Accident Analysis and Prevention 42 (2010) 84–92

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Modeling signalized intersection safety with corridor-level spatial correlations

Feng Guo[a,*], Xuesong Wang[b,1], Mohamed A. Abdel-Aty[c,2]

[a] Department of Statistics, Virginia Tech Transportation Institute, Virginia Tech, 406A Hutcheson Hall, Blacksburg, VA 24061-0439, United States
[b] School of Transportation Engineering, Tongji University, 4800 Cao'an Road, Shanghai 201804, China
[c] Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816, United States

ARTICLE INFO

Article history:
Received 23 February 2009
Received in revised form 10 July 2009
Accepted 13 July 2009

Keywords:
Bayesian approach
Signalized intersection
Safety analysis
Corridor
Spatial model
Conditional autoregressive model

ABSTRACT

Intersections in close spatial proximity along a corridor should be considered as correlated due to interacted traffic flows as well as similar road design and environmental characteristics. It is critical to incorporate this spatial correlation for assessing the true safety impacts of risk factors. In this paper, several Bayesian models were developed to model the crash data from 170 signalized intersections in the state of Florida. The safety impacts of risk factors such as geometric design features, traffic control, and traffic flow characteristics were evaluated. The Poisson and Negative Binomial Bayesian models with non-informative priors were fitted but the focus is to incorporate spatial correlations among intersections. Two alternative models were proposed to capture this correlation: (1) a mixed effect model in which the corridor-level correlation is incorporated through a corridor-specific random effect and (2) a conditional autoregressive model in which the magnitude of correlations is determined by spatial distances among intersections. The models were compared using the Deviance Information Criterion. The results indicate that the Poisson spatial model provides the best model fitting. Analysis of the posterior distributions of model parameters indicated that the size of intersection, the traffic conditions by turning movement, and the coordination of signal phase have significant impacts on intersection safety.

1. Introduction

Signalized intersections are among the most dangerous locations of a roadway network due to the complex traffic conflicting movements and frequently changing traffic signals. In the United States, more than 2.8 million intersection-related crashes occurred in the year 2000 and among those, 1.3 million crashes and 445,000 injuries occurred at signalized intersections (FHWA, 2005). Crashes related to signalized intersections also tend to be more severe: 30% of intersection-related fatalities occurred at signalized intersections while only 10% of intersections are signalized (Rice, 2007). Improving intersection safety has been considered as a top priority by federal, state, and local agencies (AASHTO, 2005; FDOT, 2007). There is an urgent need to investigate and improve traffic safety at signalized intersections.

One main goal of intersection safety studies is to identify high risk factors among intersection geometric design features, traffic control and operational features, and traffic flow characteristics. Wang et al. (2006) have shown that the traffic volume per lane has a significant impact on safety. Furthermore, there is a significant association between through/left-turning movements and the rear-end, right-angle, and left-turn crashes (Wang and Abdel-Aty, 2007, 2008; Wang et al., 2008). Intersection geometric design features (i.e., number of through lanes, right-turn lanes, left-turn lanes, etc.) and traffic control and operational features (i.e., signal phase, speed limit, etc.) were also found to have significant influence on crash occurrence (Abdel-Aty and Wang, 2006; Poch and Mannering, 1996; Wang et al., 2006, 2008; Wang and Abdel-Aty, 2007, 2008).

Generalized linear models (GLM), including Poisson and Negative Binomial (NB) regression models, are widely used to relate the number of accidents to risk factors (Abdel-Aty and Radwan, 2000; Greibe, 2003; Poch and Mannering, 1996). The assumption of independent observations for ordinary GLM is often violated for traffic safety data. This is especially true for intersections along a corridor. Various mechanisms can lead to this dependency. First, adjacent signalized intersections along a certain corridor will share a high percentage of through-traffic. Therefore, the driving behavior and driver characteristics for those intersections tend to be similar. Second, signals within 0.5 mile of each other along a corridor would be coordinated in most circumstances (Rodegerdts et al., 2004); and this coordination in signals will promote platoon of vehicles crossing intersections. The platoon of vehicles would lead to similar traffic flow patterns along intersections and thus similar crash patterns. Finally, adjacent intersections along a corridor tend to share similar land use and roadway design characteristics. For the

* Corresponding author. Tel.: +1 540 231 7933; fax: +1 540 231 3863.
E-mail addresses: feng.guo@vt.edu (F. Guo), wangxs@tongji.edu.cn (X. Wang), mabdel@mail.ucf.edu (M.A. Abdel-Aty).
[1] Tel.: +86 21 69583946.
[2] Tel.: +1 407 823 5657; fax: +1 407 823 3315.

0001-4575/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2009.07.005

Table 1
Summary of selected corridors and intersections.

County	Corridor	Direction	Selected intersection node (with sequence)	Total
Hillsborough County	56th St	NS	1102, 1103, 1397, 1104, 1105, 1106, 1107, 1108	8
	Florida Ave	NS	1130, 1126	2
	Fowler Ave	WE	1101, 1120, 1121	3
	Nebraska Ave	NS	1143, 1144, 1152	3
	SR 39 (North)	NS	1356, 1377, 1381	3
	SR 574	WE	1008, 1010, 1011, 1013, 1088, 1014, 1015, 1016, 1328, 1322, 1324, 1325, 1321, 1326	14
	SR 580	WE	1348, 1190, 1189, 1188, 1184, 1181	6
	SR 597	NS	1318, 1225, 1222, 1221, 1220, 1162, 1163, 1164, 1166	9
	SR 60	WE	1083, 1027, 1028, 1031, 1086, 1033, 1034, 1036, 1037, 1038, 1039, 1040, 1367, 1385, 1366, 1422, 1382, 1384, 1380	20
	SR 600 (East)	WE	1001, 1002, 1003, 1004, 1005, 1006, 1007, 1365, 1337	9
	SR 600 (West)	WE	1109, 1112, 1114	3
	SR 674	WE	1309, 1289, 1394, 1388, 1386, 1371, 1387	7
	SR 676	WE	1307, 1085, 1048	3
	US 301	NS	1296, 1305, 1327, 1376, 1045, 1046, 1048, 1332, 1064, 1066, 1068, 1439, 1301	13
	US 41	NS	1135, 1133	2
	US 41 (South)	NS	1370, 1307, 1362, 1375, 1369, 1302, 1288, 1378, 1309	9
Orange County	SR 15	NS	77, 18, 76	3
	SR 423	WE	101, 173	2
	SR 434	NS	337, 53, 242, 7, 6, 260, 3	7
	SR 436	NS	14, 230, 29, 212, 68	5
	SR 438	WE	211, 34, 218, 131, 217, 159, 232, 82, 185	9
	SR 50 (East)	WE	68, 63, 64, 57, 59, 58, 66, 237, 3, 36, 336, 417, 69, 360	14
	SR 50 (West)	WE	24, 386, 89, 181, 176, 35, 288, 124, 386, 344, 269, 134, 184, 183, 65, 67	16
	SR 551	NS	123, 64, 121, 78, 122	5
	US 441	NS	395, 37	2
Total	25		170	

above reasons, it is expected that the safety of spatially proximate intersections is correlated instead of independent.

There is limited research considering the dependency among signalized intersections. To avoid the spatial correlation, Poch and Mannering (1996) used a subset of intersections that were considered to be independent. Abdel-Aty and Wang (2006) are among the first to consider corridor-level correlations. The correlations for 476 signalized intersections from 41 corridors in the state of Florida were incorporated using the Generalized Estimating Equations (GEE) approach (Abdel-Aty and Wang, 2006; Wang and Abdel-Aty, 2006). The analyses for both total crashes and rear-end crashes confirmed the presence of significant correlation for successive intersections along a corridor. The main limitation of these studies is that the GEE model assumes the same correlation matrix for different corridors. However, the signal spacing, i.e., distance between adjacent signalized intersections, is unlikely to be identical among corridors. This will violate the assumption for the GEE model.

The Bayesian statistical method has been adopted in many traffic safety studies. The Bayesian approach treats model parameters as random and the inference is based on the posterior distributions, which combine information from both observed data and prior distributions. This information infusion is one major advantage of the Bayesian approach. The prior can be either objective (e.g., non-informative prior and Jeffrey's prior) or elicited from historical data or expert opinions. Hauer et al. (2002) used the Empirical Bayes (EB) method to increase the precision of estimates for small samples and to correct for the regression-to-the-mean bias. A number of studies have shown the advantages of using full Bayesian models, especially hierarchical models, in modeling traffic safety data (Carriquiry and Pawlovich, 2008; Miranda-Moreno and Fu, 2007). In recent years, Bayesian spatial models have been applied for traffic safety studies under different contexts (Guero-Valverde and Jovanis, 2006; MacNab, 2004; Miaou et al., 2003; Quddus, 2008; Song et al., 2006).

Ignoring the spatial correlation would lead to invalid statistical inference. In many cases, assuming independency will cause underestimation of standard errors for model parameters (Wang et al., 2006), thus overly optimistic significant levels. The focus of this paper is to model signalized intersection safety and incorporate the dependency among intersections along corridors in a full Bayesian framework. The dependency can be incorporated through either appropriate likelihood or prior structure. Two levels of correlation were considered. At the first level, intersections on the same corridor are assumed to be equally correlated regardless of the relative spatial distances among them. At the second level, the magnitude of the correlation among intersections within a corridor is affected by the spatial distance among intersections. The remainder of the paper is structured as follows: the intersection data are introduced in Section 2; five full Bayesian models with alternative treatments for the corridor-level correlations are introduced in Section 3; application and comparison are presented in Section 4; and Section 5 summarizes the main results and their implications.

2. Data

A total of 170 four-legged signalized intersections along 25 principle and minor arterials were selected from Orange and Hillsborough counties in the Central Florida area. Generally three-legged intersections tend to exhibit lower crash rates than four-legged intersections (Abdel-Aty and Wang, 2006). Since the safety mechanisms for three- and four-legged intersections are different, only four-legged signalized intersections were included in this study.

The Geographic Information System (GIS) and Google Earth (Google Inc., 2008) were used to visualize and explore spatial data to assist in signalized intersection selection. The geocoding procedure in ArcGIS was used to locate intersections on a GIS base map and identify the projected map coordinates of each intersection. The coordinates were used to calculate the spatial distance among intersections, which is the basis for evaluating spatial correlations. Each intersection is located on a primary corridor. The primary corridor is defined as a multi-lane highway with high speed limit and serving relatively long trips between major points. The number of intersections on each corridor varies from 2 to 19 as shown in Table 1. The road network and geocoded intersections for the two counties are presented in Fig. 1a and b.

The safety of intersections was measured by the total number of crashes. Crashes that occurred at the selected intersections

for 6 years (2000–2005) were retrieved from the Crash Analysis Reporting system maintained by the Florida Department of Transportation (FDOT) Safety Office. The total number of crashes for an intersection was calculated by combining both at-intersection and influenced-by-intersection crashes. The crashes that are influenced-by-intersection are those that occurred in road segments close to an intersection (the safety influence area). Instead of using an arbitrary fixed-length influence area (e.g., 250 feet), a dynamic safety influence areas (DSIA) was adopted in this study (Wang et al., 2008). The basic premise of DSIA is that the upstream safety influence area is mainly affected by the attributes of that approach, i.e., approach through volume, speed limit, jurisdiction, number and length of right-turn lanes, and approach left-turn protection. The DISA incorporates those factors and has been demonstrated to be appropriate in intersection safety studies (Wang et al., 2008).

Detailed information was collected regarding intersection geometric design features, traffic control and operational features, and traffic flow characteristics. The geometric design features include number of through lanes, number of left-turn lanes/exclusive left-turn lanes, presence of median, presence of exclusive right-turn lanes, types of left-turn lane offset (negative, zero, or positive offset), direction of each intersection roadway, and angle of intersecting roadways.

Traffic control and operational features were collected by inspecting signal plans provided by county traffic engineering

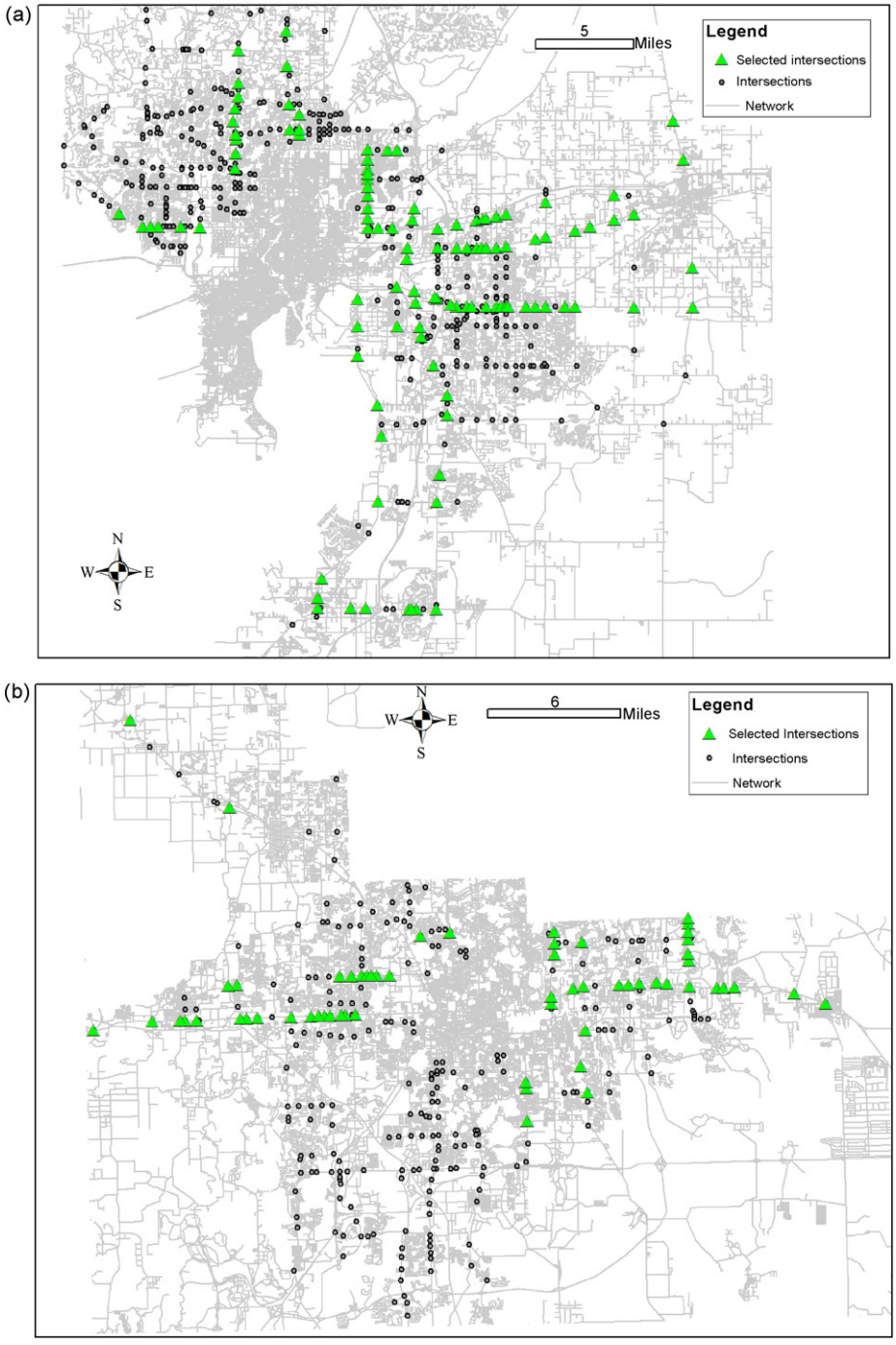

Fig. 1. (a) Road network and geocoded signalized intersections for Hillsborough County. (b) Road network and geocoded signalized intersections for Orange County.

Table 2
Summary of independent variables.

Variable name	Description	Summary statistics
Exposure[a]	Sum of total entering traffic for the entire intersection	Mean: 51.9; std: 25.5
County	0 for Orange County; 1 for Hillsborough County	60 intersections in Orange County; 110 intersections in Hillsborough County
ADTMJth[a]	ADT per lane through-traffic on major road, standardized	Mean: 6.9; std: 3.0[b]
ADTMJlt[a]	ADT per lane left-turn traffic on major road, standardized	Mean: 1.5; std: 1.0[b]
ADTMNth[a]	ADT per lane through-traffic on minor road, standardized	Mean: 2.0; std: 2.2[b]
ADTMNlt[a]	ADT per lane left-turn traffic on minor road, standardized	Mean: 2.5; std: 3.3[b]
CoorMJ	Signal coordination along corridor: 0 for isolated intersection; 1 for coordinated intersection	Isolated: 50 intersections; coordinated: 120 intersections
IntSize	Intersection size: small or medium intersections have less than 19 lanes (coded as 0); large intersections have equal or more than 19 lanes (coded as 1)	Small and medium size: 155 intersections; Large size: 15 intersections
Landuse	Intersection surrounding land use types, 1 for urban and 0 for rural	Urban: 97; rural: 73
SpeedMJ	Speed limit along major road, 1 for ≥45mph, 0 for <45mph	<45 mph: 25 intersections; ≥45 mph: 145 intersections
SpeedMN	Speed limit along minor road, 1 for ≥45mph, 0 for <45mph	<45mph: 104; >45mph: 66
Corridor	Unique identification number for each corridor	25 corridors were selected

[a] The unit for ADT is thousand-vehicle per day.
[b] The summary statistics for ADT per lane is before standardization.

departments. A number of variables were extracted including speed limit, types of left-turn control (permissive, compound or protected), key factors for signal phases (i.e., yellow time and all-red time for through and left-turn—if protected—movements), and signal coordination information. The average daily traffic (ADT) for each intersection approach was also collected including the total traffic, through-traffic, and left-turn traffic ADTs on major and minor roads.

Although a relative large number of independent variables were collected, only a few were actually included in final models as the collected variables often provide redundant information and can be highly correlated. For example, the correlation coefficient between through-traffic and total traffic on major roads is as high as 0.96. This multicollinearity will lead to identifiability problems of model coefficients and highly inflated variance estimation. The selection of independent variables thus follows three principles: (1) the variable should have a sound engineering interpretation; (2) the variable should represent different aspects of properties of an intersection; and (3) there should be a weak/moderate correlation among the selected variables.

Traffic volume has a significant effect on safety and can be considered as both exposure measure and risk factor (Qin et al., 2004). Therefore, the total traffic volume from all approaches was included to indicate the overall exposure level. Besides exposure information, the intensity of turning traffic movement is directly related to traffic level of service and thus would affect intersection safety (Wang et al., in press). The intensity can be conveniently measured by through-traffic and left-turn ADT per lane on major and minor lanes. Those four variables were standardized (subtracted by their means and divided by the standard deviation) to make them comparable. The standardized variables show moderate correlation with the total ADT (ranging from 0.18 to 0.60) and can be used in the same model.

Based on the above analysis, 12 variables representing various aspects of intersection characteristics were selected as listed in Table 2.

3. Model structure

A full Bayesian framework was used in this study. The inference is based on the posterior distribution of model parameters:

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{L(\boldsymbol{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\boldsymbol{D})}, \quad (1)$$

where $\boldsymbol{\theta}$ is the vector of parameters, $\mathbf{D}$ is the set of observed data, $\pi(\boldsymbol{\theta}|\mathbf{D})$ is the posterior distribution, $L(\mathbf{D}|\boldsymbol{\theta})$ is the likelihood function, $\pi(\boldsymbol{\theta})$ is prior distribution of $\boldsymbol{\theta}$, and $m(\mathbf{D})$ is the marginal distribution of data $\mathbf{D}$. The key to a Bayesian model is the selection of appropriate likelihood function and prior distributions. The posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{D})$ combines information from both the data (through likelihood function) and prior. As will be illustrated later, the correlation among intersections can also be incorporated through either likelihood function or prior distribution. Five alternative models were developed with difference treatment for corridor-level spatial correlation.

3.1. Poisson and NB models

The likelihood function $L(\mathbf{D}|\boldsymbol{\theta})$ is the distribution function of data $\mathbf{D}$ given parameter $\boldsymbol{\theta}$. Similar to most traffic safety studies, the Poisson and NB regression models were used as the base models. To develop the notation, let Y_{ij} represent the number of crashes at intersection j on corridor i ($i = 1,\ldots,I$, $j = 1,\ldots,n_i$, n_i is the total number of intersections on corridor i) and $\mathbf{Y}$ be the vector of Y_{ij}. The Poisson/NB model assumes

$$y_{ij} \sim \text{Poisson}(\lambda_{ij}),$$

or

$$y_{ij} \sim \text{Negbin}(\lambda_{ij}, k),$$

where λ_{ij} is the expectation of y_{ij} and k is an overdispersion coefficient. For the Poisson model, the variance is equal to the expectation: $\text{var}(y_{ij}) = \lambda_{ij}$. Overdispersion, which refers to the situation where the variance is greater than that allowed by the Poisson model, is commonly presented in safety studies (Liu and Dey, 2007; Mitra and Washington, 2006). In a NB regression model, the expectation of number of crashes is the same as Poisson model but allows larger variance through parameter k: $var(y_{ij}) = \lambda_{ij} + \lambda_{ij}^2/k$.

In a GLM framework, the expectation of y_{ij} is considered to be related to a set of independent variables. For both Poisson and NB models, a logarithm link function can be used to connect the expectation with covariates,

$$\log(\lambda_{ij}) = \rho \log(E_{ij}) + \psi_{ij}, \quad (2)$$

where E_{ij} is the total traffic volume and $\exp(\psi_{ij})$ is the crash rate per unit of exposure. It is easy to see that the expected number of crashes at intersection j on corridor i is: $\lambda_{ij} = E_{ij}^{\rho} e^{\psi_{ij}}$. It is well known

that the relationship between the expected number of crashes and traffic volume is nonlinear (Qin et al., 2004). In Eq. (2), ρ is an exposure coefficient and E_{ij}^{ρ} is the actual exposure measure. The crash rate is connected with covariates through a linear relationship:

$$\psi_{ij} = \mathbf{X}_{ij}'\boldsymbol{\beta}, \tag{3}$$

where $\mathbf{X}_{ij}$ is the covariate matrix and $\boldsymbol{\beta}$ is the vector of regression parameters.

A prior distribution is assigned for each parameter to represent *a priori* information without consulting the observed data. The priors can be elicited from either historical data or expert opinions. When no such information is available, vague/non-informative priors can be used which are typical distributions with large variance. The first two proposed models assume Poisson and NB likelihood and use vague priors for the model parameters. Specifically, the prior distributions are as follows:

$$\rho \sim N(0, 10^5) \tag{4}$$

$$\boldsymbol{\beta} \sim N_p(0, 10^5\mathbf{I}) \tag{5}$$

where N and N_p represent normal and p-dimensional normal distribution respectively; $\mathbf{I}$ is the identity matrix. For NB model, a log-normal prior was used for the dispersion parameter k:

$$\log(k) \sim N(0, 0.01).$$

This completes the model setup for the simple Poisson and NB full Bayesian models. The corridor-level correlation was not considered in these two models.

3.2. Mixed effect model

Both Poisson and NB models assume that data are independent, but they can be extended to accommodate correlated data. There are potentially three levels of correlation for intersection safety. On the first level, crashes that occurred within the same county are more "similar" comparing to those in other counties. Ideally a random effect should be used to model this county level correlation. Since only two counties are available in this study, a random effect cannot be estimated satisfactorily and a fixed effect dummy variable was used instead. The second level of correlation is based on the tenet that intersections within the same corridor are correlated with each other. This effect is represented by a random effect b_i, where $I = 1,\ldots,I$ is the index for corridor. The last level of correlation is a micro-level spatial correlation. Within a corridor, the intersections close to each other are expected to be more similar than those far apart. This spatial correlation was incorporated into the model through a conditional autoregressive (CAR) prior.

Based on the above arguments, two extensions to the basic Poisson/NB models were proposed: a mixed effect model that incorporates within corridor correlation and a CAR model that incorporates the spatial correlations. The mixed effect model setup is as follows:

$$\psi_{ij} = \mathbf{X}_{ij}'\boldsymbol{\beta} + b_i, \tag{6}$$

where $\mathbf{X}_{ij}'$ is a vector of covariates associated with intersection j on corridor i, $\boldsymbol{\beta}$ is a vector of regression parameters as specified in Eq. (3); and b_i is a corridor-specific random effect. It is assumed that b_i's follow independent and identically distributed normal distribution

$$b_i \sim N(0, 1/\tau_b),$$

where τ_b is the precision parameter. In this mixed effect model, intersections on the same corridor are correlated through b_i and those on different corridors are independent of each other. To complete the Bayesian setup, a vague gamma prior was assigned to τ_b, i.e., $\tau_b \sim \text{Gamma}(0.001, 0.001)$. The rest of model setup is identical to the simple models as introduced in Section 3.1 (Eqs. (2) and (3)).

3.3. Spatial CAR model

The mixed effect model described above treats intersections on the same corridor equally regardless of the distance between intersections. However, it is believed that there exists a micro-level spatial correlation: intersections in close spatial proximity tend to be more similar than those far apart. This micro-level correlation represents a second order variation that cannot be sufficiently explained by covariates. Fig. 2 shows the residuals of an ordinary NB regression for the selected intersections in Hillsborough County. The larger circles in the plot indicate larger residuals. As can be seen, there are certain levels of micro-level spatial correlation depending on the distance between intersections. To quantitatively evaluate the spatial autocorrelation the Moran's I index was used as shown in the following equation:

$$I = \left(\frac{n}{\sum_i\sum_j w_{ij}}\right)\left(\frac{\sum_i\sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right) \tag{7}$$

where i and j are indexes for spatial units which are the intersections in this analysis; x is the variable to be examined; $\bar{x}$ is the mean of x; and w_{ij} is the spatial weight between intersection i and j. The weights represent the spatial relationship between intersections. When there is no spatial autocorrelation, the expected value of I is $E[I] = 1/(n-1)$.

For this study, there were 170 intersections so the expected value is $-1/169 = -0.00059$. The calculated Moran's I for the residuals of the NB model is 0.15 and p value is 0.01, which indicates significant positive spatial autocorrelation. The residuals represents the spatial variance cannot be explained by the independent variables. This result confirms that spatially proximate intersections tend to have similar crash patterns and more sophisticated models are needed to incorporate this spatial autocorrelation.

A conditional autoregressive model was adopted to incorporate the spatial autocorrelation. A random effect term ϕ_{ij} was introduced to capture the variation associated with intersection ij. This extends the crash rate model in Eq. (3) to the following form:

$$\psi_{ij} = \mathbf{X}_{ij}'\boldsymbol{\beta} + \phi_{ij}. \tag{8}$$

A CAR prior was assigned to ϕ_{ij} to incorporate the spatial correlations among intersections, which reflects the expectation that for two spatially close intersections indexed by ij and ij', ϕ_{ij} and $\phi_{ij'}$ will be of similar magnitude. The formulation in Eq. (8) applies to the main structure of both Poisson and NB models. The spatial relationship among intersections can be represented by a proximity matrix $\mathbf{W}$ with entry $w_{ij,i'j'}$ indicating the spatial relationship between intersections ij and $i'j'$. The $w_{ij,i'j'}$ was defined as a function of the distance between intersections:

$$w_{ij,i'j'} = \begin{cases} c(d_{ij,i'j'}) & \text{if } ij \neq i'j' \\ 0 & \text{if } ij = i'j' \end{cases}, \tag{9}$$

where $d_{ij,i'j'}$ is the distance between intersections ij and $i'j'$. Since intersections on different corridors were treated as independent, $w_{ij,i'j'}$ is set to 0 when $i \neq i'$. Here $c(d_{ij,i'j'})$ is a decreasing function of the "distance" d_{ij} so that intersections closer to each other are more similar than those far apart. In this analysis, a commonly used inverse distance function was adopted, i.e., $c(d_{ij,i'j'}) = 1/d_{ij,i'j'}$.

The conditional distribution of CAR prior has the following form:

$$\phi_{ij}|\phi_{(-ij)} \sim N\left(\sum_{j' \neq j}\frac{w_{ij',ij}}{w_{i+}}\phi_{ij'}, \frac{1}{\tau_c w_{i+}}\right), \tag{10}$$

Fig. 2. Residuals of NB model for intersections in Hillsborough County.

where $\phi_{(-ij)}$ is the collection of $\phi_{i'j'}, \forall i'j' \neq ij$, τ_c is a precision parameter, and $w_{i+} = \sum_{j'=1}^{n_i} w_{ij,ij'}$. The joint prior distribution for ϕ_{ij}'s is

$$\pi(\boldsymbol{\varphi}) \propto \exp\left\{-\frac{\tau_c}{2}\sum_{i'j' \neq ij} \omega_{ilj,i'j'}(\phi_{ij} - \phi_{i'j'})^2\right\},$$

where $\pi(\boldsymbol{\varphi})$ is the joint distribution of $\boldsymbol{\varphi} = \{\phi_{ij}: i = 1,\ldots I, j = 1,\ldots,i_i\}$, $\propto$ represents that the likelihood function is proportional (up to a constant) to the right-hand side of the equation. Note that this is a pair-wise difference model and is not a proper distribution and the ϕ_{ij}'s are non-identifiable. As usual in such models, the constraint $\sum\phi_{ij} = 0$ is sufficient to guarantee the identifiability.

As has been well known, the NB distribution is actually a composite Poisson distribution with a non-constant mean parameter (Liu and Dey, 2007). When a random effects model is assigned to each observation, it is expected that the overdispersion can be sufficiently modeled by a Poisson model. Therefore, both NB and Poisson CAR models were tested in the application.

4. Modeling results

4.1. Model comparison

All together, five full Bayesian models were considered with different complexity (1) a fixed effect Poisson regression model; (2) a fixed effect NB model; (3) a mixed effect NB model with intersections on the same corridor as a cluster; (4) a CAR NB model for corridor effect incorporating the relative distance among intersections, and (5) a CAR Poisson model. A model comparison was conducted to evaluate the fitting of models and identify the model that provides the best fitting for the data and representation for the underlying stochastic process.

As there is no closed-form solution for the proposed model, simulation-based MCMC was used to sample from the posterior distributions. During the fitting of the CAR model, it was observed that the coefficients for county dummy variable, the intercept, and the spatial effect ϕ_{ij} were confounded and convergence could not be reached. Therefore, the county effect was excluded from the two CAR models.

The model comparison was conducted using Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), which assesses models on the marginal space and is defined as

$$\text{DIC} = \bar{D} + p_D,$$

where D is the Bayesian deviance, $D = -2\log(p(y|\theta)) + 2\log(f(y))$ and $\bar{D}$ is the posterior mean of D. $\bar{D}$ is a measure of fitting of the model and p_D is effective number of parameters. A smaller value of DIC is preferred and greater than 5 difference in DIC value indicates a substantial difference. The results of the DIC are listed in Table 3.

As can be seen, there is a general trend that for a more complex model, a better model fitting can be achieved as indicated by smaller posterior deviance $\bar{D}$. At the same time the model complexity, as indicated by the effective number of parameters, will increase. The Poisson model has the largest DIC value of 3209. A fixed effect NB model provides a substantial improvement with a DIC value of 1630. It should be noted that the effective number of parameters for the Poisson model is 12, which exactly equals the number of coefficients used in the model. With the cost of one extra parameter (13 vs. 12), the NB model provides significantly better model fitting than the Poisson model. This result again implies that the overdispersion

Table 3
Model comparison using DIC.

Models	$\bar{D}$	p_D	DIC
Poisson model	3197	12	3209
NB model	1617	13	1630
NB mixed effect model	1606	20	1626
NB CAR model	1577	51	1628
Poisson CAR model	1200	178	1378

cannot be sufficiently modeled by the Poisson model but can be accommodated by the NB model.

The NB model with the corridor-specific random effect model is more complex with an effective number of parameters of 20 but improved modeling fitting as indicated by a smaller $\bar{D}$ compared to the fixed effect NB model. Overall, the NB mixed effect model shows a marginal improvement over the fixed effect NB model (1626 vs. 1630). The DIC of the NB CAR model is between the simple NB model and the mixed effect NB and there is no significant difference between them.

The Poisson CAR model shows considerable complexity with 178 effective number of parameters and substantially improved model fitting with $\bar{D}$ of 1200. The overall DIC evaluation is significantly better than the rest of the models. This is not a surprise result as the NB distribution is essentially a compound Poisson distribution with random parameters. In the CAR model, the random effect ϕ_{ij} can accommodate the overdispersion as sufficiently as the NB model. Based on these results, the Poisson CAR model is considered the preferred model.

The DIC was used to compare the performance of the alternative models. The Bayesian model validity, i.e., if the model sufficiently explained the variation from the data, was checked by the Bayesian residuals and there is no evidence of lack of fit. The NB simple Bayesian model was compared with the classical non-Bayesian NB model and they basically provided identical results. The model validity check for the non-Bayesian NB model using deviance show good model fitting. As the CAR models performed better than the NB simple model, the model fitting is considered adequate.

Table 4
Posterior summary of Bayesian model fitting.

Factors	Poisson model	NB model	NB mixed effect model	NB CAR model	Poisson CAR model
Intercept	3.51[a] 0.04[b] (3.43,3.59)[c]	3.332 0.171 (3.01,3.7)	3.34 0.18 (2.98,3.68)	3.60 0.17 (3.26,3.94)	3.50 0.18 (3.12,3.84)
Exposure	0.44 0.03 (0.38,0.51)	0.290 0.121 (0.05,0.52)	0.24 0.12 (−0.01,0.48)	0.14 0.16 (−0.18,0.43)	≅0 0.18 (−0.35,0.37)
County	0.40 0.02 (0.35,0.44)	0.441 0.090 (0.26,0.61)	0.42 0.11 (0.19,0.64)		
ADTMJth	−0.14 0.01 (−0.17,−0.113)	−0.130 0.059 (−0.24,−0.009)	−0.14 0.06 (−0.25,−0.023)	−0.22 0.07 (−0.35,−0.081)	−0.18 0.07 (−0.3,−0.046)
ADTMJlt	0.11 0.01 (0.08,0.13)	0.145 0.057 (0.04,0.26)	0.15 0.06 (0.04,0.27)	0.15 0.06 (0.03,0.26)	0.20 0.06 (0.09,0.31)
ADTMNth	0.10 0.01 (0.08,0.12)	0.136 0.056 (0.03,0.25)	0.14 0.06 (0.03,0.25)	0.13 0.06 (0.01,0.26)	0.18 0.06 (0.06,0.31)
ADTMNlt	−0.09 0.01 (−0.12,−0.06)	−0.060 0.044 (−0.14,0.02)	−0.05 0.05 (−0.14,0.04)	0.01 0.06 (−0.11,0.14)	0.03 0.07 (−0.11,0.16)
CoorMJ	0.17 0.03 (0.12,0.23)	0.252 0.113 (0.03,0.47)	0.29 0.12 (0.05,0.53)	0.45 0.15 (0.16,0.73)	0.43 0.18 (0.08,0.79)
Intsize	0.30 0.03 (0.24,0.36)	0.391 0.149 (0.11,0.68)	0.38 0.15 (0.09,0.67)	0.40 0.16 (0.1,0.72)	0.51 0.15 (0.21,0.81)
Landuse	0.09 0.02 (0.04,0.13)	0.107 0.093 (−0.08,0.29)	0.10 0.09 (−0.09,0.29)	0.02 0.13 (−0.24,0.27)	0.01 0.15 (−0.29,0.31)
SpeedMJ	0.05 0.03 (0,0.1)	0.116 0.118 (−0.12,0.35)	0.09 0.12 (−0.14,0.32)	0.03 0.13 (−0.23,0.29)	0.09 0.14 (−0.19,0.36)
SpeedMN	0.28 0.02 (0.24,0.33)	0.316 0.094 (0.13,0.5)	0.29 0.09 (0.11,0.48)	0.24* 0.11 (0.04,0.45)	0.19 0.11 (−0.03,0.4)
NB coefficient		4.45 0.54 (3.46,5.59)	4.74 0.62 (3.61,6.05)	5.771 1.146 (4.08,8.55)	
Random effect			165.90 346.40 (11.13,1030)		
CAR effects				16.8 16.21 (2.54,62.77)	0.79 0.11 (0.58,1.04)

[a] Posterior means.
[b] Posterior standard deviations.
[c] 95% credible intervals.

4.2. Variable estimation and interpretation

The model fitting was conducted using MCMC simulation and the posterior distributions were constructed from the simulation output. The posterior summary from the MCMC output for all the five models is shown in Table 4, which includes the posterior mean, posterior standard deviation, and 95% credible interval (CI). The 95% CI, which represents the interval within which lie 90% of the possible values of the quantity of interest, was used to evaluate whether a parameter is significant: when the 95% CI includes zero, the corresponding factor is not significant at the 95% level and vice versa.

The posterior statistics of the simple Poisson and NB models were compared with the classical non-Bayesian models and they are essentially identical. These results illustrate one advantage of the Bayesian approach: for most classical methods, the Bayesian alternative can provide similar results with vague (non-informative) priors.

The standard deviations tend to increase for more sophisticated models. This is especially true when compared to the simple Poisson model in which every parameter is statistically significant (CI does not include zero). With the increased dispersion from complex model, several parameters become non-significant which is expected. As discussed in Section 1, ignoring overdispersion and spatial correlations among observed data will lead to overly optimistic and invalid statistical inference.

The simple NB and mixed effect NB models provide very similar results. The mean of the posterior of the precision parameter of the random effect is 166, which corresponds to a very small variance (1/166) comparing to the values of other regression coefficients. This implies that the difference among the corridors is not substantial. Therefore, it is not surprising to observe similar parameter estimations as well as similar DIC values.

Comparing the two CAR models, the posterior distributions of parameters for the Poisson CAR model show larger variance. The major difference is in the estimation of the CAR model precision parameters. The Poisson model provides a rather large variance (1/0.79) when compared to that of the NB CAR model (1/16.8). This is an expected result since the NB model contains an extra overdispersion parameter k. For the NB model, the variation from the data is explained by the combination of parameter k and the CAR precision parameter. In the Poisson model, the overdispersion can only be accommodated by the CAR precision parameter, which naturally shows larger variation than the CAR precision parameter in the NB CAR model. This result confirms the well-known fact that the NB can be generated by a random effect Poisson model and implies that the random effect Poisson model can be valuable in modeling the safety data.

Three out of the four standardized ADT per lane variables are identified to be significant in both CAR models. The through-traffic per lane on major road has a negative impact on the crash rate with a value of −0.18. This coefficient is the logarithm of multiplicative effect for one unit increase of an independent variable on crash rate as determined by Eqs. (2) and (3). Specifically, for one standardized unit of increase in major through-traffic (3000 vehicles per lane, i.e., one standard deviation), the expected crash rate as measured by the number of crashes per thousand vehicles will drop by a multiplicative factor of $\exp(-0.18) = 0.83$. That is for two intersections with identical conditions for all other factors, the crash rate for the one with 3000 more through-vehicles per lane on major road is 0.83 time of the crash rate for the other. This is an expected result from an engineering point of view as the major through-traffic has less conflicts with the turning traffic because of signal setting and the nature of traffic movement. The left-turn ADT per lane on major roads showed a significant result: with one unit increase (1000 vehicles per day) the crash rate will increase by a factor of $\exp(0.2) = 1.22$. Similarly, with one unit increase for through-traffic per lane on minor roads (equivalent to 2200 vehicles per day), the crash rate will increase by a factor of $\exp(0.18) = 1.2$. The exposure parameter ρ for total ADT is relatively small and not significant for both CAR models. To explore the possible causes, a model including only total ADT was fitted and the ρ was significant. This result indicates that the correlation between the total ADT and four ADT per lane variables is the potential reason for non-significance ρ.

Another interesting result is that coordinated intersections are more unsafe than the isolated ones with an increased crash rate by a factor of $\exp(0.43) = 1.53$. That is 53% more crashes per thousand vehicles are likely to happen on coordinated intersection if all other variables are kept constant. Spatially proximate intersections along a corridor would be coordinated in most circumstances (Rodegerdts et al., 2004). There is a possibility that the travel speed is higher for coordinated intersections because of the green wave, i.e., vehicle travels through several intersections without stop. Another possible reason is that the relative short distance between coordinated intersections could lead to more traffic interactions among those interactions thus more crashes. For both reasons, the signal coordination could have been applied on a specific group of intersection that could be more dangerous than the isolated ones. Therefore, it is more a consequence than the cause. This phenomenon is known as the endogeneity problem and Kim and Washington (2006) have tackled the problem using a limited information maximum likelihood method. The incorporation of this approach in a Bayesian framework will be worthy of future exploration.

Larger intersections (greater than 19 lanes) are more dangerous than smaller intersections. The crash rate is $\exp(0.51) = 1.66$ times higher for larger intersections than the smaller or medium size intersections. That is 66% more crashes per thousand vehicles could occur on larger intersections if flow and all other variables are kept constant. This result is consistent with the findings from Porter and England (2000) and Wang et al. (2006). The land use types and speed limit on both major and minor roads did not show significant impact on intersection safety.

5. Summary and discussion

The number of crashes for multiple intersections on a same corridor should not be considered as independent because of the similar traffic flow patterns, signal control, as well as geographic and design features. In this paper, the correlations of intersections on a corridor were studied at two levels. First, the corridor-level effect was modeled using a shared random effect. This mixed effect model treats all intersections on the same corridor with equal level of correlation. Second, two spatial CAR models were introduced to reflect the tenet that intersections in closer proximity on the same corridor have higher correlation levels than those far apart. The CAR models advance the mixed effect model by taking distance-related micro-level spatial correlation into consideration.

The analyses were conducted in a full Bayesian framework. The flexibility of the Bayesian method allows sophisticated models, such as CAR models, to be constructed. The inference is based on the posterior distribution from the MCMC simulation. Five alternative models including ordinary Poisson and NB models, mixed effect models, and spatial models were compared using the DIC criterion. Following conclusions were reached through model comparison.

1. Consistent with many traffic safety studies, the NB model is superior to the Poisson model due to overdispersion.
2. Mixed effect NB model improved model fitting over the simple NB model.

3. The NB spatial CAR model provides similar results as the mixed NB model.
4. The Poisson CAR model outperforms all the alternative models as evaluated by the DIC criterion.

The model comparison results indicate that models incorporated spatial correlations are superior to the ordinary GLM. This implies that the spatial models provide a better representation of the stochastic processes underneath the observed safety data. From an engineering point of view, the spatial models will produce valid statistical inference as reflected in the estimation of the regression coefficients and the credible intervals. As can be seen from the comparison of the posterior inference, the parameters for the spatial model show larger variance than non-spatial models. One explanation is that the spatial correlation will reduce the actually effective sample size thus leads to larger variance. This is analog to the well-known overdispersion in Poisson regression models, which can be accommodated by NB models. The Poisson model tends to provide smaller variance, which is not valid in the presence of overdispersion. Similarly, ignoring spatial correlation for the spatially correlated intersections will lead to biased inference for model parameters and incorrect conclusions.

The detailed information collected for the 170 signalized intersections provides a unique opportunity for evaluating factors that are associated with high crash frequencies. The results provide insight into the causation of intersection safety and valuable information for traffic management and improvement. The following conclusions were reached by examining the model output.

1. Traffic conditions as measured by the standardized ADT per lane by turning movement on the major and minor roads have a significant impact on the safety of signalized intersections.
2. Intersection size is closely related to intersection safety. In general, larger intersections are more dangerous than smaller intersections.
3. Signal coordination shows negative impact on safety. However, it is considered as a consequence of self-selection mechanism in which dangerous intersections are more likely to be coordinated.

There are several possible future developments for this research. The current paper used aggregated traffic count from six years and a natural extension is to use a spatial-temporal model to incorporate the time correlations for consecutive years. The level of corridor correlation can be evaluated quantitatively. The performance of the spatial model is closed related to the proximity function (Eq. (8)). An inverse function was conveniently chosen but it will be of interest to investigate the performance of alternative functions, e.g., exponential function. Furthermore, to consider the endogeneity problem in a Bayesian work is also the direction worth further investigation.

Acknowledgements

The authors would like to thank the Florida Department of Transportation (FDOT) for providing the crash data and the help from Orange and Hillsborough Counties' traffic departments for sharing the valuable traffic data used in this study. This paper benefited from the comments and suggestions of three anonymous reviewers. This study is partially supported by the Program for Young Excellent Talents in Tongji University (1600219099).

References

AASHTO, 2005. AASHTO Strategic Highway Safety Plan: A Comprehensive Plan to Substantially Reduce Vehicle-Related Fatalities and Injuries on the Nation's Highways (Revised). AASHTO, Washington, DC.

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.

Abdel-Aty, M., Radwan, E., 2000. Modeling traffic crash occurrence and involvement. Accident Analysis & Prevention 32, 633–642.

Carriquiry, A., Pawlovich, M., 2004. From Empirical Bayes to Full Bayes: Methods for Analyzing Traffic Safety Data, www.dot.state.ia.us/crashanalysis/pdfs/eb_fb_comparison_whitepaper_october2004.pdf (accessed 17.02.08).

FDOT, 2007. Florida Strategic Highway Safety Plan, http://www.dot.state.fl.us/Safety/StrategicHwySafetyPlan.htm (accessed 15.10.07).

FHWA, 2005. National Agenda for Intersection Safety. Federal Highway Administration [FHWA], U.S. Department of Transportation (Publication No. FHWA-SA-02-007).

Google Inc., 2008. Google Earth [Computer Software], http://earth.google.com/ (accessed 17.02.08).

Greibe, P., 2003. Crash prediction models for urban roads. Accident Analysis & Prevention 35 (2), 273–285.

Guero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis & Prevention 38 (3), 618–625.

Hauer, E., Harwood, D.W., Council, F.M., Griffith, M., 2002. Estimating safety by the Empirical Bayes method, A tutorial. In: Transportation Research Record: Journal of the Transportation Research Board, No. 1784. Transportation Research Board of the National Academies, Washington, DC, pp. 126–131.

Kim, D.G., Washington, S., 2006. The significance of endogeneity problems in crash models: an examination of left-turn lanes in intersection crash models. Accident Analysis & Prevention 38 (6), 1094–1100.

Liu, J., Dey, D.K., 2007. Hierarchical overdispersed Poisson model with macrolevel autocorrelation. Statistical Methodology 4 (3), 354–370.

MacNab, Y.C., 2004. Bayesian spatial and ecological models for small-area crash and injury analysis. Accident Analysis & Prevention 36 (6), 1019–1028.

Miaou, S., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: a space–time modeling approach. Journal of Transportation and Statistics 6 (1), 33–57.

Miranda-Moreno, L.F., Fu, L., 2007. Traffic safety study: Empirical Bayes or Full Bayes? Presented at 86th Annual Meeting of the Transportation Research Board, Washington, DC.

Mitra, S., Washington, S.P., 2006. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis & Prevention 39 (3), 459–468.

Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-crash frequencies. Journal of Transportation Engineering 122, 105–113.

Porter, B.E., England, K.J., 2000. Predicting red-light running behavior: a traffic safety study in three urban settings. Journal of Safety Research 31 (1), 1–8.

Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis & Prevention 36 (2), 183–191.

Quddus, M.A., 2008. Modeling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accident Analysis & Prevention 40 (4), 1486–1497.

Rice, E., 2007. Taking action to reduce intersection fatalities. Safety Compass 1 (2), 1–3.

Rodegerdts, L., Nevers, B., Robinson, B., Ringert, J., Koonce, P., Bansen, J., Nguyen, T., McGill, J., Stewart, D., Suggett, J., Neuman, T., Antonucci, N., Hardy, K., Courage, K., 2004. Signalized Intersection: Information Guide. FHWA, U.S. Department of Transportation (Publication FHWA-HRT-04-091).

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97, 246–273.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B 64 (4), 583–616.

Wang, X., Abdel-Aty, M., 2008. Modeling left-turn crash occurrence at signalized intersections by conflicting pattern. Accident Analysis & Prevention 40, 76–88.

Wang, X., Abdel-Aty, M., Almonte, A., Darwiche, A., in press. Incorporating traffic operation measures in safety analysis at signalized intersections. Transportation Research Record: Journal of the Transportation Research Board.

Wang, X., Abdel-Aty, M., Nevarez, A., 2008. Investigation of safety influence area for four-legged signalized intersections: nationwide survey and empirical inquiry. In: Transportation Research Record: Journal of the Transportation Research Board, No. 2083. TRB, National Research Council, Washington, DC, pp. 86–95.

Wang, X., Abdel-Aty, M., 2007. Right-angle crash occurrence at signalized intersections. In: Transportation Research Record: Journal of the Transportation Research Board, No. 2019. TRB, National Research Council, Washington, DC, pp. 156–168.

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, X., Abdel-Aty, M., Brady, P., 2006. Crash estimation at signalized intersections: significant factors and temporal effect. In: Transportation Research Record: Journal of the Transportation Research Board, No. 1953. TRB, National Research Council, Washington, DC, pp. 10–20.