alternatives (29 randomly sampled time periods in the month and 1 actual crash outcome time period). Several researchers have employed the sampling approach successfully in the transportation area (Faghih-Imani and Eluru, 2015; Scott and He, 2012; Waddell et al., 2007). The reader would note that the second component is only applicable for segments with crashes. For segments with multiple crashes in a segment, we consider multiple crash records within a repeated observation structure.

The two components mentioned in the modeling framework can potentially be considered as sequential or simultaneous. A sequential assumption indicates modeling monthly crash risk and conditional on the crash risk modeling real-time crash risk. The simultaneous approach, on the other hand, postulates that there are common unobserved factors influencing the two components. In our study, we explicitly posit that the two components are interconnected by different observed and unobserved attributes and thus adopt the simultaneous framework. For instance, we observe that a high speed limit (HSL) roadway, in general, has lower crash risk than a low speed limit (LSL) location. However, higher variance of vehicle speed within a specified time interval substantially increases the likelihood of crash risk for the HSL compared to the crash risk of LSL road. This is an example of how certain combinations of real-time traffic conditions (speed variance in the interval) and static roadway attributes (speed limit) influence static and real-time

[2] The aggregate level crash prediction models are generally developed by using count-based regression models. But in our estimation sample, we have only 371 records with multiple crashes (among 6913 records) over a month for different roadway segments. Therefore, we adopt a BL model based approach for examining monthly crash risk component. However, in case of higher number of crash events in future research, examining static crash risk by employing an ordered or count based regression based approach is straightforward within the proposed joint framework.

crash risk. In this example, these attributes are observed to the analysts and accommodating the impact of these observed variables is straightforward within the two components.

At the same time, several unobserved factors may also contribute to the interconnectedness of these two crash risk dimensions. For example, a backward shockwave resulting from an aggressive lane changing behaviour in a certain time interval is likely to increase crash risk on a HSL road (compared to LSL road). However, such driver behavior attributes are difficult to observe. Thus, accommodating for the impact of such unobserved heterogeneity necessitates for the consideration of interconnectedness between the two components. To the best of the authors' knowledge, this is the first attempt to employ such a joint framework for examining micro-level crash count events.

In summary, the current research effort contributes to safety literature on micro-level crash risk analysis both methodologically and empirically. In terms of methodology, we formulate and estimate a joint econometric framework to analyze the monthly crash risk and real-time crash risk components simultaneously, while also modeling the real-time crash risk component based on a "sampling of alternatives" approach within a multinomial logit based structure. Empirically, we demonstrate the application of the proposed approach by using data on roadway segments from three expressways in Central Florida (State Roads 408, 417, and 528) for 29 months. The monthly crash risk component is examined by using binary logit model (month with zero crash and at least one crash) employing different static roadway attributes (roadway geometry and operational attributes). While the real-time crash risk component is examined by using a random utility model employing different time varying traffic attributes (volume, speed, lane occupancy and environmental conditions) and their interactions with static attributes.

The rest of the paper is organized as follows. Section 4 provides details of the econometric model framework used in the analysis. In Section 5, the data, dependent and independent variables' formation procedures are described. Model comparison results and estimation results are presented in Section 6. Section 7 concludes the paper.

4. ECONOMETRIC FRAMEWORK

4.1 Model Structure

The focus of our study is to jointly model "monthly crash risk" and "real-time crash risk". Let us assume that i $(i = 1,2,3, \dots, N)$ be the index to represent road segment, k represent the crash states, and t $(t = 1,2,3, \dots, T)$ represent different months. In this empirical study, k take the values of 'no crash' $(k = 0)$ and 'at least one crash' $(k = 1)$. The binary logit formulation can be incorporated in an ordered logit structure and hence is formulated by using ordered outcome structure in current study context (Train, 2009). In the ordered outcome framework, the actual crash state (y_{itk}) are assumed to be associated with an underlying continuous latent variable (y^*_{itk}).

For the joint approach, the equation system for modeling the monthly crash risk component takes the familiar binary logit in an ordered outcome formulation as follows:

$$y^*_{it} = \left((\boldsymbol{\beta} + \boldsymbol{\gamma}_{it})\boldsymbol{x}_i + \varepsilon_{it} + \boldsymbol{\eta}_{it}\right), y_{it} = k \; if \; \tau_{it,k-1} < y^*_{it} < \tau_{itk} \quad (1)$$

The latent propensity y^*_{it} is mapped to the actual crash state y_{it} by τ thresholds $(\tau_0 = -\infty \; and \; \tau_K = +\infty)$ as presented in equation 1. $\boldsymbol{x}_i$ is a vector of static roadway attributes

(including constant) associated with segment i. $\boldsymbol{\beta}$ is the vector of corresponding mean effects. $\boldsymbol{\gamma}_{it}$ is a vector of unobserved factors on monthly crash risk propensity of segment i for month t and its associated roadway characteristics assumed to be realization from standard normal distribution: $\boldsymbol{\gamma}_{it} \sim N(0, \boldsymbol{\sigma}_{it}^2)$. ε_{it} is an idiosyncratic error term assumed to be identically and independently standard logistic distributed across segment i. $\boldsymbol{\eta}_{it}$ captures unobserved factors that simultaneously impact monthly crash risk and subsequent real-time crash risk for segment i.

In the joint framework, we assume a categorical discrete outcome structure for modeling the real-time crash risk component (following (McFadden, 1973). Let l (l = 1, 2) represent the l^{th} crash in the month and and j ($j = 1,2,3, \dots, J$) be the index to represent a 5-minute interval among a set of C_{itlj} alternatives of road segment i for month t specific to time interval j. Thus the real-time crash risk component takes the familiar discrete outcome formulation in linear form as follows[3]:

$$u_{itlj}^{*} = \left((\boldsymbol{\delta} + \boldsymbol{\rho}_{it})\boldsymbol{z}_{itlj} + \xi_{itlj} \pm \boldsymbol{\eta}_{it}\right) \quad (2)$$

where u_{itlj}^{*} is the latent variable of crash risk of time alternative j for l^{th} crash on segment i for month t. Within the traditional discrete outcome framework as presented in equation 2, segment i for month t will have possibility of crash outcome j if $u_{itlj}^{*} > \max_{\substack{d=1,2,3,\dots,J \\ d \neq j}} u_{itld}^{*}$. $\boldsymbol{z}_{itlj}$ is a vector of observed attributes corresponding to crash unit j. $\boldsymbol{\delta}$ is a vector of coefficients to be estimated. $\boldsymbol{\rho}_{it}$ is a vector of unobserved factors on real-time crash risk propensity of segment i for month t for l^{th} crash for time unit j and its associated observed characteristics assumed to be realization from standard normal distribution: $\boldsymbol{\rho}_{it} \sim N(0, \boldsymbol{\pi}_{it}^2)$. ξ_{itlj} is an idiosyncratic error term assumed to be identically and independently standard logistic distributed across roadway segment i for month t, crash l in crash time period j. $\boldsymbol{\eta}_{it}$ term generates the correlation between equations for monthly crash risk and real-time crash risk components. The $\pm$ sign in front of $\boldsymbol{\eta}_{it}$ in equation 2 indicates that the correlation in unobserved individual factors between the monthly crash risk and the subsequent real-time crash risk may be positive or negative. To determine the appropriate sign, one can empirically test the models with both '$+$' and '$-$' signs independently. The model structure that offers the superior data fit is considered as the final model.

It is important to note here that the unobserved heterogeneity between two components of the joint system may vary across observations. Therefore, in the current study, the correlation parameter $\boldsymbol{\eta}_{it}$ is parameterized as a function of observed attributes separately for monthly crash risk and real-time crash risk as follows:

$$\boldsymbol{\eta}_{it} = \boldsymbol{\lambda}_{it}\boldsymbol{w}_{it} \quad (3)$$

[3] Based on recent research by Guevara and Ben-Akiva (2013) there is evidence to suggest that the naïve estimator (i.e. employing random sampling based estimation) offers reasonable accuracy in model estimation for mixed multinomial logit (MMNL) model. Our joint approach builds on this research finding.

$$\boldsymbol{\eta}_{it} = \boldsymbol{\lambda}_{it}\boldsymbol{w}_{itlj} \tag{4}$$

where, $\boldsymbol{w}_{it}$ and $\boldsymbol{w}_{itlj}$ represent exogenous variables that are closely related at the aggregate (i.e. monthly) and disaggregate (time-period) levels, $\boldsymbol{\lambda}_{it}$ is a vector of unknown parameters to be estimated. The reader would note that we allow for different resolutions of $\boldsymbol{w}$ because the aggregate resolution cannot be directly employed in the real-time crash model as there is no variability across time periods (j).

4.2 Model Estimation

In examining the model structure of monthly crash risk and real-time crash risk, it is necessary to specify the structure for the unobserved vector $\boldsymbol{\gamma}_{it}, \boldsymbol{\rho}_{it}\ and\ \boldsymbol{\lambda}_{it}$ represented by Ω. In this paper, it is assumed that all these vectors are independent realizations from normal population distributions. Thus, conditional on $\boldsymbol{\gamma}_{it}\ and\ \boldsymbol{w}_{it}$, the probability of segment i corresponding to crash state k for month t is given by:

$$\begin{aligned} P_{itk}|(\boldsymbol{\gamma}_{it}, \boldsymbol{\lambda}_{it}) = {} & \varphi\{\tau_{itk} - ((\boldsymbol{\beta} + \boldsymbol{\gamma}_{it})\boldsymbol{x}_i + \boldsymbol{\eta}_{it})\} \\ & - \varphi\{\tau_{it,k-1} - ((\boldsymbol{\beta} + \boldsymbol{\gamma}_{it})\boldsymbol{x}_i + \boldsymbol{\eta}_{it})\} \end{aligned} \tag{5}$$

where, φ(·) is the standard logistic cumulative distribution function. Similarly, the probability of roadway segment i representing the real-time crash risk is given by (conditional on $\boldsymbol{\rho}_{it}\ and\ \boldsymbol{w}_{it}$):

$$R_{itlj}|(\boldsymbol{\rho}_{it}, \boldsymbol{\lambda}_{it}) = \frac{exp\big((\boldsymbol{\delta} + \boldsymbol{\rho}_{it})\boldsymbol{z}_{itlj} \pm \boldsymbol{\eta}_{it}\big)}{\sum_{j \in C_{itj}} exp\big((\boldsymbol{\delta} + \boldsymbol{\rho}_{it})\boldsymbol{z}_{itlj} \pm \boldsymbol{\eta}_{it}\big)} \tag{6}$$

Thus the likelihood function for the joint probability can be expressed as:

$$L_i = \int_{\Omega} \left(\prod_{t=i}^{T} \left[\prod_{k=0}^{1} \big(P_{it}|(\boldsymbol{\gamma}_{it}, \boldsymbol{\lambda}_{it})\big)^{d_{ik}} \text{x} \prod_{l} \prod_{j \in C_{itlj}} \left[\Big(R_{itlj}|(\boldsymbol{\rho}_{it}, \boldsymbol{\lambda}_{it})\Big)^{d_{itlj}} \right]^{\varpi_{it}} \right] \right) d\Omega \tag{7}$$

where, d_{ik} is a dummy with $d_{ik} = 1$ for the observed segment crash level, d_{itlj} is a dummy with $d_{itlj} = 1$ for the crash time period and 0 elsewhere; ϖ_{it} is a dummy with $\varpi_{it} = 1$ if roadway segment i has at least one crash in a given month t and 0 otherwise. Finally, the log-likelihood function is:

$$LL = \sum_{i} Ln(L_i) \tag{8}$$

All the parameters in the model are estimated by maximizing the logarithmic function *LL* presented in equation 8. The parameters to be estimated in the model are: $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$. To estimate the proposed model, we apply Quasi-Monte Carlo simulation techniques based on the scrambled Halton sequence to approximate this integral in the likelihood function and maximize the logarithm of the resulting simulated likelihood function across individuals (see Bhat, 2001; Eluru et al., 2008; Yasmin and Eluru, 2013 for examples of Quasi-Monte Carlo approaches in literature). The model estimation routine is coded in GAUSS Matrix Programming software (Aptech, 2015).

5. DATA

5.1 Study Area and Data Sources

Our study draws data from three different expressways from Central Florida: State Roads 408, 417, and 528. For these expressways, data were collected and compiled for 29 months from July, 2013 through December, 2015. However, due to the absence of traffic data in April, 2014, the data from other months are used (traffic detector systems were under maintenance in April, 2014). Four different types of roadway influence areas, as defined by the Highway Capacity Manual (HCM, 2010), are explored in current study context: (1) merge, (2) diverge, (3) weaving and (4) basic influence area. Figure 1 represents the illustrations for merge, diverge and weaving areas. The basic segments are segments on the roadways which are not impacted by merge, diverge, and weaving operations. Traffic and crash related data were collected for different road segments of these four influence areas of expressways other than toll-plaza related segments (toll plazas, and their upstream and downstream segments), segments with less than 500 feet, and segments with no traffic data. Finally, 247 segments were used for data analysis including: 45 merge segments, 48 diverge segments, 25 weaving segments and 129 basic segments.

Data for the empirical study is compiled from four different categories: crash data, geometry data, traffic data and weather data. The crash data is collected from the Signal Four Analytics (S4A) crash database. The database provides detailed information for each crash; such as, crash time, location, and crash type. The geometry data are mainly obtained from the Roadway Characteristic Inventory operated by the Florida Department of Transportation (FDOT). Length of segments are generated by using ArcGIS tool. The traffic data are collected from the Microwave Vehicle Detection System (MVDS) installed by Central Florida Expressway Authority. The MVDS detectors record traffic data including vehicle count, speed and lane occupancy for each lane in one-minute interval. Additionally, the MVDS detectors categorize vehicle into four groups based on vehicles' lengths and provide vehicle count for each group. In current study context, the vehicles with length greater than 24 feet are defined as truck. For weather category, we consider time-of-day attributes gathered from the United States Naval Observatory (USNO) dataset. It documents the time of sunrise and sunset for every day of different cities in the United States. Daytime is assigned to a crash if the crash time is reported to be in between sunset and sunrise time, otherwise, the crash is considered to be occurred during nighttime. We have also considered weather condition (rain and clear) collected from Florida Automated Weather Network (FAWN) records.

5.2 Experimental Design and Data Description

From the crash database, we identify the location and time of the reported crashes for the study period. The dependent variable for monthly crash risk component is defined from crash database for each segment. We develop the monthly crash risk model by using roadway geometry data that includes types of influence area, outer shoulder width, segment length, median width, speed limit and number of lanes. These variables are generated for roadway segment where the crash occurred and also for the closest upstream and closest downstream segments. On the other hand, the real-time crash risk component is designed based on unmatched case-control scheme with a ratio 1:29 for alternative generation. Each set of case-control is designed for multiple crashes (if recorded) in a month over 29-month study period. Controls are the non-crash event or conditions which did not result in a crash or were impacted by a crash. Hence, all conditions which are within five hours before or after a crash were excluded while selecting controls. Thus, for each crash event of a segment for a given month, 29 controls are randomly selected from all possible non-crash events of the same segment and in the same month of the respective crash event. The real-time crash risk model is estimated by using traffic data and weather data. Time-of-day and weather condition are the weather related variables considered in our study. Traffic data included are: traffic count, proportion of trucks, average vehicular speed, standard deviation of vehicular speed and average lane occupancy. These data are aggregated information over 5-minutes intervals. For the crash events, we extracted the traffic data which were 5–10 min prior to crash occurrence (see Xu et al., 2013; Yu and Abdel-Aty, 2013 for similar approach). We have also incorporated traffic data in the form of interactions with the roadway geometry data.

The final dataset, after removing records with missing information for essential attributes, consisted of 6,913 records in the monthly crash risk component. Among these records, 1,401 records have at least one crash for that specific month. The associated data records for real-time crash model component is 58,470 with 1,949 (56,521) number of cases (controls). Table 1 offers a summary of the sample characteristics of the exogenous factors in the final estimation dataset. The table represents the definition of variables considered for final model estimation along with the minimum, maximum and average values for continuous/ordinal variables; and frequency and percentages for indicator variables. The final specification of the model development was based on removing the statistically insignificant variables in a systematic process based on statistical significance (90% significance level). The specification process was also guided by prior research and parsimony considerations. In estimating the models, several functional forms and variable specifications were explored. The functional form that provided the best result was used for the final model specifications and, in Table 1, the variable definitions are presented based on these final functional forms.

6. EMPIRICAL ANALYSIS

6.1 Model Specification and Overall Measures of Fit

The empirical analysis involves estimation of two different models: 1) an independent binary logit (BL) and multinomial logit (MNL) model system, and 2) joint BL-MNL model with correlation parameterization. The independent models (separate BL and MNL models) were estimated to establish a benchmark for comparison. Prior to discussing the estimation results, we compare the performance of these models in this section. We employ the likelihood-ratio (LR) test to determine

the best model between independent and joint models. The LR test statistic for a given empirical model is computed as:

$$LR = 2[LL_U - LL_R] \quad (9)$$

where LL_U and LL_R are the log-likelihood of the unrestricted and the restricted models, respectively.

The log-likelihood values at convergence for the models estimated are as follows: (1) Independent BL-MNL (with 21 parameters) is -8772.67 and (2) joint BL-MNL model with correlation parameterization (with 25 parameters) is -8767.41. The computed value of the LR test is compared with the $\aleph^2$ value for the corresponding degrees of freedom (*dof*). The resulting LR test values for the comparison of independent BL-MNL and joint BL-MNL model is 10.51 (4 *dof*). The LR test values indicate that the joint model outperforms the independent indicating that joint model offers superior fit. The comparison exercise clearly highlights the superiority of the joint model with the correlation parameterization in terms of data fit compared to independent model.

6.2 Estimation Results

In presenting the effects of exogenous variables, we will restrict ourselves to the discussion of the joint model with the correlation parameterization. For the ease of presentation, the monthly crash risk (BL model) and real-time crash risk (MNL model) components are presented and discussed separately. Table 2 presents the estimation results of the joint BL-MNL model with BL component results in the first row panel of the table, and MNL component results in the second row panel. The correlation parameters within joint model specification are presented in the last row panel of Table 2.

6.2.1 Monthly Crash Risk Component – (BL Model)

In the BL model, the positive (negative) coefficient corresponds to increased (decreased) crash risk propensity. The positive sign on the threshold term does not have any substantive interpretation.

In terms of roadway segment characteristics, the result associated with length of segment, a surrogate for exposure, indicates that as segment length increases, the likelihood of crash risk also increases (see Anastasopoulos and Mannering, 2009 for similar results). The results associated with median width and outer shoulder width are found to have significant impact on crash likelihood of roadway segments. Increased median width of roadway segment is negatively associated with crash risk, perhaps indicating higher scope of driving error correction and/or scope of accommodation for vehicles in the event of impending crashes in the presence of wider median (Haleem et al., 2013; Shi et al., 2016). The model estimation results indicate an expected negative correlation of higher outer shoulder width with higher likelihood of crash risk, a result also observed in several previous studies (Noland and Oh, 2004; Xu et al., 2013). Wider shoulder provides spaces for errant vehicle and in turn may reduce the likelihood of crash risk (Bonneson and Pratt, 2009).

In BL component of the joint model system, a higher number of lanes of the mainline segment has significant impact on crash risk. We find that in presence of higher number of lanes, the possibility of crash risk increases. This is potentially because higher number of lanes also

serves as surrogate for higher exposure. With regards to average vehicular speed, higher vehicular speed is found to be negatively associated with crash risk in the BL component. At the same time, higher variation of traffic speed is found to increase the likelihood of crash risk (see Taylor et al., 2000 for similar result).

The results for influence area reveal that the likelihood of crash risk is higher for merge and diverge influence areas of roadway segment relative to basic and weaving influence areas. It is interesting to note that, within these two indicator variables, diverge influence area has a larger impact relative to merge influence area. In general, the results can be explained by high competition of spaces for merging and diverging operations in traffic streams at these influence areas relative to other roadway areas (Mergia et al., 2013; Wu et al., 2013). The result for posted speed limit indicates that the likelihood of crash risk is higher for segments with speed limit 55 and 65 mph relative to segments with 70 mph, plausibly indicating higher interactions of vehicles at a lower speed environment compared to the roadway environment with higher posted speed limits.

Characteristics of the closest upstream and downstream section are also found to have significant impact on the likelihood of static crash risk in the study. With regards to closest upstream section, median width has positive impact indicating that the higher median width of the closest upstream section is likely to increase crash risk for the studied road section. It is possible drivers upstream with higher median width tends to drive less cautiously and are likely to take longer to slow down in an impending crash situation. Increasing distance from the nearest merging upstream section is found to be negatively associated with crash risk of the studied roadway sections. The result is perhaps indicative of the potential reduction in vehicle weaving conflicts in the absence of merging section. In terms of the characteristics of the closest downstream section, as expected a larger merging influence area of the downstream section is found to have negative association with the aggregate level crash risk of the studied road locations.

6.2.2 Real-time Crash Risk Component – (MNL Model)

In MNL model component, the positive (negative) coefficient corresponds to increased (decreased) crash risk.

From the estimation results of real-time crash risk component, we can observe that potential crash risk increases with increasing average volume of traffic streams (see Christoforou et al., 2011 for similar results). Proportion of heavy vehicle volume is found to be a significant determinant of crash prone condition. The estimate for heavy vehicle proportion has a positive coefficient suggesting that presence of more heavy vehicles in traffic stream are likely to incur disruptive condition leading to crash occurrences.

As found in previous studies (Xu et al., 2014), we also find that the likelihood of crash prone condition decreases with increasing average vehicular speed. The result can be explained by smooth flow of traffic in high vehicular speed environment. On the other hand, standard deviation of average vehicular speed has positive impact on real-time crash risk component. The result for average occupancy reveals that higher average occupancy is a significant indicator of hazardous traffic condition which may lead to crash occurrence. A similar positive relationship between average occupancy and crash prone condition is documented by several previous studies (Abdel-Aty et al., 2004; Zheng et al., 2010).

The potential crash risk is higher during daytime relative to nighttime period. In real-time crash risk component, we also find that the daytime indicator variable results in a parameter that

is normally distributed with mean 0.175 and standard deviation 0.587, which indicates that the crash risk for daytime is positive for 61.79% of the cases and negative for 38.21% of the cases.

Several interaction terms between dynamic traffic and static segment-specific attributes are found to have significant impact on the real-time crash risk component of the joint model. The interaction terms reveal interesting results. From Table 2, we can see that interaction of merge area with higher vehicular speed increases the likelihood of crash occurrence. The result for interactions of posted speed limit with dynamic traffic attributes indicate that the likelihood of crash prone condition is higher for segments with speed limit 55 and 65 mph and with higher average volume of traffic stream. On the other hand, higher vehicle proportion on a roadway with speed limit 55 and 65 mph are likely to reduce the crash risk.

6.2.3 Common Unobserved Effects

Significance of the unobserved heterogeneity parameters presented in the last row panel of Table 2 highlights the presence of common unobserved factors affecting monthly crash risk and real-time crash risk components. The reader would note that we cannot have the same variable across the two model components. Hence, we considered variables with different resolutions from different components of the joint model. From estimation results, we observe that the two crash risk components are correlated based on observed exogenous attributes. In terms of exogenous variables, we find that the correlation between the two dimensions of the joint model system are moderated by: (1) average vehicular speed over a month from static component and average vehicular speed over 5 minutes from dynamic component; and (2) speed limit 55 and 65 mph from static component and standard deviation of vehicular speed over 5 minutes from real-time component. This supports our hypothesis that static and real-time crash risks components are correlated in nature. The correlation parameters are introduced with a " $+$ " sign before $\boldsymbol{\eta}_{it}$ in the real-time crash risk component (as described in econometric framework section) since these provided a substantially better fit compared to introducing them with a " $-$ " sign. Overall, the results highlight that accommodating for common unobserved effects across the two model components improves the model fit substantially.

7. CONCLUSIONS

The paper proposed, formulated and estimated a joint reactive and proactive crash modeling framework by coupling the static monthly crash risk and dynamic real-time crash risk in a unified econometric framework for a microscopic analysis unit. The research effort bridges the gap between traditional crash risk and real-time crash risk models by developing a joint model that accommodates for both dimensions in developing crash risk analysis models. In the joint modeling approach, we estimated an alternative to the case-control binary logit based real-time crash risk analysis by employing a multinomial logit based approach where time periods serve as alternatives and the chosen alternative is the time period in which crash occurs. The joint model also allowed us to accommodate for the common unobserved factors that increase the likelihood of a crash in microscopic unit to affect the real-time crash risk propensity. To the best of the authors' knowledge, this is the first attempt to employ such a joint framework for examining micro-level crash count events.

We demonstrated the application of the proposed approach by using data on roadway segments from three expressways in Central Florida (State Roads 408, 417, and 528) for 29

months. The monthly crash risk component was examined by using binary logit model employing different static roadway attributes (roadway geometry and operational attributes). The real-time crash risk component was examined by using a random utility model employing different dynamic traffic attributes (volume, speed, lane occupancy and environmental conditions). The real-time crash risk component was designed based on unmatched case-control scheme with a ratio 1:29 for alternative generation. The empirical analysis involved estimation of two different models: 1) an independent binary logit (BL) and multinomial logit (MNL) model system, and 2) joint BL-MNL model with correlation parameterization. The independent models (separate BL and MNL models) were estimated to establish a benchmark for comparison. The comparison exercise based on Bayesian Information Criterion clearly highlighted the superiority of the joint model with the correlation parameterization in terms of data fit compared to independent model. From estimation results, we observed that two crash risk components are correlated based on different observed exogenous attributes. The outcome of the proposed approach allows us to predict both the static and dynamic crash risk simultaneously in a single econometric framework.

The study is not without limitation. In examining, the real-time crash risk component, we did not consider the traffic characteristics of the closest downstream or upstream sections. It might be beneficial in future to consider those attributes in developing real-time crash risk model given the availability of the information. Moreover, we examined the monthly crash risk component by employing a binary logit model. It would be interesting to examine the aggregate level crash risk component by using either count regression or ordered logit modeling based approach.

Acknowledgements

The authors would also like to gratefully acknowledge Signal Four Analytics (S4A) for providing access to Florida crash data. The authors would also like to thank research funding from Florida Department of Transportation.

References

Abdel-Aty, M., and Pande, A., 2007. Crash data analysis: Collective vs. individual crash level approach. *Journal of Safety research* 38, 581-587.

Abdel-Aty, M., Cunningham, R., Gayah, V., and Hsia, L., 2008. Dynamic variable speed limit strategies for real-time crash risk reduction on freeways. *Transportation Research Record: Journal of the Transportation Research Board*, 108-116.

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., and Santos, C.D., 2007. Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways. *Journal of Intelligent Transportation Systems* 11, 107-120.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*, 88-95.

Ahmed, M.M., and Abdel-Aty, M.A., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. *IEEE Transactions on Intelligent Transportation Systems* 13, 459-468.

Ahmed, M.M., and Abdel-Aty, M.A., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. *IEEE Transactions on Intelligent Transportation Systems* 13, 459-468.

Anastasopoulos, P.C., and Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention* 41, 153-159.
Aptech 2015. 2015. Aptech Systems Inc, accessed from http://www.aptech.com/ on September 19th 2015.
Basso, F., Basso, L.J., Bravo, F., and Pezoa, R., 2018. Real-time crash prediction in an urban expressway using disaggregated data. *Transportation Research Part C* 86, 202-219.
Ben-Akiva, M.E., and Lerman, S.R., 1985. *Discrete choice analysis: theory and application to travel demand*. MIT press.
Bhat, C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B* 35, 677-693.
Bonneson, J.A. and Pratt, M.P., 2009. Roadway safety design workbook (No. FHWA/TX-09/0-4703-P2). Texas Transportation Institute, Texas A & M University System.
Bruce, N., Pope, D., and Stanistreet, D., 2008. *Quantitative methods for health research: a practical interactive guide to epidemiology and statistics*. John Wiley & Sons.
Calabrese, R., and Osmetti, S.A., 2013. Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. *Journal of Applied Statistics* 40, 1172-1188.
Chang, I., and Kim, S.W., 2012. Modelling for identifying accident-prone spots: Bayesian approach with a Poisson mixture model. *KSCE Journal of Civil Engineering* 16, 441-449.
Chen, C., Wang, Y., Ma, C., and Zhang, W., 2016. How Expressway Geometry Factors Contribute to Accident Occurrence? A Binary Logistic Regression Study. *Periodica Polytechnica. Transportation Engineering* 44, 215.
Christoforou, Z., Cohen, S., and Karlaftis, M.G., 2011. Identifying crash type propensity using real-time traffic data on freeways. *Journal of Safety research* 42, 43-50.
Ding, C., and Gou, C., 2016. How Expressway Characteristic Factors Contribute to Accident Counts, *CICTP 2016*, pp. 1714-1729.
Dinu, R.R., and Veeraragavan, A., 2011. Random parameter models for accident prediction on two-lane undivided highways in India. *Journal of Safety Research* 42, 39-42.
Eluru, N., Bhat, C.R., and Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis & Prevention* 40, 1033-1054.
Faghih-Imani, A., and Eluru, N., 2015. Analysing bicycle-sharing system user destination choice preferences: Chicago's Divvy system. *Journal of transport geography* 44, 53-64.
Geurts, K., and Wets, G., 2003. Black spot analysis methods: Literature review. Accessed from http://hdl.handle.net/1942/5004, July 16th, 2016.
Guevara, C. A. and Ben-Akiva, M. E., 2013. Sampling of alternatives in Logit Mixture models. *Transportation Research Part B*58, 185-198.
Haleem, K., Gan, A., and Lu, J., 2013. Using multivariate adaptive regression splines (MARS) to develop crash modification factors for urban freeway interchange influence areas. *Accident Analysis & Prevention* 55, 12-21.
Hariharan, B., Hong, J., Shankar, V., Venkataraman, N., Milton, J.C. and Van Schalkwyk, I., 2016. Roadside Geometry Effects on the Overdispersion Parameter (No. 16-5892). *Transportation Research Board 95th Annual Meeting*.
HCM. (2010). Highway Capacity Manual 2010 (HCM2010). Transportation Research Board, National Research Council, Washington, DC.
King, G., and Zeng, L., 2001. Logistic regression in rare events data. *Political analysis* 9, 137-163.

Lee, C., Saccomanno, F., and Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. *Transportation Research Record: Journal of the Transportation Research Board*, 1-8.
Lord, D., and Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A* 44, 291-305.
Mann, C., 2003. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emergency Medicine Journal* 20, 54-60.
Mannering, F.L., and Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research* 1, 1-22.
McFadden, D., 1973. *Conditional logit analysis of qualitative choice behavior*. University of California at Berkeley. Berkeley, California.
Mergia, W.Y., Eustace, D., Chimba, D., and Qumsiyeh, M., 2013. Exploring factors contributing to injury severity at freeway merging and diverging locations in Ohio. *Accident Analysis & Prevention* 55, 202-210.
Noland, R.B., and Oh, L., 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: A case study of Illinois county-level data. *Accident Analysis & Prevention* 36, 525-532.
Oh, C., Oh, J.-S., Ritchie, S., and Chang, M., 2001. Real-time estimation of freeway accident likelihood, *80th Annual Meeting of the Transportation Research Board, Washington, DC*.
Park, B.J., Lord, D. and Wu, L., 2016. Finite mixture modeling approach for developing crash modification factors in highway safety analysis. Accident Analysis & Prevention, 97, pp.274-287.
Qu, X., Wang, W., Wang, W., Liu, P., and Noyce, D.A., 2012. Real-time prediction of freeway rear-end crash potential by support vector machine, *Transportation Research Board 91st Annual Meeting*.
Roshandel, S., Zheng, Z., and Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident Analysis & Prevention* 79, 198-211.
Saccomanno, F.F., Grossi, R., Greco, D., and Mehmood, A., 2001. Identifying black spots along highway SS107 in southern Italy using two models. *Journal of Transportation Engineering* 127, 515-522.
Savolainen, P.T., Mannering, F.L., Lord, D., and Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accident Analysis & Prevention* 43, 1666-1676.
Scott, D.M., He, and S.Y., 2012. Modeling constrained destination choice for shopping: a GIS-based, time-geographic approach. *Journal of transport geography* 23, 60-71.
Shi, Q., Abdel-Aty, M., and Lee, J., 2016. A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. *Accident Analysis & Prevention* 88, 124-137.
Sun, J., and Sun, J., 2015. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C* 54, 176-186.
Taylor, M.C., Lynam, D., and Baruya, A., 2000. *The effects of drivers' speed on the frequency of road accidents*. Transport Research Laboratory Crowthorne.
Theofilatos, A., Yannis, G., Kopelias, P., Papadimitriou, F., 2018. Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accident Analysis & Prevention*.
Thompson, W.D., Kelsey, J.L., and Walter, S.D., 1982. Cost and efficiency in the choice of matched and unmatched case-control study designs. *American journal of epidemiology* 116, 840-851.

Train, K.E., 2009. *Discrete choice methods with simulation*. Cambridge university press.
Waddell, P., Bhat, C., Eluru, N., Wang, L., and Pendyala, R., 2007. Modeling interdependence in household residence and workplace choices. *Transportation Research Record: Journal of the Transportation Research Board*.
Wang, L., Abdel-Aty, M., and Lee, J., 2017a. Safety analytics for integrating crash frequency and real-time risk modeling for expressways. *Accident Analysis & Prevention* 104, 58-64.
Wang, L., Abdel-Aty, M., Lee, J., and Shi, Q., 2017b. Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors. *Accident Analysis & Prevention forthcoming*.
Wang, L., Abdel-Aty, M., Shi, Q., and Park, J., 2015. Real-time crash prediction for expressway weaving segments. *Transportation Research Part C* 61, 1-10.
Wooldridge, J.M., 2010. *Econometric analysis of cross section and panel data*. MIT press.
Wu, Y., Abdel-Aty, M., and Lee, J., 2018. Crash risk analysis during fog conditions using real-time traffic data. *Accident Analysis & Prevention* 114, 4-11.
Wu, Y., Nakamura, H., and Asano, M., 2013. A comparative study on crash-influencing factors by facility types on urban expressway. *Journal of Modern Transportation* 21, 224-235.
Xie, M., Cheng, W., Gill, G.S., Falahati, R., Jia, X., and Choi, S., 2017. Predicting Likelihood of Hit-and-Run Crashes Using Real-Time Loop Detector Data and Hierarchical Bayesian Binary Logit Model with Random Effects. *Transportation Research Board 96th Annual Meeting*.
Xu, C., Liu, P., and Wang, W., 2016. Evaluation of the predictability of real-time crash risk models. *Accident Analysis & Prevention* 94, 207-215.
Xu, C., Liu, P., Wang, W., and Li, Z., 2014. Identification of freeway crash-prone traffic conditions for traffic flow at different levels of service. *Transportation Research Part A* 69, 58-70.
Xu, C., Tarko, A.P., Wang, W., and Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention* 57, 30-39.
Xu, C., Wang, W., Liu, P., and Li, Z., 2015. Calibration of crash risk models on freeways with limited real-time traffic data using Bayesian meta-analysis and Bayesian inference approach. *Accident Analysis & Prevention* 85, 207-218.
Yasmin, S., and Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis & Prevention* 59, 506-521.
Yasmin, S., and Eluru, N., 2016. Latent segmentation based count models: analysis of bicycle safety in Montreal and Toronto. *Accident Analysis & Prevention* 95, 157-171.
You, J., Zhang, L., Fang, S., and Guo, J., 2017. Real-time crash prediction based on high definition monitoring systems, *Intelligent Transportation Engineering (ICITE)*, 2017 2nd IEEE International Conference on. IEEE, 208-211.
Yu, R., and Abdel-Aty, M., 2013. Multi-level Bayesian analyses for single-and multi-vehicle freeway crashes. *Accident Analysis & Prevention* 58, 97-105.
Zhang, J., and Kai, F.Y., 1998. What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *Jama* 280, 1690-1691.
Zheng, Z., Ahn, S., and Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention* 42, 626-636.

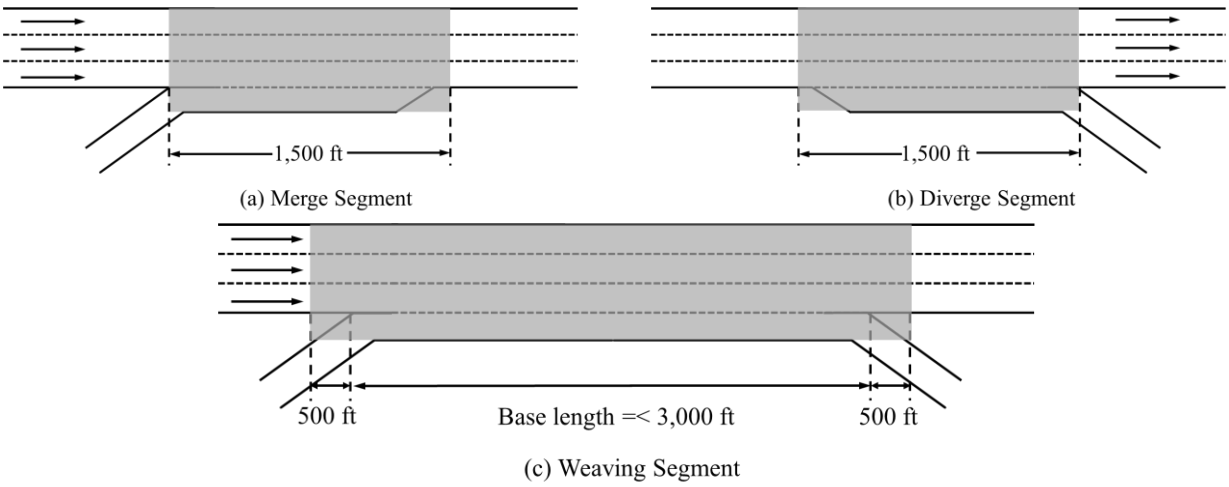

Figure 1: Influence Area of Merge (a), Diverge (b), and Weaving (c) Segments

Table 1: Sample Statistics for the Expressways

SAMPLE STATISTICS FOR MONTHLY CRASH RISK COMPONENT				
Continuous/Ordinal Variables				
Variables	**Variable definitions**	**Minimum**	**Maximum**	**Average**
Characteristics of the roadway section				
Segment length	**Ln(Segment length in feet)**	6.428	10.227	7.761
Median width	**Median width in feet**	16.000	64.000	46.747
Outside shoulder width	**Outer shoulder width in feet**	2.000	12.000	9.641
Inner shoulder width	**Inner shoulder width in feet**	4.000	24.000	7.277
Number of lanes	**Count of total number of lanes**	2.000	5.000	2.668
Average vehicular speed over a month	**Average vehicular speed over a month in mph**	15.956	78.337	66.851
Standard deviation of vehicular speed over a month	**Standard deviation of vehicular speed of over a month in mph**	0.370	34.612	4.235
Characteristics of the closest upstream section				
Median width	**Ln(Median width of the closest upstream segment in feet)**	2.773	4.159	3.784
Outer should width	**Outer should width of the closest upstream section in feet**	2.000	12.000	9.698
Inner should width	**Inner should width of the closest upstream section in feet**	4.000	36.000	7.290
Number of lanes	**Count of total number of lanes of the closest upstream section**	2.000	5.000	2.668
Distance to merging ramp	**Ln(Distance of the roadway segment from the closest merging ramp in upstream in feet)**	0.000	10.474	4.826
Distance to diverging ramp	**Ln(Distance of the roadway segment from the closest diverging ramp in upstream in feet)**	0.000	9.733	3.052
Characteristics of the closest downstream section				
Median width	**Ln(Median width of the closest downstream segment in feet)**	2.773	4.159	3.782
Outer should width	**Outer should width of the closest downstream section in feet**	2.000	13.000	9.698
Inner should width	**Inner should width of the closest downstream section in feet**	0.000	36.000	7.251
Number of lanes	**Count of total number of lanes of the closest downstream section**	2.000	5.000	2.669
Indicator Variables				
Variables	**Frequency**	**Percentage**		

Characteristics of the roadway segment		
Influence area		
Merge area	1275.000	18.444
Diverge area	1338.000	19.355
Weaving area	667.000	9.648
Basic area	3633.000	52.553
Speed limit in mph		
Speed limit 55 and 65 mph	750.000	10.849
Speed limit 70 mph	2172.000	31.419
Characteristics of the closest upstream section		
Upstream influence area		
Merge section	4136.000	59.829
Diverge section	2777.000	40.171
Characteristics of the closest downstream section		
Downstream influence area		
Merge section	3006.000	43.483
Diverge section	3878.000	56.097

SAMPLE STATISTICS FOR REAL-TIME CRASH RISK COMPONENT

Continuous/Ordinal Variables

Variables	Variable definitions	Minimum	Maximum	Average
Traffic count	**Ln(Traffic count)**	0.693	6.887	4.464
Proportion of trucks	**Count of trucks/Total traffic counts**	0.000	1.000	0.135
Average vehicular speed over 5 minutes	**Average vehicular speed over 5 minutes in mph**	1.437	108.167	63.407
Standard deviation of vehicular speed over 5 minutes	**Standard deviation of vehicular speed over 5 minutes in mph**	0.000	36.403	2.547
Average lane occupancy	**Average lane occupancy (%)/10**	0.000	5.723	0.445
Interaction terms				
Merge area*Average vehicular speed over 5 minutes		0.000	84.667	6.445
Speed limit 55 and 65 mph*Traffic count		0.000	6.887	2.137

Speed limit 55 and 65 mph*Proportion of trucks	0.000	1.000	0.055
Indicator Variables			
Variables	**Frequency**	**Percentage**	
Time of day			
Daytime	31293.000	53.520	
Nighttime	27177.000	46.480	
Weather condition			
Rain	3916.000	6.715	
Clear	54554.000	93.285	

Table 2: Joint Monthly Crash Risk-Real Time Crash Risk Model Results

MONTHLY CRASH RISK COMPONENT – BL MODEL		
Variables	**Estimate**	**t-stat**
Threshold (between zero and non-zero crash states)	13.940	7.680
Characteristics of the roadway section		
Segment length	1.471	7.838
Median width	-0.023	-3.300
Outside shoulder width	-0.064	-1.847
Number of lanes	0.324	5.316
Average vehicular speed over a month	-0.026	-2.037
Standard deviation of vehicular speed over a month	0.213	7.274
Influence area (Base: Weaving and Basic area)		
Merge area	0.626	4.154
Diverge area	1.142	6.342
Speed limit in mph (Base: Speed limit 70 mph)		
Speed limit 55 and 65 mph	0.577	3.960
Characteristics of the closest upstream section		
Median width	0.564	2.129
Distance to merging ramp	-0.046	-3.946
Characteristics of the closest downstream section		
Downstream influence area		
Merge section	-0.226	-2.497
REAL-TIME CRASH RISK COMPONENT – MNL MODEL		
Variables	**Estimate**	**t-stat**
Traffic count	0.530	8.934
Proportion of trucks	2.149	6.837
Average vehicular speed over 5 minutes	-0.066	-8.029
Standard deviation of vehicular speed over 5 minutes	0.117	8.776
Average lane occupancy	0.309	4.036
Time of day		
Daytime	0.175	2.053
SD for Daytime	0.587	2.641
Interaction terms		
Merge area*Average vehicular speed over 5 minutes	0.014	1.846
Speed limit 55 and 65 mph*Traffic count	0.299	4.171
Speed limit 55 and 65 mph*Proportion of trucks	-1.351	-2.393
CORRELATION PARAMETER		
Variables (Monthly crash risk component/Real-time crash risk component)	**Estimate**	**t-stat**
Average vehicular speed over a month/Average vehicular speed over 5 minutes	0.015	2.273
Speed limit 55 and 65 mph/Standard deviation of vehicular speed over 5 minutes	0.064	1.833