Use of Multivariate Dirichlet Process Mixture Spatial Model to Estimate Active Transportation-related Crash Counts

Wen Cheng, Ph.D. P.E. (corresponding author)
Associate Professor, Department of Civil Engineering
California State Polytechnic University, Pomona
3801 W. Temple Ave., Pomona, CA 91768
Fax: (909) 869-4342; Tel: (909) 869-2957; Email:wcheng@cpp.edu

Gurdiljot Singh Gill, M.S.C.E.
Research Associate, Department of Civil Engineering
California State Polytechnic University, Pomona
3801 W. Temple Ave., Pomona, CA 91768
Fax: (909) 869-4342; Tel: (909) 869-4312; Email: gurdiljotg@cpp.edu

Tom Vo, M.S.C.E.
Regional Planner/GIS Analyst, Southern California Association of Governments
818 West 7th Street, 12th Floor, Los Angeles, CA 90017
Fax: (213) 236-1825; Telephone: (213) 236-1800; Email: Vo@scag.ca.gov

Jiao Zhou, M.S.C.E
Research Associate, Department of Civil Engineering
California State Polytechnic University, Pomona
3801 W. Temple Ave., Pomona, CA 91768
Fax: (909) 869-4342; Tel: (909) 869-4312; Email:jiaozhou@cpp.edu

Taha Sakrani
Graduate Research Assistant,
Department of Civil Engineering, California State Polytechnic University, Pomona
3801 W. Temple Ave., Pomona, CA 91768
Tel: (909) 869-4312; Email:tsakrani@cpp.edu

Submitted for presentation at TRB 2018 Annual Meeting and Potential Publication in TRR

Word count: 5,822words + 4 tables+2 figure =7,322 equivalent words

ABSTRACT

The current study contributes to the safety literature by presenting a dedicated research for comprehensive analysis of multivariate Dirichlet process mixture spatial model for estimation of pedestrian and bicycle crash counts. This study focuses on the active transportation at Traffic Analysis Zone (TAZ) level by developing a semi-parametric model that accounted for the unobserved heterogeneity by combining the strengths of incorporating multivariate specification to accommodate correlation among crash modes, spatial random effects for the impact of neighboring TAZs, and Dirichlet process mixture for random intercept. Three alternate models, one Dirichlet while two parametric, were also developed for comparison based on different criteria.

Bicycle and pedestrian crashes shared three influential variables: the positive correlation of K12 student enrollment, the bike-lane density, and the percentage of arterial roads. 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 the slightly inferior performances associated with the spatial models. The Dirichlet models were consistently superior to non-Dirichlet ones under all evaluation criteria. Moreover, the Dirichlet models exhibited the capability to identify the latent distinct subpopulations and suggested 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.

Keywords: Dirichlet process, Multivariate, Spatial Correlation, Cross Validation.

INTRODUCTION

Non-motorists are defined as road users not in or upon a motor vehicle and generally consist of walking pedestrians, bicyclists, individuals in wheel chairs or motorized personal conveyances, skateboarders and others (1). They are a vulnerable segment of the traveling public due to the lack of a protective structure and difference in body mass between them and motor vehicles, which renders them prone to heightened injury susceptibility in case of a collision (2). On the other hand, active transportation provides enormous benefits for addressing the issues of congestion, health, and environment (3-8). Therefore, 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. In response, fairly extensive research (9-14) has been dedicated to the investigations into factors impacting non-motorist safety on roadways. While these studies are useful for identifying safety risks contributing to cyclist, pedestrian, and motor-vehicle injury occurrence, these modal crashes have been modeled separately, and few attempts have been made to combine them into a multimodal approach, which allows the flexibility to simultaneously determine the injury risk of different travel modes. Plus, the multimodal approach may also ease the task of selecting sites for safety improvement interventions as well as potentially provide a more economically viable solution and interventions for pedestrians and cyclists.

A central issue to the successful implementation of the multimodal approach is the development of multivariate crash frequency models which can jointly estimate the crash risk of different modes which share some of the unobserved heterogeneity. Ignorance of such correlation among the multiple modes has been illustrated to reduce the efficiency of the models due to lesser precise parameters (15-17). In comparison with the large number of univariate models dedicated to various mode users, very few studies have used the joint models to analyze the interaction between different modes of active transportation. Recently, Convay et al. (18) performed a bivariate correlation analysis to find the locations of conflict occurrence among bicycles and pedestrians, freight, passenger cars, and cabs in an urban area. The conflict was defined as the obstructions parked in or across the bicycle lane. The characteristics which influenced the conflicts for between these modes were also explored. In order to simultaneously analyze the injury and traffic flow outcomes for different modes, Strauss et al. (19) subsequently employed Bayesian multivariate Poisson models for studying safety outcomes for motor-vehicle, cyclist and pedestrian flow at intersections. Safety performance functions were developed and crash contributing factors were identified for each mode.

One common limitation associated with above two studies lies in the lack of consideration for spatial correlations within crash data. The significance of incorporating spatial correlations was highlighted by many studies (20-22) with the consistently superior performance of the spatial models over those accounting for heterogeneity random effect only. The study by Narayanamoorthy et al. (23) jointly analyzed the pedestrian and cyclist injury-severities while accounting for the spatial correlation at the census tract level using generalized ordered-response models. This study recommended the use of multivariate modeling and spatial dependency of injury counts. Similarly, Nashad et al. (24) employed a copula-based approach for simultaneous

estimation of crash counts for bicyclists and pedestrian crashes aggregated at the macro-level of traffic analysis zones (TAZs). The incorporation of spatial term facilitated the identification of hotspots at the zonal level which may prove beneficial for policy analysis.

Similar to the incorporation of spatial correlation structures, some studies in traffic safety addressed the unobserved heterogeneity by employing nonparametric and/or semiparametric models and observed their superiority at various fronts such as robustness and goodness-of-fit (25-26). In terms of research dedicated to active transportation, the recent study by Heydari et al. (27) proposed the Dirichlet process mixture (28) to develop flexible latent class model for joint analysis of pedestrian and cyclist injuries at the micro-level of intersections. It was 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. The results also demonstrated the need for consideration of such flexible structure as the assumption of homogeneity (in case of parametric models) among roadway entities was observed to be false.

The literature review illustrated the limited use of semi- or non-parametric models for simultaneous analysis of active transportation mode crashes. In effect, to the knowledge of the authors, the research for comprehensive analysis of flexible multivariate spatial models focusing on active transportation is non-existent in the safety literature. To fill this research gap, the authors adopted semi-parametric formulation that accounts for the unobserved heterogeneity by combining the strengths of incorporating multivariate specification of dependency among 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. LPML (log pseudo marginal likelihood) was calculated for cross-validation utilizing leave-one-out technique which makes this criterion less prone to selection bias associated with other cross-validation measures. Five other evaluation criteria were employed, namely: mean absolute deviations (MAD), mean-squared predictive error (MSPE), the R_p^2 statistic, the G^2 statistic, and residual sum of squares (RSS), which compare the alternate models on the basis of their performance to accurately predict the crash counts for both modal crashes. The benefits of the flexible model structure were also explored in terms of identification of latent clusters and accommodation of random distribution for parameters.

METHODOLOGY

Model Specification

The Full Bayesian (FB) framework was employed for estimation of six-year bicyclist and pedestrian crashes aggregated at the Traffic Analysis Zone (TAZ) level. The FB approach was chosen due to its capability to account for the unobserved heterogeneity from different perspectives, such as incorporation of complex potential correlation structures that exist within the hierarchical structure of crash data. The FB approach is deemed to be more precise for estimation of crashes with regard to its capability to generate a posterior distribution of parameters from Markov-chain Monte Carlo (MCMC) simulation where the variable samples are

random, rather than the point estimates generated by other traditional modeling approaches based on maximum likelihood estimation. This approach has been widely used for crash prediction models due to the multilevel and correlated nature of data (29). 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 specifications.

Model 1: Multivariate

This model assumes that crash count of certain modal crash $\boldsymbol{j}$ at a given location $\boldsymbol{i}$, $\mathbf{y_{ij}}$, obeys Poisson distribution, while the corresponding observation specific error term $\boldsymbol{\varepsilon_{ij}}$ follows a multivariate normal distribution:

$$y_{ij}|\lambda_{ij} \sim Poisson\ (\lambda_{ij}) \tag{1}$$

$$\ln(\lambda_{ij}) = X'_{ij}\beta + \varepsilon_{ij} \tag{2}$$

$$\varepsilon_{ij} \sim MVN\ (0,\ \sum) \tag{3}$$

Where $\boldsymbol{y_{ij}} = \begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix}$, $\boldsymbol{\lambda_{ij}} = \begin{pmatrix} \lambda_{i1} \\ \lambda_{i2} \end{pmatrix}$, $\boldsymbol{\varepsilon_{ij}} = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix}$, $\sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$ (4)

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

$$\sum^{-1} \sim Wishart(I, J) \tag{5}$$

Where I is the **J x J** identity matrix (30), and **J** is the degree of freedom, J=2 herein representing two crash outcomes corresponding to bicyclist and pedestrians crashes.

Model 2: Multivariate Spatial

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

$$\ln(\lambda_{ij}) = X'_{ij}\beta + \varepsilon_{ij} + u_{ij} \tag{6}$$

Where u_{ij} is the spatially structured random effect which follows the MCAR (multivariate conditional autoregressive) (31) formulation to incorporate the spatial correlation among crashes occurring at neighboring TAZs.

$$u_i|u_k,\ \sum i\ \sim N_j(\textstyle\sum_{k\sim i} C_{ik}\ , u_k, \sum i\ \ ) \tag{7}$$

Where each $\sum_{\boldsymbol{i}}$ is a positive definite matrix representing the conditional variance matrix, and the adjacency matrix $\mathbf{C_{ij}}$ is of the same dimension with $\sum_{\boldsymbol{i}}$ (32). 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 N_i where 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 (33-34), 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:

$$w_{ij} = \frac{1}{d_{ij}} \tag{8}$$

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 closer.

Model 3: Multivariate 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. But the nonparametric specification removes such constrains by employing a flexible approach of 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 (26):

$$\ln(\lambda_{ij}) = \beta_{0rj} + X_i'\beta \tag{9}$$

$$\beta_{0rj} \approx \sum_{n=1}^{C} p_n I_{\theta_{z_i}} \sim TDP\left(kG_{0j}\right),\ \ z_i = n\ with\ probability\ of\ p_n \tag{10}$$

$$G_0 \sim MVN\ (\mu_{G_{0j}},\ \sum) \tag{11}$$

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 multivariate 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 (28, 35). 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 $5k+2$ (28). 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 (36), 10 (37), and 52 (27).

Model 4: Multivariate 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:

$$\ln(\lambda_{ij}) = \beta_{0rj} + X_i'\beta + u_{ij} \quad (12)$$

Where all terms are defined as previously.

Comparison of Models based on Cross Validation

Many traditional approaches of cross-validation are prone to overestimation due to double usage of data, once during model development and then again for model checking. The approach of cross-validatory predictive densities was proposed to tackle this issue (38) where the full set of data was divided into two subsets (one subset for development and the other for checking). However, the splitting of two subsets posed a major problem as the selection of different subsets provides varying results. This was resolved by implementing a CV-1 (leave-one-out) technique to estimate the cross-validatory conditional predictive ordinate (CPO) (39) which removed the selection bias by employing a continuous approach of selecting all data points, except one, for model development and the left out data point to verify the prediction accuracy of the calibrated model. Under the MCMC framework, the estimate of CPO for each observation *i* can be calculated as:

$$CPO = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{f(Y_i|\beta^{(t)})}\right)^{-1} \quad (13)$$

Where Y_i is the *i*th observation (i = 1, 2, 3, . . ., n) for all 203 TAZs and β is the vector of estimated model parameters. This harmonic mean of density (CPO) may be extended to calculate the goodness-of-fit of models by computing the product of CPOs over all observations, which is known as the pseudo marginal likelihood. For computational convenience, the log pseudo marginal likelihoods (LPML) is calculated:

$$LPML = \sum_{i=1}^{n}\log(CPO_i) \quad (14)$$

The larger LPML value indicates a superior performance associated with the candidate model.

Evaluation Criteria for Predictive Accuracy

In this study, the four competing models were also evaluated based on some criteria used from previous studies: MSPE (mean-squared predictive error, 23), the G^2 statistic (40) , the R_p^2 statistic (41), the Chi-squared Residual Sum of Square (RSS, 42). The reader wishing more detail on these measures can refer to these studies. The details of each criterion are shown in the following subsections.

MSPE

As indicated by the name, such criterion is related with the average squared deviations, or, the predictive errors. Specifically, the MSPE was calculated as follows:

$$\mathrm{MSPE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\lambda_i - y_i)^2 \quad (15)$$

Where λ_i is the Bayesian estimated crash frequency for zone i while y_i is the observed crash counts of the same zone. The smaller MSPE is preferred which indicates a better prediction performance.

RSS

MSPE is based on the deviations. A potential issue is a larger estimated counts of one zone might mask the smaller ones of multiple TAZs. To address this issue, we also calculated the chi-squared residual sum of squares to determine the deviation standardized by the the estimated number of crash counts:

$$RSS = \sum_{i=1}^{n} \frac{(\lambda_i - y_i)^2}{\lambda_i} \tag{16}$$

The model with a smaller value of RSS tends to have more predictive capabilties.

The R_p^2 statistic

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, we adopted an equivalent measure, R_p^2, which is based on standardized residuals:

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n}\left[\frac{y_i - \lambda_i}{\sqrt{\lambda_i}}\right]^2}{\sum_{i=1}^{n}\left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}}\right]^2} \tag{17}$$

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

The G^2 statistic

The sum of model deviances, G^2, is zero for a model with perfect fit. The G^2 statistic is given as:

$$G^2 = 2\sum_{i=1}^{n} y_i LN\left(\frac{y_i}{\lambda_i}\right) \tag{18}$$

A large G^2 deviating from zero indicates that the model fits poorly as compared to the saturated model.

DATA PREPARATION

Pedestrian and bicyclist crashes which occurred in the City of Irvine in the period of 2007–2012 were analyzed for the study. Like many other research studies (43-45), 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).

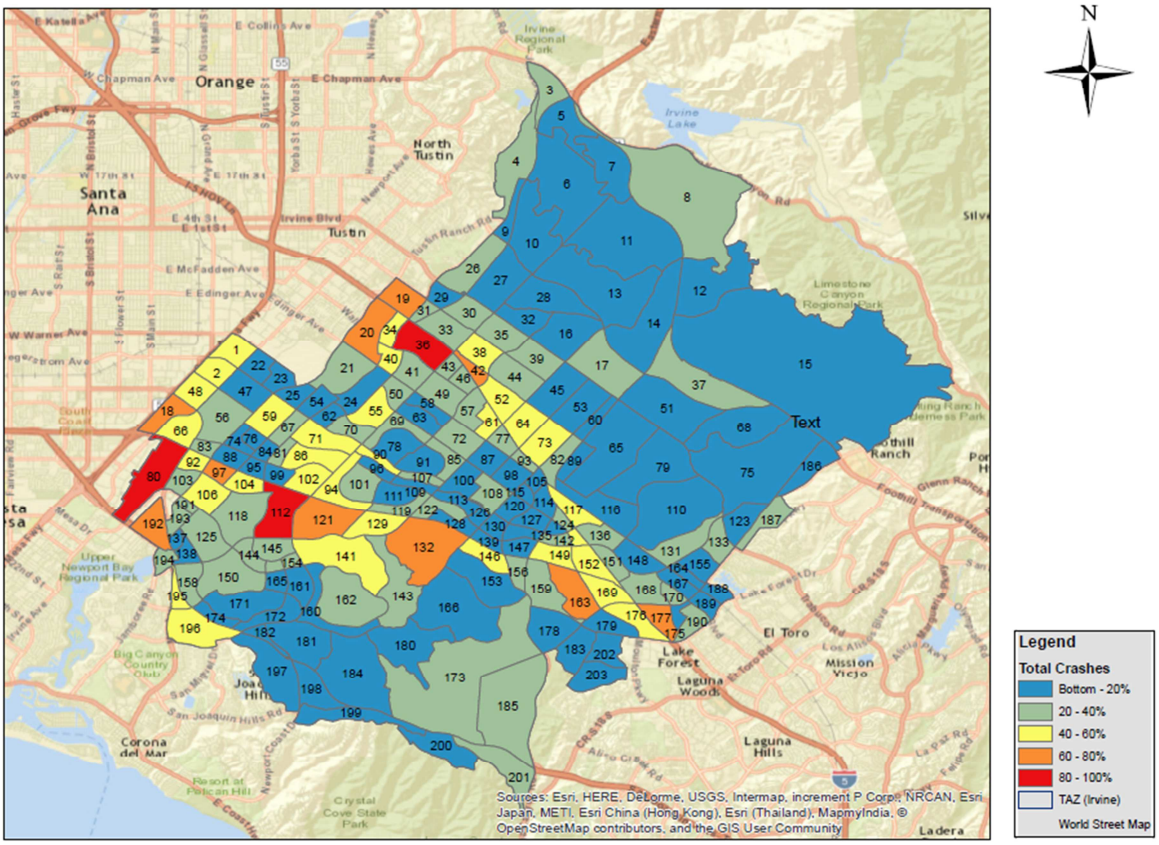


FIGURE 1 TAZ Map with Crash distributions in the City of Irvine, California.

The variables used for model development and the associated descriptive statistics are shown in Table 1. The numbers of pedestrian and bicyclist crashes aggregated from 6 years were used as the dependent variables. DVMT was utilized as the exposure variable. 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. In addition, the distance matrix containing distances among various TAZ centroids were also collected from SCAG for the estimation of distance-based spatial random effect. Since there are 203 TAZs in the city, the matrix includes 203x202 distances. Their descriptive statistics can be found in Table 1 as well.

TABLE 1 Summary Statistics of Variables for TAZ's of the City of Irvine

Variables	Description	Mean	Std. Dev.	Min.	Max.
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

RESULTS

The crash prediction models were estimated with the freeware statistical package WinBUGS (46) which sampled the estimates by employing Markov Chain Monte Carlo (MCMC) method. Two out of four models were semiparametric which utilized the Dirichlet process to allow flexibility of the intercept for all entities. Such specification supported the capability to incorporate infinite parameters based on the desired flexibility, but the truncated Dirichlet formulation was utilized to limit the number of parameters in an effort to reduce immense computational complexity associated with modeling of infinite parameters. A total of 10,000 MCMC iterations were utilized for parameter estimation after discarding first 1,000 iterations as burn-in. The crash data

for the two concerned modes fit the model specification reasonably well as reflected by the small number of iterations to reach convergence. The MCMC convergence was ensured by employing different approaches such as visual inspection of history plots, trace plots, and Gelman-Rubin diagram (47). Moreover, the accuracy of posteriors was ensured by recording the sample MC errors to be less than 5% of the associated standard deviations. In an effort to reduce the bias induced in the model estimates due to the incorporation of correlated covariates, the Harrell Miscellaneous package in R software was employed for calculation of Pearson correlation coefficient. The covariates observed to be correlated at a significance level of 0.05 were subsequently eliminated with due consideration to prevent exclusion of any potential influential variables which would result in loss of precision of estimated parameters.

Modeling Results

This study developed flexible models that accounted for the unobserved heterogeneity by combining the strengths of incorporating multivariate specification of dependency among crash modes, spatial random effects for the impact of neighboring TAZs, and Dirichlet process mixture for random intercept. For comparison purposes, all models were developed over the multivariate model which may be regarded as the base model to observe the potential advantages of Dirichlet models or inclusion of spatial random effects. As shown in Table 2, 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 filtered out similar significant factors that affect crash frequency for a particular mode. In the case of bicycle crashes, three variables were 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 above expectation of lower perceived risk which may encourage bicyclists to ride more in such areas, but the corresponding higher chances of the exposure of bicyclists to vehicular traffic increase the crash risk. 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 was also observed to be influential, along with other three factors shared with bicycle crashes. The increase in student population in the colleges of TAZs was observed 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 influenced 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 3, the heterogeneity error term demonstrated the presence of statistically significant correlation among the bicycle and pedestrian crashes which further justifies the employment of multivariate structure for joint estimation of crashes. However, the spatial random effect term exhibited the absence of a significant correlation as indicated by the covariance matrix. It may be possible that the explanatory variables incorporated for model development were sufficiently robust to account for the spatial characteristics that influence crash occurrence for the particular modes.

TABLE 2 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 Table1 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: Multivariate; Model 2: Multivariate Spatial; Model 3: Multivariate Dirichlet process mixture; Model 4: Multivariate Dirichlet process mixture spatial

TABLE 3 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: Multivariate; Model 2: Multivariate Spatial; Model 3: Multivariate Dirichlet process mixture; Model 4: Multivariate Dirichlet process mixture spatial

Evaluation Results

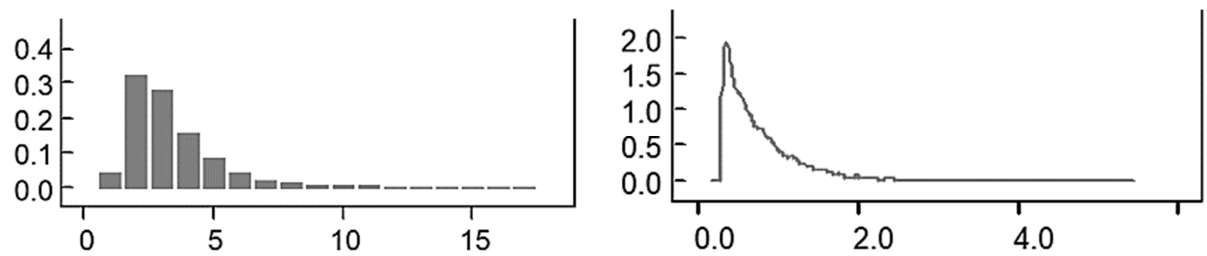
As previously stated, the four models were evaluated from different perspectives using five evaluation criteria. The conditional predictive ordinate (CPO) was calculated to cross-validate the crash estimates and eventually obtain LPML for comparison of model fit. Unlike the traditional parametric models which usually employ DIC (deviance information criterion) for model comparison, LPML was adopted in this study as DIC is not generated by the WinBUGS due to its sensitivity to different parameterizations (28,48**)**. 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 (49). As shown in Table 4, the LPML values of all four models were close enough to not cross the threshold of 5 points for identification of model of interest with superior fit. 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 demonstrated the best fit based on relative large LPML (-474.433), closely followed by Model 4. A Similar trend was 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 structured heterogeneity (Models 1 and 3). Such phenomenon suggests that the inclusion of spatial correlation structures and the resultant increased model complexity were not compensated by expected advantage at crash prediction. The potential reason might be due to the insignificant spatial dependency among the two modal crashes as shown in Table 3. Clearly, the Dirichlet models (Models 3 and 4) outperformed the non-Dirichlet ones (Model 1 and 2) based on all five criteria suggesting the use of such flexible framework. Apart from the better predictive accuracy and model fit, another advantage associated with Dirichlet models is the capability to identify the presence of distinct subpopulations. As shown in Figure 2, the kernel posterior density plots of Dirichlet precision parameter k illustrates 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. 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 justifies the use of Dirichlet process mixture with 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.

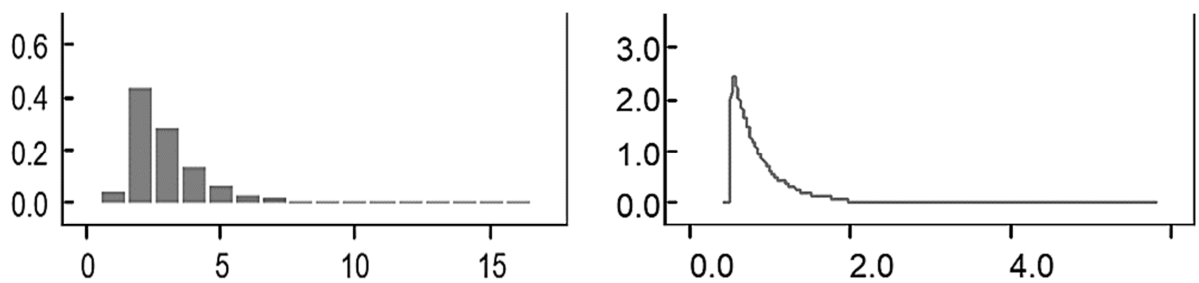
TABLE 4 Evaluation Results for Alternate Models

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

Note: Model 1: Multivariate; Model 2: Multivariate Spatial; Model 3: Multivariate Dirichlet process mixture; Model 4: Multivariate Dirichlet process mixture spatial



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



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

FIGURE 2 Kernel Density Plots for Precision Parameter and Latent Clusters

CONCLUSIONS AND RECOMMENDATIONS

The current study contributes to the safety literature by proposing a multivariate Dirichlet process mixture spatial model and comparing its performance for crash predictions with other three competing models. This study focuses on the active transportation at TAZ level by developing a semi-parametric model that accounted for the unobserved heterogeneity by combining the strengths of incorporating multivariate 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. Moreover, such sophisticated structure also facilitates for identification of latent subpopulations which may escape the traditional parametric models. The FB framework allowed the flexibility to accommodate the hierarchical structure and complex correlations in the crash data to jointly model pedestrian and bicycle crashes while accounting for the spatial correlations among TAZs.

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. Similar correlation among the concerned factors and pedestrian crashes was observed which indicated the advantage of joint modeling due to similar influential factors to the crash risk for the vulnerable modes. 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 possible explanation is that the lower perceived risk may encourage bicyclists to ride more in such areas and therefor 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 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. Such fact may be justified by the policies implemented in these areas for reduced or null 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 multivariate 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 the accommodation of such correlation structures 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. Finally, the crash dataset utilized for model development was aggregated for a six-year time period and future studies may incorporate temporal correlations and adopt disaggregated crash counts.

REFERENCES

1. National Highway Traffic Safety Administration, (2012b). Traffic Safety Facts: Bicy-clists and Other Cyclists. *United States Department of Transportation* (DOT HS 811 624).
2. Williams, A.F., (2013). Protecting Pedestrians and Bicyclists: Some Observations and Research Opportunities. *Insurance Institute for Highway Safety*.
3. Berrigan, D., Troiano, R.P., McNee, T., Disogra, C., Ballard-Barbash, R., (2006). Active transportation increases adherence to activity recommendations. *Am. J. Prev. Med.* 31 (3 (Sep)), 210–216.

4. Frank, L.D., Greenwald, M.J., Winkelman, S., Chapman, J., Kavage, S., (2010). Carbonless footprints: promoting health and climate change stabilization through active transportation. *Prev. Med. 50* (Suppl 1 (Jan)), S99–S105
5. Furie, G.L., Desai, M.M., (2012). Active transportation and cardiovascular dis-ease risk factors in U.S. adults. *Am. J. Prev. Med. 43* (6 (Dec)), 621–628
6. Giles-Corti, B., Foster, S., Shilton, T., Falconer, R., (2010). The co-benefits for health of investing in active transportation. *N. S. W. Public Health Bull. 21* (5–6 (May–Jun)), 122–127
7. Insall, P., (2013). Active travel: transport policy and practice for health. *Nutr. Bull. 38* (1), 61–69,
8. Wanner, M., Götschi, T., Martin-Diener, E., Kahlmeier, S., Martin, B.W., (2012). Active transport, physical activity, and body weight in adults: a systematic review. *Am. J. Prev. Med. 42* (5 (May)), 493–502
9. Lee, C., Abdel-Aty, M., (2005). Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida. *Accid. Anal. Prev. 37* (4 (Jul)), 775–786.
10. Moudon, A.V., Lin, L., Jiao, J., Hurvitz, P., Reeves, P., (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. *Accid. Anal. Prev. 43* (1 (Jan)), 11–24
11. Vivoda, J.M., Eby, D.W., St Louis, R.M., Kostyniuk, L.P., (2008). Cellular phone use while driving at night. *Traffic Inj. Prev. 9* (1 (Mar)), 37–41
12. Beck, L.F., Dellinger, A.M., O'Neil, M.E., (2007). Motor vehicle crash injury rates by modeof travel United States: using exposure-based methods to quantify differences. *Am. J. Epidemiol.*, 166.
13. Wardlaw, M., (2002). Assessing the actual risks faced by cyclists. *Traffic Eng. Control* 352–356.
14. Cai, Q., Lee, J., Eluru, N., & Abdel-Aty, M. (2016). Macro-level pedestrian and bicycle crash analysis: incorporating spatial spillover effects in dual state count models. *Accident Analysis & Prevention*, *93*, 14-22.
15. Bijleveld, F. D. (2005). The Covariance between the Number of Accidents and the Number of Victims in Multivariate Analysis of Accident Related Outcomes. *Accident Analysis and Prevention*, Vol. 37, No. 4, pp. 591–600.
16. Park, E., & Lord, D. (2007). Multivariate poisson-lognormal models for jointly modelling crash frequency by severity. *Transportation Research Record*, *2019*(2019), 1-6
17. Congdon, P., (2001). Bayesian Statistical Modeling. *John Wiley & Sons, West Sussex, United Kingdom*
18. Conway, A., Cheng, J., Peters, D., & Lownes, N. (2013). Characteristics of Multimodal Conflicts in Urban On-Street Bicycle Lanes. *Transportation Research Record: Journal of the Transportation Research Board*, (2387), 93-101.
19. Strauss, J., Miranda-Moreno, L. F., & Morency, P. (2014). Multimodal injury risk analysis of road users at signalized and non-signalized intersections.*Accident Analysis & Prevention*, *71*, 201-209.
20. Guo, F.,Wang, X., Abdel-Aty,M. A., 2010. Modeling signalized intersection safety with corridor-level spatial correlations. Accid.Anal.Prev. 42(1), 84-92.
21. Abdel-Aty, M., Wang, X., 2006. Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. In: Transportation Research Record: Journal of the Transportation Research Board, No. 1953. Transportation Research Board of the National Academies, Washington, DC, pp. 98–111.

22. Gill, G., W. Cheng, M. Xie, T. Vo, X. Jia, J. Zhou. (2017) Evaluating the Influence of Neighboring Structures on Spatial Crash Frequency Modeling and Site Ranking Performance, Transportation Research Record: Journal of the Transportation Research Board, No. 2659 (DOI: 10.3141/2659-13).
23. Narayanamoorthy, S., Paleti, R., & Bhat, C. R. (2013). On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transportation research part B: methodological, 55, 245-264.
24. Nashad, T., Yasmin, S., Eluru, N., Lee, J., & Abdel-Aty, M. A. (2016). Joint modeling of pedestrian and bicycle crashes: copula-based approach. Transportation Research Record: Journal of the Transportation Research Board, (2601), 119-127.
25. Shively, T. S., Kockelman, K., & Damien, P. (2010). A Bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics. Transportation research part B: methodological, 44(5), 699-715.
26. Heydari, S., Fu, L., Lord, D., & Mallick, B. K. (2016). Multilevel Dirichlet process mixture analysis of railway grade crossing crash data. *Analytic methods in accident research*, *9*, 27-43.
27. Heydari, S., Fu, L., Miranda-Moreno, L. F., & Joseph, 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.
28. Ohlssen, D. I., Sharples, L. D., & Spiegelhalter, D. J. (2007). Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. Statistics in medicine, 26(9), 2088-2112.
29. Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X. and Zhou, J.. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. *Accident Analysis & Prevention*, *99*, 2017, pp.330-341.
30. Congdon, P., (2006). Bayesian Statistical Modeling. *2nd ed. Wiley, New York.*
31. Mardia, K.V., (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *J. Multivar. Anal.* 24 (2), 265–268.
32. Jonathan, Aguero-Valverde, Kun-Feng Ken Wu, and Eric T. Donnell. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis & Prevention* 87 (2016): 8-16.
33. Aguero-Valverde, J., & Jovanis, P. (2009). Bayesian multivariate poisson lognormal models for crash severity modeling and site ranking. *Transportation Research Record*, *2136*(2136), 82-91.
34. Xu, P., Huang, H., (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accid. Anal. Prev*. 75, 16-25.
35. Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161-173.
36. Ghosh SK, Norris JL. Bayesian capture-recapture analysis and model selection allowing for heterogeneity and behavioral effects. Journal of Agricultural Biological and Environmental Statistics 2005; 10:35–49.
37. Erkanli A, Sung M, Jane Costello E, Angold A. Bayesian semi-parametric roc analysis. Statistics in Medicine, 2006.

38. Gelfand, A. E., Dey, D. K., & Chang, H. (1992). *Model determination using predictive distributions with implementation via sampling-based methods* (No. TR-462). STANFORD UNIV CA DEPT OF STATISTICS.
39. Gelfand, A. E. (1996). Model determination using sampling-based methods. *Markov chain Monte Carlo in practice*, 145-161.
40. Cheng, W., Washington, S.P., (2005). Experimental evaluation of hotspot identification methods. *Accident Analysis & Prevention* 37 (5), 870–881.
41. Washington, S., M. G. Karlaftis, and F. L. Mannering (2003). Statistical and Econometric Methods for Transportation Data Analysis, Chapman & Hall.
42. Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., & Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International journal of health geographics*, *6*(1), 54.
43. Ladron de Guevara, F., S. P. Washington, and J. Oh. (2004). Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Models Applied in Tucson, Arizona. In Transportation Research Record: Journal of the Transportation Research Board, No. 1897, TRB, National Research Council, Washington, D.C., pp. 191–199.
44. Hadayeghi A., A. Shalaby, and B. Persaud (2007). Safety Prediction Models: Proactive Tool for Safety Evaluation in Urban Transportation Planning Applications. In Transportation Research Record: Journal of the Transportation Research Board, No. 2019, TRB, National Research Council, Washington, D.C., pp. 225–236.
45. Abdel-Aty, M., C. Siddiqui, H. Huang, and X. Wang (2011). Integrating Trip and Roadway Characteristics to Manage Safety in Traffic Analysis Zones. In Transportation Research Record: Journal of the Transportation Research Board, No. 2213, TRB, National Research Council, Washington, D.C., pp. 20–28.
46. Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., (2003). WinBUGS Version 1.4 User Manual. *MRC Biostatistics Unit, Cambridge, http://www.mrc-cam.ac.uk/bugs*
47. Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457-472.
48. Geedipally, S. R., Lord, D., & Dhavala, S. S. (2014). A caution about using deviance information criterion while modeling traffic crashes. *Safety science*, *62*, 495-498.
49. Ntzoufras, I. (2009). Bayesian modeling in WinBugs. Hoboken.