Crash frequency modeling for signalized intersections in a high-density urban road network

Kun Xie a,b,c,*, Xuesong Wang d, Kaan Ozbay a,b,c, Hong Yang a,b,c

a Department of Civil and Urban Engineering, New York University, Brooklyn, NY 11201, USA
b Center for Urban Science and Progress (CUSP), New York University, Brooklyn, NY 11201, USA
c Urban Mobility and Intelligent Transportation Systems (UrbanMITS) Laboratory, New York University, Brooklyn, NY 11201, USA
d Department of Transportation Engineering, Tongji University, Shanghai, China

Article info

Article history:
Received 12 October 2013
Received in revised form 10 June 2014
Accepted 10 June 2014

Keywords:
Crash frequency model
Signalized intersection
Hierarchy
Spatial correlation
Hierarchical conditional autoregressive model
High-density network

Abstract

Conventional crash frequency models rely on an assumption of independence among observed crashes. However, this assumption is frequently proved false by spatially related crash observations, particularly for intersection crashes observed in high-density road networks. Crash frequency models that ignore the hierarchy and spatial correlation of closely spaced intersections can lead to biased estimations. As a follow-up to our previous paper (Xie et al., 2013), this study aims to address this issue by introducing an improved crash frequency model. Data for 195 signalized intersections along 22 corridors in the urban areas of Shanghai was collected. Moran’s I statistic of the crash data confirmed the spatial dependence of crash occurrence among the neighboring intersections. Moreover, Lagrange Multiplier test was performed and it suggested that the spatial dependence should be captured in the model error term. A hierarchical model incorporating a conditional autoregressive (CAR) effect term for the spatial correlation was developed in the Bayesian framework. A deviance information criterion (DIC) and cross-validation test were used for model selection and comparison. The results showed that the proposed model outperformed traditional models in terms of the overall goodness of fit and predictive performance. In addition, the significance of the corridor-specific random effect and CAR effect revealed strong evidence for the presence of heterogeneity across corridors and spatial correlation among intersections.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Statistical modeling of the inter-relationship between crash frequencies and contributing factors associated with intersections is of great interest in intersection safety studies. Traditional crash frequency models mainly assume that crash observations are independent from each other. However, this assumption is often proved false by crash observations occurring at signalized intersections in high-density urban road networks. Take the urban areas of Shanghai as an example, where the average intersection spacing is only about 200 m. Such short spacing leads to potential, and sometimes unavoidable, dependence among crash observations at adjacent intersections. This dependence can be attributed to two factors: (a) hierarchy – signalized intersections located along the same corridor share similar traffic flow, geometric design
and land use, and thus characteristics of the intersection can be described by a two-level hierarchical data structure consisting of the intersection-level information and corridor-level information; and (b) spatial correlation — neighboring intersections typically interact with one another in operation, for example, signals of adjacent intersections are usually coordinated and queue spillovers may frequently occur in dense road networks. Modeling aggregated crash frequencies without properly addressing the issues of hierarchy or spatial correlation can lead to unreliable findings.

This study is a follow-up to our recent paper (Xie et al., 2013) which focuses on the analysis of contributing factors to crashes in view of the unique traffic characteristics of Chinese cities. Different from the previous one, the main objective of this study is to develop an improved crash frequency model for closely spaced signalized intersections by accounting for the hierarchy and spatial correlation of the crash observations.

2. Literature review

Poisson models (Jones et al., 1991; Miaou and Lum, 1993) and negative binomial (NB) models (Miaou, 1994; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000) have been widely used to capture the relationship between traffic crashes and contributing factors. It is widely recognized that Poisson models outperform the standard regression models in handling the nonnegative, random and discrete features of crash counts (Joshua and Garber, 1990; Maher and Summersgill, 1996). Despite the improved performance, however, the constraint of the mean being equal to the variance in Poisson models is often violated by over-dispersed crash data. Alternatively, NB models are used to accommodate this over-dispersion issue by incorporating an independently distributed error term. However, with the assumption of independent observations, neither the Poisson models nor the NB models address any inherent correlation of crash data.

To complement the Poisson models and NB models, random effect models have been proposed in previous studies to account for the potential heterogeneity across homogeneous groups (Shankar et al., 1998; Chin and Quddus, 2003; Wang et al., 2014). In addition, random parameter models which can be viewed as extensions of random effect models are developed to incorporate the variability of both the intercept and the variable coefficients across observations and thus provide more flexibility for handling the heterogeneity (Anastasopoulos and Mannering, 2009; El-Basyouny and Sayed, 2009a; Venkataraman et al., 2011, 2013, 2014). More recently, hierarchical models have become the preferred method to accommodate a multilevel data structure (Jones and Jorgensen, 2003; Lenguerrand et al., 2006; Kim et al., 2007; Huang and Abdel-Aty, 2010; Ahmed et al., 2011; Xie et al., 2013; Chen and Persaud, 2014). Hierarchical models can accommodate the heterogeneity among different groups and have the ability to incorporate variables at the specific levels where impacts of specific variables occur (Gelman and Hill, 2007).

Spatial dependence is primarily modeled in two ways: using a spatially lagged dependent variable and using an error term. (Anselin, 1988b). The former way is denoted the spatial lag specification which allows spatial dependence through both spatial spillover effects (observed variables at one location can affect the dependent variable of itself and its neighboring locations) and spatial error correlation effects (omitted variables at one location can affect the dependent variable of itself and its neighboring locations) (Narayanamoorthy et al., 2013; Chiou et al., 2014). The latter is referred to as the spatial error specification that assumes the spatial dependence is only due to spatial error correlation effects. To model spatially correlated observations, generalized estimating equations (GEEs) are often the go-to method (Lord and Persaud, 2000; Hutchings et al., 2003; Abdel-Aty and Wang, 2006; Wang and Abdel-Aty, 2006). However, GEEs have the limitation of setting the same correlation matrix for all intersection groups, and thus cannot reflect the discrepancies in correlations among different groups of intersections. Conditional autoregressive (CAR) models can provide more flexibility in specifying the magnitude of correlation and have been recommended in many recent studies (Song et al., 2006; Aguero-Valverde and Jovanilis, 2008; El-Basyouny and Sayed, 2009b; Guo et al., 2010). The CAR models capture the spatial dependence using the spatial error specification (Narayanamoorthy et al., 2013). To address the hierarchy and spatial correlation of crash data, a hierarchical model incorporating CAR effect terms is proposed in this study.

3. Data description

For this study, a total of 195 signalized intersections located along 22 corridors was selected in the urban areas of Shanghai. These intersections were limited to either 3-legged or 4-legged designs so that the analysis of geometric features could be simplified. To ensure the independence of intersections across corridors, efforts were made to avoid choosing intersecting corridors.

Variables at both intersection and corridor levels were extracted from different sources. First, the crash data for the year of 2009 were geocoded on a GIS map using the crash location description, and then linked to selected intersections. Second, traffic volumes for each intersection were acquired from loop detectors. Third, the Sydney coordinated adaptive traffic system (SCATS) provided the geometric design and traffic control data. In addition, for each intersection, the presence or absence of an elevated road over the intersection was observed and distances among adjacent intersections were measured from the projected coordinates of intersections on the GIS map.

In addition, floating car data (FCD) was obtained from over 40,000 GPS equipped taxis in Shanghai, which provided a continuous position data of each taxi operating in the city every ten seconds. The locations of the taxis were matched with the GIS road network, and only taxi samples that passed through the study corridors were chosen. The mean speed and speed variance of each corridor were acquired from the FCD of taxis traveling along the corridor.
Table 1
Intersection related variables with their description and descriptive statistics.
Source: Xie et al. (2013).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection type</td>
<td>0 for 4-leg, 1 for 3-leg</td>
<td>0</td>
<td>1</td>
<td>0.17</td>
<td>0.38</td>
</tr>
<tr>
<td>Minimum angle</td>
<td>Minimum angle of intersecting roadways</td>
<td>30.00</td>
<td>90.00</td>
<td>84.25</td>
<td>13.12</td>
</tr>
<tr>
<td>Maximum angle</td>
<td>Maximum angle of intersecting roadways</td>
<td>90.00</td>
<td>150.00</td>
<td>95.66</td>
<td>12.25</td>
</tr>
<tr>
<td>Total number of lanes</td>
<td>Total number of lanes for all entering approaches</td>
<td>3</td>
<td>25</td>
<td>10.74</td>
<td>4.34</td>
</tr>
<tr>
<td>Number of through lanes</td>
<td>Number of through lanes for all entering approaches</td>
<td>0</td>
<td>14</td>
<td>4.87</td>
<td>3.35</td>
</tr>
<tr>
<td>Number of right-turn lanes</td>
<td>Number of right-turn lanes for all entering approaches</td>
<td>0</td>
<td>6</td>
<td>1.12</td>
<td>1.34</td>
</tr>
<tr>
<td>Number of left-turn lanes</td>
<td>Number of left-turn lanes for all entering approaches</td>
<td>0</td>
<td>9</td>
<td>2.23</td>
<td>2.42</td>
</tr>
<tr>
<td>Number of through-right lanes</td>
<td>Number of through-right mixed use lanes for all entering approaches</td>
<td>0</td>
<td>4</td>
<td>1.24</td>
<td>1.21</td>
</tr>
<tr>
<td>Ratio of turning lanes</td>
<td>The ratio of the total number of turning lanes to the total number of lanes</td>
<td>0</td>
<td>0.92</td>
<td>0.21</td>
<td>0.24</td>
</tr>
<tr>
<td>Under elevated roads or not</td>
<td>1 for intersection under elevated roads, 0 for not</td>
<td>0</td>
<td>1</td>
<td>0.23</td>
<td>0.42</td>
</tr>
<tr>
<td>Proximity of intersections</td>
<td>Distance to the nearest intersection along corridors (km)</td>
<td>0.09</td>
<td>1.17</td>
<td>0.35</td>
<td>0.23</td>
</tr>
<tr>
<td>Total number of phases</td>
<td>0 for 2–3 phases, 1 for 4–6 phases</td>
<td>0</td>
<td>1</td>
<td>0.69</td>
<td>0.46</td>
</tr>
<tr>
<td>Cycle length</td>
<td>Traffic signal cycle length (s)</td>
<td>40.00</td>
<td>240.00</td>
<td>171.11</td>
<td>42.23</td>
</tr>
<tr>
<td>Saturation degree</td>
<td>Saturation degree of an intersection</td>
<td>0.00</td>
<td>166.00</td>
<td>85.71</td>
<td>26.22</td>
</tr>
<tr>
<td>Intersection ADT</td>
<td>Average daily traffic entering the entire intersection (in 10^4 vehicles)</td>
<td>0.77</td>
<td>14.03</td>
<td>4.62</td>
<td>2.35</td>
</tr>
<tr>
<td>Intersection ADTPL</td>
<td>Average daily traffic per lane of entire intersection (in 10^4 vehicles)</td>
<td>0.10</td>
<td>1.36</td>
<td>0.51</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Table 2
Corridor related variables with their description and descriptive statistics.
Source: Xie et al. (2013).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of a median</td>
<td>0 for without median, 1 for with median</td>
<td>0</td>
<td>1</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>One-way road or not</td>
<td>0 for two-way road, 1 for one-way road</td>
<td>0</td>
<td>1</td>
<td>0.09</td>
<td>0.29</td>
</tr>
<tr>
<td>ADT along corridor</td>
<td>Average daily traffic volume along corridors in two directions (in 10^4 vehicles)</td>
<td>0.75</td>
<td>5.64</td>
<td>2.62</td>
<td>1.21</td>
</tr>
<tr>
<td>ADTPL along corridor</td>
<td>Average daily traffic volume per lane in two directions along corridors (in 10^4 vehicles)</td>
<td>0.21</td>
<td>0.82</td>
<td>0.54</td>
<td>0.23</td>
</tr>
<tr>
<td>Mean speed</td>
<td>Mean speed of taxi samples on each studied corridor (km/h)</td>
<td>10.78</td>
<td>42.01</td>
<td>22.61</td>
<td>6.94</td>
</tr>
<tr>
<td>Speed variance</td>
<td>Speed variance of taxi samples on each studied corridor</td>
<td>5.18</td>
<td>98.01</td>
<td>46.12</td>
<td>29.41</td>
</tr>
</tbody>
</table>

For a detailed description of the dataset, please refer to Xie et al. (2013). Listed in Tables 1 and 2 are the 17 intersection-level variables and 6 corridor-level variables, along with brief descriptions and statistics for each variable.

4. Diagnostic tests for spatial dependence

4.1. Moran’s I test

Global Moran’s I statistic proposed by Moran (1948) has been widely used to measure the spatial autocorrelation of observations (Goodchild, 1986; Anselin, 1988b; Getis and Ord, 1992). The Moran’s I statistic is defined as

\[ I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} z_i z_j}{S_0 \sum_{i=1}^{n} z_i^2} \]  

(1)

where \( z_i \) and \( z_j \) are deviations of observations \( i \) and \( j \) from the mean, \( W_{ij} \) represents the spatial weight between observations \( i \) and \( j \), \( n \) is the total number of observations, and \( S_0 \) is the aggregation of spatial weights \( \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} \). The z-score of Moran’s I can be computed by the following equation:

\[ z_I = \frac{I - E[I]}{\sqrt{V[I]}} \]  

(2)

where \( E[I] \) and \( V[I] \) are the expectation and variance of \( I \)

\[ E[I] = -\frac{1}{n-1} \]  

(3)

\[ V[I] = E[I^2] - E[I]^2 \]  

(4)

A statistically significant positive z-score suggests that the distribution of the observations are spatially clustered, whereas a negative z-score implies that the observations tend to be more dissimilar than one might expect. A z-score close
to zero indicates that observations are likely to be randomly and independently distributed in space (Goodchild, 1986). By assuming z-score is from a standard normal distribution, its associated p-value can be obtained. If the z-score follows a skewed distribution instead of standard normal distribution, the p-value obtained can be unreliable. Anselin et al. (2006) suggested a more robust metric called pseudo p-value based on a random permutation test. The pseudo p-value is computed as \( \frac{M+1}{S+1} \), where \( M \) is the number of instances with Moran’s I equal to or greater than that of the observed data and \( S \) is the total number of permutations.

The spatial statistics tool in ArcGIS was used to test whether or not intersection crashes are spatially correlated (Mitchell, 2005). GeoDa was used to perform 9999 permutations to compute the pseudo p-value (Anselin, 2003). The spatial weight \( w_{ij} \) was defined by the inverse distance between intersections \( i \) and \( j \) if they were within the threshold distance. To perform the test, the minimum threshold distance should be long enough to ensure all intersections would have at least one neighbor (Mitchell, 2005). The minimum threshold distance (909 m) and additional longer threshold distances (1000 m, 1500 m and 2000 m) were used to test the spatial autocorrelation of crashes.

The summary of Moran’s I statistics is reported in Table 3. According to Table 3, all the z-scores are significantly positive at the confidence level of 95% (p-values < 0.05). Moreover, the small pseudo p-values suggest highly significant Moran’s I statistic. These results confirm the clustering characteristics of crashes at adjacent intersections. Ignoring the spatial correlation of crash data would reduce the precision of the estimates.

### 4.2. Lagrange Multiplier test

Spatial dependence can be caused by different kinds of spatial correlation effects (i.e., spatial error and spatial lag). Lagrange Multiplier (LM) test provides a good basis for the indication of selecting an appropriate spatial regression models (Anselin and Florax, 1995). The LM test was used by Burridge (1980) to diagnose spatial error dependence and later used for spatial lag dependence by Anselin (1988a). An adjusted LM test later proposed by Bera and Yoon (1993) has good finite sample properties and is proved to be more suitable for the identification of the source of dependence than the original LM test (Anselin et al., 1996). The results from the adjusted LM test are robust to the presence of a nuisance parameter representing spatial autocorrelation (Anselin et al., 1996).

GeoDa was used to perform LM and adjusted LM tests (Anselin, 2003). All estimations were conducted using the spatial weight matrix generated based on the minimum threshold distance. The summary of LM statistics is presented in Table 4. For both the spatial lag and spatial error specifications, their LM statistics are found to be highly significant with their p-values < 0.0001. When both LM statistics reject the null hypothesis, the adjusted LM statistics should be considered (Anselin, 2003). The adjusted LM statistic for the spatial error specification is found to be significant (p-value = 0.0366) but not for the spatial lag specification (p-value = 0.5474). The results suggest that a spatial error model should be developed to account for the spatial dependence of crash data.

### 5. Statistical modeling methodology

#### 5.1. Model specification

Let \( Y_{ij} \) denote the crash frequency at \( i \)th intersection on \( j \)th corridor \((i=1,...,n_i, \ n_i \text{ is the total number of intersections on } j\text{th corridor}; \ j=1,...,J, J \text{ is the total number of corridors}) \) in a given time period. Assuming \( Y_{ij} \) follows Poisson distribution with
the mean $\lambda_{ij}$, the probability of observing $y_{ij}$ crashes in the intersection can be given by

$$P(Y_{ij} = y_{ij}) = \frac{e^{-\lambda_{ij}y_{ij}}}{y_{ij}!}$$

To specify the Poisson parameter $\lambda_{ij}$, site-specific explanatory variables $X_{p ij}$ ($p = 1, \ldots, P$, $P$ is the total number of explanatory variables) are incorporated into the model

$$\ln(\lambda_{ij}) = \beta_0 + \sum_{p=1}^{P} \beta_{p}X_{p ij}$$

where $\beta_0$ and $\beta_p$ ($p = 1, \ldots, P$) are the regression coefficients to be estimated. Eqs. (5) and (6) constitute the Poisson model which serves as a basis for modeling crash frequency. In this section, five types of crash frequency models derived from the Poisson model are proposed to identify the most appropriate one which can accommodate the specific data structure used in this study.

Model 1. Negative binomial (NB) model
To address the over-dispersion issue, the negative binomial model is obtained by introducing an error term $\epsilon_{ij}$ into Eq. (6)

$$\ln(\lambda_{ij}) = \beta_0 + \sum_{p=1}^{P} \beta_{p}X_{p ij} + \epsilon_{ij}$$

where $\exp(\epsilon_{ij})$ is assumed to be gamma-distributed with mean 1 and variance $\alpha$.

Model 2. Random effect negative binomial (RENB) model
Random effect models have the ability to deal with the hierarchy of data. To account for unobserved heterogeneity across corridors, a corridor-specific random effect term $\nu_j$ is incorporated into Eq. (7)

$$\ln(\lambda_{ij}) = \beta_0 + \sum_{p=1}^{P} \beta_{p}X_{p ij} + \epsilon_{ij} + \nu_j$$

where $\nu_j$ is a normally distributed term with mean 0 and variance $\sigma_c^2$. The variance component $\sigma_c^2$ represents the variation across corridors.

Model 3. Random parameter (RPNB) model
Random parameter models are capable of allowing some or all estimated parameters to vary across corridors. Random parameter models are equivalent to random effect models, if only the constant term is allowed to vary. The specification of the random parameter models is given in Eqs. (9) and (10) (Anastasopoulos and Mannering, 2009).

$$\ln(\lambda_{ij}) = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj}X_{p ij} + \epsilon_{ij} + \nu_j$$

$$\beta_{pj} = \beta_{p} + \kappa_j$$

where $\kappa_j$ is a normally distributed term with mean 0 and variance $\sigma_c^2$.

Model 4. Hierarchical negative binomial (HNB) model
As mentioned above, the data structure used in this study consists of a two-level hierarchy, namely the intersection level and the corridor level. In the hierarchical model proposed, parameters at the intersection level are expressed by probability models at the corridor level. The HNB model, which was also used in our previous study (Xie et al., 2013), can be specified as follows:

Intersection-level model:

$$\ln(\lambda_{ij}) = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj}X_{p ij} + \epsilon_{ij} + \nu_j$$

Corridor-level model:

$$\beta_{0j} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q}Z_{q j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\vdots$$

$$\beta_{pj} = \gamma_{p0}$$

where $Z_{q j}$ are the corridor-level variables ($q = 1, \ldots, Q$, $Q$ is the total number of corridor-level variables), $\gamma_{00}, \gamma_{10}, \gamma_{20}, \ldots, \gamma_{p0}$ and $\gamma_{0q}$ are the regression coefficients to be estimated. Like in Eq. (8), $\nu_j$ is the random effect at the corridor level where $\nu_j \sim N(0, \sigma_c^2)$. 
Model 5. Hierarchical conditional autoregressive (HCAR) model

The Moran’s I statistic confirmed the clustering characteristics of crashes in the neighboring intersections and results of the Lagrange Multiplier test indicated the spatial dependence should be captured using spatial error specification. The proposed hierarchical conditional autoregressive (HCAR) model cannot only account for the heterogeneity across different corridors but also can capture the spatial dependence among different intersections in the error term. The HCAR model assumes that within a corridor, intersections are spatially correlated and the correlation of intersections close to each other is expected to be greater than those far apart. We derive the HCAR model from the HNB model, with Eq. (11) modified as

\[
\ln(\lambda_{ij}) = \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} X_{pj} + \epsilon_{ij} + \phi_{ij}
\]

In Eq. (13), a conditional autoregressive (CAR) effect term \(\phi_{ij}\) was introduced to capture the spatial correlation among intersections located along the same corridor. The joint probability distribution of \(\phi\) under the Gaussian CAR model is as follows (Ghosh et al., 1999; Miaou and Song, 2005):

\[
p(\phi_{1j}, \phi_{2j}, \ldots, \phi_{nj}) \propto \exp \left[-\frac{1}{2\sigma_{\phi}^2} \sum_{i=j}^{n} w_{ij} (\phi_{ij} - \rho \phi_{ij})^2 \right]
\]

where \(\rho\) represents the direction and magnitude of the spatial autocorrelation, and \(\sigma_{\phi}^2\) is a parameter controlling the variance of the Gaussian CAR distribution. A proximity matrix \(W\) was used to describe the spatial relationship, and the correlation \(w_{ij}\) between intersections \(ij\) and \(i'j'\) was defined as the inverse function of their distance \(d_{ij}\):

\[
w_{ij} = \begin{cases} \frac{1}{\sigma_{w}} & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases}
\]

When the corridor index \(j \neq j'\), \(w_{ij} = 0\) represents that intersections located along different corridors are independent from each other. Based on Eq. (15), the full conditional distribution of \(\phi_{ij}\) can be obtained as follows (Miaou and Song, 2005):

\[
p(\phi_{ij}|\phi_{-ij}) \propto \exp \left[ -\frac{w_{ij}}{2\sigma_{\phi}^2} (\phi_{ij} - \rho \sum_{i' \neq i} w_{ij} \phi_{ij} + \phi_{ij})^2 \right]
\]

where \(\phi_{-ij}\) is the set of \(\phi_{ij}\) for any \(ij \neq i'j\) and \(w_{ij} = \sum_{i' \neq i} w_{ij} \phi_{ij}\) representing the sum of weights on \(j\)th corridor. An intrinsic version of the CAR model proposed by Besag et al. (1991) is adopted in this study. The parameter \(\rho\) is set to be 1 and the conditional distribution of \(\phi_{ij}\) given \(\phi_{-ij}\) turns out to be a normal distribution with mean \(\sum_{i' \neq ij} w_{ij} \phi_{ij}/w_{ij} + \phi_{ij}\) and variance \(\sigma_{\phi}^2/w_{ij}\):

\[
\phi_{ij}|\phi_{-ij} \sim N \left( \frac{\sum_{i' \neq ij} w_{ij} \phi_{ij}}{w_{ij}}, \frac{\sigma_{\phi}^2}{w_{ij}} \right)
\]

In the Bayesian procedure, Eq. (17) is used to assign CAR priors to \(\phi_{ij}\). The Eqs. (5), (12), (13) and (17) represent the specification of the HCAR model.

5.2. Bayesian procedure

All model parameters were estimated using the Bayesian method that combines prior distributions with a likelihood function obtained from the observed data to estimate posterior distributions. Several researchers have shown the advantages of Bayesian methods in accommodating complex model structures when compared to classical statistical methods (Mitra and Washington, 2007; Persaud et al., 2010; Xie et al., 2013; Yang et al., 2013). Bayesian inference is usually implemented by a Markov Chain Monte Carlo (MCMC) algorithm (Gilks et al., 1998). MCMC is a classic method that utilizes independent and identically distributed simulations of a random process to approximate the desired distribution. The primary technique of MCMC is Gibbs sampling (Geman and Geman, 1984), each iteration of which draws a new value for each unobserved stochastic node from its full conditional distribution given the current values of all the other quantities in the model (Lunn et al., 2000). The WinBUGS statistical software package was used to provide a computing approach for the calibration of Bayesian models using Gibbs sampling (Spiegelhalter et al., 2002).

As mentioned above, the priors of the CAR effect term \(\phi_{ij}\) are generated from Eq. (17). Without credible prior information for other parameters, uninformative priors were assumed. Uninformative priors express vague and general information about parameters. All regression coefficients were assumed with the Normal distribution \((0, 10^6)\); the variance of the randomly distributed terms \(\sigma_i^2\) and the variance of the Gaussian CAR distribution \(\sigma_{\phi}^2\) were assumed to follow the Inverse-Gamma distribution \((10^{-1}, 10^{-3})\). The logarithm of the variance of the NB error term \((\ln \alpha)\) was assumed with Normal distribution \((0, 10^2)\). The Brooks–Gelman–Rubin (BGR) diagnostic proposed by Brooks and Gelman (1998) was used to assess the convergence of multiple chains. Convergence was assumed to occur when the BGR statistic is less than 1.2. Considering
convergence and time of updating, two MCMC chains of 150,000 iterations were run, and the first 50,000 samples were discarded as burn-in.

5.3. Model assessment

5.3.1. Deviance information criterion

The deviance information criterion (DIC) is widely used as a Bayesian measure of model fitting and complexity (Spiegelhalter et al., 2002). Specifically, DIC is calculated as follows:

$$\text{DIC} = D(\hat{\theta}) + p_D$$

where $D(\hat{\theta})$ is the Bayesian deviance of the estimated parameter $\theta$. $D(\hat{\theta})$ denotes the posterior mean of $D(\theta)$ and can be used to indicate how well the model fits the data. $p_D$ defines the effective number of parameters and can be taken as a measure of model complexity. A difference in DIC that is larger than 5 suggests that the model with a smaller DIC should be favored.

5.3.2. Cross-validation test

To compare the predictive performance of different models, the cross-validation procedure proposed by Gelfand (1996) was used. A recent study by Yang et al. (2013) showed the practical use of such a procedure in traffic safety modeling. Unlike the original procedure that leaves out a single observation at each step, the modified procedure was designed to leave out all observations for each corridor at the same time at each step. A data set of size $n$ from specific models, CPO $Y$, where $D$ predictive density for observation $\theta$ is calculated by using the following equation:

$$\text{CPO}_{ij} = f(Y_{ij,\text{obs}} | Y_{-j,\text{obs}}, M_a)$$

where $Y_{-j,\text{obs}}$ means that all observations other than that of $j$th corridor are used to estimate a model $M_a$, and $\text{CPO}_{ij}$ is the predictive density for observation $Y_{ij,\text{obs}}$ for $i$th intersection at $j$th corridor by applying the model $M_a$ to the remaining training data set, the conditional predictive ordinate (CPO) of each intersection in the test data set was calculated by using the following equation:

$$\text{CPO}_{ij} = f(Y_{ij,\text{obs}} | Y_{-j,\text{obs}}, M_a)$$

where $Y_{ij,\text{obs}}$ supports model $M_a$, whereas $\text{CPO}_{ij}$ supports model $M_b$. The model supported by more observations is preferred.

The CR for all the observations is the cumulative product of the $\text{CR}_{ij}$ of all associated intersections. The cumulative CR (CCR) can be calculated by Eq. (21)

$$\text{CCR} = \prod_{i=1}^{n} \prod_{j=1}^{J} \text{CR}_{ij}$$

$\text{CCR} > 1$ indicates that overall the observations support model $M_a$, whereas $\text{CCR} < 1$ indicates the observations support model $M_b$.

6. Results and discussion

The NB, RENB, RPNB, HNB and HCAR models for intersection crash frequency were developed based on both the intersection and the corridor related variables listed in Tables 1 and 2. To conduct comparisons, all the explanatory variables included in the five models were kept the same. It should be noted that the HNB and HCAR models included corridor related variables into the corridor-level models, while the NB, RENB and RPNB treated corridor related variables in the same way as did the intersection related variables. In the RPNB model, if the estimated standard deviation of a parameter distribution was not statistically different from 0, the parameter was fixed across the corridors. Consequently, the intercept and the parameter of intersection ADT were allowed to vary randomly from one corridor to another in the RPNB model. Bayesian posterior estimations of the five models were presented in Table 5. The 95% Bayesian Credible Interval (95% BCI) was used to examine the significance of estimations. Estimations can be regarded as significant at the 95% level if the BCIs do not cover 0 and vice versa (Gelman, 2004). The ratio of turning lanes in the HCAR model was the only variable not found to be significant.

6.1. Discussion on explanatory variables

According to the variable coefficients of the HCAR model presented in Table 5, the intersections associated with higher crash frequencies were found to be under elevated roads, in close proximity to each other, with larger total number of phases, with higher ratios of turning lanes, with greater ADT, and along corridors with higher mean speeds. The proximity of intersections indicates the road network density. Box-plots of the intersection crash rate for the 22 corridors, ordered by
Table 5
Posterior summary of Bayesian model fitting.

<table>
<thead>
<tr>
<th>Variable</th>
<th>NB</th>
<th>RENB</th>
<th>RPNB</th>
<th>HNB</th>
<th>HCAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>95% BCI</td>
<td>Mean</td>
<td>95% BCI</td>
<td>Mean</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.163</td>
<td>(0.668, 1.750)</td>
<td>1.232</td>
<td>(0.721, 1.779)</td>
<td>1.527</td>
</tr>
<tr>
<td>Intersection level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under elevated roads or not</td>
<td>0.492</td>
<td>(0.155, 0.848)</td>
<td>0.488</td>
<td>(0.051, 0.981)</td>
<td>0.470</td>
</tr>
<tr>
<td>Proximity of intersections</td>
<td>-0.719</td>
<td>(-1.531, -0.049)</td>
<td>-0.722</td>
<td>(-1.478, -0.013)</td>
<td>-0.739</td>
</tr>
<tr>
<td>Ratio of turning lanes</td>
<td>1.000</td>
<td>(0.292, 1.749)</td>
<td>0.906</td>
<td>(0.188, 1.595)</td>
<td>0.968</td>
</tr>
<tr>
<td>Total number of pases</td>
<td>0.368</td>
<td>(0.061, 0.670)</td>
<td>0.386</td>
<td>(0.107, 0.685)</td>
<td>0.389</td>
</tr>
<tr>
<td>Intersection ADT</td>
<td>0.054</td>
<td>(0.003, 0.113)</td>
<td>0.051</td>
<td>(0.005, 0.112)</td>
<td>0.060</td>
</tr>
<tr>
<td>(SD of parameter distribution)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.040</td>
</tr>
<tr>
<td>Corridor level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean speed</td>
<td>0.090</td>
<td>(0.067, 0.110)</td>
<td>0.087</td>
<td>(0.061, 0.112)</td>
<td>0.072</td>
</tr>
<tr>
<td>Dispersion $\alpha$</td>
<td>0.743</td>
<td>(0.598, 0.915)</td>
<td>0.663</td>
<td>(0.524, 0.837)</td>
<td>0.670</td>
</tr>
<tr>
<td>Random effect $\sigma^2$</td>
<td>–</td>
<td>–</td>
<td>0.098</td>
<td>(0.018, 0.253)</td>
<td>0.071</td>
</tr>
<tr>
<td>CAR effect var($\phi_{ij}$)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td>1825</td>
<td></td>
<td>1817</td>
<td></td>
<td>1812</td>
</tr>
<tr>
<td>CCR$^a$</td>
<td>2.662</td>
<td>1.719</td>
<td>1.459</td>
<td>1.225</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ CCR is calculated using Eqs. (20) and (21) where the HCAR model is used as $M_b$ while the other models are used as $M_a$. The model with lower CCR has better predictive performance.
proximity of intersections, are presented in Fig. 1. In the box-plot, crash rate values more than three interquartile ranges from the end of a box are denoted with asterisks (*), and values more than 1.5 interquartile ranges but less than 3 interquartile ranges are shown in circles (o). Overall, increases in the proximity of intersections are associated with decreases in the crash rates. Marginal effects represent an expected change in crash frequency given a unit change in the explanatory variables (Mitra and Washington, 2012). The estimated marginal effect of the proximity of intersections was \(-7.69\) for the HCAR model with all other variables set at their means. This indicates that a one-kilometer increase in the distance between two intersections is expected to reduce the crash frequency by 7.69, annually. This appears to be counterintuitive for saturated arterials, where the traffic slows down due to the queue spillback between close intersections. When two intersections are separated by a larger distance, cars speed up and cause more crashes. The possible reason for this finding is that the shorter distance between intersections limits the gaps to implement safe lane changes, and ultimately results in more traffic conflicts. This finding of the impact of the distances between two intersections is important for designing safe road networks in urban areas.

According to the marginal effect of mean speed, one unit increase of mean speed is associated with an expected increase in crash frequency by 1.00, annually. Somewhat unexpectedly, we did not find speed variance to be significantly related to the frequency of crashes. The possible reason for this finding is that the speed features extracted from the taxi FCD data could be different from those of other vehicles due to the taxis’ stop-and-go behavior as they pick up or drop off clients. This difference may introduce the issue of selectivity bias which can cause the inaccurate statistical inferences on speed features. The results regarding the speed features should thus be used with caution. Additionally, it should be noted that some factors affecting crash frequency such as bus traffic, pedestrian volume, median width and pavement conditions were not available and thus were unable to be included in the models. Due to these limitations of data, the number of explanatory variables that were found to be statistically significant is relatively small. The omission of other contributing factors may introduce unobserved heterogeneity and has risk of leading to biased estimations of explanatory variables in the models (Mannering and Bhat, 2014).

### 6.2. Discussion on model structure and performance

The estimated dispersion values (0.743 in the NB model, 0.663 in the RENB model, 0.670 in the RPNB model, 0.656 in the HNB model and 0.668 in the HCAR model) validate the assumption of over-dispersion. The random effect variances (0.098 in the RENB model, 0.071 in the RPNB model, 0.098 in the HNB model and 0.155 in the HCAR model) provide strong evidence for the presence of heterogeneity across corridors. The CAR effect variance (0.498 in the HCAR model) is found to be statistically significant and confirms the spatial correlation of intersections along the same corridor. Ignoring the spatial correlation would have resulted in the biased estimations of parameters in the models without the CAR effect term.

Table 5 presents the DIC values for the three Bayesian models. The DIC value of the HCAR model (1795) is 30, 22, 17 and 15 less than that of the NB model (1825), the RENB model (1817), the RPNB model (1812) and the HNB model (1810) respectively. The HCAR model is shown to be superior to the other four models in terms of the overall goodness of fit with the lowest DIC value. By including the random effect term, the RENB model is found to outperform to the NB model. The model performance is further improved in RPNB by allowing the effects of intersection ADT to vary across corridors. The HNB model has a DIC value close to that of the RPNB model and shows substantial improvement compared to the NB and RENB models. It should be noted that the only difference between the RENB and HNB models is that the HNB model includes the variable mean speed into the corridor-level model instead of mixing it with other intersection related variables. The HNB
model shows its advantage over the RENB model by incorporating the mean speed at the level where its actual impact occurs. The HCAR model yields an improved result compared with the HNB model by incorporating the CAR effect term. The DIC value is reduced significantly by 15 (from 1810 for the HNB model to 1795 for the HCAR model) when the spatial correlation among intersections is properly addressed.

A recent study by Geedipally et al. (2014) suggests not using DIC as the sole model selection criterion if the likelihood does not remain the same across all the models under consideration. Therefore, the cross-validation test described in the previous section was conducted to further compare the HCAR model with the other four crash frequency models alternatively. By applying Eqs. (20) and (21), the CR for each intersection and the CCR for all the observations were estimated. The CCRs of the HCAR model versus the NB, RENB, RPNB, HNB are reported in Table 5. Referring that CCRs of the NB model (2.662), the RENB model (1.719), the RPNB model (1.459) and the HNB model (1.225) are all greater than 1, the HCAR model is confirmed to have better predictive performance than any other models. Moreover, according to CRs of the HCAR model versus the NB model shown in Fig. 2a, the HCAR model is preferred in 166 (denoted with green circles) out of the 195 cases, whereas the NB model is supported by only 29 intersection observations (denoted with red triangles). Similarly, as shown in Fig. 2b–d, the HCAR model performs better than the RENB, RPNB and HNB models because 149, 109 and 103 out of 195 CRs respectively are greater than 1.

The aforementioned Moran’s I statistic confirms the necessity of considering spatial dependence of crash data and the results from the Lagrange Multiplier test advocates the needs of developing a model with a spatial error specification. The proposed HCAR model can capture the spatial dependence in the CAR effect term $\phi_{ij}$. As discussed in the model specification, the conditional distribution of $\phi_{ij}$ is assumed to follow a normal distribution with the weighted average of CAR effects of neighboring intersections as its mean. The advantage of including CAR effects over random effects is that the

![Fig. 2. Model comparison based on CPO ratio (CR), (a) HCAR vs NB, (b) HCAR vs RENB, (c) HCAR vs RPNB and (d) HCAR vs HNB.](image-url)
distribution mean can vary from one entity to another and is related to the weight matrix which can be specified flexibly. In the HCAR model, the variation of the crash frequency is accommodated by the combination of dispersion parameter, the corridor-specific random effect term and the CAR effect term. According to the DIC and cross-validation test, the HCAR model performs better than other models in terms of the overall goodness of fit and predictive performance, by taking corridor-level heterogeneity and distance-related spatial dependence into account.

7. Conclusions

As a follow-up to our previous paper (Xie et al., 2013), this study proposed an advanced crash frequency model that can account for the hierarchy and spatial correlation of crash data of signalized intersections in a high-density urban road network. The proposed model can serve as a useful complement to the previous safety analysis methods in terms of the clustering characteristics of crash data in high-density road networks.

Closely spaced signalized intersections in the urban areas of Shanghai were used as a case study. Detailed crash data as well as related traffic and geometric data were collected. The Moran’s I statistic was computed for crash observations and the results confirmed the clustering characteristics of crashes in the neighboring intersections. This finding indicates the necessity of accounting for spatial effects when modeling crash observations. Moreover, we conducted the Lagrange Multiplier tests and found that the spatial dependence should be captured in the model error term. To address the spatial correlations of crashes among different intersections, a hierarchical conditional autoregressive (HCAR) model incorporating both intersection-level and corridor-level variables and CAR effects was developed in the Bayesian framework. To examine the performance of the proposed model, a negative binomial (NB) model, a random effect negative binomial (RENB) model, and a random parameter negative binomial model (RPNB) and a hierarchical negative binomial model were also estimated for comparison purposes.

The significance of the corridor-specific random effects in the RENB, RPNB, HNB and HCAR models confirmed the hierarchy of the crash data. In addition, the significance of the CAR effects in the HCAR model indicated that intersections along the same corridor were spatially correlated. These findings affirm the needs to capture the unique structure of crash data for intersections in high-density networks. Besides, the safety effect of intersection proximity as associated with road network density was investigated. A negative relation was found between the crash frequency and the distance of two adjacent intersections. This finding is important for planning, designing, and managing high-density urban road networks.

The deviance information criterion (DIC) and cross-validation test were used for model comparison. Considering the drawbacks of using DIC as the sole model criterion (Geedipally et al., 2014), we emphasized the application of less frequently used cross-validation test to measure the predictive performance of models. Overall, the proposed HCAR model was found to outperform the NB, RENB, RPNB and HNB models. These results should be attributed to the two advantages of the HCAR model: first, the HCAR model is able to construct multilevel (i.e., intersection-level and corridor-level) model structure and to account for the safety effects of intersection related and corridor related variables at specific levels of models; second, the HCAR model can capture the spatial dependence of intersection crashes by specifying the correlation matrix for intersections along each corridor. However, an advantage of the RPNB model over the HCAR model is that it can account for the unobserved heterogeneity and thus the effects of explanatory variables can be adjusted. The preference of the HCAR and RPNB models can be related to the site specific characteristics of the data used in this study. To further understand the performance of those models, additional studies based on other datasets are suggested.

Despite the performance of the proposed approach, we should mention the limitation of missing other potential explanatory variables due to their unavailability in this study. This may alter the estimations of certain parameters and result in unreliable inferences. To further improve the performance of the HCAR model, additional procedures, for instance, incorporation of random parameters, are suggested to further account for the unobserved heterogeneity. Moreover, the temporal correlations and correlations among different crash types at the same site can be considered simultaneously as well as spatial correlations among different sites.

Acknowledgments

This study was jointly supported by the Center for Urban Science and Progress (CUSP) at New York University, the Urban Mobility and Intelligent Transportation Systems (UrbanMITS) Laboratory, and the Chinese National Science Foundation (No. 51008230). The contents of this paper only reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents of the paper do not necessarily reflect the official views or policies of the agencies.

References


Anselin, L. 2003. GeoDa™ 0.9 user's guide. Urbana 51, 61801.


El-Basyouny, K., Saidy, T. 2009b. Urban arterial accident prediction models with spatial effects. Transportation Research Record 2102, 27–33.


Yang, H., Ozbay, K., Ozturk, O., Yildirimoglu, M., 2013. Modeling work zone crash frequency by quantifying measurement errors in work zone length. Accident Analysis and Prevention 55, 192–201.