STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

TR0003 (REV 10/98)

Lock Data on Form

1. REPORT NUMBER CA20-3290	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER

4. TITLE AND SUBTITLE

Development of Active Transportation Safety Performance Functions in California

5. REPORT DATE

January 31, 2021

6. PERFORMING ORGANIZATION CODE

7. AUTHOR

Wen Cheng, Ph.D. ; Xudong Jia, Ph.D.; Hairui Tang, MSCE; Mankirat Singh, MSCE.

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Cal Poly Pomona Foundation, Inc.
3801 W. Temple Avenue
Pomona, CA 91768-2557

10. WORK UNIT NUMBER

11. CONTRACT OR GRANT NUMBER

65A0705

12. SPONSORING AGENCY AND ADDRESS

California Department of Transportation
1727 30th Street 1227 O Street
Sacramento CA 95816

13. TYPE OF REPORT AND PERIOD COVERED

Final Report

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Encouraging individuals to indulge in active transportation, involving walking and bicycling, brings with it a societal obligation to protect commuters as they engage in these modes of travel. Investigating the factors impacting non-motorist safety on roadways is essential in this regard. The most popular method is to develop the safety performance function (SPF). The project is dedicated to the active transportation-oriented SPF development in California, which can not only illustrate the significant influential factors, but also paves the way for other activities such as accurate identification of hot spots of active transportation, countermeasure treatment selection, benefit-cost analysis, etc. The active transportation-related SPFs were developed for the micro-level, county-level, and traffic analysis zone-level crash counts, respectively.

17. KEY WORDS

Active Transportation; Safety Performance Function; Micro-level; County; Traffic Analysis Zone.

17. DISTRIBUTION STATEMENT

No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. SECURITY CLASSIFICATION *(of this report)* Unclassified	20. NUMBER OF PAGES 65	21. COST OF REPORT CHARGED N/A

Reproduction of completed page authorized.

DISCLAIMER STATEMENT

This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. This publication does not constitute a standard, specification or regulation. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in Braille, large print, audiocassette, or compact disk. To obtain a copy of this document in one of these alternate formats, please contact: the Division of Research and Innovation, MS-83
California Department of Transportation, P.O. Box 942873, Sacramento, CA 94273-0001

Development of Active Transportation Safety Performance Functions in California

Final Technical Report

Prepared by California State Polytechnic University Pomona

for the

California Department of Transportation

January 31, 2021

ACKNOWLEDGMENTS

Financial support for this study was provided by Caltrans (Grant number: 65A0705). We are indebted to Safe Transportation Research and Education Center (SafeTREC) from UC Berkeley for the provision of estimated pedestrian counts. Other data provided by Caltrans, Southern California Association of Governments (SCAG), and Highway Safety Information System (HSIS) are also greatly appreciated.

TABLE OF CONTENTS

1. INTRODUCTION

Walking activity has been considered as an important travel mode due to its immense benefits. However, the percentage of total trips undertaken by pedestrians compared to other modes is very low. In the United States, the National Household Travel Survey (NHTS, 2018) reported that trips made by walking accounted for only 0.6% of total person-miles travel (PMT). Past studies reveal that out of many reasons for the relatively low distance traveled by walking, one main reason is that the pedestrians are among the most vulnerable and unsafe road users from the viewpoint of traffic crashes (De Hartog et al., 2010; Retting et al., 2003). According to the national statistics, (Governor Highway Safety Association, 2019), a total of 6590 pedestrian fatalities and around 70,000 injuries were estimated in pedestrian-vehicle crashes in 2019. These consequences create an urgent need to prevent pedestrian-involved crashes by the implementation of better policies and strategies to provide safe traffic environment for active transportation commuters.

Given such context, many researchers have investigated various factors such as roadway-built characteristics (Mansfield et al., 2018, Miranda-Moreno et al., 2011), pedestrian behavior (Jing Xu et al., 2018; Dommes et al., 2014; Mwakalonge et al., 2015), driver behavior (Baker et al., 1974; Geruschat and Hassan, 2005; Schroeder and Rouphail, 2011), traffic characteristics (Shi et al., 2007; Barton et al., 2011), drug/alcohol use (Li et al., 2019; Plurad et al., 2006), social and demographic attributes (LaScala et al., 2000; Tabibi et al., 2012; Ryb et al., 2007) to highlight the crucial insights related to pedestrian-related crashes. Among the distinct strategies, the development of safety performance functions (or, crash frequency models) is one of the most popular strategies to address traffic safety which not only aids in screening out the significant influential factors, but also predicting the crash counts for various purposes (Ukkusuri et al., 2011; Wu et al., 2018; Harwood et al., 2008). Poisson regression models were initially widely adopted due to popularity of Poisson distribution for discrete outcomes (Miranda-Moreno, 2006). However, Poisson regression models are not able to provide reliable and unbiased results in the case of over-dispersion (i.e., variance greater than the associated mean). In response, researchers employed alternate model formulations such as Poisson gamma or negative binomial (Hauer, 2001), Poisson lognormal (Park and Lord, 2007), zero inflated models (Aguero-Valverde, 2013), and others, which can better address over-dispersion issues and hence provide more valid inferences.

Among the above-mentioned models, univariate model framework has seen widespread applications in traffic safety studies due to its ease of implementation with only one dependent variable being involved (Anarkooli et al., 2019). However, univariate model is not capable to address the unobserved heterogeneities shared by various crash types or severities occurring in the same locations or situations (Mannering and Bhat, 2014). To overcome this issue, multivariate models have been proposed owing to their enhanced capabilities to tackle the common heterogeneity among different crash types via the explicit consideration of correlated random effects (Lee et al., 2015; Park and Lord, 2007). As a special case of multivariate setting, bivariate model is dedicated to the crashes of two categories and also enjoys frequently applications. For example, Russo et al. (2014) used bivariate framework to examine the factors pertaining to crash injury severities involved in angled collisions. The results demonstrate that bivariate models provide more insightful findings related with the factors influencing the propensity of crashes. Zheng and Sayed (2019) developed bivariate models to integrate the traffic conflict indicators for crash estimation. The finding showed that bivariate model improved the crash estimation precision and accuracy.

Another dimension of model classification resides in the transportation modes. For instance, as previously stated, the SPF can be divided into motor vehicle-oriented, non-motorist-centered, and so on. Overall, the vehicle-related SPF dominates the SPF development given the largest proportion of such mode in the current transportation system. However, with constant promotion of active transportation provided by various levels of government agencies in the past decades, ever-increasing interests were directed toward the SPF development for pedestrians or bicyclists (Wier et al., 2009; Rasciute and Downward, 2010). For example, Thomas (2013) investigated road crashes involving pedestrians and bicyclists based on a very short period of volume count data available for these two modes. The results showed that the risk for an individual pedestrian/bicyclist to be involved in a crash decreases with an increase in the number of pedestrians/bicyclists. Subsequently, McArthur et al. (2014) conducted a study to develop SPF to estimate the pedestrian crashes over a five-year data including socioeconomic and demographic characteristics. Gates et al. (2016) developed SPFs for pedestrian and bicyclist crashes at road segments and intersections. The results demonstrated that the pedestrian and bicycle crashes tend to increase with vehicle traffic volumes being increased. A common among these papers is the lack of or very limited exposure information directly related to the active transportation modes such as pedestrian counts. The potential reasons for such data scarcity are due to lack of definite paths or routes followed by pedestrian and bicyclists, limited use of emerging technologies (e.g., crowdsourcing), expensive data collection devices, and so on. To address these issues, different strategies have been used in the past. Some studies employed daily vehicle miles traveled as the proxy for active transportation modes based on the assumption that most nonmotorist-pertinent collisions are related with vehicles (Cheng et l., 2018). Others relied on the formulation of the active transportation volume models using predictors such as land use, transportation system attributes, and neighborhood socioeconomic characteristics. One recent example is the pedestrian count model develop by UC Berkeley researchers (Griswold et al., 2019), which can be used in the pedestrian SPF as a major estimation of pedestrian exposure, rather than other proxy information.

Thus far, the previously mentioned models follow into the paradigm of the parametric ones. Some studies in traffic safety observed the superiority of nonparametric and/or semiparametric models to address the unobserved heterogeneity (Heydari et al., 2016; Shirazi et al., 2016). Regarding research dedicated to active transportation, the recent study by Heydari et al. (2017) proposed the Dirichlet process mixture (Ohlssen at al., 2007) to develop a flexible latent class model for joint analysis of pedestrian and cyclist injuries at the micro-level of intersections. The authors observed that the flexible approach was advantageous as it demonstrated superior predictive performance and better capability to capture the correlated crash data which eventually provided more accurate interpretation of influential factors for improvement of safety environment.

Building upon the previous research studies, this project aims to utilize various model frameworks to develop Safety Performance Functions (SPFs) of different modes for both motorized and non-motorized ones. In addition, given the different characteristics associated with the various units, the crash frequency models for macro-level consisting of 58 counties in CA and 202 traffic analysis zones (TAZ) from City of the Irvine as the example, and micro-level including intersection and ramps. Finally, to demonstrate the strengths and weaknesses of different modeling frameworks, distinct methodologies and evaluation techniques are employed across the various transportation modes and spatial units.

Given the data availability, the SPFs of the micro-level are developed based on pedestrian- and vehicle-involved collisions at intersections and ramps. First, bivariate models were used to account for the common unobserved heterogeneity shared by the pedestrian- and vehicle-related crashes at the same intersections and ramps. Second, in consideration of the crucial impact of pedestrian volume to SPF of pedestrian, this project investigated pedestrian volume at 6,000 intersections, but still lacks sufficient data for other remaining 15,000 intersections and the whole ramps. To overcome this issue, an adjustment factor was introduced in this research, which is able to replace the impact of pedestrian volume in the statistical models. Third, both variable importance ranking technique and correlation analyses were employed to determine the features to be fed into the models, which are different for each of the statistical models. Such practice leads to the proper covariates not only striking a balance between multi-collinearity and omitted variable bias issues, but also enhancing model flexibility with different inputs to specific transportation mode. Fourth, in comparison with the typical Bayesian hierarchical models based on the Markov chain Monte Carlo (MCMC) algorithm, the integrated nested Laplace approximation (INLA) approach was selected due to faster calculation and more robust results (Taylor and Diggle, 2014). Finally, for a comprehensive comparison of the predictive accuracy of the models, distinct goodness-of-fit measurements which include DIC (deviance information criterion), Dbar (posterior mean deviance), Pd (effective number of parameters) and LPML (log pseudo marginal likelihoods) were employed.

For SPFs of the county level, the authors proposed two multivariate spatial-temporal models to analyze the modal crash data: one with fixed time trend applied to all modes; the other with mode-varying time trend coefficients. These models were then compared with three types of multivariate models used in the past including multivariate without temporal and spatial random effects, multivariate spatial, and multivariate temporal. The major objective of this macro-level SPFs is to examine the benefits of the relatively newly proposed models which have substantially increased computational cost since both dimensions of time and space are considered, as well as their interactions. Moreover, the relative goodness-of-fit or prediction performance among the alternate models were also evaluated with different evaluation criteria of varying complexity, namely: deviance information criterion (DIC), mean absolute deviations (MAD), mean-squared predictive error (MSPE), the G^2 statistic, residual sum of squares (RSS), and total rank difference (TRD). Overall, four different modes (motor-vehicle only, pedestrian-involved, bicyclist-involved and motorcyclist-related) were investigated at the macro level of counties of California.

Lastly, for SPFs of the macro level of Irvine TAZs, the authors adopted semi-parametric formulation that accounts for the unobserved heterogeneity by combining the strengths of incorporating bivariate specification of dependency among the two active transportation crash modes (pedestrian and bicyclists), spatial random effects for the impact of neighboring areas, and Dirichlet process mixture for random intercepts. Four alternate models were developed for comparison based on the goodness-of-fit and predictive accuracy. The models were evaluated by employing different criteria, namely: LPML (log pseudo marginal likelihood), MSPE (mean-squared predictive error), the Rp^2 statistic, the G^2 statistic, and RSS (residual sum of squares).

2. METHODOLOGY

As stated above, a large number of modeling methods and evaluation techniques are utilized for SPFs of the disparate spatial units and transportation modes. For ease of description, the presentation of SPF methodologies are provided in the order of micro level, macro level of California counties, and macro level of Irvine TAZs.

2.1 SPF of the Micro Level

For this type of SPF development, this project employed both negative binomial model and Bayesian joint hierarchical model. The following subsections cover the corresponding methodological details in order.

2.1.1 Bayesian Joint Hierarchical Model Specification

This project employed Poisson lognormal model which assumes the crash count to follow Poisson distribution with the logarithm of poison rate following normal distribution. The model formulation is shown as follows (Cheng et al., 2018).

$$\boldsymbol{y} \sim Poisson\ (\boldsymbol{\lambda}) \tag{1}$$

$$\ln(\boldsymbol{\lambda}) = \boldsymbol{\beta_0} + \boldsymbol{\beta X} + \boldsymbol{\varepsilon} \tag{2}$$

Where $\mathbf{y}$ is a matrix consisting of crash counts of both modes at different intersections, $\boldsymbol{\lambda}$ is a matrix consisting of the corresponding Poisson rates of different modes and intersections, $\boldsymbol{\beta_0}$ represents a global intercept vector for the two modes, $\boldsymbol{\beta}$ is a coefficient vector,   is the covariate matrix, and $\boldsymbol{\varepsilon}$ represents the white noise matrix.

To better describe the joint models with different predictor input, let $\boldsymbol{\lambda_p}$ and $\boldsymbol{\lambda_v}$ denote the pedestrian- and vehicle-involved Poisson rate vector, respectively. The model framework for each of the transportation modes can be expressed using the following equations.

$$\ln(\boldsymbol{\lambda_p}) = \beta_{0p} + \boldsymbol{\beta_{cp}}\mathbf{X_c} + \boldsymbol{\beta_{dp}}\mathbf{X_{dp}} + \boldsymbol{\varepsilon_p} \tag{3}$$

$$\ln(\boldsymbol{\lambda_v}) = \beta_{0v} + \boldsymbol{\beta_{cv}}\mathbf{X_c} + \boldsymbol{\beta_{dv}}\mathbf{X_{dv}} + \boldsymbol{\varepsilon_v} \tag{4}$$

Where the subscripts v and p represent the vehicle and pedestrian modes, β_0 is the global intercept, $\boldsymbol{\beta_c}$ is the vector of coefficients for the independent variables common to both modes, $\mathbf{X_c}$ is the matrix of covariates common to both modes, $\boldsymbol{\beta_d}$ is the vector of coefficients for the independent variables which are different between the two modes, $\mathbf{X_d}$ is the corresponding covariate matrix, and $\boldsymbol{\varepsilon}$ is the vector of error terms. The two models are developed simultaneously with the two error vectors, $\boldsymbol{\varepsilon_p}$ and $\boldsymbol{\varepsilon_v}$, following the bivariate normal distribution:

$$\boldsymbol{\varepsilon} \sim \boldsymbol{MN}\ (\boldsymbol{\mu}, \textstyle\sum) \tag{5}$$

$$\text{Where: } \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_p \\ \boldsymbol{\varepsilon}_v \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_p \\ \mu_v \end{pmatrix}, \ \textstyle\sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \tag{6}$$

In above equations, **MN**( ) represents multivariate (or, bivariate in the present study) normal distribution, $\boldsymbol{\varepsilon}$ is the random effect matrix which capture the extra-Poisson heterogeneity among intersections, $\boldsymbol{\mu}$ is the vector of the mean values for the bivariate normal distribution, and $\sum$ is the variance-covariance matrix where the diagonal elements (ie., σ_{11} and σ_{22}) in the matrix represent the variances of the random effects, while the off-diagonal element represent the covariance. The inverse of the variance-covariance matrix represents the precision matrix, which can be formulated using the Wishart distribution:

$$\textstyle\sum^{-1} \sim Wishart(I, J) \tag{7}$$

Where I is the J identity matrix (Congdon, 2006), and J is the degree of freedom, J=2 herein representing two transportation modes. The non-informative specifications (Heydari et al., 2017) for various coefficients were specified with a normally distributed vague priors N (0,100). Such diffused normal distribution with zero mean and a large variance is commonly employed as a vague prior of posterior estimates due to the absence of sufficient knowledge of priori distribution (Cheng et al., 2018).

2.1.2 Random Forest (RF) and Variable Importance Ranking

Decision tree (Wu et al., 2020) is one of the predictive models which come up with an item's target value (leaves) via the observations about the item (branches). Based on the nature of the target value, the decision tree model can be used for both regression and classification purposes. Compared with the other typical regression techniques, the decision tree gained its popularity as it closely mirror human decision-making process. As implied by the name, random forest (Fatholahzade et al., 2018) consists of a collection of individual decision trees that operate as an ensemble. The method combines bagging and the random selection of features to construct different decision trees with controlled variation. Using ensembles of predictors has proved to give more accurate results than using a single predictor. This technique has an advantage over the traditional decision trees in obtaining unbiased error estimates without separating cross-validation test dataset. When a particular tree in the RF is grown from a bootstrap sample, usually one third of the training cases are left out (also called out-of-bag, OOB, data) from the tree-growing. The OOB data are then used later for determination of the optimum number of predictors for each tree and the optimum number of trees in the RF which result in the minimum OOB error rate.

As a robust data mining technique with the implicit wisdom: "a large number of uncorrelated individuals operating as a committee will make better decision than do these individuals", RF has seen wide applications in various fields including traffic safety (Abdel-Aty et al., 2008; Harb et al., 2008; Ahmed and Abdel-Aty, 2012). For the classical purposes of regression and classification, RF has been frequently utilized for determining the importance of various response variables, based on the mean decrease of either prediction accuracy or node purity with a specific variable being excluded from the model (Jiang, 2016). The multiple steps are involved when the former metric is implemented. First, the prediction accuracy on the OOB sample is estimated. Second, the values of the variable in the OOB sample are randomly shuffled, with all other variables remaining the same. Third, the decreased prediction accuracy on the shuffled data is calculated. Finally, the average drop of accuracy across all trees is reported for the variable.

The more decreased prediction accuracy in the OOB data, the more predictive power the variable tends to have. The second method follows the similar process as does the first one, except that the prediction accuracy is replaced with the node purity (or, Gini), which has the largest value when only one single class or value is involved in the node. Compared with the predictive accuracy-oriented metric, the node-purity one has the advantage of faster computation and is therefore chosen in the study (Nicodemus, 2011).

2.1.3 Various Model Evaluation Criteria

As a hierarchical modeling generalization of the Akaike Information Criterion (AIC) which uses maximum likelihood estimates (Hurvich et al., 1998), Deviance Information Criterion (DIC) has been used extensively to assess the complexity and goodness of fit of the Bayesian models based on the posterior means. The calculation of DIC can be done via the following expression (Spiegelhalter et al., 2003):

$$\text{DIC} = \overline{D} + P_D \tag{8}$$

Where, $\overline{D}$ is the posterior mean deviance which measures the closeness of the fitted data to the original observations, P_D denotes the effective number of parameters in a model representing the model complexity. The effective number is used since the number of independent parameters in a Bayesian hierarchical model is not clearly defined (Meyer, 2014). In general, models with more parameters tend to over-fit the data resulting in smaller deviance. Therefore, the P_D term can be considered as a compensation for this effect by favoring models with a smaller number of parameters. Even though models with smaller DIC values are preferred, it is important to note the general rule for model comparison suggested by Lunn et al. (2012): the models with DIC score less than 5 of the 'best' model are also strongly supported (provided they do not make very different inferences), values within 5 and 10, slightly inferior, and models with a DIC greater than 10 points are obviously worse. Overall, while $\overline{D}$ is regarded as an approximation to in-sample error, the DIC can be treated as the adjusted training error with the asymptotic bias corrected for the model complexity (James et al., 2013).

Different than DIC and $\overline{D}$, which are based on within-sample predictive errors, other alternatives are based on the test data using cross-validation techniques. Nonthless, the typical approaches of cross-validation are prone to selection bias related with data-splitting into subsets. To circumvent such bias, a robust conditional predictive ordibate (CPO) based on CV-1 (leave-one-out) was employed in this study (Pettit, 1990). Under this condition, an iterative process was performed where, for each step, one data point was left out for the validation of prediction accuracy of the calibrated model based on all other data (Ross and Held, 2011). Within the INLA framework, the estimate of CPO for each observation i can be calculated as (Gelfand, 1996; Liu and Sharma, 2017):

$$CPO_i = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{f\left(Y_i|\beta^{(t)}\right)}\right)^{-1} \tag{9}$$

Where Y_i is the i^{th} observation (i = 1, 2, 3, . . ., n) for all intersections and β represents the estimated model parameters.

Based on the CPO, the Log pseudo marginal likelihoods (LPML) can be calculated and have been employed in recent safety literature (Heydari et al., 2017; Cheng et al., 2018). The computation for LPML can be performed using the following equation:

$$\text{LPML} = \sum_{i=1}^{n} log(CPO_i) \tag{10}$$

Where i, n and CPO are denoted as those in Equation 9.

The larger value of LPML signifies a better predictive capability related with the candidate model. In comparison with DIC, LPML may be regarded as a measure for direct assessment of test errors (out-of-sample) as it is generated by employing the leave-one-out cross-validatory approach.

2.1.4 Negative Binomial Model

The log-linear model generally has two types, dependent on the assumption of distribution for the crash count, either Poison or Negative Binomial. For Poisson regression model, the framework for probability P(r$_i$) of a number of crashes i over different time periods r$_i$ (hour, weekday, month) can be defined using the following equation (Poch and Mannering, 1996):

$$\text{P(r}_\text{i}\text{)} = \frac{\exp(-\lambda\ )\lambda}{r_i!} \tag{10}$$

Where λ is the Poisson parameter for approach i, which is equal to approach i's expected r_i). One restriction of Poisson regression is the requirement that the mean and variance of the number of crashes are equal to each other, or, E(r_i) = Var (r_i). Investigation of crash frequency of the study reveals that the variance is significantly larger than the mean value, that is, E(r_i) < Var (r_i). To address the over-dispersed issue in the data, the commonly used NB model (Lord and Mannering, 2010; Poch and Mannering, 1996) is also used in the research. The corresponding equation for the variance of the crash count is shown as follows.

$$\text{Var } (r\ ) = \text{E}(r\ )[1+\alpha\text{E}(r\ )] = \text{E}(r_i) + \alpha\text{E}(r_i)^2 \tag{11}$$

Where α can be estimated from the maximum likelihood function. When α is zero, the model becomes Poisson regression, and if α is found to be significantly different from zero, then the NB regression can be used instead of the Poisson.

2.2 SPF of the Macro Level of Counties

For this type of SPF development, this study analyzed four different transportation mode users-involved crashes occurring at the 58 counties of California. The process involved the development of multivariate spatial-temporal models and compared their modeling performance with three other competing multivariate models assuming a Poisson-Lognormal distribution for crash counts. All models in the study were developed using the Full Bayes approach. Similar to the Empirical Bayesian (EB) method (Cheng et al, 2017c), the Full Bayesian (FB) method has been widely used in safety analysis (Davis & Yang, 2001; Washington & Oh, 2006; Cheng et al., 2017a). Even though numerous studies have illustrated favorable results yielded by the EB method (Maher & Mountain, 1988; Higle & Hecht, 1989; Cheng & Washington, 2005), an FB was chosen due to some of its advantages over EB: its capability to seamlessly integrate prior information and all available data into a posterior distribution (rather than point estimates), its capability to provide more valid safety estimates in smaller data samples, and its capability to allow more complicated model specifications. In addition to the normal Poisson-Gamma distribution, the FB models are also capable of accommodating the Poisson-Lognormal distribution and various hierarchical Poisson distributions that can address the serial and spatial correlations among the sites (Pawlovich et al., 2006; Miranda-Moreno, 2006; Cheng et al., 2017b). The details of various models are presented as follows in the order of complexity.

2.2.1 Different Modeling Formulations

Model 1: Multivariate Poisson-Lognormal Model (MVPLN)

This model assumes that crash count of certain modal crash *j* at a given location *i* in time *t* (in years), y_{ijt}, obeys Poisson distribution, while the corresponding observation specific error term **ε**

$$y_{ijt} \sim Poisson\ (\lambda_{ijt} e_{it}) \tag{12}$$

$$\ln(\lambda_{ijt}) = X'_{ijt}\beta + \varepsilon_{ijt} \tag{13}$$

$$\varepsilon_{ijt} \sim Normal\ (0,\ \Sigma) \tag{14}$$

Where $y_{ijt} = \begin{pmatrix} y_{it}^1 \\ y_{it}^2 \\ y_{it}^3 \\ y_{it}^4 \end{pmatrix}$, $\lambda_{ijt} = \begin{pmatrix} \lambda_{it}^1 \\ \lambda_{it}^2 \\ \lambda_{it}^3 \\ \lambda_{it}^4 \end{pmatrix}$, $\varepsilon_{ijt} = \begin{pmatrix} \varepsilon_{it}^1 \\ \varepsilon_{it}^2 \\ \varepsilon_{it}^3 \\ \varepsilon_{it}^4 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{14} \\ \vdots & \ddots & \vdots \\ \sigma_{41} & \cdots & \sigma_{44} \end{pmatrix}$ (15)

In above equations, λ_{ijt} is the Bayesian estimated Poisson crash rate for a mode *j* of year *t* at location *i* obtained by using offset of traffic exposure (e_{it}) at county *i* for year *t*, *X'* is the matrix of risk factors, β is the vector of model parameters, ε_{ijt} is the independent random effect which captures the extra-Poisson heterogeneity among locations. Σ is called the covariance matrix. The diagonal element σ_{jj} in the matrix represents the variance of ε_{ij}, where the off-diagonal elements represent the covariance of crash counts of different modes. The inverse of the covariance matrix represents the precision matrix and has the following distribution:

$$\Sigma^{-1} \sim Wishart(I, J) \tag{16}$$

Where *I* is the *J* x *J* identity matrix (Congdon, 2006), and *J* is the degree of freedom, *J*=4 herein representing 4 crash outcomes corresponding to four different modes.

It is important to note that in this study, the Daily Vehicle Miles Traveled (DVMT) was utilized as the exposure term (e_{it}) for the calculation of crash rate. This approach of employing DVMT as an offset to generate the crash rate has been implemented by previous studies (Miaou et al., 2003; Eksler & Lassarre, 2008; Huang et al., 2009; Flask et al., 2014; Dong et al., 2014; Dong et al., 2016a; Gill et al., 2017a, Gill et al., 2017b). This study preferred the crash rate method over the traditional crash count approach which treats traffic exposure as an explanatory variable as the former provided better and quicker convergence than the crash count models. More importantly, the plots of crash counts and DVMT across all transportation modes depicted the existence of a linear relationship, which satisfies the assumption for implementing the crash rate approach. The rationale behind this linear relationship is the large spatial scale of this study where the crash data are aggregated at the county level. Moreover, the use of crash rate helps normalize the dependent variable across counties as the crash data demonstrated very significant variability due to the differences in geographic area and population across 58 counties.

Model 2: Multivariate Poisson-Lognormal with Time Trend (MVPLNT)

Under this model, a yearly trend term t is added to Equation 13 resulting in the new expression:

$$\ln(\lambda_{ijt}) = X'_{ijt}\beta + \varepsilon_{ij} + \gamma t_j * T \tag{17}$$

Where γt_j is the trend coefficient vector for various crash modes, and T is yearly trend. Various types of trend were explored in previous studies (Lawson, Browne, & Rodeiro, 2003). This study assumes a linear yearly trend for various crash modes with a non-informative prior N (0, 100^2).

Model 3: Multivariate Poisson-Lognormal Spatial Model (MVPLNS)

In this model, a spatially structured error term u_{ij} is added to Equation 13 which leads to the following expression:

$$\ln(\lambda_{ijt}) = X'_{ijt}\beta + \varepsilon_{ijt} + u_{ij} \tag{18}$$

Where u_{ij} is fit by a zero-centered multivariate conditional auto-regressive model (Mardia, 1988) which has a conditional normal density shown as follows:

$$u_i | u_k, \ \textstyle\sum_i \ \sim N_j(\sum_{k\sim i} C_{ik}, u_k, \sum_i) \tag{19}$$

Where each $\sum_{\boldsymbol{i}}$ is a positive definite matrix representing the conditional variance matrix, and the adjacency matrix C_{ij} is of the same dimension with $\sum_{\boldsymbol{i}}$ (Aguero-Valverde, Wu, & Donnell, 2016). The precision matrix $\sum^{-1}$ follows the Wishart distribution as shown in Equation 16.

As we can see from the above equations, estimation of the risk in any site is conditional on risks in neighboring locations. Subscripts *i* and *k* refer to a county and its neighbor, respectively, and *k* belongs to N_i where N_i represents the set of neighbors of county *i*. Besides the identification of neighbors, the assigned weights also affect the risk estimation. In the past studies (Wang & Abdel-Aty, 2006; Guo et al., 2010; Aguero-Valverde & Jovanis, 2006; Xu & Huang, 2015; Gill et al., 2017a), weight structures such as various adjacency-based, distance-based models, and semi-parametric geographically weighted, have been explored. As the current study is focused on the evaluation of alternate spatiotemporal multimodal models, the commonly used distance-based structure was employed as an example to explore the spatial correlations with the following formulation:

$$w_{ij} = \frac{1}{d_{ij}} \tag{20}$$

Where w_{ij} is the weight between counties *i* and *j*, and d_{ij} is the distance between counties *i* and *j*. With this weight structure, it is known that more weightage was assigned to counties which are relatively close. It should be noted that an array of approaches has been employed to generate the spatial weights for CAR specification, but this study chose the inverse distance as a representation of a large body of existing research which implemented similar approach to accommodate the spatial correlation (Song et al.,, 2006; Guo et al., 2010; Dong et al., 2014; Dong et al., 2015; Cheng et al., 2017a; Gill et al., 2017a,b). The introduction of complex approaches is avoided as the primary focus is on the comparison of multivariate multimodal models which differ on the basis of incorporation of spatial and temporal terms.

Model 4: Multivariate Poisson-Lognormal Spatial-Temporal Model (MVPLNST) with Fixed Time Coefficient

This model represents the first multivariate space-time model with the assumption of a fixed yearly trend for various crash types. The corresponding formula is shown as follows:

$$\ln(\lambda_{ijt}) = X'_{ijt}\beta + \varepsilon_{ij} + u_{ij} + (\gamma t + \delta_{ij}) * T \quad (21)$$

Where γt is the fixed yearly trend coefficient for all crash modes, and δ_{ij} is an interaction random effect between space and time which allows different temporal trends in crash risk for different spatial locations. γt was assigned a non-information prior of N (0, 100^2) and δ_{ij} was assumed to have the same prior with u_{ij}.

Model 5: Multivariate Poisson-Lognormal Spatial-Temporal Model (MVPLNST) with Varying Time Coefficients for Crash Types

This model represents the second multivariate space-time model under the premise that the yearly trends for various crash types are different. The model takes the following form:

$$\ln(\lambda_{ijt}) = X'_{ijt}\beta + \varepsilon_{ij} + u_{ij} + (\gamma t_j + \delta_{ij}) * T \quad (22)$$

Where γt_j has the same definition and prior distribution as shown in Equation 17. The time-varying coefficient specification may be regarded as a limiting case of random parameters approach (Cheng et al., 2017b), which has been employed in safety literature to account for the unobserved heterogeneity (Dong et al., 2014; Dong et al., 2016b; Dong et al., 2017).

2.2.2 Goodness-of-Fit of the Models

For the county-level SPFs, the above five alternate models were evaluated based on some criteria used from previous studies: DIC (Spiegelhalter, Thomas, Best, & Lunn, 2003), MAD (Konno & Koshizuka, 2005), MSPE (Gill et al., 2017a), the G^2 statistic (Cheng & Washington, 2008), the Chi-squared RSS (Earnest et al., 2007), and TRD (Cheng & Washington, 2008). The readers wishing more detail on these measures can refer to these studies. Descrition about DIC can be found in the section of 2.1.3. The details of other criteria are shown as follows.

Mean Absolute Deviation (MAD)

MAD is based on the model deviation or residue and frequently used by the researchers across different fields to check fitness of data as it is not limited to a particular distribution. MAD aims to estimate the average difference between estimated and observed crash counts for each county, and it can be calculated with the following equation:

$$\mathrm{MAD} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{Y_i} - \mathrm{O_i}| \quad (23)$$

Where Y_i is the Bayesian-estimated crash frequency and O_i is the observed crash count for county i by a model during the same time period. The smaller the MAD value, the better fitness to the data.

Mean Squared Predictive Error (MSPE)

This criterion differs from MAD by considering the square of deviation rather than the absolute values. To assess the prediction capability of models, the mean-squared predictive error (MSPE) was calculated as follows:

$$\mathrm{MSPE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{Y} - \mathrm{O_i})^2 \qquad (24)$$

Where Y_i is the Bayesian-estimated crash count of county i by a model, and O_i is the observed crash count of county *i* by the same model. It is expected that MSPE assigns more penalties to the counties whose deviations are larger due to the squared deviation. Again, the larger MSPE indicates an inferior performance.

The G^2 statistic

The sum of model deviances, G^2, gives a test of whether the model gives an adequate explanation of the data relative to the saturated model (Washington et al., 2003). The G^2 statistic is given as:

$$G^2 = 2\sum_{i=1}^{n} y_i LN\left(\frac{O_i}{Y_i}\right) \qquad (25)$$

Where terms are as defined previously. G^2 is zero for a model with perfect fit. A large G^2 deviating from zero indicates that the model fits poorly as compared to the saturated model.

Residual Sum of Squares (RSS)

The model comparisons based on MAD and MSPE calculations may be biased as the larger counties are expected to subject to more penalties due to greater counts and residues. To address this issue, this study employed the chi-squared RSS which tends to remove such bias by calculating the squared residual relative to estimated number of crashes. RSS is defined as:

$$RSS = \sum_{i=1}^{n} \frac{(O_i - Y_i)^2}{O_i} \qquad (26)$$

The model with a smaller value of RSS tends to have better predictive capabilities.

Total Rank Difference (TRD)

The aforementioned criteria utilized the magnitude of residual value for assessment of model fit based on prediction accuracy. TRD introduces a different perspective for model comparison as it accounts for the rank deviations based on the observed and estimated crash counts. The rank difference is calculated by using the following equation.

$$TRD = \sum_{i=1}^{n} |R(i_o) - R(i_Y)| \qquad (27)$$

Where $R(i_o)$ is the observed data rank at county i and $R(i_Y)$ is the rank based on estimated crash counts for the same time period. A model is considered superior if smaller TRD value is revealed, which signifies that the specific model assigns rankings close to the observed crash counts.

2.3 SPF of the Macro Level of TAZs

For this type of SPPs, the Full Bayesian (FB) framework was employed for estimation of six-year bicyclist and pedestrian crashes aggregated at the Traffic Analysis Zone level. Four crash frequency models were developed. The general functional form of the models is given in the following subsections while progressing from simple to sophisticated.

2.3.1 Different Modeling Formulations

Model 1: Bivariate

This model assumes that crash count of certain modal crash ***j*** at a given location ***i***, **y_{ij},** obeys Poisson distribution, while the corresponding observation specific error term ε_{ij} follows a bivariate normal distribution:

$$y_{ij}|\lambda_{ij} \sim Poisson\ (\lambda_{ij}) \tag{28}$$

$$\ln(\lambda_{ij}) = X'_{ij}\beta + \varepsilon_{ij} \tag{29}$$

$$\varepsilon_{ij} \sim MVN\ (0,\ \Sigma) \tag{30}$$

Where $$y_{ij} = \begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix}, \qquad \lambda_{ij} = \begin{pmatrix} \lambda_{i1} \\ \lambda_{i2} \end{pmatrix}, \qquad \varepsilon_{ij} = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \tag{31}$$

In above equations, X*'* is the matrix of risk factors, β is the vector of model parameters, ε_{ij} is the independent random effect which captures the extra-Poisson heterogeneity among locations. ∑ is called the covariance matrix. The diagonal element σ_{jj} in the matrix represents the variance of ε_{ij}, where the off-diagonal elements represent the covariance of crash counts of different modes. The inverse of the covariance matrix represents the precision matrix and has the following distribution:

$$\Sigma^{-1} \sim Wishart(I, J) \tag{32}$$

Where I is the J x J identity matrix (Congdon et al., 2006), and J is the degree of freedom, J=2 herein representing two crash outcomes corresponding to bicyclist and pedestrians crashes.

Model 2: Bivariate Spatial

Under Model 2, the spatial random effects were incorporated over the model represented in Equation 29. The final model takes the following form to account for spatial correlations among the TAZs:

$$\ln(\lambda_{ij}) = X'_{ij}\beta + \varepsilon_{ij} + u_{ij} \tag{32}$$

Where u_{ij} is the spatially structured random effect which follows the MCAR (multivariate conditional autoregressive) (Mardia, 1988) formulation to incorporate the spatial correlation among crashes occurring at neighboring TAZs.

$$u_i|u_k,\ \sum i\ \sim N_j(\textstyle\sum_{k\sim i} C_{ik}\ , u_k, \sum i\ ) \tag{33}$$

Where each $\sum_i$ is a positive definite matrix representing the conditional variance matrix, and the adjacency matrix $\mathrm{C_{ij}}$ is of the same dimension with $\sum_i$ (Jonathan et al., 2016). The precision matrix $\sum^{-1}$ follows the Wishart distribution as shown in Equation 5.

As we can see from the above equations, estimation of the risk in any site is conditional on risks in neighboring locations. Subscripts i and k refer to a TAZ and its neighbor, respectively, and k belongs to $\mathrm{N_i}$ where $\mathrm{N_i}$ represents the set of neighbors of TAZ i. Besides the identification of neighbors, the assigned weights also affect the risk estimation. In the past studies (Aguero-Valverde and Jovanis, 2009; Xu and Huang, 2015), weight structures including various adjacency-based, distance-based models, and semi-parametric geographically weighted, and so on, have been explored. The current study employs the commonly used distance-based structure to explore the spatial correlations with the following formulation:

$$w_{ij} = \frac{1}{d_{ij}} \tag{34}$$

Where w_{ij} is the weight between TAZ i and j, and d_{ij} is the distance between TAZ i and j. With this weight structure, it is known that more weightage was assigned to TAZs which are relatively close.

Model 3: Bivariate Dirichlet Process Mixture

The parametric model specification of the aforementioned models assumed the distribution of the parameters to be specific (normal in this study) across all concerned sites. However, the nonparametric specification removes such constraints by employing a flexible approach of the Dirichlet process that allows the incorporation of unknown random density for the parameters. The current study employs a semi-parametric approach which relaxes the restrictive distributional assumption for the intercept only, instead of all of the parameters. The removal of constraints for the intercept to follow a specific distribution represents a plausible scenario where the TAZs are not expected to have a normal distribution. This flexible approach is expected to capture the extra variability which may escape the error terms introduced in parametric models. Equation 2 was modified to use Dirichlet process mixture over the intercept as follows (*Heydari et al., 2016*):

$$\ln(\lambda_{ij}) = \beta_{0rj} + X_i'\beta \tag{35}$$

$$\beta_{0rj} \approx \textstyle\sum_{n=1}^{C} p_n I_{\theta_{z_i}} \sim TDP\ (kG_{0j}),\ \ z_i = n\ with\ probability\ of\ p_n \tag{36}$$

$$G_0 \sim MVN\ (\mu_{G_{0j}},\ \Sigma) \tag{37}$$

Where β_{0rj} is the intercept for cluster r (r ranges from 1 to C) of mode j, k is the precision parameter, and G_0 is the baseline distribution for β_{0r}which follows a bivariate normal distribution with mean μ_{G_0} and variance $\sum$, which also follows the Wishart distribution. β_{0rj} essentially represents a vector of probabilities over the space of concerned entities (203 TAZs) and follows a Truncated Dirichlet Process (TDP) with a vector of parameters represented by kG_{0j}. The precision parameter k indicates the variability of the Dirichlet process around $G_{0j.}$ The intercept draws random points (θ_{Z_i}) and the associated probabilities (p_n) can be obtained through the stick-breaking procedure (Shirazi et al., 2016; Ohlsse at al., 2007; Ishwaran and James, 2001). If one cluster is occupied, the indicator function $(I_{\theta_{Z_i}})$ at θ_{Z_i} will take the value of 1, otherwise it would be 0. The number of latent clusters (r) in β_{0rj} could range from 1 to infinity, which requires immense computational effort. To reduce the computational complexity by obtaining finite dimensional approximation, a truncated Dirichlet process is utilized to fix the maximum number of possible clusters to C, where C is governed by the precision parameter k and is estimated by $5k$+2 (Ohlssen et al., 2007). As the prior distribution for precision parameter k was assumed to be $k \sim$ uniform (0.3, 9), so eventually the number of clusters were limited to be maximum of 47. The value of C used in the study can be considered in a normal range given the different C values utilized previously such as 5 (Ghosh and Norris, 2005), 10 (Erkanli et al., 2006), and 52 (Heydari et al., 2017).

Model 4: Bivariate Dirichlet Process Mixture Spatial

Model 4 is distinct from Model 3 by incorporating the spatial random effects to account for the correlation among the neighboring TAZs. The model in Equation 9 takes the following form:

$$\ln(\lambda_{ij}) = \beta_{0rj} + X_i'\beta + u_{ij} \tag{38}$$

Where all terms are defined as previously.

2.3.2 Goodness-of-Fit of the Models

To compare the performance of the TAZ-pertinent SPFs, the similar evaluation criteria including cross-validatory conditional predictive ordinate (CPO), the log pseudo marginal likelihoods (LPML), MSPE, the G^2 statistic, the Chi-squared Residual Sum of Square (RSS) and the Rp^2 statistic. The first two criteria, CPO and LPML, are based on the cross validation and the associated details are presented in the section of 2.1.2. The last four rely on in-sample validation and the description of MSPE, the G^2 statistic, and RSS is provided in the section of 2.2.2. The typical R-square in ordinary linear regression cannot be directly applied to the crash frequency model due to the nonlinearity of conditional mean (E[y|X]) and heteroscedasticity associated with the Poisson models. Therefore, the research also adopted an equivalent measure, R_p^2, which is based on standardized residuals:

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n}\left[\frac{y_i-\lambda_i}{\sqrt{\lambda_i}}\right]^2}{\sum_{i=1}^{n}\left[\frac{y_i-\bar{y}}{\sqrt{\bar{y}}}\right]^2} \tag{39}$$

Where $\bar{y}$ represents the mean value of the observed counts. Similar to R-square, a smaller R_p^2value indicates the inferior performance.

3. DATA DESCRIPTION

Even though the research focuses on the development of active transportation-oriented SPFs, the data are unavailable for all transportation modes for the different spatial units. In specific, the micro level of count data consist of pedestrian and vehicles, while the county-level crashes cover the four modes (vehicle, motorcycle, bicycle and pedestrian) and the TAZ-related counts are collected for pedestrians and bicyclists. The data descriptions are presented in order as follows.

3.1 Data for the Development of SPF of the Micro Level

The data employed to this type of SPFs are from various database. The first analysis was based on the data derived from California Traffic Accident Surveillance and Analysis System (TASAS). TASAS is a traffic records system which includes crash database and infrastructure database consisting of highway segments, intersections, ramps, and other data. The study focused on crashed occurring at the intersections which have 73 variables available in the raw file in TASAS. Nonetheless, some of these variables were not associated with pedestrian or vehicle collisions like intersection location information (district, county, route and milepost), date of intersection update (begin date of intersection update, end date of intersection update), and so on. After data cleaning, 21 covariate variables were selected from a total of 6,198 intersections in the state routes, where the estimated annual pedestrian volume at each intersection was available through the pedestrian count model developed by Griswold et al. (2019). Overall, a total of 43705 pedestrian and vehicle collisions spanning over six years (2012 to 2017) were aggregated for the research purpose. The detailed information for all data including variable names, description, and other descriptive statistics are illustrated in Table 1.

Table 1. Descriptive Statistics of the First Database for the Micro Level

Numerical Variables					
Variables	**Description**	**Minimum**	**Maximum**	**Mean**	**S.D.**
MNL	Mainline - number of lanes	2	8	3.33	1.39
MOL	Mainline - override length (buffer)	15	350	187.90	61.94
X-NL	Cross street - number of lanes	0	6	2.13	0.55
X-OL	Cross street - override length	0	250	2.25	22.81
MADT	Mainline - average daily traffic	180	125000	20198	15254.28
X-ADT	Cross street - average daily traffic	0	77000	1911	4272.05
IRG	Intersection rate group	1	29	17.91	7.61
APV	Estimated annual pedestrian volume (2016)	520	9400000	116636	481942.50
Veh counts	Vehicle related accidents counts	0	137	6.88	11.79
Ped counts	Pedestrian related accidents counts	0	6	0.09	0.39
Categorical Variables					
Variables	**Description**	**Details of categories (frequency, percentage)**			
Highway Group	Highway group of mainline in the intersection	1-Divided Highway (3294, 53.15%); 2-Undivided Highway (2881, 46.48%); 3-Right or Left Independent Alignment (23, 0.37%)			
Population Group	Population code of the intersection	-Urban (1539, 24.83%); 2-Rural (1278, 20.62%); 3-Urbanized (3381, 54.55%)			

Intersection Design	Intersection design	1-Four-legged (2328, 37.56%); 2->four-legs (67, 1.08%); 3-Offset (349, 5.63%); 4-Tee (3182, 51.34%); 5-Wye (206, 3.32%); 6-Other (66, 1.06%)
Light Condition	Presence of light condition at Intersection	1-No Lighting (1561, 25.19%); 2-Lighted (4637, 74.81%)
Mastarm	Presence of signal mastarm on the mainline of the intersection	1-No Mastarm (4901, 79,07%); 2-Yes, Mastarm (1297, 20.93%)
Left Turn	Left turn channelization on mainline at the intersection	1-Curbed Median Left Turn Channelization (808, 13.04%); 2-No Left Turn Channelization (3005, 48.48%); 3 - Painted Left Turn Channelization (2355, 37.00%); 4 - Others (30, 0.48%)
Right Turn	Right turn channelization on mainline at the intersection	1-No Right Turn Channelization (5579, 90.01%); 2-Others (617, 9.99%)
Traffic Flow	Traffic flow on the mainline of the intersection	1-Two-Way Traffic, No Left Turns Permitted (297, 4.81%); 2-Two-Way Traffic, Left Turn Permitted (5839, 94.21%); 3 - Others (61, 0.98%)
X-Mastarm	Presence of signal mastarm on the cross-street of the intersection	1-No Mastarm (5341, 86.17 %); 2-Yes, Mastarm (857, 13.83%)
X-Left Turn	Left turn channelization on the cross-street	1-Curbed Median Left Turn Channelization (131, 2.11%); 2-No Left Turn Channelization (5421, 87.47%); 3-Painted Left Turn Channelization (622, 10.04%); 4-Others (24, 0.39%)
X-Right Turn	Right turn channelization on the cross-street.	1-No Right Turn Channelization (5631, 90.85%); 2-Others (567, 9.15%)
X-Traffic Flow	Traffic flow on the cross-street of the intersection	1–Two Way Traffic, No Left Turns Permitted (269, 4.34%); 2–Two-Way Traffic, Left Turn Permitted (5846, 94.32%); 3-Others (83, 01.34%)
Intersection Control Condition	Intersection control condition	1-No Control (210, 3039%); 2-Stop signs on Cross Street Only (4514, 72.83%); 3-Signals Pretimed (2 Phase) (152, 2.45%); 4-Signals Semi-Traffic Actuated, Two-phase (125, 2.02%); 5 - Signals Full Traffic Actuated, Multi-Phase (993, 16.02%); 6-Others (204, 3.29%)

Note: S.D. represents standard deviation.

The data used in second analysis were obtained from California Department of Transportation, which included infrastructure level and accident level. The second database contains the same input variables from infrastructure file as the first database except pedestrian volume. In addition to independent variables this analysis selected exposure information of vehicle accidents occurred from 2015 to 2017 to be the output variable. Overall, 20 covariates pertaining to intersection information are available to 18,562 intersections from infrastructure file. As there are exactly the same 3,162 intersections matched to the first database, after removing the matched observations, the second database utilized 15,401 intersections with 20 infrastructure-related variables crossing vehicle accident number of each intersection. The detailed descriptive statistics for all variables are shown in Table 2. It’s important to note that the difference between the first and second database since the it doesn’t have pedestrian volume for intersections.

Table 2. Descriptive Statistics of the Second Database for the Micro Level

Numerical Variables					
Variables	**Description**	**Minimum**	**Maximum**	**Mean**	**S.D.**
MNL	Mainline - number of lanes	2	9	3.01	1.32

MOL	Mainline - override length (buffer)	0	350	207.40	61.08
X-NL	Cross street - number of lanes	0	8	2.07	0.51
X-OL	Cross street - override length	0	250	4.577	32.73
MADT	Mainline - average daily traffic	95	116817	16200	15515.53
X-ADT	Cross street - average daily traffic	0	980119	1518	8886.67
IRG	Intersection rate group	1	30	16.16	7.65
Veh counts	Vehicle related accidents counts	0	83	2.89	5.91

Categorical Variables		
Variables	**Description**	**Details of categories (frequency, percentage)**
Highway Group	Highway group of mainline in the intersection	1-Divided Highway (6030, 40.17%); 2-Undivided Highway (8748, 58.28%); 3-Right or Left Independent Alignment (233, 1.55%)
Population Group	Population code of the intersection	1-Urban (2578, 17.17 %); 2-Rural (7133, 47.52%); 3-Urbanized (5300, 35.31%)
Intersection Design	Intersection design	1-Four-legged (4835, 32.21%); 2->four-legs (124, 0.83%); 3-Offset (545, 3.63%); 4-Tee (8489, 56.55%); 5-Wye (885, 5.90%); 6-Other (133, 0.89%)
Light Condition	Presence of light condition at Intersection	1-No Lighting (7608, 50.68%); 2-Lighted (7403, 49.32%)
Mastarm	Presence of signal mastarm on the mainline of the intersection	1-No Mastarm (12903, 85.96%); 2-Yes, Mastarm (2108, 14.04%)
Left Turn	Left turn channelization on mainline at the intersection	1-Curbed Median Left Turn Channelization (1327, 9.14%); 2-No Left Turn Channelization (8994, 59.92%); 3 - Painted Left Turn Channelization (4579, 30.50%); 4 - Others (66, 0.44%)
Right Turn	Right turn channelization on mainline at the intersection	1-No Right Turn Channelization (13444, 89.56%); 2-Others (1567, 10.44%)
Traffic Flow	Traffic flow on the mainline of the intersection	1-Two-Way Traffic, No Left Turns Permitted (682, 4.54%); 2-Two-Way Traffic, Left Turn Permitted (14049, 93.59%); 3 - Others (280, 1.87%)
X-Mastarm	Presence of signal mastarm on the cross-street of the intersection	1-No Mastarm (13545, 90.23%); 2-Yes, Mastarm (1466, 9.77%)
X-Left Turn	Left turn channelization on the cross-street	1-Curbed Median Left Turn Channelization (270, 1.80 %); 2-No Left Turn Channelization (13624, 90.76%); 3-Painted Left Turn Channelization (1068, 7.11%); 4-Others (49, 0.33%)
X-Right Turn	Right turn channelization on the cross-street.	1-No Right Turn Channelization (13750, 91.60%); 2-Others (1261, 8.40%)
X-Traffic Flow	Traffic flow on the cross-street of the intersection	1–Two Way Traffic, No Left Turns Permitted (606, 4.04%); 2–Two-Way Traffic, Left Turn Permitted (14023, 93.42%); 3-Others (382, 2.54%)
Intersection Control Condition	Intersection control condition	1-No Control (1453, 9.68%); 2-Stop signs on Cross Street Only (11012, 73.36%); 3-Signals Pretimed (2 Phase) (220, 1.47%); 4-Signals Semi-Traffic Actuated, Two-phase (106, 0.71%); 5 - Signals Full Traffic Actuated, Multi-Phase (1609, 10.72%); 6-Others (571, 3.80%)

Given the importance of ramps, the third analysis focused on the crashes occurring on the ramp of highway. The data were also provided by California Department of Transportation from separate files. The average daily traffic from 2015 to 2017, as a main exposure-related factor, was extracted from the infrastructure file. It also provided the data for other independent variables linked with ramp information such as the highway group of the ramp, type of ramp, On/off Indicator. In addition, the vehicle- and pedestrian-related accidents occurred on ramp over the corresponding three years were obtained from the accident file. Overall, this analysis would utilize six input variables and two output variables crossing 15226 ramp observations to develop the Bayesian hierarchical joint model. Then, acquire the ramp SPF of pedestrian and vehicle from the model results. Table 3 shows the summary information for variables used in this research.

Table 3. Descriptive Statistics of the Third Database for the Micro Level

Numerical Variables					
Variables	**Description**	**Minimum**	**Maximum**	**Mean**	**S.D.**
ADT	Ramp - average daily traffic (2015~2017)	1	97301	7359	9021.80
Veh counts	Vehicle-related accidents counts	0	173	8.31	12.22
Ped counts	Pedestrian related accidents counts	0	8	0.06	0.28

Categorical Variables		
Variables	**Description**	**Details of categories (frequency, percentage)**
Ramp Highway Group	Highway group of the ramp	1-Divided Highway (14820, 97.33%); 2-Undivided Highway (98, 0.64%); 3-Right or Left Independent Alignment (308, 2.03%)
Ramp Design Code	Type of Ramp	1- Frontage Road (34, 0.22%); 2- Collector Road(142, 0.93%); 3- Dir/Semi Lft Rmp(650, 4.27%); 4- Diamond Interchg(6714, 44.10%); 5- Slip Ramp(341, 2.24%); 6- Dir/Semi Rgt Rmp (2206, 14.49%); 7- Loop W/Lft Trn(616, 4.05%); 8- Buttonhook Ramp (1167, 7.66%); 9- Scissors Ramp (320, 2.10%); 10- Split Ramp (985, 6.47%); 11- Loop W/O Lff Trn (1316, 8.64%); 12- Two-Way Ramp (38, 0.25%); 13- Dummy Paired Rmp (125, 0.82%); 14- Rest Area/Vista(332, 2.17%); 15- Dummy-Volume Onl (84, 0.55%); 16- Other(156, 1.02%)
Ramp On/Off Code	On/Off Ramp Indicator	1- Off (7369, 48.40%); 2- On (7684, 50.47%); 3- Other (173, 1.13%)
Ramp Area 4 Indicator	Whether the ramp in question is associated with “area 4” for accident location purposes.	1- No (2979, 19.57%); 2- Yes (12247, 80.43%)
Ramp Population Group	Population group of the ramp	1- Urban (859, 5.64%); 2- Rural (2459,16.15%); 3- Urbanized (1190878.21%)

3.2 Data for the Development of County-Level SPF

The development of county-level SPFs is based on the crashes of different modes occurring in the 58 counties of California over a six-year period (2008-2013). The segregated collision counts over a relatively long period were considered to closely assess the impact of different temporal treatments for crash prediction models. Four different transportation mode-related crashes were collected for each year of the period 2008-2013 from SWITRS (California Statewide Integrated Traffic Records System) which include pedestrian, bicyclist, motorcyclist and vehicle only crashes. In cases of crashes pertaining to two or more crash modes, the relatively vulnerable mode is designated as the crash victim and hence that crash is counted towards that specific mode. For example, in case of a crash involving a motorcycle and a vehicle, that crash would count towards a motorcycle crash as such crashes are usually underrepresented. Also, this crash would only be counted once towards motorcycle crashes and not for vehicle crashes otherwise one crash would result in two counts. Similarly, in case of a crash between a pedestrian and a bicyclist, the crash would be counted as a pedestrian crash. The mode-based crash counts were to develop crash prediction models for the estimation of crash rate, where the crash counts were offset by Daily Vehicle Miles Travel (DVMT), which is a main exposure-related factor at the macro-level (Miaou et al., 2003) and was collected from Highway Performance Monitoring System (HPMS) for the corresponding six years. HPMS also provided the data for independent variables linked with roadways and traffic conditions such as maintain miles and travel time for work trips, respectively. The other independent variables comprised of various demographic, socioeconomic, and land use data which were expected to impact the multimodal activity in the counties and influence the collisions. The main demographic factor, population, along with other factors depicting the socioeconomic activity such as retail sales, household income, per capita income, and percent of people in poverty, employment, and land area were obtained from the California Department of Finance and the US Census Bureau, respectively. In addition, the data for the geometric centroid distance among the counties were provided by Southern California Association of Governments (SCAG), which was utilized for calculation of distance-based weights for accommodating the spatial aspect of models.

Table 4 illustrates the summary information for all dependent and independent variables considered for model development. It should be noted that this study incorporated a mix of time-varying (yearly) and constant variables which account for the temporal trends and spatial-only covariates, respectively. This dataset replicates the real-world scenario where the possibility for collection of a continuous set of some variables is not feasible at the macro-level. The continuous data for the given time period were available for multimodal crashes, DVMT, population, and roadway miles, while rest of the variables were mostly obtained from average of data over some time period. The studies which focus on crash counts on yearly-basis may be prone to erroneous inferences due to the bias induced in the model estimates by the excessive amount of zero crash counts present in the data. As evident from the nature of independent variables, some variables were observed to be correlated and filtered out using two techniques before incorporating during the model development for crash estimation. First, the correlation tests were conducted using the Harrell Miscellaneous package in R software which allowed the calculation of Pearson correlation coefficient. The variables observed to be correlated at a significance level of 0.05 were then eliminated in multiple steps using engineering judgment to prevent exclusion of any potential influential variables which would result in loss of precision of estimated parameters.

Table 4. Descriptive Statistics of Collected Data of Various Counties

Variables	Description	Year	Minimum	Maximum	Median	Mean	S.D.
Collision	Motor Vehicle	2008	15	41,794	631	2,389	5,841
		2009	20	40,197	611	2,289	5,612
		2010	18	39,560	537.5	2,249	5,531

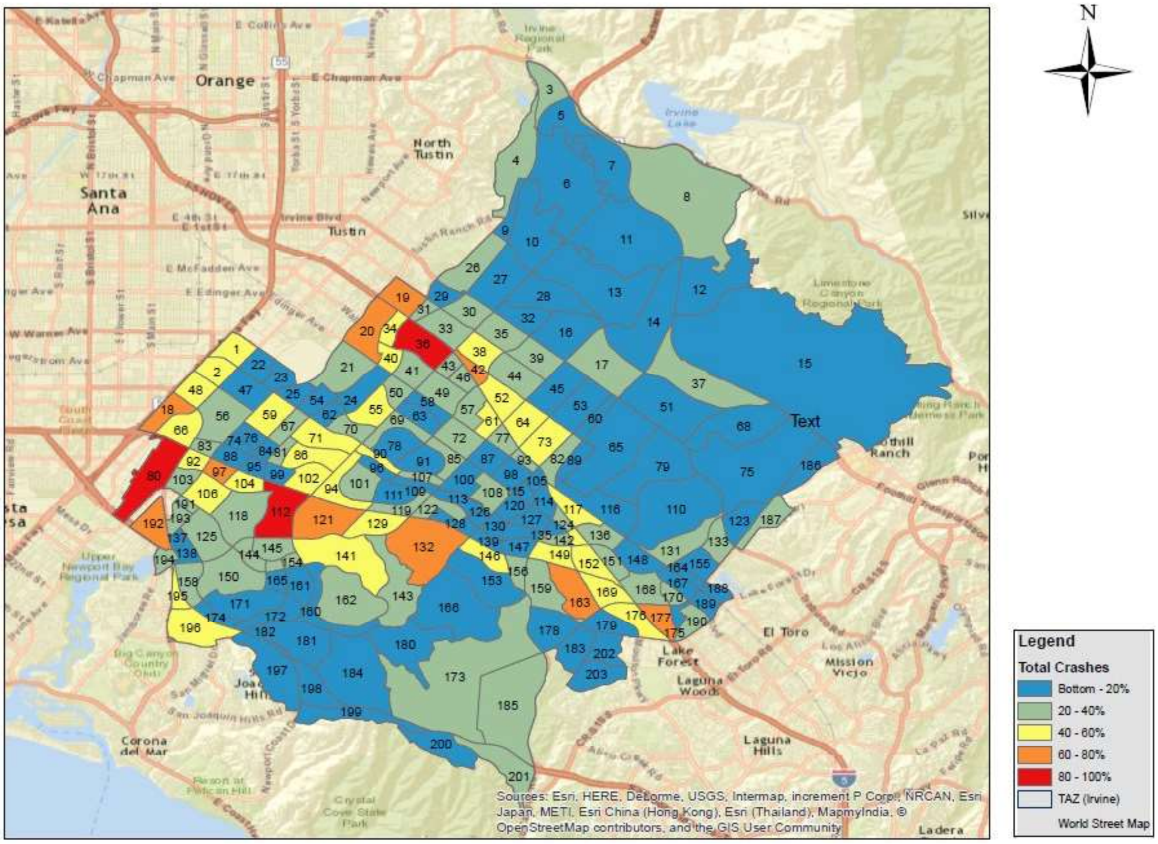
		2011	14	38,933	576	2,184	5,430
		2012	16	38,477	560	2,171	5,388
		2013	21	38,855	544	2,140	5,436
Collision	Pedestrian	2008	0	5,199	41.5	231	702
		2009	0	5,097	41.5	224	687
		2010	0	4,730	36.5	218	641
		2011	0	4,748	37	218	644
		2012	0	5,024	35	228	684
		2013	0	4,932	38.5	213	667
Collision	Bicycle	2008	0	3,348	46.5	203	481
		2009	0	3,747	48	208	531
		2010	1	4,226	49.5	219	587
		2011	0	4,788	51.5	236	662
		2012	0	4,955	44.5	241	685
		2013	0	4,682	51.5	230	647
Collision	Motorcycle	2008	7	3,048	66.5	205	436
		2009	3	2,802	60	181	399
		2010	2	2,711	60.5	171	388
		2011	5	3,112	57.5	189	443
		2012	4	3,349	54.5	200	483
		2013	3	3,614	65.5	208	516
DVMT	Daily Vehicle Miles Traveled (miles)	2008	168.265	214,971	5,005	15,387	31,617
		2009	170.69	214,236	4,836	15,317	31,469
		2010	169.42	211,876	5,448	15,482	31,148
		2011	164.587	214,458	4,761	15,353	31,594
		2012	166.923	214,482	4,551	14,768	31,478
		2013	165.18	215,817	4,462	14,924	31,747
Pop	Population	2008	1,214	10,347,422	180,923	656,696	1,469,310
		2009	1,194	10,398,067	182,519	662,962	1,478,749
		2010	1,177	9,840,555	179,588	644,265	1,408,182
		2011	1,113	9,866,172	179,134	647,470	1,413,526
		2012	1,088	9,923,806	180,800	652,028	1,422,391
		2013	1,078	10,002,804	181,150	657,967	1,434,566
MM	Maintained Miles	2008	287	21,686	2,009	2,974	3,329
		2009	266	21,678	2,012	2,963	3,333
		2010	266	21,746	2,012	2,967	3,341
		2011	266	360,857	2,008	9,128	47,113

		2012	270	21,694	2,021	3,026	3,432
		2013	265	21,858	1,921	3,017	3,428
RS	Total Retail Sales ($1,000)	2012	576	121,389,378	1,859,337	8,306,904	18,251,399
TT	Mean Travel Time to Work (minutes)	2014	13	34	25	24	4
HI	Median Household Income (dollars)	2014	35,997	35,997	35,997	35,997	35,997
PCI	Per Capita Income for past year (dollars)	2014	16,409	58,004	26,190	27,604.34	8,198
PP	Persons in Poverty (percentage)	2014	7	28	16	16	5
TE	Total Employment	2014	211	3,932,904	44,911	232,458	573,978
AF	All Firms	2012	125	1,146,701	13,613	61,845	160,522
LA	Land Area (Square miles)	2010	46	20,056	1,535	2,685	3,102
Distance	Distance among centroids of counties (miles)	N/A	25	962	227	273	176

Note: S.D. represents standard deviation; N/A means Not Applicable

3.3 Data for the Development of TAZ-Level SPF

The development of TAZ-level SPFs is based on the Pedestrian and bicyclist crashes which occurred in the City of Irvine from 2007 to 2012. Like many other research studies (Ladron de Guevara et al., 2004; Hadayeghi et al., 2007; Abdel-Aty et al., 2011), TAZs were selected as the base units, and the crash data were aggregated at the TAZ-level. Overall, there are 203 TAZs in the City. The map in Figure 1 displays the distribution of all TAZs and associated crash counts. The two transportation mode-related crashes were collected from SWITRS (California Statewide Integrated Traffic Records System) Shape file of TAZ boundary and TAZ characteristics were provided by SCAG (Southern California Association of Governments).



Sources: Esri, HERE, DeLorme, USGS, Intermap, increment P Corp., NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Figure 1. TAZ Map with Crash Distributions in the City of Irvine, California.

The variables used for model development and the associated descriptive statistics are shown in Table 4. The six-year aggregated pedestrian and bicyclist crashes were used as the dependent variables. DVMT acted as a measure of exposure. The explanatory variables were the predictors commonly used in previous regional safety analyses which include socioeconomic, transportation-related, and environment-related factors, and so on. It is worth mentioning that the data from 2008 were available for explanatory variables due to less frequent collection by the agencies and hence it is used for model development. Also, the distance matrix containing distances among various TAZ centroids were also collected from SCAG for the estimation of distance-based spatial random effect. Their descriptive statistics can be found in Table 4 as well.

Table 4. Summary Statistics of Variables for TAZ's of the City of Irvine

Variables	Description	Mean	SD	Minimum	Maximum
Bike	Total bike-involved crashes (2007-2012)	1.82	2.45	0	12
Ped	Total pedestrian-involved crashes (2007-2012)	0.81	1.33	0	8
DVMT	Daily vehicle miles traveled	5,4262.44	56,156.84	112.57	276,079.90

Acre	TAZ Area in acre	282.90	431.75	0.69	5,062.95
Median	Median house income ($)	48,440.78	50,635.10	0	183,347
Pop_den	Population density by area	6.18	7.96	0	32.40
HH_den	Household density by area	2.34	3.15	0	13.62
Emp_den	Employment density by area	10.34	17.43	0	121.10
Ret_den	Retail job density	0.79	2.02	0	17.45
% age 5_17	% of population age 5-17	8.64%	8.78%	0	27%
% age 18_24	% of population age 18-24	5.79%	7.42%	0	40%
% age 24_64	% of population age 24-64	38.35%	36.12%	0	95%
% age 65+	% of population age 65 or older	6.25%	10.21%	0	83%
K12	K12 student enrollment	0.39	1.00	0	5.52
College	College student enrollment	0.11	1.00	0	12.59
Int34_den	Intersection density (3- and 4- legs)	0.12	0.12	0	0.62
BKlnACC	Bike lane access (1=if a TAZ has bike lane)	0.92	0.28	0	1
BL_den	Bike lane density	3.40	1.80	0	7.26
Rail	1=at least one rail station in a TAZ	0.01	0.10	0	1
TTbus_D	Total Bus Stop Density	0.05	0.09	0	0.53
Exbus_D	Stop density for Express Bus and BRT	0.002	0.007	0	0.06
HFLbus_D	High-Frequency Bus Stop Density (local bus headway <= 20 mins)	0.001	0.004	0	0.03
WalkAcc	Walk Accessibility	3.87	9.46	0	74.53
% Arterial	Percent of main arterial (45-55mph) of TAZ	10.61%	17.33%	0	80%
Distance	Distance among TAZ centroids (in miles)	4.06	2.09	0.16	11.78

Note: S.D. represents standard deviation.

4. RESULTS

Similar to the methodology and data description, the results are also presented in the order of the micro level, county level and the TAZ level.

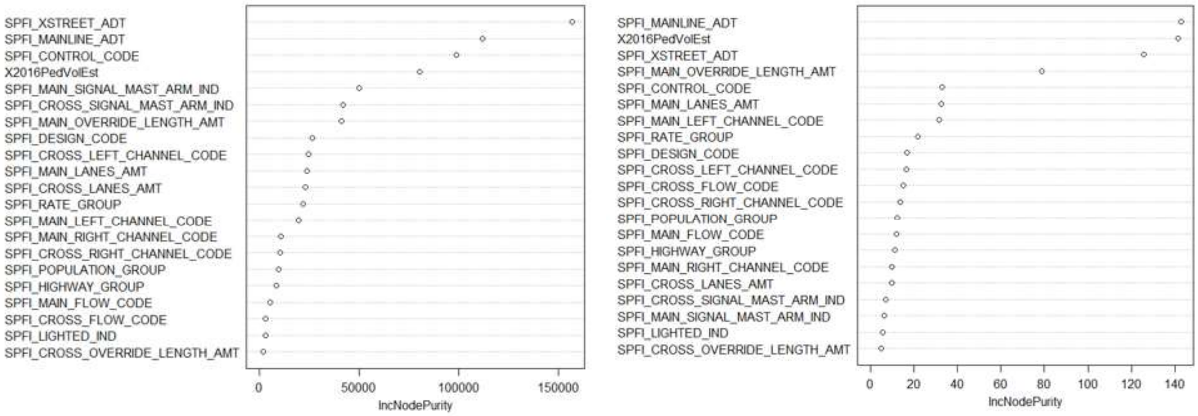
4.1 Results for the Micro-Level SPF

The micro level data consist of intersection and ramp data. For the intersection-related SPF, the SPFs for both pedestrians and vehicles were developed jointly via the Bayesian joint hierarchical model for those intersections with both pedestrian and vehicle volumes are available. However, lots of other intersections where the pedestrian volumes are unavailable, for these intersections, the Bayesian joint models results cannot be used due to the lack of the pedestrian volumes. To address this issue, the SPF(P)'s were first developed for the intersections where the pedestrian volumes are available and included in the model. The SPF(V)'s were then developed for the same intersections where the pedestrian volumes are available but excluded from the model, or, only the vehicle volume is included in the mode. The ratio of SPF(P) to SPF(V) can then be used as a rough adjustment factor to estimate the crash count at the intersections where only vehicle volume is available. It is important to note that both SPF(P) and SPF(V) were developed using the Negative Binomial model. Finally, due to ramp data availability, the SPFs for both pedestrians and vehicles were also developed jointly via the Bayesian joint hierarchical model for the ramps.

4.1.1 Bayesian Joint Hierarchical Model Results for Intersections

To develop the bivariate joint model, the distinct covariates were selected for pedestrians and vehicles by using random forest metric and correlation analysis. Under the INLA framework, models were developed with the posterior mean serving as the estimate for the model parameters Different evaluation criteria were used to assess the predictive accuracy of the models.

a) Feature Importance Ranking by Random Forest



(a) Vehicle **(b) Pedestrian**

Figure 2. A Variable Importance Plot for (a) Vehicle Crash Counts and (b) Pedestrian Crash Counts

The importance of variables was reported and ranked using the R package “randomforest” (Cutler et al., 2012). When estimating the random forest model, m = 4 variables were randomly sampled as candidate at each split, with the OOB error rate reaching a minimum value of 0.132 and 59.24% of data variability being explained by the model. The variable importance plots for both pedestrian and vehicles are shown in Figure 1 with the decreasing order of “IncNodePurity”, which represents the mean decrease of node purity in predictions on OOB samples with a given variable being excluded from the model (James et al., 2017).

b) Correlation Analysis for Covariates

Table 5. Correlation Coefficients and P-Value for the Numerical Variables

	MNL	MOL	X-NL	X-OL	MADT	X-ADT	IRG	APV
MNL	1.000	-0.116	0.230	0.001	**0.722**	0.273	0.118	0.254
MOL	0.000	1.000	0.046	-0.001	-0.088	0.032	-0.139	-0.152
X-NL	0.000	0.000	1.000	0.118	0.227	**0.600**	-0.087	0.112
X-OL	0.925	0.917	0.000	1.000	0.005	0.187	-0.034	0.194
MADT	0.000	0.000	0.000	0.683	1.000	0.269	0.154	0.257
X-ADT	0.000	0.014	0.000	0.000	0.000	1.000	-0.118	0.184
IRG	0.000	0.000	0.000	0.008	0.000	0.000	1.000	-0.046
APV	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

Notes: 1. The upper triangle of the matrix shows the correlation coefficients of the variables, and the gray grids in the lower triangle of the matrix shows the p-values. 2. Highly correlated estimate with correlation coefficient greater than 0.6 are marked as bold font. 3. Refer to Table 1 for details of variable definition.

Variable importance ranking was used along with the correlation of numerical variables for determination of the covariate inputs to the model development. The correlation tests were conducted using the Harrell Miscellaneous package in R software which allowed the calculation of Pearson's correlation coefficient and the accompanying p-values. The variables observed to be correlated by using the popular cut line of 0.6 for the correlation coefficient and with a significance level of 0.05 were eliminated in multiple steps using engineering judgment to choose the minimum subset of variables while maintaining the maximum data variability. In other words, the selection procedure strived to strike the balance between omitted variable bias and multi-collinearity issues. As shown in Table 5, the upper portion values are the Pearson's correlation coefficient magnitudes and lower shaded cells represent the associated p-values. Based on the results of correlation test, out of eight numerical variables, six of them which include MOL, X-OL, MADT, X-ADT, IRG and APV were retained.

Combining both results from random forest and correlation analysis, the final list of predictors to be fed into subsequent model development can be found in Table 6. It is important to note that variables of "Right Turn" and "X-OL" were retained only for pedestrians, while "X-Mastarm" and "X-Right Turn" were included only for vehicles. They were not considered for both modes at the same time since they had little influence on one of the modes according to the variable importance ranking results via RF.

c) **Joint Model Estimates**

Table 6. Description of Model Parameter Estimates

Variables		**β1 (Pedestrian)**		**β2 (Vehicle)**	
		Mean	**SD**	**Mean**	**SD**
		Fixed Effects			
(Intercept)		**-5.222**	**0.772**	**3.788**	**0.735**
Highway Group	Highway Group 1 (Base)				
	Highway Group 2	-0.008	0.053	-0.082	0.141
	Highway Group 3	0.494	0.37	-0.941	0.831
Population Group	Population Group 1 (Base)				
	Population Group 2	**-0.834**	**0.092**	-0.38	0.321
	Population Group 3	**0.247**	**0.051**	0.088	0.141
Intersection Design	Intersection Design 1 (Base)				
	Intersection Design 2	-0.215	0.176	-0.34	0.448
	Intersection Design 3	**-0.341**	**0.085**	**-0.401**	**0.21**
	Intersection Design 4	**0.391**	**0.157**	**-1.761**	**0.485**
	Intersection Design 5	**0.485**	**0.189**	**-3.634**	**1.107**
	Intersection Design 6	**0.681**	**0.246**	**-1.994**	**0.779**
Light Condition	Light Condition 1 (Base)				
	Light Condition 2	**0.307**	**0.054**	**0.747**	**0.213**
Mastarm	Mastarm 1 (Base)				

	Mastarm 2	-0.207	0.131	-0.087	0.271
Left Turn	Left Turn 1 (Base)				
	Left Turn 2	**-0.264**	**0.078**	-0.041	0.183
	Left Turn 3	0.048	0.063	-0.152	0.14
	Left Turn 4	**0.572**	**0.284**	**1.682**	**0.559**
Right Turn	Right Turn 1				
	Right Turn 2	N/A	N/A	**-0.246**	**0.153**
Traffic Flow	Traffic Flow 1 (Base)				
	Traffic Flow 2	**0.871**	**0.148**	-0.025	0.336
	Traffic Flow 3	0.338	0.256	0.784	0.455
X-Mastarm	X-Mastarm 1 (Base)				
	X-Mastarm 2	**0.251**	**0.077**	N/A	N/A
X-Right Turn	X-Right Turn 1 (Base)				
	X-Right Turn 2	0.038	0.068	N/A	N/A
X-Left Turn	X-Left Turn 1 (Base)				
	X-Left Turn 2	0.118	0.14	0.182	0.264
	X-Left Turn 3	0.159	0.136	0.232	0.258
	X-Left Turn 4	0.079	0.353	-1.311	1.125
X-Traffic Flow	X-Traffic Flow 1 (Base)				
	X-Traffic Flow 2	0.143	0.144	0.169	0.345
	X-Traffic Flow 3	0.292	0.224	0.234	0.473
Intersection Control Condition	Intersection Control Condition 1 (Base)				
	Intersection Control Condition 2	**1.134**	**0.137**	0.708	0.503
	Intersection Control Condition 3	**1.975**	**0.201**	0.922	0.554
	Intersection Control Condition 4	**2.218**	**0.219**	**1.446**	**0.59**
	Intersection Control Condition 5	**1.899**	**0.192**	**1.168**	**0.566**
	Intersection Control Condition 6	**1.611**	**0.186**	0.662	0.578
MOL		**0.002**	**0.001**	-0.001	0.001
X-OL		N/A	N/A	**0.002**	**0.001**
MADT		**2.997**	**0.215**	**1.845**	**0.497**
X-ADT		**3.581**	**0.388**	**2.598**	**0.718**
IRG		**-0.076**	**0.01**	0.048	0.031
APV		-0.146	0.451	0.429	0.775
		Random Effects			
Observation. ID		**0.643**	**0.018**	**2.297**	**0.462**
Goodness-of-fit Criteria					
DIC		29113.63			
$\overline{D}$		24753.94			

P_D	43596.89
LPML	-32668.16

Notes: Notes: 1. S.D. represents standard deviation; DIC represents deviance information criterion; $\overline{D}$ represents posterior mean deviance; P_D represents effective number of parameters; LPML represents log pseudo marginal likelihood; NA means Not Applicable. 2. Refer to Table 1 for details of variable definition. 3. The bold fonts represent the variables with statistically significant impact.

The posterior model estimates of model parameters across pedestrian and vehicle crash counts are shown in Table 6. The estimated coefficients for ten influential variables including 'Intersection Design 2' (>four legs), 'Intersection Design 3'(offset), 'Intersection Design 4' (tee), 'Intersection Design 5' (wye), 'Intersection Design 6' (others), 'Light Condition 2' (Lighted), 'Left turn 4' (No left turn channelization), 'Intersection Control Condition 4' (signals semi-traffic actuated, two phase), 'Intersection Control Condition 5' (signals full traffic actuated, multi-phase), 'MADT' (mainline-average daily traffic), and 'X-ADT' (crossline-average daily traffic), appeared to be statistically significant across both pedestrian and vehicle crash counts. Interestingly, among these significant covariates, five variables were found to have a negative impact, in which one variable ('Intersection Design 3') was common for both modes and other four variables were observed for vehicles crash count only (or, 'Intersection Design 4', Intersection Design 5', Intersection Design 6', 'Right Turn 2'). It follows that, compared with the base condition of four-legged intersection, offset intersection seems to be more advantageous in terms of traffic safety for both pedestrians and vehicle drivers. For vehicle drivers only, the tee and wye intersections and those without right-turn channels tend to provide more safety benefits compared with the base conditions of four-leg and intersection with right turn channels, respectively. The better safety performance associated with those without right turn channels is somewhat counterintuitive, which warrants further verifications from other studies.

At the individual mode level, ten covariates which contains 'Population Group 2' (rural), 'Population Group 3' (urbanized), 'Left Turn 2' (no left turn channelization), 'Traffic Flow 2' (two-way traffic, left turn permitted), 'X-Mastarm 2' (presence of signal mastarm), 'Intersection Control Condition 2' (stop signs on cross street only), 'Intersection Control Condition 3' (signals pretimed), 'Intersection Control Condition 6' (others), 'MOL'(mainline - overide length), and 'IRG' (intersection rate group) observed to be statistically significant for pedestrians. Similarly, for vehicles, there are two statistically significant variables which include 'Right Turn 2' (right turn channelization) and 'X-OL' (cross street - override length). Such phenomenon indicates that, relative to drivers, pedestrians are not only subject to more injury severities, but also sensitive to more intersection features such as left turn channelization, intersection control, and so on.

This study also employed various types of evaluation criteria including ($\overline{D}$) (measure of training errors), DIC (indirect measure of test errors), and LPML (measure of test errors) to assess the models from different perspectives. Under close review of the evaluation results, it is obvious that DIC is the sum of $\overline{D}$ and P_d, where P_d serves as the correction term to the in-sample error so that DIC can approximate the out-of-sample error. Different from DIC, the LPML provide a direct cross validation-oriented error. Both values have relatively large magnitude (or, 29113.63 and -32668.16) due to the large sample size of the intersections used in the study, that is, 6,198.

Table 7. Correlation and Covariance Matrix between the Random Effects of Pedestrian and Vehicle Counts

Observation. ID	β1 (Pedestrian)	β2 (Vehicle)
β1 (Pedestrian)	1.000	**0.899**
β2 (Vehicle)	0.344	1.000

Notes: 1.The lower triangle of the matrix shows the covariance, while the upper triangle of the matrix shows the associated correlation coefficient. 2. The bold font indicates the statistics is statistically significant at the significance level of 0.05. 3. Correlation coefficients are listed in the diagonal of the matrix at the same time. 4. Refer to Equation 6 for definition of covariance of the two random effects.

To better explore the suitability of using the bivariate setting, the random effects of the two transportations modes were also collected. Their correlation and covariance are shown in Table 7. The statistically significant correlation coefficient signifies the strong positive correlation between the two types of crashes, corroborating the importance of developing the joint models where the correlation between the two response variables was explicitly considered.

4.1.2 Development of SPF(P) & SPF(V) for Intersections Via Negative Binomial Model

As previously mentioned, both SPF(P) and SPF(V) were developed separately using the Negative Binomial models where the ratio of SPF(P) to SPF(V) can be utilize to estimate the pedestrian-related counts at the intersections where only the vehicle volumes are available. The calculation of such ratio is important given the difficulty to collect the pedestrian volumes and most intersections do not have such volume information. The two SPFs function are presented as follows. For any intersections, once the values for the covariates are known beforehand, the ratio can then be easily calculated to estimate the pedestrian crash counts.

SPF (P) = exp(-33.45 – 0.6959 * Highway_Group(L) – 2.012 * Highway_Group(R) – 0.1614 * Highway_Group(U) – 0.2883 * Population_Group(R) + 0.1643 * Population_group(U) – 0.3789 * Design_Code(M) **- 0.5059 * Design_Code(S) - 1.935 * Design_Code(T) - 3.655 * Design_Code(Y) – 2.082 * Design_Code(Z) + 0.6229 * Lighted_Ind(Y)** + 0.5730 * Main_Signal_Mast_Arm_Ind(Y) + 0.1055 * Main_Left_Channel_Code(N) – 0.06056 * Main_Left_Channel_Code(P) + 1.599 * Main_Left_Channel_Code(R) + 1.108 * Main_Left_Channel_Code(Y) - 0.7116 * Main_Right_Channel_Code(N) - 0.4238 * Main_Right_Channel_Code(P) - 0.9187 * Main_Right_Channel_Code(Y) + 0.4112 * Main_Flow_Code(P) + 1.136 * Main_Flow_Code(R) + 1.788 * Main_Flow_Code(W) +0.8833 * Main_Flow_Code(Z) – **0.4608 * Cross_Signal_Mast_Arm_Ind(Y) -** 0.2015 * Cross_Left_Channel_Code(N) + 0.04076 * Cross_Left_Channel_Code(P) – 0.3690 * Cross_Left_Channel_Code(R) - 34.71 * Cross_Left_Channel_Code(Y) – 1.465 * Cross_Right_Channel_Code(N) – 1.130 * Cross_Right_Channel_Code(N) – 1.827 * Cross_Right_Channel_Code(Y) + 0.1194 * Cross_Flow_Code(P) + 1.440 * Cross_Flow_Code(R) +0.04587 Cross_Flow_Code (W) – 0.2921 * Cross_Flow_Code(Z) + 0.4822 * Control_Code(B) – 32.94 Control_Code(C) +1.024 * Control_Code(D) + 1.086 * Control_Code(E) -33.69 * Control_Code(F) – 33.90 * Control_Code(G) – 32.21 * Control_Code(H) + 0.2521 * Control_Code(J) + 0.7080 * Control_Code(K) – 0.7541 * Control_Code(L) + 0.2734 * Control_Code(M) + 0.6973 * Control_Code(N) + 0.3742 * Control_Code(P) + 1.514 * Control_Code(Z) + 0.08851 * Main_Lanes_AMT **- 0.002416 * Main_Override_Length_AMT** – 0.1326 * Cross_Lane_AMT +

0.002418 * Cross_Override_Length_AMT + 3.079 * 10^{-6} * Mainlane_ADT + **2.721 * 10^{-5} Cross_ADT** + 0.06172 * Rete_Group + 7.727 * 10^{-8} * Ped_Volume)

SPF(V) = exp(-0.5155 + 1.234 * Highway_Group(L) + **1.700 * Highway_Group(R)** – 0.04159 * Highway_Group(U) – **0.6334 * Population_Group(R)** + **0.2074 * Population_group(U)** – **0.3364 * Design_Code(M) - 0.3521 * Design_Code(S)** + 0.1232 * Design_Code(T) + 0.2753 * Design_Code(Y) + 0.4219 * Design_Code(Z) + **0.2221 * Lighted_Ind(Y)** + 0.04814 * Main_Signal_Mast_Arm_Ind(Y) - 0.1247 * Main_Left_Channel_Code(N) + 0.1073 * Main_Left_Channel_Code(P) + 0.08698 * Main_Left_Channel_Code(R) + 0.6138 * Main_Left_Channel_Code(Y) – 1.148 * Main_Right_Channel_Code(N) – 0.9084 * Main_Right_Channel_Code(P) – 1.028 * Main_Right_Channel_Code(Y) + **0.7688 * Main_Flow_Code(P)** + 0.5570 * Main_Flow_Code(R) – 0.8873 * Main_Flow_Code(W) +0.2613 * Main_Flow_Code(Z) + 0.03921 * Cross_Signal_Mast_Arm_Ind(Y) + 0.09065 * Cross_Left_Channel_Code(N) + 0.1435 * Cross_Left_Channel_Code(P) - 0.1662 * Cross_Left_Channel_Code(R) - 0.05379 * Cross_Left_Channel_Code(Y) + 0.9430 * Cross_Right_Channel_Code(N) + 0.8555 * Cross_Right_Channel_Code(N) + 0.9924 * Cross_Right_Channel_Code(Y) + 0.06795 * Cross_Flow_Code(P) – 0.1321 * Cross_Flow_Code(R) + 0.2146 * Cross_Flow_Code (W) + 0.1416 * Cross_Flow_Code(Z) + **1.085 * Control_Code(B)** + **1.713 * Control_Code(C)** + **1.522 * Control_Code(D)** + **1.539 Control_Code(E)** + **2.030 * Control_Code(F)** + 1.140 * Control_Code(G) – 0.3689 * Control_Code(H) + 1.625 * Control_Code(J) + 0.8943 * Control_Code(K) + **1.254 * Control_Code(L)** + **1.584 * Control_Code(M)** + **1.835 * Control_Code(N)** + **1.497 * Control_Code(P)** + **0.9204 * Control_Code(Z)** + 0.01029 * Main_Lanes_AMT + **0.002098 * Main_Override_Length_AMT** – 0.02036 * Cross_Lane_AMT + 1.444 * 10^{-4} * Cross_Override_Length_AMT + **2.311 * 10^{-5} * Mainlane_ADT** + **4.971 * 10^{-5} Cross_ADT** + **0.05554 * Rete_Group**)

Note: The bold fonts represent the variables which are statistically significant at the significance level of 0.05 in the negative binomial model.

As displayed above, SPFs consist of various categorical and numerical variables which contain plenty of intersection characteristics (number of lanes, traffic control type, average daily traffic, etc). Equations revealed the impacts of each variable on corresponding target variable, and it's able to predict average number of crashes at certain location. In both equations, there are several categorical variables. To substitute suitable categorical values into SPF, each categorical level was listed in the equations, so that we can directly substitute observed categorical value into equation. The following tables illustrate the base condition and Crash Modification Factors(CMFs) of each categorical variable.

Table 8. Base Conditions of SPFs based on Negative Binomial Models

Variable	**Base Condition**	**Description of Base Condition**
Highway Group	Highway Group 1	Divided highway
Population Group	Population Group 1	Urban
Design Code	Design Code 1	4-Legged intersection
Lighted Ind	Lighted Ind 1	No lighting

Signal Mast Arm Ind	Signal Mast Arm Ind 1	No mast arm
Main Left Channel Code	Main Left Channel Code 1	Curbed median left turn channelization
Main Right Channel Code	Main Right Channel Code 1	Curbed median right turn channelization
Main Flow Code	Main Flow Code 1	2-Way traffic, no left turns permitted
Cross Mast Arm Ind	Cross Mast Arm Ind 1	No mast arm
Cross Left Channel Code	Cross Left Channel Code 1	Curbed median left turn channelization
Cross Right Channel Code	Cross Right Channel Code 1	Curbed median right turn channelization
Cross Flow Code	Cross Flow Code 1	2-Way traffic, no left turns permitted
Traffic Control Code	Traffic Control Code 1	No control

Table 9. Crash Modification Factors Associated with the Base Conditions

Variables	**Crash Modification Factor (Pedestrian)**	**Crash Modification Factor (Vehicle)**
Highway Group 1 (Base)	CMF = 0	CMF = 0
Highway Group 2	CMF = exp (-0.6959) =0.4986	CMF = exp (1.234) =3.4349
Highway Group 3	CMF = exp (-2.012) =0.1337	CMF = exp (1.700) =5.4739
Highway Group 4	CMF = exp (-0.1614) =0.851	CMF = exp (-0.04159) =0.9593
Population Group 1 (Base)	CMF = 0	CMF = 0
Population Group 2	CMF = exp (-0.2883) =0.7495	CMF = exp (-0.6334) =0.5308
Population Group 3	CMF = exp (0.1643) =1.1786	CMF = exp (0.2074) =1.2305
Design Code 1 (Base)	CMF = 0	CMF = 0
Design Code 2	CMF = exp (-0.3789) =0.6846	CMF = exp (-0.3364) =0.7143
Design Code 3	CMF = exp (-0.5059) =0.603	CMF = exp (-0.3521) =0.7032
Design Code 4	CMF = exp (-1.935) =0.1444	CMF = exp (0.1232) =1.1311
Design Code 5	CMF = exp (-3.655) =0.0259	CMF = exp (0.2753) =1.3169

Design Code 6	CMF = exp (-2.082) =0.1247	CMF = exp (0.4219) =1.5249
Lighted Ind 1 (Base)	CMF = 0	CMF = 0
Lighted Ind 2	CMF = exp (0.6229) =1.8643	CMF = exp (0.2221) =1.2487
Main Signal Mast Arm 1 (Base)	CMF = 0	CMF = 0
Main Signal Mast Arm 2	CMF = exp (0.5730) =1.7736	CMF = exp (0.04814) =1.0493
Main Left Channel Code 1 (Base)	CMF = 0	CMF = 0
Main Left Channel Code 2	CMF = exp (0.1055) =1.1113	CMF = exp (-0.1247) =0.8828
Main Left Channel Code 3	CMF = exp (0.06056) =1.0624	CMF = exp (0.1073) =1.1133
Main Left Channel Code 4	CMF = exp (1.108) =3.0283	CMF = exp (0.08698) =1.0909
Main Left Channel Code 5	CMF = exp (1.599) =4.9481	CMF = exp (0.6138) =1.8474
Main Right Channel Code 1 (Base)	CMF = 0	CMF = 0
Main Right Channel Code 2	CMF = exp (-0.7116) = 0.4909	CMF = exp (-1.148) =0.3173
Main Right Channel Code 3	CMF = exp (-0.4238) =0.6546	CMF = exp (-0.9084) =0.4032
Main Right Channel Code 4	CMF = exp (-0.9187) =0.3990	CMF = exp (-1.028) =0.3577
Main Flow Code 1 (Base)	CMF = 0	CMF = 0
Main Flow Code 2	CMF = exp (0.4112) =1.5086	CMF = exp (0.7688) =2.1572
Main Flow Code 3	CMF = exp (1.136) =3.1143	CMF = exp (0.5570) =1.7454
Main Flow Code 4	CMF = exp (1.788) =5.9775	CMF = exp (-0.8873) =0.4118
Main Flow Code 5	CMF = exp (0.8833) =2.4189	CMF = exp (0.2613) =1.2986
Cross Signal Mast Arm 1 (Base)	CMF = 0	CMF = 0
Cross Signal Mast Arm 2	CMF = exp (-0.4608) =0.6308	CMF = exp (0.03921) =1.04

Cross Left Channel Code 1 (Base)	CMF = 0	CMF = 0
Cross Left Channel Code 2	CMF = exp (-0.2015) =0.8175	CMF = exp (0.09065) =1.0949
Cross Left Channel Code 3	CMF = exp (0.04076) =1.0416	CMF = exp (0.1435) =1.1543
Cross Left Channel Code 4	CMF = exp (-0.3690) =0.6914	CMF = exp (-0.1662) =0.8469
Cross Left Channel Code 5	CMF = exp (-34.71) =0	CMF = exp (-0.05379) =0.9476
Cross Right Channel Code 1 (Base)	CMF = 0	CMF = 0
Cross Right Channel Code 2	CMF = exp (-1.465) =0.2311	CMF = exp (0.9430) =2.5677
Cross Right Channel Code 3	CMF = exp (-1.130) =0.323	CMF = exp (0.8555) =2.3526
Cross Right Channel Code 4	CMF = exp (-1.827) =0.1609	CMF = exp (0.9924) =2.6977
Cross Flow Code 1 (Base)	CMF = 0	CMF = 0
Cross Flow Code 2	CMF = exp (0.1194) =1.1268	CMF = exp (0.06795) =1.0703
Cross Flow Code 3	CMF = exp (1.440) =4.2207	CMF = exp (-0.1321) =0.8763
Cross Flow Code 4	CMF = exp (0.04587) =1.0469	CMF = exp (0.2146) =1.2394
Cross Flow Code 5	CMF = exp (-0.2921) =0.7467	CMF = exp (0.1416) =1.1521
Control Code 1 (Base)	CMF = 0	CMF = 0
Control Code 2	CMF = exp (0.4822) =1.6196	CMF = exp (1.085) =2.9594
Control Code 3	CMF = exp (-32.94) =0	CMF = exp (1.713) =5.5456
Control Code 4	CMF = exp (1.024) =2.7843	CMF = exp (1.522) =4.5814
Control Code 5	CMF = exp (1.086) =2.9624	CMF = exp (1.539) =4.6599
Control Code 6	CMF = exp (-33.69) =0	CMF = exp (2.030) =7.6141
Control Code 7	CMF = exp (-33.90) =0	CMF = exp (1.140) =3.1268
Control Code 8	CMF = exp (-33.21) =0	CMF = exp (-0.3689) =0.6915
Control Code 9	CMF = exp (0.2521) =1.2867	CMF = exp (1.625) =5.0784

Control Code 10	CMF = exp (0.7080) =2.0299	CMF = exp (0.8943) =2.4456
Control Code 11	CMF = exp (-0.7541) =0.4704	CMF = exp (1.254) =3.5043
Control Code 12	CMF = exp (0.2734) =1.3144	CMF = exp (1.584) =4.8744
Control Code 13	CMF = exp (0.6973) =2.0083	CMF = exp (1.835) =6.2651
Control Code 14	CMF = exp (0.3742) =1.4538	CMF = exp (1.497) =4.4683
Control Code 15	CMF = exp (1.154) =3.1709	CMF = exp (0.9204) =2.5103
Main lanes AMT	CMF = exp (0.08851) =1.0925	CMF = exp (0.01029) =1.0103
Main Override Length AMT	CMF = exp (-0.002416) =0.9976	CMF = exp (0.002098) =1.0021
Cross Lanes AMT	CMF = exp (-0.1326) =0.8758	CMF = exp (-0.02036) =0.9798
Cross Override Length	CMF = exp (0.002418) =1.0024	CMF = exp ($-1.444*10^{4}$) =1.0001
Mainline ADT	CMF = exp ($3.079*10^{-6}$) =1.0000	CMF = exp ($2.311*10-^{5}$) =1.0000
Cross ADT	CMF = exp ($2.721*10^{-5}$) =1.0000	CMF = exp ($4.971*10-^{5}$) =1.0000
Rate Group	CMF = exp (0.06172) =1.0637	CMF = exp (0.05554) =1.0571
Ped Volume	CMF = exp ($7.727*10^{-7}$) =1.0000	NA

Note: In order to fit statistical model more accurate, the original value of three variables (Mainline ADT, Cross ADT, Ped Volume) is reduced by 1000 times when imported data into the model. Therefore, when using SPF to predict the number of accidents, the substituted value should also be reduced by 1000 times.

4.1.3 Bayesian Joint Hierarchical Model Results for Ramps

Given the importance of ramp, this project also developed ramp SPF of pedestrian and vehicle using Bayesian joint hierarchical model. The model inputs contain 5 categorical variables and ramp ADT across pedestrian and vehicle crash counts shown in Table 10. The estimated coefficients for 'Ramp On/Off Code 2' (On ramp), 'Ramp Area 4 Indicator 2' (No), 'Average Daily Traffic', seems to be statistically significant with both pedestrian and vehicle. It appears that, compared to base condition of off ramp, on ramp has negative impact on both traffic model. In addition, ramps which are associated with "Area 4" seems more advantageous in terms of traffic safety for both modes. For vehicle crashes only, 'Highway Group 4'(Undivided Highway), 'Ramp Design Code 3'(Dir/Semi Left Ramp), 'Ramp Design Code 5' (Slip Ramp), 'Ramp Design Code' (Dir/Semi Right Ramp), 'Ramp Design Code 8'(Buttonhook ramp), 'Ramp Design Code 9' (Scissors Ramp), 'Ramp Design Code 10' (Split Ramp), 'Ramp Design Code 14' (Rest Area/Vista), 'Ramp Design Code 16' (Others), 'Ramp On/ Off Code 3'(Others), 'Ramp Population Group 2' (Rural), tend to provide more safety benefits compared with related base condition. Compared with rural and urban, urbanized area provides less safety performance to vehicle.

Table 10. Description of Model Parameter Estimates for Ramps

Variables		**β1 (Pedestrian)**		**β2 (Vehicle)**	
		Mean	**SD**	**Mean**	**SD**
		Fixed Effects			
(Intercept)		**-5.731**	**1.191**	**6.519**	**1.167**
Highway Group	Highway Group 1 (Base)				
	Highway Group 2	0.477	0.321	-0.159	0.107
	Highway Group 3	-0.386	0.463	-0.206	0.117
	Highway Group 4	-0.013	0.727	**-0.605**	**0.162**
Ramp Design Code	Ramp Design Code 1(Base)				
	Ramp Design Code 2	0.339	1.094	-0.414	0.245
	Ramp Design Code 3	0.466	1.186	**-0.549**	**0.270**
	Ramp Design Code 4	0.981	1.173	-0.133	0.266
	Ramp Design Code 5	-0.427	1.215	**-1.335**	**0.274**
	Ramp Design Code 6	0.401	1.175	**-0.542**	**0.266**
	Ramp Design Code 7	0.175	1.193	-0.213	0.270
	Ramp Design Code 8	-0.324	1.188	**-0.577**	**0.268**
	Ramp Design Code 9	-0.547	1.242	**-0.934**	**0.275**
	Ramp Design Code 10	-0.610	1.190	**-1.295**	**0.270**
	Ramp Design Code 11	-0.008	1.182	-0.324	0.267
	Ramp Design Code 12	-4.547	13.396	-0.031	0.312
	Ramp Design Code 13	-7.023	11.564	-12.717	8.716
	Ramp Design Code 14	0.302	1.256	**-0.874**	**0.278**
	Ramp Design Code 15	-7.140	11.517	-13.122	8.601
	Ramp Design Code 16	0.595	1.206	**-0.619**	**0.280**
Ramp On/Off Code	Ramp On/Off Code 1(Base)				
	Ramp On/Off Code 2	**-0.415**	**0.075**	**-0.473**	**0.021**
	Ramp On/Off Code 3	-1.150	0.651	**-1.017**	**0.173**
Ramp Area 4 Indicator	Ramp Area 4 Indicator 1 (Base)				
	Ramp Area 4 Indicator 2	**0.692**	**0.128**	**0.300**	**0.030**
Ramp Population Group	Ramp Population Group 1 (Base)				
	Ramp Population Group 2	**-1.340**	**0.342**	**-0.634**	**0.056**
	Ramp Population Group 3	**1.132**	**0.233**	**1.024**	**0.049**
Average Ramp ADT (15~17)		**$5.1*10^{-5}$**	**$5*10^{-6}$**	**$7.3*10^{-5}$**	**$1*10^{-6}$**
		Random Effects			
Observation. ID		**0.870**	**0.014**	**2.873**	**0.399**
Goodness-of-fit Criteria					
DIC		4102463			
$\overline{D}$		2078761			
P_D		2023702			
LPML		-46863.44			

To better explore the suitability of using the bivariate setting, the random effects of the two transportations modes were also collected. Their correlation and covariance are shown in Table 11. The statistically significant correlation coefficient signifies the strong positive correlation between the two types of crashes, corroborating the importance of developing the joint models where the correlation between the two response variables was explicitly considered.

Table 11. Correlation and Covariance Matrix between the Random Effects of Pedestrian and Vehicle Counts

Observation. ID	β1 (Pedestrian)	β2 (Vehicle)
β1 (Pedestrian)	1.000	**0.995**
β2 (Vehicle)	0.448	1.000

Notes: 1. The lower triangle of the matrix shows the covariance, while the upper triangle of the matrix shows the associated correlation coefficient. 2. The bold font indicates the statistics is statistically significant at the significance level of 0.05. 3. Correlation coefficients are listed in the diagonal of the matrix at the same time. 4. Refer to Equation 6 for definition of covariance of the two random effects.

4.2 Results for the County-Level SPF

4.2.1 Model Estimates

In general, the larger the effective number of parameters is, the easier it is for the model to fit the data. To obtain a parsimonious model and avoid the risk of inclusion of collinear variables, backward stepwise and multi-collinearity methods were employed in selecting covariates, respectively. Besides, a correlation matrix for the variables entered in the final models has been checked to avoid multi-collinearity issues. Results of parameter estimation and associated uncertainty estimates of significant variables in the final models are presented in Table 12. It is known that mostly the same significant variables are identified for all five models across different crash modes. The robustness of results indicates that the models yield mostly consistent inferences by selecting the influential factors of crashes. The variable coefficients change little across the five multivariate models. Comparatively speaking, the change of coefficients for the smaller sample size outcomes (motorcyclists, bicyclists, and pedestrians) is larger than that of the larger one (vehicle-only). It is also worth mentioning that some coefficients exhibit varying relationship (positive and negative) for the same crash mode across different models. For example, the positive coefficient of Median House Income in case of Model 1 motorcycle crashes changes to a negative coefficient for rest of the models. This may be explained by the inclusion of additional spatial and temporal specifications for rest of the models which account for the unobserved heterogeneity that may have escaped the explanatory variables, which are account for the variability in case of Model 1. The consistent negative relationship between Median House Income and motorcycle crashes for four models may suggest that the inclusion of space-time specifications generated accurate posterior estimates and should be incorporated for obtaining more informed inferences. Another example of reversing relationship between dependent and independent variable across five models is the case of yearly trend for bike and motorcycle crashes. The contradictory coefficient signs for Model 2 and Model 4 may be explained by the different treatment of temporal correlation. Model 2 attempts to address the temporal changes by introducing a linear trend for each of four crash modes while Model 4 adopts a more restrictive approach of a fixed coefficient across all modes. The fixed approach seems to be limited in accounting for the temporal instability for multiple modes as it intuitive to perceive that different crash modes would experience a varying impact of time (Readers are referred to Mannering (2018) for complete review of temporal instability in crash data). The model results also demonstrate consistency in terms of variable significance as except for maintain miles of roadways, rest of the explanatory variables were observed to be consistently significant across different modes and models. As for the yearly trend, if the fixed trend is assumed for all crash types, the coefficient is significant with a negative value. However, if the varying trends are expected for different crash outcomes, the coefficient was not significant for bicycle crashes.

Table 12. Estimates of Regression Coefficients Obtained by various Multivariate Models

Crash Types	Variables	Model 1: MVPLN	Model 2: MVPLNT	Model 3: MVPLNS	Model 4: MVPLNST (fixed time coefficient)	Model 5: MVPLNST (varying time coefficients)
Motorcycle	Intercept	**-4.136 (0.3174)**	**-4.788 (0.2594)**	**-2.904 (0.1903)**	**203.3 (12.51)**	**-3.773 (0.157)**
	Population	**-0.1246 (0.01406)**	**-0.1595 (0.01351)**	**-0.3 (0.03681)**	**-0.07534 (0.01261)**	**-0.1509 (0.009309)**
	Maintain Miles	**0.01771 (0.01096)**	0.008537 (0.01257)	0.01726 (0.01005)	0.007134 (0.01105)	0.003428 (0.01105)
	Mean Travel Time	**0.7519 (0.05739)**	**1.013 (0.03292)**	**1.219 (0.04727)**	**1.312 (0.1081)**	**1.243 (0.05635)**
	Median House Income	**0.01055 (0.03558)**	-0.0183 (0.0247)	**-0.08794 (0.02847)**	**-0.1458 (0.03375)**	**-0.1632 (0.03605)**
	Land Area	**-0.1773 (0.04202)**	-0.09556 (0.06168)	**-0.1182 (0.03334)**	-0.03841 (0.04562)	**-0.1168 (0.04535)**
	Yearly trend	NA	**0.0118 (0.00229)**	NA	**-0.02315 (0.001936)**	**-0.0268 (0.003868)**
Bike	Intercept	**38.4 (0.2814)**	**35.8 (0.1714)**	**37.11 (0.1224)**	**1611 (89.29)**	**37.73 (0.1686)**
	Population	**-1.206 (0.0161)**	**-1.08 (0.01198)**	**-0.9042 (0.02227)**	**-0.1633 (0.08462)**	**-0.8376 (0.01467)**
	Maintain Miles	0.01595 (0.008597)	0.009649 (0.00803)	0.01801 (0.01002)	**0.02482 (0.009951)**	0.01493 (0.009377)
	Mean Travel Time	**-3.975 (0.05043)**	**-4.552 (0.1291)**	**-4.13 (0.04431)**	**-3.891 (0.07165)**	**-4.136 (0.1293)**
	Median House Income	**-0.6162 (0.0205)**	**-0.5001 (0.02759)**	**-0.8102 (0.02916)**	**-1.1 (0.02324)**	**-0.9162 (0.02111)**
	Land Area	**-1.287 (0.02203)**	**-1.083 (0.04495)**	**-1.284 (0.02056)**	**-0.8865 (0.02032)**	**-1.179 (0.02715)**

	Yearly trend	NA	**0.02759** **(0.002135)**	NA	**-0.02315** **(0.001936)**	-0.003148 (0.004556)
Pedestrian	Intercept	**0.8672** **(0.2519)**	0.3048 (0.4678)	**1.801** **(0.147)**	**-203.6** **(10.21)**	-0.1208 (0.6806)
	Population	0.008662 (0.03157)	**0.122** **(0.01273)**	**0.0923** **(0.02112)**	**0.0387** **(0.01525)**	**0.04489** **(0.01151)**
	Maintain Miles	0.004058 (0.008569)	0.0045 (0.01035)	0.003207 (0.01223)	-0.0002712 (0.008451)	0.003542 (0.008898)
	Mean Travel Time	**-0.6804** **(0.09504)**	**-1.016** **(0.05308)**	**-0.9415** **(0.118)**	**-1.555** **(0.1048)**	**-0.4644** **(0.1235)**
	Median House Income	**-0.06005** **(0.02965)**	**-0.08702** **(0.02407)**	**-0.1553** **(0.01316)**	**0.04646** **(0.01498)**	**-0.1021** **(0.03049)**
	Land Area	**-0.4166** **(0.01479)**	**-0.3362** **(0.03252)**	**-0.4307** **(0.02115)**	**-0.4611** **(0.02627)**	**-0.3457** **(0.06915)**
	Yearly trend	NA	**-0.01135** **(0.002134)**	NA	**-0.02315** **(0.001936)**	**-0.0245** **(0.004889)**
Vehicle	Intercept	**39.03** **(0.08003)**	**39.35** **(0.06195)**	**40.08** **(0.2003)**	**-84.76** **(7.271)**	**39.86** **(0.09192)**
	Population	**-0.8921** **(0.00555)**	**-0.9098** **(0.005664)**	**-0.8547** **(0.003782)**	**-0.8458** **(0.006263)**	**-0.8649** **(0.01119)**
	Maintain Miles	-0.004897 (0.002662)	0.0006527 (0.003382)	-0.006827 (0.002789)	0.00002946 (0.003023)	0.0006173 (0.003312)
	Mean Travel Time	**-3.838** **(0.01695)**	**-3.827** **(0.03526)**	**-4.161** **(0.02898)**	**-4.105** **(0.0119)**	**-4.06** **(0.03589)**
	Median House Income	**-1.001** **(0.007701)**	**-1.114** **(0.01056)**	**-1.054** **(0.006385)**	**-1.022** **(0.005542)**	**-0.9793** **(0.007107)**
	Land Area	**-0.9818** **(0.01169)**	**-0.8436** **(0.0161)**	**-0.9957** **(0.006574)**	**-0.9722** **(0.005265)**	**-0.9735** **(0.01605)**
	Yearly trend	NA	**-0.02453** **(0.0006668)**	NA	**-0.02315** **(0.001936)**	**-0.02508** **(0.001537)**

Note: 1. Refer to Table1 for detailed description of variables
2. Numbers in parentheses represent uncertainty estimates, or, posterior standard deviations
3. The statistically significant correlation coefficients are shown in bold.

4.2.2 Correlation among Crash Types

Similar to previous literature (Park & Lord, 2007; Aguero-Valverde & Jovanis, 2009; Aguero-Valverde et al., 2016), the variance estimates of all four crash types for all multivariate models are statistically significant at the 0.05 level of significance which indicate the presence of over-dispersion in all modal crashes. In addition, a correlation analysis of error terms for all models was conducted. The result for multivariate space-time model with varying time trend (Model 5) is shown in Table 13 for the illustrative purpose. Table 13 depicts the variance-covariance and correlation of heterogeneity error term, the spatial random effect, and the space-time interacted term, among various crash modes.

In case of heterogenous residuals, the results demonstrate the presence of strong posterior correlation between some crash modes. For example, the bicycle crashes exhibit the highest correlation coefficient of 0.989 with vehicular crashes, followed by pedestrians and bikes (0.84) and pedestrians and vehicles (0.816). These findings suggest that some pairs of crash modes have closer behaviors and the heterogeneity error term helps account for the variability due to common factors which may not be incorporated as explanatory variables during model development. Interestingly, the correlation results also depict a negative relationship (though non-significant and relatively weak as shown by small coefficients) of motorcycle crashes with bicycle and vehicular crashes. This finding suggests that unaccounted factors may have a reversed impact on the crash risk of such modes. In other words, an increase in the quantity of a specific common factor may increase the crash risk of one mode and decrease it for another mode.

In case of spatial and interacted spatial, the variance is observed to be statistically significant. This demonstrates the expected spatial clustering of similar crash types among neighboring entities. It should be noted that the statistical significance of variance demonstrates the clustering of crashes for a particular mode only. For example, the increased vehicular crash risk at one site is positively correlated with increased vehicular crash risk at neighboring site. However, the non-significance of covariance results suggest that the crashes of different modes may not be correlated across neighboring sites. For example, the increased vehicular crash risk at a site may not be correlated with an increased pedestrian crash risk at a neighboring site. Overall, the correlation results demonstrate that existence of common factors among different crash modes, but the crashes of multiple modes may not be correlated across neighboring sites.

Table 13. The Estimate of the Correlation among Crash Types and associated p-values for Model 5

Modes	Heterogeneity (εij)				Independent Spatial (uij)				Interacted Spatial (δij)			
	MC	Bike	Ped	Veh	MC	Bike	Ped	Veh	MC	Bike	Ped	Veh
MC	**0.21**	-0.14	0.03	-0.17	**0.001**	9.20E-05	9.14E-05	8.79E-05	**0.001**	9.23E-05	9.33E-05	9.49E-05
Bike	-0.117	**6.8**	**1.2**	**6.23**	0.092	**0.001**	8.83E-05	9.04E-05	0.0923	**0.001**	9.34E-05	9.33E-05
Ped	0.119	**0.840**	**0.3**	**1.08**	0.0914	0.088	**0.001**	8.77E-05	0.0933	0.093	**0.001**	9.12E-05
Veh	-0.153	**0.989**	**0.816**	**5.83**	0.0879	0.0904	0.0877	**0.001**	0.0949	0.093	0.0912	**0.001**

Notes: 1. The shaded cells represent covariance matrix
2. Unshaded cells represent correlation coefficient
3. Bold text represents statistical significance at 0.05.

4.2.3 Goodness-of-fit

The goodness-of-fit measures reveal the notable variations in performance of models. As exhibited in Table 14, the spatiotemporal models (Model 4 and Model 5) were observed to exhibit substantially lower $\overline{\text{D}}$ value, indicating that the interaction of space and time while comprising the multivariate nature fits the data very well. However, such benefit is accompanied by the largest values of P_D, showing the relatively large effective number of parameters. On the contrary, the model without consideration for spatial correlations or temporal trend (Model 1) enjoyed the lowest value of P_D, which was roughly half of the sophisticated spatiotemporal models. However, this significant difference of model complexity failed to compensate for poor model fit as the overall fit (DIC) for Model 1 remained among the highest. Since the DIC differences are more than 7 points among all models, it can be concluded that MVPLNST, especially the one with varying time coefficient for different transportation mode users, significantly improves the model-fitting performance by borrowing strength from neighbors as well as considering the time trend. On the other hand, the MVPLN model, which doesn't consider either temporal or spatial effect, has the inferior modeling performance. In case of rest of the evaluation criteria, Model 5 consistently performed the best due to least discrepancy between estimated and observed crash counts, followed by Model 4. The results clearly demonstrated that the spatiotemporal models were remarkably superior in all respects while the traditional MVPLN performed the worst. This was expected as this model did not have the benefit of accommodating the heterogeneity which escaped from the covariates. Model 2 shows slightly better results which may be accredited to the inclusion temporal trend. However, the MVPLNS model failed to enhance crash prediction performance indicating that the spatial correlations may have been addressed by the explanatory variables and additions of such structures adversely impact model performance due to the increased complexity associated with a potentially higher number of effective parameters employed for spatial random effects during model development. Such negative impact of spatial random effects seems to have been compensated by the interaction with time trend for the spatiotemporal models which enabled them to have a superior edge at predictive accuracy and goodness-of-fit. Model 5 was noted to have the best scores across all evaluation criteria, closely followed by Model 4, while Model 1 and Model 3 exhibited the worst performance, which was more pronounced in case of MSPE with four times higher score. This repetitive trend across different evaluation criteria corroborates the previous observation that the consideration of only spatial correlations over a multivariate specification raises model complexity while the interaction of space and time substantially benefits the model. This benefit at model fit may be attributed to the inclusion of temporal trend since Model 2 (which only had temporal trend over multivariate) also exhibited significant improvement at model fit, compared to Model 3 with spatial, indicating that the crash data at the macro level of county is distributed preferably more across time. It is noteworthy that the overall trend between the models for predictive accuracy is similar to the trend for $\overline{\text{D}}$, which is the measure of in-sample error. It may be inferred from the results of both the measures that there is a correlation between the posterior deviance ($\overline{\text{D}}$) and the prediction capability of a model.

Table 14. Evaluation Results for Alternate Models

Models	$\overline{\text{D}}$	P_D	DIC	MAD	MSPE	G^2	RSS	TRD
Model 1	15804.9	**222.76**	16027.7	40.92	21290.61	6838.81	6640.32	37420

Model 2	14243.2	225.002	14468.3	32.4	7935.69	5242.59	5093.35	37352
Model 3	15878.7	256.34	16135.0	41.03	22092.09	6852.41	6671.17	37468
Model 4	12888.8	529.68	13418.5	25.97	5438.21	3929.03	3499.34	32133
Model 5	**12716.7**	527.42	**13244.2**	**25.31**	**5194.17**	**3342.29**	**3319.39**	**31708**

Note: Bold text represents best performance for the particular criterion.

4.3 Results for the TAZ-Level SPF

The crash prediction models for the 203 TAZs in the City of Irvine were estimated with the freeware statistical package WinBUGS (*Abdel-Aty et al., 2011*). A total of 10,000 MCMC iterations were utilized for parameter estimation after discarding first 1,000 iterations as burn-in. The convergence was ensured by employing different approaches such as visual inspection of history plots, trace plots, and Gelman-Rubin diagram (Gelman and Rubin, 1992). The Pearson correlation coefficient was calculated and the covariates correlated at a significance level of 0.05 were subsequently eliminated.

4.3.1 Modeling Results

As shown in Table 15, the posterior inferences for influential factors for all four models demonstrate their robustness to fit the multimodal crash data at the TAZ spatial scale. All four models identify similar significant factors that affect crash frequency for a particular mode. In the case of bicycle crashes, three variables are observed to be statistically significant, namely: K12 student enrollment, percentage of arterials, and bike-lane density for the TAZ. The TAZs with higher K12 student enrollment increases the crash risk as the instances of interaction of bicyclists with other modes increases due to more exposure. However, the similar positive correlation for bike-lane density seems counter-intuitive since the presence of bike lanes is expected to facilitate more usage of bicycles due to lower perceived risk of interaction with other modes. The possible rationale for this finding may be explained by the lower perceived risk which may encourage bicyclists to ride more in such areas, while conversely increasing the crash risk due to higher exposure of bicyclists to vehicular traffic. The negative relationship among percentage of arterial roads and bicycle crashes indicates that maybe the bicyclists tend to travel less in areas with more arterials. For the crashes pertaining to pedestrians, the college enrollment is also observed to be influential, along with other three factors shared with bicycle crashes. The increase in student population in the colleges of TAZs is noted to be negatively linked with pedestrian crashes, though the increased pedestrian activity usually associated with the presence of college students was expected to increase crash occurrence. The probable justification may be that the known presence of students influences the vehicle drivers to be more cautious and drive sensitively, or the vehicular activity may be minimal in such areas which may help significantly reduce the possibility of interaction with pedestrians. The common significant factors (K12 student enrollment, percentage of arterials, and bike-lane density) responsible for bicycle and pedestrian crashes support the joint estimation of such modes which are most vulnerable and impacted by similar characteristics. As shown in Table 16, the heterogeneity error term demonstrates the presence of statistically significant correlation among the bicycle and pedestrian crashes which further justifies the employment of bivariate structure for joint estimation of crashes. However, the spatial random effect term exhibits the absence of a significant correlation, as indicated by the covariance matrix. It may be possible that the explanatory variables incorporated for model development are sufficiently robust to account for the spatial characteristics that influence crash occurrence for the particular modes.

Table 15. Posterior Inference for Bicyclist and Pedestrian-involved Crash Counts

Count Type	Variables	Model 1	Model 2	Model 3	Model 4
Bicyclist	Intercept	**-10.860 (0.243)**	**-10.880 (0.246)**	**-10.780 (0.248)**	**-10.790 (0.234)**
	% age 65+	1.532 (0.922)	1.467 (0.895)	1.413 (0.830)	1.401 (0.798)
	K12	**0.203 (0.088)**	**0.203 (0.091)**	**0.213 (0.079)**	**0.211 (0.074)**
	College	-0.013 (0.078)	-0.015 (0.077)	-0.014 (0.079)	-0.012 (0.075)
	WalkAcc	-0.007 (0.010)	-0.008 (0.010)	-0.006 (0.010)	-0.007 (0.010)
	% Arterial	**-3.517 (0.674)**	**-3.529 (0.685)**	**-3.472 (0.691)**	**-3.399 (0.655)**
	BL_den	**0.260 (0.056)**	**0.271 (0.057)**	**0.245 (0.056)**	**0.246 (0.056)**
Pedestrian	Intercept	**-12.390 (0.326)**	**-12.430 (0.357)**	**-12.360 (0.340)**	**-12.380 (0.346)**
	% age 65+	1.205 (1.145)	1.192 (1.101)	1.097 (1.074)	1.131 (1.009)
	K12	**0.280 (0.104)**	**0.280 (0.106)**	**0.291 (0.095)**	**0.291 (0.094)**
	College	**-0.976 (0.567)**	**-0.968 (0.563)**	**-0.962 (0.562)**	**-0.957 (0.558)**
	WalkAcc	0.009 (0.010)	0.008 (0.010)	0.010 (0.010)	0.009 (0.010)
	% Arterial	**-3.826 (0.989)**	**-3.805 (0.985)**	**-3.727 (0.991)**	**-3.658 (0.996)**
	BL_den	**0.384 (0.068)**	**0.397 (0.075)**	**0.374 (0.069)**	**0.375 (0.074)**

Notes: 1. Intercept for Dirichlet Process models indicates the intercept mean from mixture points.
2. Refer to Table 4 for detailed description of variables.
3. Numbers in parentheses represent uncertainty estimates, or, posterior standard deviations.
4. The statistically significant variable coefficients are shown in bold.
5. Model 1: bivariate; Model 2: bivariate spatial; Model 3: bivariate dirichlet process mixture; Model 4: bivariate dirichlet process mixture spatial.

Table 16. Covariance Matrices for the Four Alternative Models

Models	Modes	Heterogeneity (ε_{ij})		Spatial (u_{ij})	
		Bicycle	Pedestrian	Bicycle	Pedestrian
Model 1	Bicycle	**0.896 (0.166)**	**0.854 (0.166)**		

	Pedestrian	**0.854 (0.166)**	**0.890 (0.237)**		
Model 2	Bicycle	**0.860 (0.168)**	**0.827 (0.153)**	**0.001 (2.2x10^{-4})**	6.7x10^{-5} (1.5x10$^{-4)}$
	Pedestrian	**0.827 (0.153)**	**0.856 (0.213)**	6.7x10^{-5} (1.5x10$^{-4)}$	**0.001 (2.2x10^{-4})**
Model 3	Bicycle	**0.602 (0.200)**	**0.538 (0.182)**		
	Pedestrian	**0.538 (0.182)**	**0.561 (0.226)**		
Model 4	Bicycle	**0.507 (0.231)**	**0.461 (0.234)**	**0.001 (2.1x10^{-4})**	7.4x10^{-5} (1.5x10$^{-4)}$
	Pedestrian	**0.461 (0.234)**	**0.503 (0.270)**	7.4x10^{-5} (1.5x10$^{-4)}$	**0.001 (2.2x10^{-4})**

Notes: 1. Numbers in parentheses represent posterior standard deviations.
2. The statistically significant covariance values are shown in bold.
3. Model 1: bivariate; Model 2: bivariate spatial; Model 3: bivariate dirichlet process mixture; Model 4: bivariate dirichlet process mixture spatial

4.3.2 Evaluation Results

As previously stated, the four models are evaluated from different perspectives using five evaluation criteria. Unlike the traditional parametric models which usually employ DIC (deviance information criterion) for model comparison, LPML is adopted in this study as DIC is not generated by the WinBUGS due to its sensitivity to different parameterizations (Ohlssen et al., 2007; Geedipally et al., 2014**)**. The higher value of LPML is desirable as it reflects relatively superior model fit property and a difference of more than 5 points among two competing models help identify the model of interest (Ntzoufras, 2012). As shown in Table 17, the LPML values of all four models are close enough to not cross the threshold of 5 points for identification of the model of interest. However, the sample size also impacts the numerical value of LPML. Hence it may be worthwhile to record the model with highest LPML value and compare the observation with other criteria. As evident from the evaluation results, Model 3 demonstrates the best fit based on relatively large LPML (-474.433), closely followed by Model 4. A similar trend is observed for all other criteria suggesting the strong correlation among the capability of a model to fit crash data and its performance at crash predictive accuracy.

Further inspection of the evaluation results reveals that the models which account for spatial correlations (Models 2 and 4) have consistently inferior performance to those with spatially unstructured heterogeneity (Models 1 and 3). Such phenomenon suggests that the inclusion of spatial correlation structures raises the model complexity without notable advantage at crash prediction, which is usually expected in such cases as reduced posterior deviance compensates the increased complexity. The potential reason might be due to the insignificant spatial dependency among the two modal crashes as shown in Table 16. Clearly, the Dirichlet models (Models 3 and 4) outperform the non-Dirichlet ones (Model 1 and 2) based on all five criteria suggesting the use of such flexible framework.

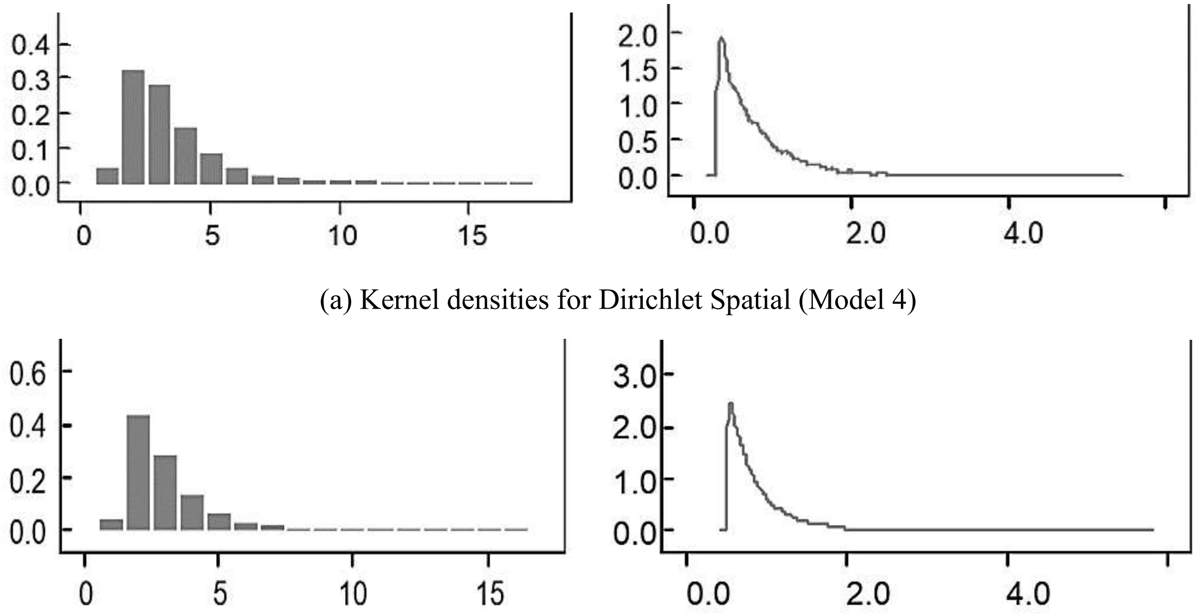
Apart from the above findings, the logical aspect of employing the flexible approach should also be given consideration. For a given crash dataset, the parametric approach assumes a restrictive stationary distribution of explanatory variables across all the sites under focus. As discussed in previous studies of semi-parametric models (Heydari et al., 2016; Shirazi et al., 2016), the Dirichlet formulation allows the examination of the adequacy of standard parametric assumption. As clearly shown in Figure 3, the kernel posterior density plots of Dirichlet precision parameter *k* illustrate the closeness of the peak towards zero which reflects that the unknown density (G) of non-parametric intercept is far from the baseline distribution (G_0). Similar plots for both Dirichlet models suggest their robustness and indicate that the normal assumption of intercept associated with traditional parametric models does not hold true for the TAZ level crash dataset of the current study. This indicates that the 203 intercepts associated with the TAZs are not normally distributed and the standard parametric approach does not hold true for the concerned pedestrian and bicycle crashes at the planning level. This finding seems plausible since the safety mechanisms which impact the pedestrian and bicycle crashes may vary across different TAZs due to diverse factors (such as driving behavior, road environment, and so forth) which may not be captured in the explanatory variables. These findings also suggest the presence of distinct subpopulations among the crash data which was confirmed from the histogram of posterior number of latent clusters with a median of 2 clusters for most of the data. This capability of Dirichlet models to identify the latent subpopulations may prove highly beneficial for the safety agencies to investigate similarities in the safety issues among different sites and allocate funding for dedicated countermeasures (Shirazi et al., 2016). This is achieved by calculating the expected probabilities of sites to fall into same clusters, which allows detection of the degree of similarity or dissimilarity among sites based on the crash risk (Heydari et al., 2016).

The aforementioned advantages justify the use of Dirichlet process mixture with a flexible intercept as such model specification helps more precise estimation leading to better inferences. Contrary to the parametric models which restrict the priors to a specific distribution fixed across all entities, the latent clusters capture the multimodality due to unconstrained nature.

Table 17. Evaluation Results for Alternative Models

Model	LPML	MSPE	R_p^2	G^2	RSS
Model 1	-476.753	0.690	0.786	177.995	272.367
Model 2	-477.492	0.691	0.781	179.544	278.749
Model 3	**-474.433**	**0.682**	**0.823**	**169.137**	**225.018**
Model 4	-474.831	0.687	**0.823**	169.998	225.291

Notes: Model 1: bivariate; Model 2: bivariate spatial; Model 3: bivariate dirichlet process mixture; Model 4: bivariate dirichlet process mixture spatial.



(a) Kernel densities for Dirichlet Spatial (Model 4)

(b) Kernel Densities for Dirichlet without Spatial (Model 3)

Figure 3. Kernel Density Plots for Precision Parameter and Latent Clusters

5. CONCLUSIONS AND RECOMMENDATIONS

Likewise, for ease of description, the conclusions and recommendations are outlined in the order of the micro-level, county-level, and TAZ-level SPFs.

5.1 Conclusions and Recommendations for the Micro-level SPF

For this type of SPF, compared with the vehicle modes, much less research has been dedicated to the development of SPF for active transportation modes such as pedestrians. There are multiple reasons behind such situation which include the dominant use of vehicle modes and the difficulty to obtain exposure information of pedestrians. For the intersection SPF exploration, bivariate models are used to account for the common unobserved heterogeneity shared by the two types of crashes at the same intersections. Then, both robust variable importance ranking technique and correlation analyses among numerical variables are employed to determine the mode-specific covariate inputs, enhancing both model flexibility and accuracy with more related variables being included for each of the modes. In addition, this project also employed negative binomial model to investigate SPFs without feature selections, and determined Crash Modification Factors (CMFs) for all variables to pedestrian and vehicle. The following conclusions were drawn based on the research results:

1. Compared with the base condition of four-legged intersection, offset intersection demonstrates better safety performance for both pedestrians and drivers.

2. In contrast of off ramp, on ramp is more advantageous in terms of traffic safety for both traffic modes.
3. For drivers only, the tee and wye intersections and those without right-turn channels tend to provide more safety benefits compared with the base conditions of intersections with four-leg and right turn channels, respectively. The better safety performance associated with the intersections without right turn channels is relatively contradictory, which warrants further investigations from other studies.
4. There are much more statistically significant variables associated with pedestrians on the intersection, suggesting that pedestrians are more sensitive to various intersection features than the vehicle drivers. In contrast, vehicles have more statistically significant variables on the ramp, indicating that vehicles are relatively unsafe on the ramp.
5. The correlation and covariance matrix between the random effects of both pedestrian and vehicle counts demonstrates existence of strong correlation for both of intersection and ramp analysis, indicating the sensibility of using the bivariate models which explicitly consider the correlation between the two modes.

The aforementioned findings from this study reflect an improvement to current SPF development with mode-specific inputs of predictors and count model-estimated pedestrian exposure being utilized. However, it is important to mention that the current findings are based on the empirical results obtained from the intersection- and ramp-related crash data in California. Some of the model findings may not hold true when employing data at a different spatial level. Moreover, only crashes of two modes are investigated. More modes involved might lead to different results given more complex interrelationships are introduced among all crash outcomes. Finally, this study considered timely aggregated crashes only. The consideration of serial correlation among various years of crashes is also worth of further investigation.

5.2 Conclusions and Recommendations for the County-level SPF

For the county-level SPF, the traffic safety field has employed separate temporal and spatial correlations for simultaneous estimation of crash outcomes. However, this is no or little research considering both dimensions of time and space, as well as the associated interactions, for the multivariate models. To this end, this study proposed two multivariate spatial-temporal models. The proposed models were developed using the Full Bayesian framework and incorporated the spatial-temporal random effects with fixed and mode-varying time coefficients for various modal crashes. This study was primarily focused on the comparison of the proposed models with the alternate multivariate models which either did not incorporate or incorporated only one correlation: spatial or temporal.

The models were compared based on the fitness of estimated and observed crash data using different evaluation criteria. The model fitness results from DIC revealed that the proposed models significantly improved the model fitting by pooling strength from the neighbors, consideration of time trend, as well as their interactions. Among the two proposed models, the model with mode-varying time coefficients was observed to be superior. The influential factors for all the models were mostly the same. The in-sample error ($\bar{D}$) was observed to be the governing factor for overall fit of the model and consistently showed a strong positive correlation between model fit and prediction accuracy. In other words, the models with closer DIC or $\bar{D}$ values tend to yield more similar prediction performance.

Although the study clearly demonstrated the advantages of the proposed models due to the capabilities of combination of spatial and temporal random effects, still there are some recommendations to further bolster the significance of these models. Firstly, this study was focused at the county level with a set of influential variables. Somewhat different results may be expected for other geographic areas, like block level, TAZ, or smaller entities like intersections. Secondly, the inverse of distance was calculated to generate the spatial weights for CAR specification. Future studies may adopt other approaches. Thirdly, the fitness of models was assessed by employing the DIC. Other techniques could be utilized for such assessment like MAPE (mean absolute percentage error), RMSE (root-mean-square error), among others. Moreover, cross-validation techniques would also help verify the expected advantages at crash prediction. The model comparison results may also exhibit deviations from this study when such multivariate space-time specifications are employed using the crash count approach, unlike the crash rate used in this study, for a smaller spatial scale. Finally, this study used a linear time-space interaction and the time-varying coefficients for development of models. The fitness and performance of other time-space relationships or random parameter models could be explored and compared with the proposed models.

5.3 Conclusions and Recommendations for the TAZ-level SPF

For the TAZ-level SPF, The current study contributes to the safety literature by proposing a bivariate Dirichlet process mixture spatial model and comparing its performance for crash predictions with other three competing models. The proposed semi-parametric model accounted for the unobserved heterogeneity by combining the strengths of incorporating bivariate specification to accommodate correlation among crash modes, spatial random effects for the impact of neighboring TAZs, and Dirichlet process mixture for random intercept. The present model structure allowed the flexibility to infer stochastic parameter from the crash data instead of a prespecified distribution.

All four models shared similar influential factors across both crash modes which indicated the robustness of the models. For crashes pertaining to bicycles, K12 student enrollment, percentage of arterials, and bike-lane density for the TAZ were observed to be statistically significant at the 95% confidence interval. The positive correlation of K12 student enrollment with crash risk suggests the increased risk due to higher chances of physical interactions of bicyclists/pedestrians with other modes due to more exposure. However, the perceived risk appears to be the governing factor in the case of positive correlation for bike-lane density, which seems counter-intuitive. The lower perceived risk may encourage bicyclists to ride more in such areas and therefore yield higher chances of the exposure of bicyclists to vehicular traffic. A negative correlation was observed for percentage of arterial roads and bicycle crashes which suggests a lesser tendency of bicyclists to travel in areas with more arterials, hence reducing the exposure to possible interactions. The pedestrian crashes were observed to reduce with an increase in student population in the colleges of TAZs. This may be justified by the policies implemented in these areas for reduced vehicular traffic which eventually reduces the possibility of interaction with pedestrians.

The heterogeneity error term demonstrated the presence of statistically significant correlation among the bicycle and pedestrian crashes while the spatial random effect term exhibited the absence of a significant correlation, which might explain why models considering the spatial random effects did not yield the expected advantages compared with their non-spatial counterparts. In the comparison between Dirichlet and non-Dirichlet models, the former ones were consistently superior to typical bivariate ones under all criteria. These findings demonstrate the advantages associated with consideration of flexible approach, Dirichlet process mixture in the current study, based on the goodness-of-fit and predictive accuracy of estimated crash counts. Moreover, the Dirichlet models exhibited the capability to identify the latent distinct subpopulations and suggested the that the normal assumption of intercept associated with traditional parametric models does not hold true for the TAZ level crash dataset of the current study. These findings justify the development of sophisticated flexible models which generate more precise estimate due to the unrestrictive approach which eventually leads to better inferences.

Based on the results, this study recommends careful consideration of spatial correlations at the macro-level of TAZs as they increased the complexity without any significant advantage at model fit or predictive accuracy. The authors also recommend exploring other spatial levels and observe if the results of the current study hold true or if the spatial random effects prove beneficial. Similar to other studies that focus on crashes pertaining to modes of active transportation, it should be noted that both the pedestrian and bicycle crashes have been modeled by utilizing the exposure of vehicles, rather than pedestrians and bikes, due to the unavailability of exposure data for the concerned modes. It is recommended that novel methods may be explored to account for the exposure data such as using bike mode share, or calibrating the exposure from socio-economic factors related to such modes (e.g. number of employees walking or cycling to work). Finally, the crash dataset utilized for model development was aggregated for a six-year period and future studies may incorporate temporal correlations and adopt disaggregated crash counts (Cheng et al., 2017).

6. REFERENCES

Abdel-Aty, M., and X. Wang. (2006). Crash Estimation at Signalized Intersections Along Corridors: Analyzing Spatial Effect and Identifying Significant Factors. Transportation Research Record: Journal of the Transportation Research Board, No. 1953, pp. 98-111.

Abdel-Aty, M., C. Siddiqui, H. Huang, and X. Wang. (2011). Integrating Trip and Roadway Characteristics to Manage Safety in Traffic Analysis Zones. Transportation Research Record: Journal of the Transportation Research Board, No. 2213, pp. 20-28.

Abdel-Aty, M., & Haleem, K. (2011). Analyzing angle crashes at unsignalized intersections using machine learning techniques. *Accident Analysis & Prevention*, *43*(1), 461-470.

Abdel-Aty, M., Lee, J., Siddiqui, C., & Choi, K. (2013). Geographical unit based analysis in the context of transportation safety planning. Transportation Research Part A: Policy and Practice, 49, 62-75.

Aguero-Valverde, J., & Jovanis, P. (2006). Spatial Analysis of Fatal and Injury Crashes in Pennsylvania. Accident Analysis and Prevention. Vol. 38, No. 3, pp. 618–625.

Aguero-Valverde, J., and P. Jovanis. (2009). Bayesian Multivariate Poisson Lognormal Models for Crash Severity Modeling and Site Ranking. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2136, pp. 82-91.

Aguero-Valverde, J. (2013). Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. *Accident Analysis & Prevention*, 50, 289-297.

Aguero-Valverde, J., Wu, K.F.K. & Donnell, E.T., (2016). A multivariate spatial crash frequency model for identifying sites with promise based on crash types. Accident Analysis & Prevention, 87, pp.8-16.

Akaike, H. (2011). Akaike's Information Criterion. In International Encyclopedia of Statistical Science (pp. 25-25). Springer Berlin Heidelberg

Anarkooli, A. J., Persaud, B., Hosseinpour, M., & Saleem, T. (2019). Comparison of Univariate and Two-Stage Approaches For Estimating Crash Frequency By Severity-Case Study For Horizontal Curves On Two-Lane Rural Roads. *Accident Analysis & Prevention*, 129, 382-389.

Andrey, J., & Yagar, S. (1993). A temporal analysis of rain-related crash risk. Accident Analysis & Prevention, 25(4), 465-472.

Bambach, M. R., Grzebieta, R. H., Olivier, J., & McIntosh, A. S. (2011). Fatality risk for motorcyclists in fixed object collisions. Journal of Transportation Safety & Security, 3(3), 222-235.

Barker, S. P., O'neill, B., Haddon, W., & Long, W. B. (1974). The Injury Severity Score: A Method for Describing Patients with Multiple Injuries and Evaluating Emergency Care. *The Journal of Trauma: Injury, Infection, and Critical Care,* 14(3), 187-196.

Barton, B. K., & Morrongiello, B. A. (2011). Examining the impact of traffic environment and executive functioning on children's pedestrian behaviors. *Developmental Psychology, 47*(1), 182–191.

Bayesian spatiotemporal crash frequency models with mixture components for space-time interactions. *Accident Analysis & Prevention*, 112, 84-93.

Beck, L., A. Dellinger, and M. O'Neil. (2007). Motor Vehicle Crash Injury Rates by Mode of Travel, United States: Using Exposure-Based Methods to Quantify Differences. American Journal of Epidemiology, Vol. 166, No. 2, pp. 212-218.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., & Songini, M. (1995). Bayesian analysis of space—time variation in disease risk. Statistics in medicine, 14(21-22), 2433-2443.

Berrigan, D., R. Troiano, T. McNeel, C. DiSogra, and R. Ballard-Barbash. (2006). Active Transportation Increases Adherence to Activity Recommendations. American Journal of Preventive Medicine, Vol. 31, No. 3, pp. 210-216.

Best N, Richardson S, & Thomson A. (2005). A comparison of Bayesian spatial models for disease mapping. Statistical methods in medical research, 14(1), 35-59.

Bijleveld, F. (2005). The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis & Prevention, Vol. 37, No. 4, pp. 591-600.

Cai, Q., J. Lee, N. Eluru, and M. Abdel-Aty. (2016). Macro-level pedestrian and bicycle crash analysis: Incorporating spatial spillover effects in dual state count models. Accident Analysis & Prevention, Vol. 93, pp. 14-22.

Cheng, W., and S. Washington. (2005). Experimental evaluation of hotspot identification methods. Accident Analysis & Prevention, Vol. 37, No. 5, pp. 870-881.

Cheng, W., & Washington, S. (2008). New criteria for evaluating methods of identifying hot spots. Transportation Research Record: Journal of the Transportation Research Board, (2083), 76-85.

Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X. & Zhou, J., (2017a). Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. Accident Analysis & Prevention, 99, pp.330-341

Cheng, W., Gill, G. S., Sakrani, T., Dasu, M., & Zhou, J. (2017b). Predicting motorcycle crash injury severity using weather data and alternative Bayesian multivariate crash frequency models. Accident Analysis & Prevention, 108, pp.172-180.

Cheng, W., Gill, G. S., Loera, L., Wang, X., & Wang, J. H. (2017c). Evaluation of the impact of traffic volume on site ranking. Journal of Transportation Safety & Security, 1-15.

Cheng, W., Gill, G. S., Choi, S., Zhou, J., Jia, X., & Xie, M. (2017d). Comparative Evaluation of Temporal Correlation Treatment in Crash Frequency Modelling. Transportmetrica A: Transport Science, (just-accepted), 1-40.

Cheng, W., Gill, G. S., Ensch, J. L., Kwong, J., & Jia, X. (2018a). Multimodal crash frequency modeling: multivariate space-time models with alternate spatiotemporal interactions. Accident Analysis & Prevention, 113, 159-170.

Cheng, W., G. S. Gill, Y. Zhang, and Z. Cao. (2018b). Bayesian spatiotemporal crash frequency models with mixture components for space-time interactions. Accident Analysis & Prevention, 112, pp.84-93.

Choi, J., Reich, B. J., Fuentes, M., & Davis, J. M. (2009). Multivariate spatial-temporal modeling and prediction of speciated fine particles. Journal of statistical theory and practice, 3(2), 407-418.

Congdon, P. (2001). Bayesian Statistical Modeling. John Wiley & Sons, West Sussex, United Kingdom.

Congdon, P. 2006. Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Comput. Statist. Data Anal.*, 50: 346–357

Conway, A., J. Cheng, D. Peters, and N. Lownes. (2013). Characteristics of Multimodal Conflicts in Urban On-Street Bicycle Lanes. Transportation Research Record: Journal of the Transportation Research Board, No. 2387, pp. 93-101.

Cunto, F. J., & Ferreira, S. (2017). An analysis of the injury severity of motorcycle crashes in Brazil using mixed ordered response models. Journal of Transportation Safety & Security, 9(sup1), 33-46.Cutler, A., Cutler, D., & Stevens, J. (2012). Random Forests. *Ensemble Machine Learning*, 157-175.

Davis, G.A., & Yang, S., (2001). Bayesian identification of high-risk intersections for older drivers via Gibbs sampling. Transportation Research Record: Journal of the Transportation Research Board,1746, 84–89

De Hartog, J., Boogaard, H., Nijland, H., & Hoek, G. (2010). Do the Health Benefits of Cycling Outweigh the Risks?. *Environmental Health Perspectives*, *118*(8), 1109-1116.

Dommes, A., Cavallo, V., Dubuisson, J., Tournier, I., & Vienne, F. (2014). Crossing a two-way street: comparison of young and old pedestrians. *Journal of Safety Research*, *50*, 27-34.

Dong, C., Xie, K., Zeng, J., & Li, X. (2018). Multivariate dynamic Tobit models with lagged observed dependent variables: an effectiveness analysis of highway safety laws. Accident Analysis & Prevention, 113, 292-302.

Dong, C., Shi, J., Huang, B., Chen, X., & Ma, Z. (2017). A study of factors affecting intersection crash frequencies using random-parameter multivariate zero-inflated models. International journal of injury control and safety promotion, 24(2), 208-221.

Dong, C., Clarke, D. B., Yan, X., Khattak, A., & Huang, B. (2014). Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accident Analysis & Prevention, 70, 320-329.

Dong, C., Clarke, D. B., Richards, S. H., & Huang, B. (2014). Differences in passenger car and large truck involved crash frequencies at urban signalized intersections: An exploratory analysis. Accident Analysis & Prevention, 62, 87-94.

Dong, C., Clarke, D. B., Nambisan, S. S., & Huang, B. (2016). Analyzing injury crashes using random-parameter bivariate regression models. Transportmetrica A: Transport Science, 12(9), 794-810.

Dong, N., Huang, H., Xu, P., Ding, Z., & Wang, D. (2014). Evaluating spatial-proximity structures in crash prediction models at the level of traffic analysis zones. Transportation Research Record: Journal of the Transportation Research Board, (2432), 46-52.

Dong, N., Huang, H., & Zheng, L. (2015). Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. Accident Analysis & Prevention, 82, 192-198.

Dong, N., Huang, H., Lee, J., Gao, M., & Abdel-Aty, M. (2016). Macroscopic hotspots identification: a Bayesian spatio-temporal interaction approach. Accident Analysis & Prevention, 92, 256-264.

Earnest, A., G. Morgan, K. Mengersen, L. Ryan, R. Summerhayes, and J. Beard. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. International Journal of Health Geographics, Vol. 6, No. 1, p. 54.

Eksler, V., & Lassarre, S. (2008). Evolution of road risk disparities at small-scale level: Example of Belgium. Journal of safety research, 39(4), 417-427.

El-Basyouny, K., Barua, S., & Islam, M. T. (2014). Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis & Prevention, 73, 91-99.

Erkanli, A., M. Sung, E. Jane Costello, and A. Angold. (2006). Bayesian semi-parametric ROC analysis. Statistics in Medicine, Vol. 25, No. 22, pp. 3905-3928.

Fatholahzade, N., Akbarizadeh, G., & Romoozi, M. (2018). Implementation of Random Forest Algorithm in Order to Use Big Data to Improve Real-Time. *Journal Of Advances In Computer Engineering And Technology*, *4*(2), 51-60.

Flask, T., Schneider IV, W.H., & Lord, D. (2014). A segment level analysis of multi-vehicle motorcycle crashes in Ohio using Bayesian multi-level mixed effects models. Safety science, 66, 47-53.

Frank, L., M. Greenwald, S. Winkelman, J. Chapman, and S. Kavage. (2010) Carbonless footprints: Promoting health and climate stabilization through active transportation. Preventive Medicine, Vol. 50, pp. S99-S105.

Furie, G., and M. Desai. (2012). Active Transportation and Cardiovascular Disease Risk Factors in U.S. Adults. American Journal of Preventive Medicine, Vol. 43, No. 6, pp. 621-628.

Gates, T., Savolainen, P., Stapleton, S., Kirsch, T., & Miraskar, S. (2016). Development of Safety Performance Functions and Other Decision Support Tools to Assess Pedestrian and Bicycle Safety. *Transportation Research Center For Livable Communities*.

Geedipally, S., D. Lord, and S. Dhavala. (2014). A caution about using deviance information criterion while modeling traffic crashes. Safety Science, Vol. 62, pp. 495-498.

Gelfand, A. E., D. K Dey, and H. Chang. (1992). Model determination using predictive distributions with implementation via sampling-based methods (No. TR-462). STANFORD UNIV CA DEPT OF STATISTICS.

Gelfand, A. E. (1996). Model determination using sampling-based methods. Markov chain Monte Carlo in practice, 145-161.

Geruschat, D. R., & Hassan, S. E. (2005). Driver Behavior in Yielding to Sighted and Blind Pedestrians at Roundabouts. *Journal of Visual Impairment & Blindness*, 286-302

Gelman, A., and D. Rubin. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statistical Science, Vol. 7, No. 4, pp. 457-472.

Ghosh, S., and J. Norris. (2005). Bayesian capture-recapture analysis and model selection allowing for heterogeneity and behavioral effects. Journal of Agricultural, Biological, and Environmental Statistics, Vol. 10, No. 1, pp. 35-49.

Giles-Corti, B., S. Foster, T. Shilton, and R. Falconer. (2010). The co-benefits for health of investing in active transportation. New South Wales Public Health Bulletin, Vol. 21, No. 6, p. 122.

Gill, G. S., Cheng, W., Xie, M., Vo, T., Jia, X., & Zhou, J. (2017a). Evaluating Influence of Neighboring Structures on Spatial Crash Frequency Modeling and Site-Ranking Performance. Transportation Research Record: Journal of the Transportation Research Board, (2659), 117-126.

Gill, G. S., Cheng, W., Zhou, J., & Park, V. S. (2017b). Comparative analysis of cost-weighted site ranking using alternate distance-based neighboring structures for spatial crash frequency modeling. Journal of Transportation Safety & Security, 1-23.Governor Highway Safety Association - GHSA, 2019. Bicyclist and Pedestrian Safety.

Griswold, J., Medury, A., Schneider, R., Amos, D., Li, A., & Grembek, O. (2019). A Pedestrian Exposure Model for the California State Highway System. *Transportation Research Record: Journal of The Transportation Research Board*, *2673*(4), 941-950.

Guo, F., X. Wang, and M. Abdel-Aty. (2010). Modeling signalized intersection safety with corridor-level spatial correlations. Accident Analysis & Prevention, Vol. 42, No. 1, pp. 84-92.

Harb, R., Yan, X., Radwan, E., & Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests. *Accident Analysis & Prevention*, *41*(1), 98-107.

Harwood, D. W., Bauer, K. M., Richard, K. R., Gilmore, D. K., Graham, J. L., Potts, I. B., ... & Hauer, E. (2008). *Pedestrian safety prediction methodology* (No. NCHRP Project 17-26).

Hauer, E., (2001). Overdispersion in modeling accidents on road sections and in empirical bayes estimation. *Accident Analysis and Prevention*. 33 (6), 799–808.

Hadayeghi, A., A. Shalaby, and B. Persaud. (2007). Safety Prediction Models: Proactive Tool for Safety Evaluation in Urban Transportation Planning Applications. Transportation Research Record: Journal of the Transportation Research Board, No. 2019, pp. 225-236.

Hay, J. L., & Pettitt, A. N. (2001). Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. Biostatistics, 2(4), 433-444.

Held, L., Schrödle, B., & Rue, H. (2010). Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA. In *Statistical modelling and regression structures* (pp. 91-110).

Heydari, S., L. Fu, D. Lord, and B. Mallick. (2016). Multilevel Dirichlet process mixture analysis of railway grade crossing crash data. Analytic Methods in Accident Research, Vol. 9, pp. 27-43.

Heydari, S., Fu, L., Miranda-Moreno, L. F., & Jopseph, L. (2017). Using a flexible multivariate latent class approach to model correlated outcomes: A joint analysis of pedestrian and cyclist injuries. *Analytic methods in accident research*, 13, 16-27.

Higle, J. L. & M. B. Hecht. (1989). A Comparison of Techniques for the Identification of Hazardous Locations. Transportation Research Record 1238, TRB, National Research Council, Washington, DC, pp. 10–19.

Huang, H., Chin, H., & Haque, M. (2009). Empirical evaluation of alternative approaches in identifying crash hot spots: naive ranking, empirical Bayes, and full Bayes methods. Transportation Research Record: Journal of the Transportation Research Board, (2103), 32-41.

Hurvich, C. M., & Tsai, C. L. (1998). A crossvalidatory AIC for hard wavelet thresholding in spatially adaptive function estimation. *Biometrika*, *85*(3), 701-710.

Insall, P. Active travel: Transport policy and practice for health. Nutrition Bulletin, Vol. 38, No. 1, 2013, pp. 61-69.

Ishwaran, H., and L. James. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. Journal of the American Statistical Association, Vol. 96, No. 453, pp. 161-173.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* (Vol. 112, pp. 3-7). New York: springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jiang, X., Abdel-Aty, M., Hu, J., & Lee, J. (2016). Investigating macro-level hotzone identification and variable importance using big data: A random forest models approach. *Neurocomputing*, 181, 53-63.

Jonathan, A., K. Wu, and E. Donnell. (2016). A multivariate spatial crash frequency model for identifying sites with promise based on crash types. Accident Analysis & Prevention, Vol. 87, pp. 8-16.

Konno, H., & Koshizuka, T. (2005). Mean-absolute deviation model. Iie Transactions, 37(10), 893-900.

Ladrón de Guevara, F., S. Washington, and J. Oh. (2004). Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona. Transportation Research Record: Journal of the Transportation Research Board, No. 1897, pp. 191-199.

Lan, B., & Persaud, B. (2012). Evaluation of multivariate Poisson log normal Bayesian methods for before-after road safety evaluations. Journal of Transportation Safety & Security, 4(3), 193-210.

LaScala, E. A., Gerber, D., & Gruenewald, P. J. (2000). Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. *Accident Analysis & Prevention*, 32(5), 651 - 658.

Lawson, A. B., Browne, W. J., & Rodeiro, C. L. V. (2003). Disease mapping with WinBUGS and MLwiN (Vol. 11). John Wiley & Sons.

Lee, C., and M. Abdel-Aty.(2005). Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accident Analysis & Prevention, Vol. 37, No. 4, pp. 775-786.

Lee, J., Abdel-Aty, M., & Jiang, X. (2015). Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accident Analysis & Prevention, 78, 146-154.

Li, D., Liang, J., Di, Y., Gong, H., & Guo, X. (2016). The spatial-temporal variations of water quality in controlling points of the main rivers flowing into the Miyun Reservoir from 1991 to 2011. Environmental monitoring and assessment, 188(1), 1-12.

Li, Z., Chen, X., Ci, Y., Chen, C., & Zhang, G. (2019). A hierarchical Bayesian spatiotemporal random parameters approach for alcohol/drug impaired-driving crash frequency analysis. *Analytic Methods in Accident Research*, 21, 44-61.

Liu, C., & Sharma, A. (2017). Exploring spatio-temporal effects in traffic crash trend analysis. *Analytic Methods in Accident Research*, *16*, 104-116.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A Practical Introduction to Bayesian Analysis*. CRC press.

Maher, M.J., & Mountain, L.J., (1988). The Identification of Accident Blackspots: A Comparison of Current Methods. Accident Analysis and Prevention, Vol. 20, No. 2, pp. 143–151.

Mannering, F., & Bhat, C. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, *1*, 1-22.

Mannering, F. (2018). Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research, 17, 1-13.

Mardia, K. V. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. Journal of Multivariate Analysis, 24(2), 265-284.

Mansfield, T. J., Peck, D., Morgan, D., McCann, B., & Teicher, P. (2018). The effects of roadway and built environment characteristics on pedestrian fatality risk: A national assessment at the neighborhood scale. *Accident Analysis & Prevention*, 121, 166-176.

Mardia, K. (1998). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. Journal of Multivariate Analysis, Vol. 24, No. 2, pp. 265-284.

McArthur, G., Chapman, P., Robert, C., Larkin, J., Haanen, J., & Dummer, R. et al. (2014). Safety and efficacy of vemurafenib in BRAFV600E and BRAFV600K mutation-positive melanoma

(BRIM-3): extended follow-up of a phase 3, randomised, open-label study. *The Lancet Oncology*, *15*(3), 323-332.

Meyer, R. (2014). Deviance information criterion (DIC). *Wiley StatsRef: Statistics Reference Online*, 1-6.

Miaou, S.P., Song, J.J., & Mallick, B.K., (2003). Roadway traffic crash mapping: a space-time modeling approach. Journal of Transportation and Statistics, 6, pp.33-58.

Miranda-Moreno, L.F., (2006). Statistical Models and Methods for Identifying Hazardous Locations for Safety Improvements. University of Waterloo.

Miranda-Moreno, L. F. (2006). *Statistical Models and Methods for The Identification of Hazardous Locations for Safety Improvements* (Doctoral dissertation, Ph. D. Thesis, Department of Civil Engineering, University of Waterloo).

Miranda-Moreno, L. F., Morency, P., & El-Geneidy, A. M. (2011). The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. *Accident Analysis & Prevention*, 43(5), 1624 - 1634.

Moudon, A., L. Lin, J. Jiao, P. Hurvitz, and P. Reeves. (2011). The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. Accident Analysis & Prevention, Vol. 43, No. 1, pp. 11-24.

Narayanamoorthy, S., R. Paleti, and C. Bhat. (2013). On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transportation Research Part B: Methodological, Vol. 55, pp. 245-264.

Nashad, T., S. Yasmin, N. Eluru, J. Lee, and M. Abdel-Aty. (2016). Joint Modeling of Pedestrian and Bicycle Crashes. Transportation Research Record: Journal of the Transportation Research Board, No. 2601, pp. 119-127.

National Highway Traffic Safety Administration. Traffic Safety Facts: Bicyclists and Other Cyclists. United States Department of Transportation, 2012, (DOT HS 811 624).

National Highway Traffic Administration – NHSTA, 2018. Pedestrian Safety

Nicodemus, K. K. (2011). Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics*, *12*(4), 369-373.

Ntzoufras, I. (2012). Bayesian modeling in WinBugs. Wiley Series in Computational Statistics, Hoboken, USA.

Ohlssen, D., L. Sharples, and D. Spiegelhalter. (2007). Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. Statistics in Medicine, Vol. 26, No. 9, pp. 2088-2112.

Park, E., & Lord, D. (2007). Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. *Transportation Research Record: Journal of The Transportation Research Board,* 2019(1), 1-6.

Prato, C. G., Kaplan, S., Rasmussen, T. K., & Hels, T. (2016). Infrastructure and spatial effects on the frequency of cyclist-motorist collisions in the Copenhagen Region. Journal of Transportation Safety & Security, 8(4), 346-360.

Pawlovich, M.D., W. Li, A. Carriquiry & T. Welch. (2006). Iowa's Experience with "Road Diet" Measures: Impacts on Crash Frequencies and Crash Rates Assessed Following a Bayesian Approach. Transportation Research Record: Journal of the Transportation Research Board 1953, TRB, National Research Council, Washington, DC, pp.163–171.

Pulugurtha, S. S., Duddu, V. R., & Kotagiri, Y. (2013). Traffic analysis zone level crash estimation models based on land use characteristics. Accident Analysis & Prevention, 50, 678-687.

Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological),* 52(1), 175-184.

Plurad, D., Demetriades, D., Gruzinski, G., Preston, C., Chan, L., Gaspard, D., . . . Cryer, H. G. (2006). Pedestrian Injuries: The Association of Alcohol Consumption with the Type and Severity of Injuries and Outcomes. *Journal of the American College of Surgeons*, 202(6), 919 - 927.

Physica-Verlag HD. Russo, B. J., Savolainen, P. T., Schneider, W. H., & Anastasopoulos, P. C. (2014). Comparison of Factors Affecting Injury Severity in Angle Collisions by Fault Status Using A Random Parameters Bivariate Ordered Probit Model. Analytic Methods in Accident Research, 2, 21 - 29.

Rasciute, S., & Downward, P. (2010). Health or Happiness? What Is the Impact of Physical Activity on the Individual?. *Kyklos*, *63*(2), 256-270.

Retting, R. A., Ferguson, S. A., & McCartt, A. T. (2003). A review of evidence-based traffic engineering measures designed to reduce pedestrian–motor vehicle crashes. *American journal of public health*, 93(9), 1456-1463.

Schroeder, B. J., & Rouphail, N. M. (2011). Event-Based Modeling of Driver Yielding Behavior at Unsignalized Crosswalks. *Journal of Transportation Engineering*, 137, 455-465.

Schwarz, Gideon. (1978) Estimating the dimension of a model. The annals of statistics 6.2: 461-464.

Serhiyenko, V., Mamun, S., Ivan, J., & Ravishanker, N. (2016). Fast Bayesian inference for modeling multivariate crash counts. *Analytic Methods In Accident Research*, *9*, 44-53.

Shi, J., Chen, Y., Ren, F., & Rong, J. (2007). Research on Pedestrian Behavior and Traffic Characteristics at Unsignalized Midblock Crosswalk: Case Study in Beijing. *Transportation Research Record*, 2038(1), 23–33.

Shirazi, M., D. Lord, S. Dhavala, and S. Geedipally. (2016). A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis & Prevention, Vol. 91, pp. 10-18.

Siddiqui, C., Abdel-Aty, M., & Huang, H. (2012a). Aggregate nonparametric safety analysis of traffic zones. Accident Analysis & Prevention, 45, 317-325.

Siddiqui, C., Abdel-Aty, M., & Choi, K. (2012b). Macroscopic spatial analysis of pedestrian and bicycle crashes. Accident Analysis & Prevention, 45, 382-391.

Song, J.J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of multivariate analysis, 97(1), 246-273.

Spiegelhalter, D., A. Thomas, N. Best, D. Lunn. (2003). WinBUGS Version 1.4 User Manual. MRC Biostatistics Unit, Cambridge, http://www.mrc-cam.ac.uk/bugs

Strauss, J., L. Miranda-Moreno, and P. Morency. (2014). Multimodal injury risk analysis of road users at signalized and non-signalized intersections. Accident Analysis & Prevention, Vol. 71, pp. 201-209.

Tabibi, Z., Pfeffer, K., & Sharif, J. T. (2012). The influence of demographic factors, processing speed and short-term memory on Iranian children's pedestrian skills. *Accident Analysis & Prevention*, 47, 87 - 93.

Taylor, B., & Diggle, P. (2013). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal Of Statistical Computation And Simulation*, *84*(10), 2266-2284.

Thomas, B., & DeRobertis, M. (2013). The safety of urban cycle tracks: A review of the literature. *Accident Analysis & Prevention*, *52*, 219-227.

Tulu, G. S., Washington, S., Haque, M. M., & King, M. J. (2017). Injury severity of pedestrians involved in road traffic crashes in Addis Ababa, Ethiopia. Journal of Transportation Safety & Security, 9(sup1), 47-66.

Ukkusuri, S., Hasan, S., & Aziz, H. M. (2011). Random Parameter Model Used to Explain Effects of Built-Environment Characteristics on Pedestrian Crash Frequency. *Transportation Research Record*, 2237 (1), 98-106.

Vapnik, V.N. & Vapnik, V., 1998. Statistical learning theory (Vol. 1). New York: Wiley.

Vivoda, J., D. Eby, R. St. Louis, and L. Kostyniuk. (2008). Cellular Phone Use While Driving at Night. Traffic Injury Prevention, Vol. 9, No. 1, pp. 37-41.

Wanner, M., T. Götschi, E. Martin-Diener, S. Kahlmeier, and B. Martin. Active Transport, Physical Activity, and Body Weight in Adults. American Journal of Preventive Medicine, Vol. 42, No. 5, 2012, pp. 493-502.

Wardlaw, M. (2002). Assessing the actual risks faced by cyclists. Traffic engineering and control, 43(11), 352–356.

Wang, X., Abdel-Aty, M., & Brady, P. (2006). Crash estimation at signalized intersections: significant factors and temporal effect. Transportation Research Record: Journal of the Transportation Research Board, (1953), 10-20.

Wang, X. & Abdel-Aty, M. (2006). Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident Analysis & Prevention 38.6: 1137-1150.

Wang, C., Quddus, M., & Ison, S. (2013). A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. Transportmetrica A: Transport Science, 9(2), 124-148.

Washington, S., M. G. Karlaftis, and F. L. Mannering. (2003). Statistical and Econometric Methods for Transportation Data Analysis, Chapman & Hall.

Washington, S., & Oh, J., (2006). Bayesian Methodology Incorporating Expert Judgment for Ranking Countermeasures Effectiveness under Uncertainty: Example Applied to at Grade Railroad Crossings in Korea. Accident Analysis & Prevention 38, 234–247.

Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., & Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications. *Accident Analysis and Prevention*, 41(1), 137-145.

Wheater, H. S., Butler, A. P., Stewart, E. J., & Hamilton, G. S. (1991). A multivariate spatial-temporal model of rainfall in southwest Saudi Arabia. I. Spatial rainfall characteristics and model formulation. Journal of Hydrology,125(3), 175-199.

Williams, A.F. (2013) Protecting Pedestrians and Bicyclists: Some Observations and Research Opportunities. Insurance Institute for Highway Safety, http://www.iihs. org/frontend/iihs/documents/masterfiledocs.ashx

Wu, J., Xu, H., Zheng, Y., & Tian, Z. (2018). A novel method of vehicle-pedestrian near-crash identification with roadside LiDAR data. *Accident Analysis & Prevention*, *121*, 238-249.

Wu, Y., Boyle, L. N., & McGehee, D. V. (2018). Evaluating variability in foot to pedal movements using functional principal components analysis. *Accident Analysis & Prevention*, 118, 146-153.

Xu, P., and H. Huang. (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. Accident Analysis & Prevention, Vol. 75, pp. 16-25.

Xu, J., Ge, Y., Qu, W., Sun, X., & Zhang, K. (2018). The mediating effect of traffic safety climate between pedestrian inconvenience and pedestrian behavior. *Accident Analysis & Prevention*, *119*, 155-161.

Yu, R., & Abdel-Aty, M. (2014). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety science*, *63*, 50-56.

Zhao, M., Liu, C., Li, W., & Sharma, A. (2017). Multivariate Poisson-Lognormal Model for Analysis of Crashes on Urban Signalized Intersections Approach. Journal of Transportation Safety & Security, (just-accepted). https://doi.org/10.1080/19439962.2017.1323059

Zheng, L., & Sayed, T. (2019). A full Bayes approach for traffic conflict-based before–after safety evaluation using extreme value theory. *Accident Analysis & Prevention*, *131*, 308-315.