Macro-level safety analysis of pedestrian crashes in Shanghai, China

Xuesong Wang a,b,∗, Junguang Yang a, Chris Lee c, Zhuoran Ji a, Shikai You a

a School of Transportation Engineering, Tongji University, Shanghai 201804, China
b The Key Laboratory of Road and Traffic Engineering, Ministry of Education, China
c Department of Civil and Environmental Engineering, University of Windsor, Windsor, Ontario N9B 3P4, Canada

A R T I C L E   I N F O

Article history:
Received 30 November 2015
Received in revised form 17 July 2016
Accepted 21 July 2016

Keywords:
TAZ-level safety analysis
Pedestrian crashes
Spatial weight features
Bayesian Conditional Autoregressive Model
Transportation safety planning

A B S T R A C T

Pedestrian safety has become one of the most important issues in the field of traffic safety. This study aims at investigating the association between pedestrian crash frequency and various predictor variables including roadway, socio-economic, and land-use features. The relationships were modeled using the data from 263 Traffic Analysis Zones (TAZs) within the urban area of Shanghai—the largest city in China. Since spatial correlation exists among the zonal-level data, Bayesian Conditional Autoregressive (CAR) models with seven different spatial weight features (i.e., (a) 0–1 first order, adjacency-based, (b) common boundary-length-based, (c) geometric centroid-distance-based, (d) crash-weighted centroid-distance-based, (e) land use type, adjacency-based, (f) land use intensity, adjacency-based, and (g) geometric centroid-distance-order) were developed to characterize the spatial correlations among TAZs. Model results indicated that the geometric centroid-distance-order spatial weight feature, which was introduced in macro-level safety analysis for the first time, outperformed all the other spatial weight features. Population was used as the surrogate for pedestrian exposure, and had a positive effect on pedestrian crashes. Other significant factors included length of major arterials, length of minor arterials, road density, average intersection spacing, percentage of 3-legged intersections, and area of TAZ. Pedestrian crashes were higher in TAZs with medium land use intensity than in TAZs with low and high land use intensity. Thus, higher priority should be given to TAZs with medium land use intensity to improve pedestrian safety. Overall, these findings can help transportation planners and managers understand the characteristics of pedestrian crashes and improve pedestrian safety.

© 2016 Published by Elsevier Ltd.

1. Introduction

Pedestrians play an important role in travel modes around the world. In Shanghai, the largest city in China, 35.1% of the residents in the urban area walked for their travel needs in 2012, with the average daily walking distance at 0.9 km in the center of Shanghai (Shanghai Urban and Rural Construction and Transportation Development Institute, 2013). Pedestrians are regarded as the most vulnerable road users due to their fragility, slow movement and lack of lighting equipment. Therefore, they have a higher risk of traffic crash involvement than the drivers and passengers of motorized vehicles (Zhang et al., 2014). In particular, pedestrian fatalities accounted for 22% of the traffic crash fatalities worldwide and 25% in China (World Health Organization, 2013). With rapid urbanization and motorization, pedestrian crashes in China have surged in recent years.

In order to discover and identify specific contributing features for pedestrian crash occurrences, crash locations are usually aggregated into spatial units such as segments, intersections, zones, and so forth (Lee et al., 2015b). Micro-level safety analysis focuses on specific roadway entities such as roadway segments, intersections, corridors, etc., and macro-level safety analysis focuses on zonal-level traffic crashes at various levels of area-aggregation such as census tract (Cottrill and Thakuriah, 2010; Ukkusuri et al., 2011; Abdel-Aty et al., 2013), Traffic Analysis Zone (TAZ) (Siddiqui et al., 2012; Abdel-Aty et al., 2013; Wang et al., 2013; Lee et al., 2015b), and block group (Noland et al., 2013). Micro-level analysis is effective in identifying and solving safety problems at one specific location, but it becomes more difficult to capture spatial trends and problems in a larger area. Compared to micro-level safety analysis, macro-level safety analysis can more effectively identify safety problems in a larger area, and thus is more useful in helping establish long-term planning policy to improve safety. Researchers have performed many safety analyses using zone-based data and

∗ Corresponding author at: School of Transportation Engineering, Tongji University, Shanghai 201804, China.
E-mail address: wangxs@tongji.edu.cn (X. Wang).

http://dx.doi.org/10.1016/j.aap.2016.07.028
0001-4575/© 2016 Published by Elsevier Ltd.
identified relationships between pedestrian crash frequency and related features (Cottril and Thakuriah, 2010; Ukkusuri et al., 2011; Abdel-Aty et al., 2013), and it has been shown that spatial correlation is an important factor when establishing the statistical model for macro-level safety analysis. Bayesian spatial analysis with CAR (Conditional Autoregressive) priors, which can effectively accommodate the spatial autocorrelations of study units, has therefore become prevalent in current research (Siddiqui et al., 2012; Noland et al., 2013; Wang and Kockelman, 2013).

Shanghai is characterized by the highest urbanization rate (88.86%) in China in 2010 (China International Urbanization Development Strategy Research Committee, 2012), high density of population (23.0 million in 2010 (Bureau of Statistic of Shanghai, 2011)), large motor vehicle ownership (3.1 million in 2010 (Bureau of Statistic of Shanghai, 2011)), high density of roadway network (8.0 km/km²) and typical mixed traffic flow. Pedestrian crashes were not equally distributed throughout the analysis area – they were spatially concentrated in the center of Shanghai. In particular, more than two thirds of pedestrian crash hot zones (ranking in the top 15% of total crash counts) were in the 54 TAZs within the Inner Ring in the Puxi area (see Fig. 1), which is more densely developed than the Pudong area. These 54 TAZs accounted for just 12.3% of the analysis area, but 38.2% of pedestrian crashes overall. The 54 TAZs consisted of three majority land use types: residential and commercial (31.5%); commercial, residential and official (27.8%); and exhibition, financial and tourism (27.8%). Commercial and residential areas were the key zones for pedestrian aggregation, which is associated with pedestrian crash occurrence (Kim et al., 2006; Loukaitou-Sideris et al., 2007; Ukkusuri et al., 2011). Traffic safety was set as the primary objective for the next decade in the Shanghai Transportation Development White Paper (Shanghai Municipal People’s Government, 2014). As the White Paper underscored the need of discovering risk factors and improving pedestrian safety, higher priority will be given to the safety of walking and biking in short-distance travel to protect these convenient and comfortable travel modes.

In summary, the main purposes of this study are to: (a) develop macro-level pedestrian crash models to investigate the relationship between pedestrian crashes and various roadway, socio-economic, and land use features in Shanghai; (b) compare the seven spatial weight features as CAR priors in Bayesian framework and analyze the spatial correlation among the analysis areas; (c) identify the features which have significant influence on pedestrian crashes; and (d) provide suggestions and references for transportation planners and managers on improving pedestrian safety in the traffic system of China.

2. Literature review

2.1. Macro-level predictors for pedestrian crashes

In most previous studies of pedestrian crashes, macro-level crash prediction models were developed to relate the crashes to a variety of explanatory zonal features (including roadway, traffic, socio-economic, and land use characteristics). Research on these features will be introduced below separately.

In examining the relationship between pedestrian crashes and roadway characteristics, traffic engineers have found that total number of intersections is linked to the number of pedestrian crashes, as intersections with higher numbers of pedestrians crossing have led to higher probabilities of vehicle-pedestrian crashes (Siddiqui et al., 2012; Abdel-Aty et al., 2013). A large percentage of these crashes have occurred on major urban arterial roads (Miles-Doan and Thompson, 1999). However, as Miles-Doan and Thompson (1999) have argued, transportation planners have ignored the needs of pedestrians along arterials. Length of minor roads have been shown to have significant effects on crashes as well (Quddus, 2008), and studies have also explored how speed limit affects pedestrian crashes (Kim et al., 2010; Abdel-Aty et al., 2013; Lee et al., 2015a,b). A recent TAZ-level analysis found that a greater number of pedestrian crashes occurred on roadway segments with 25 and 35 mph PSL (Posted Speed Limit) as compared to PSLs higher than 35 mph (Abdel-Aty et al., 2013). Additionally, a greater proportion of roadway segments with PSLs higher than or equal to 55 mph was found to significantly decrease pedestrian crashes. In other words, pedestrian crashes are more likely to occur in areas with many low-speed local roads (Lee et al., 2015a,b).

Significant correlations have been found between pedestrian crashes and traffic characteristics (Loukaitou-Sideris et al., 2007; Wier et al., 2009; Abdel-Aty et al., 2013; Lee et al., 2015b). Vehicle Miles Traveled (VMT) has been positively associated with pedestrian crashes in TAZ-level analysis (Abdel-Aty et al., 2013; Lee et al., 2015b). Researchers also found that the Average Annual Daily Traffic (AADT) is a statistically significant predictor of pedestrian crashes (Loukaitou-Sideris et al., 2007; Wier et al., 2009). Kim et al. (2010) investigated the safety effects of different vehicle types on pedestrian crashes, and found that trucks tended to increase the likelihood of pedestrian crash occurrence by 370%.

Another set of studies has examined the relationship between pedestrian crashes and socio-economic characteristics. The most frequently analyzed features, such as population (Kim et al., 2006; Ukkusuri et al., 2011; Lee et al., 2015b), population density (Loukaitou-Sideris et al., 2007; Siddiqui et al., 2012), employment population (Wier et al., 2009; Siddiqui et al., 2012), employment density (Loukaitou-Sideris et al., 2007) and proportion of unemployed population (LaScala et al., 2000), all had positive effects on pedestrian crashes in previous studies. The number of vulnerable road users, including children and older people, has been found to be significantly correlated with pedestrian crashes. A higher number of pedestrian crashes have been associated with a higher density of children (Abdel-Aty et al., 2013) and a lower percentage of resident population aged 65 and older possibly due to the decrease of physical activity (Wier et al., 2009). As for economic features, median household income had negative effect on pedestrian crashes (Siddiqui et al., 2012; Lee et al., 2013), while percentage of residents living below the poverty line (Wier et al., 2009; Lee et al., 2015a) and proportion of household without vehicles (Noland et al., 2013; Lee et al., 2015b) all had positive effects on pedestrian crashes. Similarly, areas with a greater proportion of uneducated residents show a positive effect on pedestrian crashes (Ukkusuri et al., 2011), while a greater proportion of high school graduates or higher shows a negative effect (LaScala et al., 2000). The TAZ-level pedestrian exposure variable is also critical in a pedestrian crash prediction model. Pedestrian exposure is an abstract concept that reflects the opportunity for a potentially harmful pedestrian-vehicle interaction to occur. However, pedestrian exposure is very difficult to measure directly since this would involve tracking the movements of all people at all times (Greene-Roesel et al., 2007). Therefore, a good alternative exposure measure is sought as surrogate. Greene-Roesel et al. (2007) summarized five common metrics used to describe pedestrian exposure including population, number of pedestrians, trips, distance traveled, and time spent traveling. In this study, population was used as a surrogate for pedestrian exposure.

Land use features could influence human activity and potentially lead to pedestrian crashes. Land use types generating pedestrian traffic (such as parks, commercial and retail facilities, high-density housing, schools) have been associated with larger numbers of pedestrian crashes (Kim et al., 2006; Wedagama et al., 2006; Loukaitou-Sideris et al., 2007; Ukkusuri et al., 2011). Ng et al. (2002) investigated 27 types of land use features in Hong Kong and found...
that the number of cinema seats, commercial areas, flatted factory floor areas, and market stalls have positive effects on the number of pedestrian-related crashes. Pedestrian crashes were fewer in areas with vacant, industrial, office, or greenbelt land uses (Ng et al., 2002; Loukaitou-Sideris et al., 2007). However, the findings have been often contradictory. High-density pedestrian activity areas can cause traffic congestion, reduce traffic speed, and therefore reduce the incidence of the pedestrian-vehicle crashes (Graham and Glaister, 2003). Wang and Kockelman (2013) introduced land use entropy as the balance parameter. They found that balanced land development had a mild positive impact on reducing severe crashes and could serve as a countermeasure for curbing pedestrian fatalities.

2.2. Macro-level pedestrian crash models

Researchers have traditionally used the Poisson, Negative Binomial (NB), and Poisson-lognormal models in pedestrian safety research. However, these models are “disaggregate” in nature as they ignore spatial autocorrelation and aggregation in crash data. This leads to bias in model estimation and statistical inference (MacNab, 2004). More recently, many safety researchers have started to account for spatial dependency in macro-level analyses. Use of Bayesian spatial analysis with CAR priors, which can effectively accommodate the spatial autocorrelations of study units, has become prevalent in current research (Siddiqui et al., 2012; Noland et al., 2013; Wang et al., 2013; Wang and Kockelman, 2013). Siddiqui et al. (2012) examined pedestrian and bicycle crashes occurring in two counties of Florida. Results indicated that in both pedestrian and bicycle crashes, the Bayesian models, which accounted for spatial correlation among TAZs, performed better than the models that did not.

Spatial correlation cannot be ignored when establishing the statistical model for macro-level safety analysis. As Dong et al. (2015) pointed out, the spatial dimensions of socio-demographic, economic or regional activities may truly represent an important aspect in model development; accounting for spatial correlation may help to improve the accuracy and robustness of crash prediction and avoid underestimation of standard errors for model parameters. To characterize the correlation among spatial units, various spatial weight features have been used in the modeling process. For instance, a majority of previous studies have employed the 0–1 first order, adjacency-based spatial weight feature (Siddiqui et al., 2012; Lee et al., 2015a,b). A further study by Dong et al. (2014) compared four spatial weight features in total crash modeling: (a) 0–1 first order, adjacency-based, (b) common boundary-length-based, (c) geometric centroid-distance-based, (d) crash-weighted centroid-distance-based. Results indicated that the

![Figure 1. TAZ boundaries and pedestrian crashes per TAZ in Shanghai.](image)
common boundary-length-based spatial weight feature performed best, followed by the crash-weighted centroid-distance-based spatial weight feature. However, a clear guideline on which spatial weight feature should be chosen in pedestrian crash modeling has not been presented. Thus, it is important to examine which spatial weight feature in Bayesian spatial analysis can better reflect spatial correlation in pedestrian modeling and improve the safety estimate results for the zones at the macroscopic level.

3. Data preparation

3.1. Data collection

This study analyzed 263 TAZs within the 645 km² downtown area of Shanghai’s Outer Ring. Fig. 1 shows the TAZ boundaries, the Outer Ring, the Inner Ring and the number of pedestrian crashes per TAZ in Shanghai. The Huangpu River divides Shanghai into Puxi and Pudong areas. ArcGIS 10.0 (ESRI) was used to integrate pedestrian crash data, roadway data, socio-economic data, and land use data at the TAZ level.

Crashes involving pedestrians in 2009 were obtained from the Shanghai Traffic Police Corps. Fig. 1 shows the pedestrian crash number per TAZ. Minor injury crashes accounted for 88.7% of all crashes, severe injury crashes accounted for 1.5% of all crashes, and fatal crashes accounted for 1.7% of all crashes. This data was imported into ArcGIS and geocoded to the road segments and intersections for use in this analysis. For crashes occurring on segments or at intersections within a TAZ, they were assigned to the corresponding TAZ. For crashes occurring on segments or at intersections on TAZ boundaries, there were two situations: (1) if the crash location could be confirmed, the crash was assigned to the exact TAZ; (2) if the crash location could not be confirmed, the crash would be shared by adjacent TAZs by pro-rating the crash with a weight equal to the reciprocal of the number of adjacent TAZs, as described by Sun and Lovegrove (2010). For example, if the crash was on the boundary of $N(N \geq 2)$ adjacent TAZs and the crash location could not be confirmed, then the crash was pro-rated by $1/N$ for these $N$ adjacent TAZs.

Roadway features were calculated for each TAZ, including total numbers of 3-legged, 4-legged, and multi-legged intersections, total number of intersections, percentage of 3-legged and 4-legged intersections, average intersection spacing, road density, and total length of major and minor arterials. The area for each TAZ was calculated using ArcGIS, and the Spatial Join function in ArcGIS was used to divide, associate, and aggregate roadway attributes among the corresponding TAZs.

2009 population data of the 263 TAZs were obtained from the Shanghai Urban Planning and Design Research Institute, and land use type and intensity were obtained from the Shanghai Construction and Management Commission. The twelve land use types in China’s National Standard Current Land Use Classification (GB/T 21010-2007) defined by the Ministry of Land and Resources of China (2007) were aggregated to eight types based on the land use condition in Shanghai. Land use intensity contains six grades (1–6, in which 6 represents the highest intensity), which were determined based on the floor area ratio and the average building density. Then, land use intensity was classified into three categories as follows: low intensity (if grade = 1–2), medium intensity (if grade = 3), and high intensity (if grade = 4–6).

Table 1 shows the descriptive statistics of the factors included in the model. Possible correlations among all explanatory variables included in the final model were checked using the Pearson correlation coefficients. The strongest correlation was found between “Population” and “Length of minor arterials” ($r = 0.406$). Thus, none of the variables in the final model have evident correlations.

3.2. Moran’s I evaluation

Moran’s I is a spatial autocorrelation index used to reflect whether observed crashes are spatially correlated among adjacent zones, and the index ranges from $-1$ to $+1$. A positive value of Moran’s I indicates a positive spatial correlation, or clustering, within the study area. Moran’s I index can be converted to a Z-score for the statistical test. In the test, values greater than 1.96 or smaller than $-1.96$ indicate that spatial autocorrelation in the region is statistically significant at a 95% confidence level.

Since the Moran’s I was 0.616 and the Z-score was 2.014 ($>1.96$) for this study, spatial autocorrelation was significant. Bayesian model without CAR prior was developed for comparison. Modeling results showed that DIC (Deviance Information Criterion) was 1819.29, which was significantly greater than DICs for Bayesian models with CAR priors. This indicates the necessity to take into consideration the spatial correlation among TAZs in the models.

4. Methodology

4.1. Model development

Bayesian spatial models were developed to relate various zone-level risk features to pedestrian crash occurrence, while accounting for possible spatial autocorrelation among analyzed zones. CAR priors account for spatial autocorrelation and reduce the influence of outlying local rates by allowing each spatial area to “borrow strength” from its neighbors (Dong et al., 2014). The models can be efficiently estimated using the freeware WinBUGS package.

Similarly to previous studies (Wang et al., 2012; Siddiqui et al., 2012), $y_i$ represents the number of crashes occurring in TAZ $i$. The NB model assumes that the dependent variable $y_i$ follows the Poisson Gamma distribution as follows:

$$y_i | \phi_i \sim \text{Negbin}(\phi_i, \gamma)$$

where $\phi_i$ is the expected value of $y_i$ and $\gamma$ is the over-dispersion coefficient.

For NB models, the logarithm was used as a function to link the expected value of $y_i$ with independent variables as follows:

$$\log(\phi_i) = X' \beta + \varepsilon_i$$

where $X$ is the covariate matrix, $\beta$ is the vector of regression coefficients, and $\varepsilon_i$ is the random effects term, which follows normal distribution.

The Bayesian CAR models were proposed to capture the spatial dependence among TAZs by incorporating a random effect variable $\psi_i$ into the model to account for the spatial correlation. The model is defined as follows:

$$\log(\phi_i) = X' \beta + \varepsilon_i + \psi_i$$

Proximity matrix $W$ with entry $w_{ij}$ reflects the spatial association of two TAZs. In the Bayesian framework, the conditional distribution of CAR prior is defined as follows:

$$\psi_i | \psi_{-i} \sim N \left( \sum_{j \neq i} w_{ij} \psi_j, \frac{1}{\tau_c w_{ii}} \right)$$

where $\psi_{-i}$ is the collection of all $\psi$ except for $\psi_i$, $\tau_c$ is a precision parameter, $w_{ii}$ is the sum of $w_{ij}$ in the TAZs that are adjacent to TAZ $i$, and the joint prior distribution for $\psi_i$ as follows:

$$\pi(\psi) \propto \exp \left\{ -\frac{\tau_c}{2} \sum_{i \neq j} w_{ij}(\psi_i - \psi_{-i})^2 \right\}$$

where $\pi(\psi)$ is the joint distribution of $\psi$. Eq. (5) denotes that the likelihood function is proportional to the right-hand side of the equation.
Table 1
Descriptive statistics of the factors (N = 263).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area (km²)</td>
<td>0.463</td>
<td>18.055</td>
<td>2.453</td>
<td>2.183</td>
</tr>
<tr>
<td>Number of 3-legged intersections</td>
<td>0</td>
<td>184</td>
<td>20.240</td>
<td>23.560</td>
</tr>
<tr>
<td>Number of 4-legged intersections</td>
<td>4</td>
<td>124</td>
<td>15.570</td>
<td>15.019</td>
</tr>
<tr>
<td>Number of multi-legged intersections</td>
<td>0</td>
<td>9</td>
<td>1.116</td>
<td>0.692</td>
</tr>
<tr>
<td>Total number of intersections</td>
<td>7</td>
<td>244</td>
<td>36.926</td>
<td>35.778</td>
</tr>
<tr>
<td>Percentage of 3-legged intersections (%)</td>
<td>0</td>
<td>90.476</td>
<td>54.812</td>
<td>15.734</td>
</tr>
<tr>
<td>Percentage of 4-legged intersections (%)</td>
<td>8.333</td>
<td>100</td>
<td>42.165</td>
<td>15.620</td>
</tr>
<tr>
<td>Length of major arterials (km)</td>
<td>0.726</td>
<td>4.428</td>
<td>3.600</td>
<td>1.615</td>
</tr>
<tr>
<td>Length of minor arterials (km)</td>
<td>0.942</td>
<td>2.452</td>
<td>1.818</td>
<td>0.769</td>
</tr>
<tr>
<td>Road density (km/km²)</td>
<td>0.473</td>
<td>53.726</td>
<td>6.961</td>
<td>6.926</td>
</tr>
<tr>
<td>Average intersection spacing (km)</td>
<td>0.226</td>
<td>0.521</td>
<td>0.387</td>
<td>0.235</td>
</tr>
<tr>
<td>Population (10⁵)</td>
<td>1.040</td>
<td>132.440</td>
<td>42.690</td>
<td>24.260</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land use types</td>
<td>0: residential land</td>
<td>22.4%</td>
</tr>
<tr>
<td></td>
<td>1: residential and commercial land</td>
<td>28.9%</td>
</tr>
<tr>
<td></td>
<td>2: commercial, residential and official land</td>
<td>11.0%</td>
</tr>
<tr>
<td></td>
<td>3: residential, cultural, educational and commercial land</td>
<td>11.0%</td>
</tr>
<tr>
<td></td>
<td>4: exhibition, financial and tourism land</td>
<td>7.6%</td>
</tr>
<tr>
<td></td>
<td>5: residential and industrial land</td>
<td>9.1%</td>
</tr>
<tr>
<td></td>
<td>6: residential and logistics land</td>
<td>7.2%</td>
</tr>
<tr>
<td></td>
<td>7: residential and ecological land</td>
<td>2.7%</td>
</tr>
<tr>
<td>Land use intensity</td>
<td>0: low land use intensity</td>
<td>22.4%</td>
</tr>
<tr>
<td></td>
<td>1: medium land use intensity</td>
<td>27.4%</td>
</tr>
<tr>
<td></td>
<td>2: high land use intensity</td>
<td>50.2%</td>
</tr>
</tbody>
</table>

Apportionment of variability in error component due to spatial correlation is calculated as follows:

\[ \alpha = \frac{sd(w_i)(sd(w_i)+sd(e))}{\sum w_{ij}w_{ji}} \]  

where sd is the empirical marginal standard deviation function. Estimates of these empirical marginal standard deviations were used to calculate the contribution of spatial correlation in the variability of the random effect models (Dong et al., 2014). The value of \( \alpha \) greater than 50% indicates that the spatial effect on crashes was greater than the random effect. Thus, the higher the value of \( \alpha \), the greater the random effect influenced by spatial autocorrelation.

4.2. Spatial weight features

Spatial weight features reflect the spatial association of analysis units. CAR models employ weight matrix \( W \) to characterize the spatial relationships of the units in analysis (Wang et al., 2012; Siddiqui et al., 2012; Dong et al., 2014, 2015).

\[ W = \begin{bmatrix} w_{11} & \cdots & w_{1J} \\ \vdots & \ddots & \vdots \\ w_{J1} & \cdots & w_{JJ} \end{bmatrix} \]  

It has been demonstrated that land use may have an impact on pedestrian crashes. Land use, along with weather effects and driver population, is a potential covariate that shows variability in space. However, these features have been rarely measured or accounted for in road safety spatial models (Aguero-Valverde and Jovanis, 2008). Considering the interaction of different land use types and intensity between adjacent TAZs, land use type and land use intensity spatial weight features were estimated in this study to reflect their influence.

Effective range, the distance beyond which there is no spatial correlation between two TAZs, is an important factor in the spatial analysis. Aguero-Valverde and Jovanis (2010) studied various spatial weight features (including adjacency-based and distance-order models) in the safety modeling of road segments, and they presented a distance-order model that considered effective range of spatial correlation among road segments. Based on their work, this study introduced a geometric centroid-distance-order spatial weight feature in macro-level safety modeling. In this case, the spatial correlation is defined in terms of distances, i.e., the first order neighbors are those within A km of TAZs' geometric center distance, the second order neighbors are those between A km and B km, and the third order neighbors are between B km and C km (A–C are constant terms). This can be considered as a hybrid adjacency model, where the order is distance-based but the weights are still the inverse of the order (i.e. 1, 1/2, and 1/3), as proposed by Aguero-Valverde and Jovanis (2010).

In this study, seven different spatial weight features were estimated by: (a) 0–1 first order, adjacency-based, (b) common boundary-length-based, (c) geometric centroid-distance-based, (d) crash-weighted centroid-distance-based, (e) land use type, adjacency-based, (f) land use intensity, adjacency-based, and (g) geometric centroid-distance-order. These spatial weight features are explained in detail as follows:

(a.) 0–1 first-order, adjacency-based spatial weight feature (0–1 swf)

The simplest form of the spatial weight feature is the 0–1 swf. Most previous studies have specified CAR priors to account for the spatial correlation with 0–1 swf. If TAZ and TAZ share a common border, \( w_{ij} \) equals 1; otherwise \( w_{ij} \) equals 0 (Wang et al., 2012; Siddiqui et al., 2012).

(b.) Common boundary-length-based spatial weight feature (CBL swf)

Crashes that occur close to TAZ boundaries may be substantially affected by the attributes of the neighboring zones, which is called “boundary effect”. Thus, common boundary length is used as an indicator to reflect the proximity of neighboring zones. If TAZ, and
TAZs share a common border, \( w_{ij} \) equals the length of the common boundary; otherwise it would be 0 (Dong et al., 2014, 2015).

(c.) Geometric centroid-distance-based spatial weight feature (GCD swf)

The spatial weight features (a) and (b) above are limited to adjacent zones only, and ignore possible spatial correlation between non-adjacent zones. To overcome this limitation, a spatial weight feature, based on distances between geometric centroids of TAZs, is formulated. Specifically, the distances between centroids are used as weights to reflect the correlation of spatial units. \( w_{ij} \) equals the geometric centroid distance between TAZ \( i \) and TAZ \( j \) (Dong et al., 2014, 2015).

\[
\begin{align*}
  w_{ij} &= \begin{cases} 
    1, & \text{if } d_{ij} \leq 0.9; \\
    1/2, & \text{if } 0.9 < d_{ij} \leq 1.8; \\
    1/3, & \text{if } 1.8 < d_{ij} \leq 6.5; \\
    0, & \text{if } d_{ij} > 6.5
  \end{cases}
\end{align*}
\] (8)

4.3. Model evaluation

DIC was used as an index to compare the goodness-of-fit of candidate models. DIC offers a Bayesian measure of model fitting and complexity, in which models with smaller DICs are preferred (Speigelhalter et al., 2003). Conventionally, differences of more than 10 might rule out the model with the higher DIC, and differences between 5 and 10 are considered substantial (El-Basyouny and Sayed, 2009). DIC is defined in Eq. (9).

\[
\text{DIC} = D(\theta) + pD
\] (9)

where \( D(\theta) \) is the Bayesian deviance of the estimated parameter, and \( D(\theta) \) is the posterior mean of \( D(\theta) \). \( D(\theta) \) can be viewed as a measure of model fitting; \( pD \) denotes the effective number of parameters and indicates model complexity.

Mean Absolute Deviation (MAD) and Bayesian R-square values were also employed to compare the model predictive performance with different spatial weight features. MAD provides a measure of the average misprediction of the model, so a value of MAD closer to zero suggests that the model predicts the observed data better than a model with a higher MAD. R-square is known as the coefficient of determination, indicating how well a particular combination of covariates explains the crash frequency. A model with higher R-square value is preferred.

5. Results and discussion

5.1. Model comparison

Table 2 presents the final estimation results for Bayesian CAR models with different spatial weight features. Models 1–7 refer to the following spatial weight features: (a) 0–1 swf, (b) CBL swf, (c) GCD swf, (d) CCD swf, (e) LUT swf, (f) LUI swf, and (g) GCO swf. The Bayesian models were developed, as is usual practice, using a Markov Chain Monte Carlo (MCMC) algorithm. The open source software WinBUGS was used to calibrate the Bayesian models with MCMC. In the process of developing the models, because of convergence and time of updating, two MCMC chains of 25,000 iterations were run, and the first 5000 samples were discarded as burn-in.

Based on the DIC values, the best model was Model 7 (DIC = 1566.530) followed by Model 1 (DIC = 1566.530) and Model 6 (DIC = 1567.010). Based on the MAD and R-square values, the best model was Model 7 (MAD = 1.201, R-square = 0.995), followed by Model 6 (MAD = 1.882, R-square = 0.987), and Model 1 (MAD = 1.999, R-square = 0.985). Thus, Model 7 (GCO swf), which was applied for the first time in macro-level safety analysis, outperformed all the other six models.

Parameter \( \alpha \) was used to estimate the proportion of the variability in the random effects owing to spatial autocorrelation. The results showed that variations attributed to spatial clustering were substantial, that is, over 79% in Models 1, 2, 5–7. However, variations attributed to spatial clustering were lower than 46% in Models 3 and 4. This indicates that the spatial clustering in the models that considered all the zones tended to be weakened. In contrast to Dong et al. (2014) who found that Models 2–4 performed better than Model 1, and Model 2 performed best among the first four
Table 2
Results for seven Bayesian NB CAR models.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model 1 (0–1swf)</th>
<th>Model 2 (CBLswf)</th>
<th>Model 3 (GCD swf)</th>
<th>Model 4 (CCD swf)</th>
<th>Model 5 (LUT swf)</th>
<th>Model 6 (LUI swf)</th>
<th>Model 7 (GCO swf)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of major arterials</td>
<td>Mean</td>
<td>0.270</td>
<td>0.278</td>
<td>0.411</td>
<td>0.410</td>
<td>0.289</td>
<td>0.287</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>(0.190, 0.360)</td>
<td>(0.194, 0.366)</td>
<td>(0.298, 0.523)</td>
<td>(0.297, 0.521)</td>
<td>(0.174, 0.391)</td>
<td>(0.202, 0.368)</td>
</tr>
<tr>
<td>Length of minor arterials</td>
<td>Mean</td>
<td>0.507</td>
<td>0.499</td>
<td>0.777</td>
<td>0.810</td>
<td>0.603</td>
<td>0.527</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>(0.311, 0.691)</td>
<td>(0.323, 0.698)</td>
<td>(0.530, 1.018)</td>
<td>(0.562, 1.044)</td>
<td>(0.415, 0.807)</td>
<td>(0.303, 0.721)</td>
</tr>
<tr>
<td>Road density</td>
<td>Mean</td>
<td>0.016</td>
<td>0.015</td>
<td>0.022</td>
<td>0.022</td>
<td>0.018</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>(0.006, 0.027)</td>
<td>(0.002, 0.027)</td>
<td>(0.007, 0.037)</td>
<td>(0.007, 0.036)</td>
<td>(0.006, 0.030)</td>
<td>(0.002, 0.030)</td>
</tr>
<tr>
<td>Average intersection spacing</td>
<td>Mean</td>
<td>−1.667</td>
<td>−1.283</td>
<td>−</td>
<td>−1.164</td>
<td>−1.347</td>
<td>−2.572</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>(−2.737, −0.731)</td>
<td>(−2.156, −0.215)</td>
<td>(−2.368, −0.206)</td>
<td>(−2.396, −0.334)</td>
<td>(−3.366, −1.519)</td>
<td></td>
</tr>
<tr>
<td>Percentage of 3-legged intersections</td>
<td>Mean</td>
<td>−0.600</td>
<td>−0.509</td>
<td>−</td>
<td>−</td>
<td>−0.471</td>
<td>−0.495</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>(−1.072, −0.162)</td>
<td>(−0.965, −0.041)</td>
<td>(−0.955, −0.040)</td>
<td>(−0.975, −0.048)</td>
<td>(−1.335, −0.282)</td>
<td></td>
</tr>
<tr>
<td>Area</td>
<td>Mean</td>
<td>0.057</td>
<td>0.057</td>
<td>−</td>
<td>−</td>
<td>0.057</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>(0.021, 0.092)</td>
<td>(0.023, 0.089)</td>
<td>(0.022, 0.089)</td>
<td>(0.019, 0.090)</td>
<td>(0.001, 0.079)</td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>Mean</td>
<td>0.013</td>
<td>0.014</td>
<td>0.017</td>
<td>0.016</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>(0.010, 0.017)</td>
<td>(0.011, 0.017)</td>
<td>(0.013, 0.022)</td>
<td>(0.012, 0.021)</td>
<td>(0.011, 0.018)</td>
<td>(0.011, 0.018)</td>
</tr>
<tr>
<td>Land use intensity</td>
<td>Medium vs low</td>
<td>−</td>
<td>−</td>
<td>0.430</td>
<td>0.438</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>−</td>
<td>−</td>
<td>(0.116, 0.768)</td>
<td>(0.123, 0.747)</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td>High vs low</td>
<td>−</td>
<td>−</td>
<td>0.446</td>
<td>0.447</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>−</td>
<td>−</td>
<td>(0.165, 0.729)</td>
<td>(0.181, 0.716)</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td>α</td>
<td>0.888</td>
<td>0.845</td>
<td>0.460</td>
<td>0.147</td>
<td>0.804</td>
<td>0.790</td>
</tr>
<tr>
<td></td>
<td>95% BCI</td>
<td>(0.738, 0.962)</td>
<td>(0.708, 0.968)</td>
<td>(0.095, 0.911)</td>
<td>(0.007, 0.595)</td>
<td>(0.652, 0.956)</td>
<td>(0.684, 0.912)</td>
</tr>
<tr>
<td>DIC</td>
<td></td>
<td>1566.530</td>
<td>1573.700</td>
<td>1749.260</td>
<td>1698.960</td>
<td>1595.010</td>
<td>1567.010</td>
</tr>
<tr>
<td>MAD</td>
<td></td>
<td>1.999</td>
<td>2.034</td>
<td>7.747</td>
<td>7.680</td>
<td>2.214</td>
<td>1.882</td>
</tr>
<tr>
<td>R-square</td>
<td></td>
<td>0.985</td>
<td>0.983</td>
<td>0.694</td>
<td>0.703</td>
<td>0.979</td>
<td>0.987</td>
</tr>
</tbody>
</table>

* Significant at 90% BCI.
* Significant at 80% BCI.
Spatial weight features, this study found that Models 2–4 were not significantly better than Model 1. The likely explanation for the discrepancy in results is that different spatial modeling results are related to the distribution characteristics of the study area. In this study, the spatial clustering was strongest (posterior mean of $\alpha = 89.9\%$) in Model 7 (GCO swf), which considered the effective range of spatial correlation between adjacent zones and all zones.

In order to further analyze the nature of model performance with different CAR priors, Fig. 2 compares the probability distributions of predicted and observed pedestrian crashes based on the seven CAR models. Crash counts in each TAZ range from 0 to 160, and more than 95% of crashes are in the range from 0 to 60. Thus, crash counts ranging from 61 to 160 were excluded in the comparison figures due to the difference between percentages of predicted and observed pedestrian crashes closing to 0%. Models 1, 2, 5–7 performed well with a bias less than 4.2% towards the difference between the percentage of predicted crashes and observed crashes. Overall, Model 7 outperformed with less bias than the other models. However, it was found that Model 3 underestimated up to 9.9% the proportion of values ranging from 0 to 5, and overestimated up to 7.6% the proportion of values ranging from 10 to 15. Model 4 underestimated up to 9.9% the proportion of values ranging from 0 to 5 and overestimated up to 6.8% the proportion of values ranging from 10 to 15. The probable reason for the underperformance of Models 3 and 4 is that spatial correlations were weaker when considering all the zones; the consideration of spatial correlations among all the zones significantly increased the model complexity, which probably resulted in the reduced model predictive performance, as explained by Dong et al. (2014). Thus, model fit was worse and the prediction bias was higher in Models 3 and 4.

### 5.2. Interpretation of predictor variables

The features significantly associated with pedestrian crashes at a 95% Bayesian Credible Interval (BCI) were analyzed based on the best model – Model 7. The explanatory variables used in the models were classified into three categories: socio-economic, roadway, and land use features. The relationships between pedestrian crashes and significant variables in each category are investigated below.

Population was designated as a surrogate measure for pedestrian exposure in this study. Population was significant in all the models. In general, larger population represents higher levels of pedestrian activity and imposes greater pedestrian crash risk. A 1000 increase in population is associated with an approximate 1.4% increase in pedestrian crashes ($e^{0.014} - 1 = 1.4\%$). The positive effect of population has also been reported by Kim et al. (2006), Wier et al. (2009), and Lee et al. (2015b).

Length of major arterials had a positive effect on pedestrian crashes, which is similar to the finding from Ukkusuri et al. (2011): zones with a greater proportion of primary roadways without access restrictions are more prone to pedestrian–vehicle crashes; and the positive effect of length of minor arterials is consistent with Quddus (2008). Average intersection spacing was negatively correlated with pedestrian crashes, while road density was positively correlated with pedestrian crashes. Both factors represent an increase in vehicle-pedestrian interactions, making pedestrian crashes more likely to occur. The percentage of 3-legged intersections had a negative effect on pedestrian crashes. Since 3-legged intersections have fewer conflict points and signal phases than 4-legged intersections, the probability of pedestrians running the red light is relatively lower and thus pedestrian crashes can be reduced. The number of pedestrian crashes had a positive correlation with medium land use intensity. Its impact on pedestrian crashes was greater than that of both low and high land use intensity. Graham and Claister (2003) used population and employment node density to reflect the extent of built development and mixed land uses when analyzing British pedestrian crashes, and they found that node density had a convex-shaped effect on pedestrian casualties. When node density reached a certain level, the number of pedestrian casualties decreased as node density increased. High land use intensity is related to higher traffic density, lower travel speed and higher pedestrian volume. Because these conditions are more likely to alert drivers and pedestrians of collision risk, they reduce pedestrian crashes. Moreover, TAZs with higher land use intensity generally have better maintained roadway systems and
safer pedestrian facilities, such as in the center of Shanghai. It is interesting to note that all the land use types were not significant in the modeling results. This is potentially due to heterogeneous influences of unobserved factors such as the combination of different land use types.

6. Summary and conclusions

This research has investigated the relationship between pedestrian crashes and various predictors in 263 TAZs in Shanghai, China. Macro-level pedestrian crash prediction models with seven spatial weight features as CAR priors in a Bayesian framework were developed, and their model fitting and predictive performances were compared. The statistical test Moran’s I and the posterior means of \( \alpha \) indicated the existence of spatial correlation among TAZs. Thus, spatial correlation must be considered when establishing TAZ-level models with spatially aggregated data.

Among the seven spatial weight features in the Bayesian model, the GCO swf (Model 7) produced the best model fit, followed by the 0–1 swf (Model 1) and CBL swf (Model 2). In this study, the characteristics of average daily walking distance and average daily trip distance in Shanghai were considered in the GCO swf (Model 7), which was introduced in macro-level safety analysis for the first time. The GCO swf (Model 7) accounted for the effective range of spatial autocorrelation between adjacent zones and all zones, which provided a more reasonable way to characterize the effect of spatial autocorrelation. Although model performances of LUF swf (Model 5) and LUI swf (Model 6) were not as good as expected, they still provide new ways of measuring the exact proximity between the adjacent TAZs and illustrate the impact of spatial heterogeneity of land use. However, the poor model fit of the GCD swf (Model 3) and CCD swf (Model 4) indicates that the consideration of spatial correlation among all zones reduces model fit and increases prediction bias. Further research on determining specific effective range of spatial correlation among analysis units and providing better spatial weight features as CAR priors in the Bayesian framework is necessary.

The results of the modeling show that several roadway, socio-economic, and land use features of TAZs contributed to pedestrian crash occurrence in this study. Population, length of major arterials, length of minor arterials, road density, area of TAZ, and land use intensity had positive effects on the probability of pedestrian crashes. Average intersection spacing and the percentage of 3-legged intersections had negative effects. Population, as the surrogate measure for pedestrian exposure, was significant in all the models. The results indicate that a 1000 increase in population is associated with an approximately 1.4% increase in pedestrian crashes \( (e^{0.014} - 1 = 1.4\%) \). Safety practitioners could use this figure to predict the change of pedestrian crashes that are likely to come with socio-economic development.

This increased understanding of the characteristics of pedestrian crashes in Shanghai will be useful to reduce pedestrian crash occurrence and improve pedestrian safety in other similar large cities. First, similar data, and the macro-level pedestrian crash models considering spatial correlation developed in this study, could be used to identify areas or zones with higher-than-expected levels of pedestrian crashes. Priorities could then be pursued and set for various area-based engineering, enforcement, and educational strategies and programs. Second, results can be used to predict TAZ-level change in pedestrian crashes associated with land use development and transportation planning decisions. Thus countermeasures could be implemented through the transportation planning process to consider road mileage, road network density, and intersection design in areas of high crash prediction. Pedestrian-related improvement measures include better protective isolation facilities on major and minor roads to reduce traffic conflicts among pedestrians, non-motorized vehicles, and motorized vehicles in typical mixed traffic flow. Other countermeasures include providing pedestrian signals that count down remaining green time and walk beacons at mid-block crossings to reduce the probability of pedestrians running the red light. Finally, this analysis leads to an important conclusion that higher priority should be given to reduce pedestrian crashes in areas with medium land use intensity.

Acknowledgements

This study was jointly sponsored by the Science and Technology Commission of Shanghai Municipality (15DJ1204800), the Chinese National Science Foundation (51522810), and the Fundamental Research Funds for the Central Universities.

References


Shanghai Urban and Rural Construction and Transportation Development Institute, 2013. Shanghai Comprehensive Transportation 2013 Annual Report.


