https://dx.doi.org/10.1016/j.aap.2018.09.010

Boundary Crash Data Assignment in Zonal Safety Analysis: An Iterative Approach based on Data Augmentation and Bayesian Spatial Model

Xiaoqi Zhai[1], Helai Huang[1*], Mingyun Gao[2], Ni Dong[3], N.N.SZE[4]

[1] School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan, China
[2] Department of Industrial and Business Management, Hunan University, Changsha, Hunan, China
[3] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China
[4]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong
[*]Correspondence: huanghelai@csu.edu.cn

Abstract

Boundary effect refers to the issue of ambiguous allocation of crashes occurred on or near the boundaries of neighboring zones in zonal safety analysis. It results in bias estimates for associate measure between crash occurrence and possible zonal factors. It is a fundamental problem to compensate for the boundary effect and enhance the model predictive performance. Compared to conventional approaches, it might be more reasonable to assign the boundary crashes according to the crash predisposing agents, since the crash occurrence is generally correlated to multiple sources of risk factors. In this study, we proposed a novel iterative aggregation approach to assign the boundary crashes, according to the ratio of model-based expected crash number in adjacent zones. To verify the proposed method, a case study using a dataset of 738 Traffic Analysis Zones (TAZs) from the county of Hillsborough in Florida was conducted. Using Bayesian spatial models (BSMs), the proposed approach demonstrated the capability in reasonably compensating for the boundary effect with better model estimation and predictive performance, as compared to three conventional approaches (i.e., half and half ratio method, one to one ratio method, and exposure ratio method). Results revealed that several factors including the number of intersections, road segment length with 35 mph speed limit,

road segment length with 65 mph speed limit and median household income, were sensitive to the boundary effect.

Keywords: Macroscopic safety analysis; Zonal-level CPMs; Boundary effect; Iterative algorithm.

1. Introduction

The prevalent zonal safety analysis attracts growing interests. It facilitates the identification of crash pattern, distinguishing possible factors to crash occurrences, and recommending targeted safety countermeasures at zonal levels. In zonal safety analysis, traffic crashes are usually aggregated as per certain finite spatial unit (Huang and Abdel-Aty, 2010). Researchers usually encounter the problem of how to reasonably allocate the boundary crashes (i.e., crashes occurred on or near the boundaries of neighboring zones) in data preparation. Since the spatial unit is finite, the data aggregation will inevitably induce boundary effect. It refers to the issue of ambiguous allocation of boundary crashes, and in turn bias estimation for zonal safety analysis.

Since crashes are spatially correlated (Huang and Abdel-Aty, 2010; Quddus, 2008; Siddiqui et al., 2012; Xu et al., 2014; Huang et al., 2016), the boundary crashes are assumed to be collectively affected by the zonal factors of the neighbor spatial units. In accordance to *Tobler's first law of geography* (Waldo, 1970), "Everything is related to everything else, but near things are more related than distant things". Crashes located on or near zone boundaries may have inter-zonal influence. The boundary crashes could be more correlated with neighboring zones due to the fact that they are closer to the adjacent units than to the interior of a zone. However, most of the previous studies aggregated boundary crashes simply according to the geocodes in crash records and related zonal attributes to all crashes assigned to the specific zones. In such cases, without accounting for the potential inter-zonal effect, modeling merely based on the characteristics of an individual zone may result in bias in estimating the safety effect of zonal factors (Siddiqui and Abdel-Aty, 2012).

The boundary effect has been recognized and investigated in several studies (Fotheringham and Wegener, 1999; Lovegrove, 2007; Siddiqui and Abdel-Aty, 2012; Wang et al., 2012; Lee et al., 2014; Cui et al., 2015). The general approach to compensate for the boundary effect is to construct buffer zones along the regional boundary and aggregate the boundary crashes to the neighboring zones based on certain simple methods, including the one-to-one ratio method and the half-to-half ratio method (Lovegrove and Sun, 2010). The one-to-one ratio method and the half-to-half ratio method give an equal weight to adjacent zones in allocating the boundary crashes, by assuming that boundary crashes are collectively affected by the risk factors in neighboring zones. However, they ignored the fact that neighboring zones hardly have equal effect on the boundary crashes. Later on, some researchers attempted to consider the variations of some basic characteristics of the adjacent zones while allocating the boundary crashes. Wei (2010) proposed the ratio of exposure method. It allocated the boundary crashes based on the ratio of the variable of vehicle kilometers travelled (VKT) or the total lane kilometers (TLKM). Results indicated that the mere consideration of the ratio of the variable of VKT or TLKM among the neighborhood didn't work well, because they failed to fully account for the complicated crash mechanism and potential risk factors. Cui et al. (2015) proposed a collision density ratio method to aggregate the boundary crashes based on crash spatial distribution. It was found that this method led to better model predictive performance, as compared to the previous methods (i.e., the half-to-half ratio method and the one-to-one ratio method), and its boundary crash aggregation results were closer to the true value from the manual inspection.

However, crash occurrence is associated with a variety of potential zonal risk factors in terms of socioeconomic and demographic status (e.g., Aguero-Valverde and Jovanis, 2006; Hadayeghi et al., 2010; Huang et al., 2010; Siddiqui et al., 2012), transportation network (e.g., Abdel-Aty et al., 2011; Siddiqui et al., 2012), road facilities and traffic flow (e.g., Abdel-Aty et al., 2011; Dong et al., 2014, 2015). It might be more reasonable to assign the boundary crashes according to the zonal crash predisposing agents by taking complicated crash causes into account. Crash Prediction Model (CPM) is an essential tool in traffic safety analysis to associate crash occurrence with confounding contributors, including crash exposure, various risk factors and the unobserved heterogeneity caused by omitted factors and data correlation (Xu et al., 2014; Yu et al., 2015; Peng et al., 2017). Bayesian spatial model (BSM) has been one of the state of the art zonal-level CPMs in modeling spatial correlation to proxy the unobserved heterogeneity (Huang and Abdel-Aty, 2010; Quddus, 2008; Siddiqui et al., 2012; Xu et al., 2014; Huang et al., 2016; Xu et al., 2017).

The present study proposes a novel iterative boundary crash allocation method to assign boundary crashes according to the BSM-based expected crash number in adjacent zones of analysis. Specifically, the main procedure of the proposed method can be summarized as follows: (a) divide each zone into boundary (buffer zones) and interior, (b) develop a BSM based on the interior crashes to calculate the initial expected crash number of each zone; (c) aggregate the boundary crashes to the adjacent zones based on the proportion of expected crash number obtained in step 2 and (d) re-run the BSM to update the expected crash number. The operation of CPM and the boundary crash aggregation process are alternated, until the

predicted crash number ends updating bounded by a given limit, and by then the process is finished. For the purpose of evaluation, a case study on a dataset from the county of Hillsborough in Florida was carried out. Using BSMs, the model performance of the proposed boundary crash aggregation method was compared with three traditional methods, i.e., half and half ratio, one to one ratio, and exposure ratio method. The standard-difference-in-means test was further employed to examine the risk factors that were sensitive to boundary effect.

2. Iterative Boundary Crash Aggregation Approach

2.1 Procedure of the iterative boundary crash allocation method

Step 1: divide zones into boundary and interior

Figure 1 presents four adjacent zones (A_a , a=1, 2, 3, 4). Buffer zones (the dark gray zones) are created along the zone boundaries. Let d denote the zone buffer size (distance between the specific boundary of interior zone and the original zone boundary), therefore the boundary size (width of buffer zone) is 2d.

Let Y denote the total number of crashes of the whole area of analysis, and Y_{A_a} denote the count of crashes in zone A_a . Y_b is the number of the boundary crashes, which have occurred within the buffer zone. And $Y_{\backslash b}$ is the number of the interior crashes. $Y_{\backslash b}$ and Y_b satisfy

$$Y_{\backslash b} \cap Y_b = \varnothing \quad (1)$$

and

$$Y = Y_{\backslash b} \cup Y_b . \quad (2)$$

Y_{A_a} is composed of the number of the interior crashes $\{Y_{\backslash b}\}_{A_a}$ and the number of the boundary crashes $\{Y_b\}_{A_a}$. The boundary crash aggregation approach aims to assign boundary crashes $\{Y_b\}_{A_a}$ to Y_{A_a}.

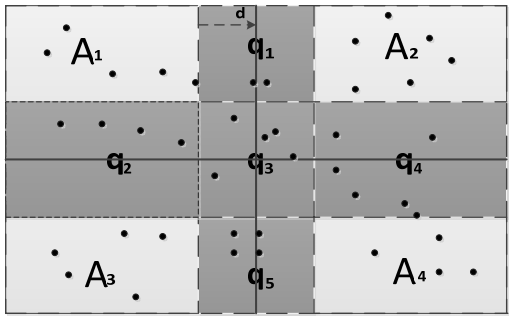

Figure 1. The Structure of a Neighborhood: Boundary Zone and Interior Zone

Step 2: develop the BSM based on interior crash data

In the first instance, the boundary crash was not considered, therefore, the CPMs can be constructed only based on the count of interior crashes. The Bayesian spatial model with conditional auto-regressive (CAR) priors is employed in zonal-level CPM development.

In this study, the basic model structure with zone i developed by Besag et al. (1991) is employed:

$$Y_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \boldsymbol{\alpha} + \log(e_i) + \boldsymbol{x_i}\boldsymbol{\beta} + \delta_i + \vartheta_i \quad (3)$$

(4)

where for zone i ($i = 1,2,\mathrm{L}\ ,N$), Y_i is the number of crashes, λ_i is the Poisson parameter, and e_i is the crash exposure. The exposure is reflected by the DVMT in each individual zone. where $\boldsymbol{x_i}$ denotes the vector of explanatory variables, $\boldsymbol{\beta}$ is the vector of fixed effect

parameters, and δ_i is the random effect to account for unstructured over-dispersion error, which is specified via an ordinary exchangeable normal prior,

$$\delta_i \sim N(0,1/\tau_h) \tag{5}$$

where τ_h is the precision parameter, which follows a prior gamma (0.5, 0.0005) as recommended by Xu et al. (2014). ϑ_i is the spatial correlation term reflecting two zones having a shared border, which is specified with a CAR prior as suggested by Besag et al. (1991),

$$\vartheta_i \sim N(\bar{\vartheta}_i,1/\tau_i) \tag{6}$$

where

$$\bar{\vartheta}_i = \frac{1}{\sum_{i\neq j}\omega_{ij}}\sum_{i\neq j}\vartheta_i\omega_{ij} \tag{7}$$

and

$$\tau_i = \frac{\tau_f}{\sum_{i\neq j}\omega_{ij}} \tag{8}$$

in which τ_f is the precision parameter, which follows a prior gamma (0.5, 0.0005) as recommended by Wakefield et al. (2000), and ω_{ij} is the entries in the proximity matrix and generally reflects the spatial correlation of two zones, and

$$\omega_{ij} = \begin{cases} 1 & if\ i,j\ are\ adjacent \\ 0 & if\ i,j\ are\ not\ adjacent \end{cases}. \tag{9}$$

Obviously, the values of τ_h and τ_f control the amount of extra-Poisson variability allocated to area-wide heterogeneity and clustering effect among adjacent zones, respectively.

Step 3: aggregate the boundary crashes to adjacent units

By not considering the boundary crash data, we obtain the initial expected crash number from

step 2. Then we assign the boundary crashes into adjacent zones according to the proportion of the expected crash number. For zone i, the boundary crashes can be allocated in an iterative process as follows:

$$Y_i = \{Y_{\backslash b}\}_i + \sum_{q=1}^{Q} \frac{\lambda_i}{\sum_{j=1}^{N} l_{jq}\lambda_j} \{Y_b\}_q \tag{10}$$

where λ_i is the expected crash number, $\{Y_{\backslash b}\}_i$ is the count of interior crashes in zone i, $\{Y_b\}_q$ is the count of boundary crashes in buffer zone q, Q is the total number of buffer zones, N is the total number of zones, and l_{jq} is the value that discriminates whether the zone j is adjacent to the buffer zone q,

$$l_{jq} = \begin{cases} 1 & if\ j\ and\ q\ are\ adjacent \\ 0 & if\ j\ and\ q\ are\ not\ adjacent \end{cases} . \tag{11}$$

For example, the adjacent matrix of the Figure 1 can be described as follow:

$$l = \begin{Bmatrix} & A_1 & A_2 & A_3 & A_4 \\ q_1 & 1 & 1 & 0 & 0 \\ q_2 & 1 & 0 & 1 & 0 \\ q_3 & 1 & 1 & 1 & 1 \\ q_4 & 0 & 1 & 0 & 1 \\ q_5 & 0 & 0 & 1 & 1 \end{Bmatrix} . \tag{12}$$

Step 4: re-run the BSM to update the crash prediction counts

Return to step 2 to run the BSMs based on the boundary data aggregation results obtained from step 3. The step 2 and the step 3 are alternated till Y_i ends updating to a given limit. That is,

$$\max_i \left| Y_i^{k+1} - Y_i^k \right| \le 0.001 , \tag{13}$$

where Y_i^{k+1} is the allocation results in the k-th iteration.

The flow chart of the iterative boundary crash allocation method is shown in Figure 2.

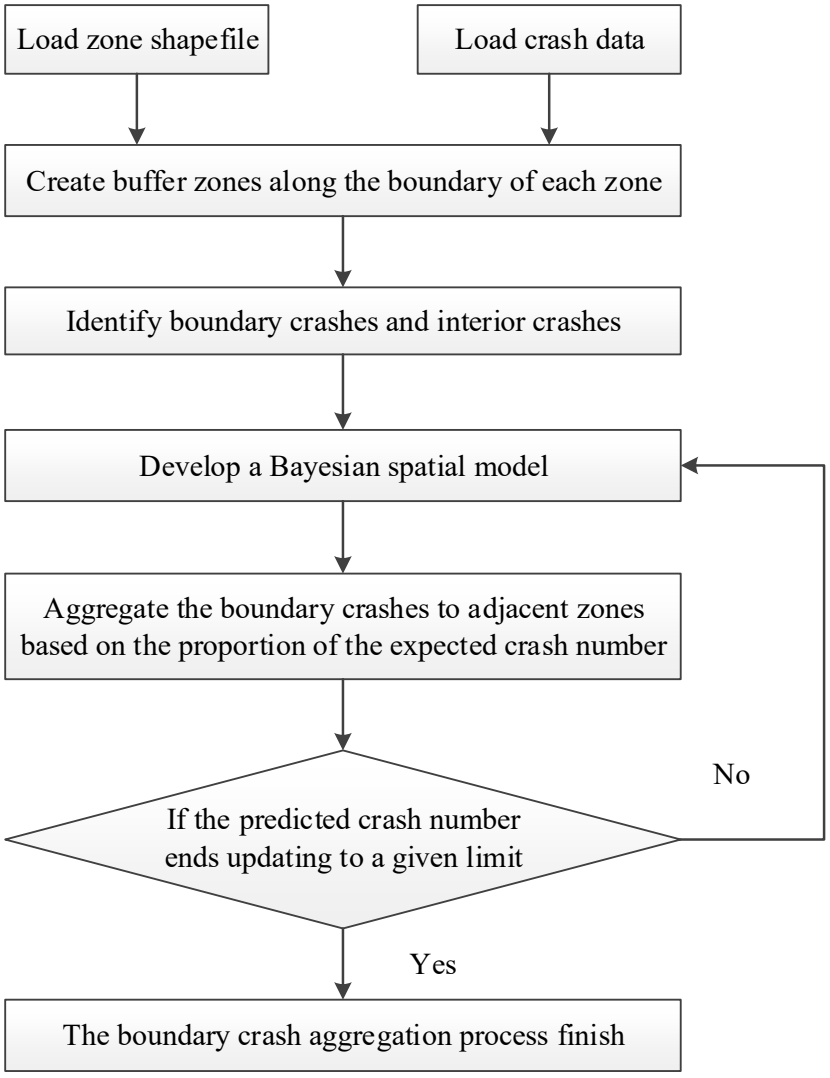

Figure 2. Flow Chart of the Iterative Boundary Crash Allocation Method

2.2 Traditional aggregation methods for comparison

Three traditional boundary crash assignment methods for comparison are described as follows.

Method 2: Half-and-Half ratio method

Half-and-Half ratio method is the simplest method to allocate the boundary crashes. It gives an equal weight while assigning the boundary crashes to the adjacent zones. It counts each crash

once and assigns the average collision number to the neighborhoods. The method could be specified as:

$$Y_i = \{Y_{\setminus b}\}_i + \sum_{q=1}^{Q} \frac{1}{\sum_{j=1}^{N} l_{jq}} \{Y_b\}_q . \quad (14)$$

Method 3: One-to-One ratio method

Similar to the half-and-half ratio method, there is another equal weight assignment method according to the one-to-one ratio. The one-to-one ratio method counts the boundary collisions once for every neighboring zone adjacent to the location of crash. It can be estimated as follows,

$$Y_i = \{Y_{\setminus b}\}_i + \sum_{q=1}^{Q} \sum_{j=1}^{N} l_{jq} \{Y_b\}_q . \quad (15)$$

Method 4: The ratio of exposure variables method

The boundary crashes can be assigned to adjacent zones according to the ratio of exposure variable represented by e_i (i.e., Daily vehicle miles traveled, DVMT). The method could be specified as:

$$Y_i = \{Y_{\setminus b}\}_i + \sum_{q=1}^{Q} \frac{e_i}{\sum_{j=1}^{N} l_{jq} e_j} \{Y_b\}_q . \quad (16)$$

2.3 Model evaluation

Compensating for the boundary effect could improve the model performance and produce more reliable model estimates (Cui et al., 2015). In this study, we compare the model performance and prediction accuracy by Deviance Information Criterion (DIC), Mean Absolute Deviation

(MAD) and Mean Squared Prediction Error (MSPE).

DIC is based on the posterior distribution of the deviance and can be seen as a generalization of the Akaike Information Criterion (AIC) (Spiegelhalter et al., 2001). Specifically, the DIC is defined as:

$$DIC = \overline{D} + p_D \tag{17}$$

where p_D is the effective number of parameters in the model, and $\overline{D}$ is the posterior mean of the deviance. Models with low er DIC are preferred. Roughly, differences in the DIC of more than 10 can definitely rule out a model with a higher DIC, and differences between 5 and 10 are considered substantial (Spiegelhalter et al., 2007).

To compare the observed data with predicted data based on the fitted model, the MAD and MSPE can be employed. MAD is used to measure the average of the absolute deviations from observed data. The MAD is given by

$$\text{MAD} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right| \tag{18}$$

where $\hat{y}_i$ is the predicted crash frequency of zone i and N is the total number of zones.

MSPE is an estimator used to measure the average of the squares of the "errors", that is, the difference between the fitted values implied by the predictive function and the observed values. It can be estimated by

$$\text{MSPE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 . \tag{19}$$

Model with lower values of DIC, MAD and MSPE implies a better fit.

2.4 Standard-difference-in-means test

In order to examine the sensitivity of the parameter estimates to boundary effect, it would be useful to determine whether the differences in parameter estimates based on different aggregation methods are statistically significant. The standard-difference-in-means test (Fotheringham and Wong, 1991) is shown in the following equation,

$$t = \frac{\hat{\beta}_{ij} - \hat{\beta}_{ik}}{S\left|\hat{\beta}_{ij} - \hat{\beta}_{ik}\right|} \tag{20}$$

Where $\hat{\beta}_{ij}$ and $\hat{\beta}_{ik}$ are estimated coefficients of independent variable i at the j th and k th boundary crash aggregation methods, respectively. S is the standard error.

3. Case Study

3.1 Data preparation

To evaluate the iterative boundary crash allocation method, a case study was conducted based on a dataset from the Hillsborough County, Florida. Hillsborough is in west-central Florida and constitutes 738 TAZs representing both rural and urban areas. We obtained the shape file of TAZs from the Florida Department of Transportation (i.e., FDOT) District's Intermodal Systems Development Unit. Among various spatial units applied in zonal safety research, traffic analysis zones (TAZs) might associate with the most serious boundary effect problem. Some researchers suggested that the boundary effect has little influence on their model results if the proportion of the boundary crashes is low (Guevara et al. 2004, Khondakar et al. 2010). However, the proportion of boundary crashes was reported more than 70% at TAZ-level (Lee et al., 2014), because the TAZs are generally delineated by arterial roads where the crash likelihoods are higher. Additionally, TAZs have been widely adopted as the basic zonal

schemes in zonal safety analysis, for being the only traffic-related zone system (Siddiqui and Abdel-Aty 2012, Wang et al. 2012, Schneider et al. 2013). The ambiguous allocation of boundary crashes in TAZ would cause unreliable estimation and therefore might mislead the development of safety countermeasures. Thus, it is crucial to deal with the boundary effect problem at TAZ level.

We obtained the crash data from Florida Department of Transportation (FDOT) Crash Analysis Reporting System. In the year 2009, there were 8,595 crashes recorded in Hillsborough County, Florida. The ArcGIS software was used to make geo-process and create buffer zones in this case. A suitable boundary zone size is critical for boundary crash assignment. According to the previous research (Lee et al., 2014; Siddiqui and Abdel-Aty, 2012; Ivan et al., 2006), three boundary size (i.e. 250ft, 300ft, and 350ft) were examined. For instances, proportion of boundary crash of these three sizes are 70.3%, 71.3% and 71.8% respectively.

As mentioned in the Step 1, the number of total crashes per TAZ is composed of the number of interior crashes and assigned boundary crashes. The variables of the number of total crashes by four boundary aggregation methods in different boundary sizes were chosen as the dependent variables. In the zonal traffic safety analysis, it is an essential part to reflect the effects of demographic and socioeconomic factors, and traffic/roadway characteristics on crash occurrence at the zonal-level. A number of exploratory variables were collected from FDOT's Roadway Characteristics Inventory and the U.S. Census using PLANSAFE Census Tool. In particular, DVMT was used to proxy the crash exposure. The number of intersections was

considered as a key contributor to crash risk. As the speed limit is one of the top priorities in safety campaigns (Yu and Abdel-Aty, 2014), the road segment length with 15, 25, 35, 45, and 65 speed limits were considered. The other roadway/traffic data, total trip productions/attraction, home-based work productions/attraction, and non-home-based work productions/attraction were also incorporated. For demographic and socioeconomic factors, yearly median income per household was taken into consideration. The descriptive statistics of the collected data are summarized in Table 1.

Prior to the establishment of crash prediction models, collinearity among the variables were examined. The variable of total trip attraction was excluded as it was highly correlated with the variable of total trip production. Therefore we retained the variables including number of intersections, road lengths with 35, 45 and 65 mph speed limits, total trip productions and median household income in the final models.

Table 1. Descriptive Statistics for Explanatory Variables

Variables	Definition	Mean	Std	Min	Max.
Total crash0	Total number of crashes per TAZ	11.65	15.69	0.00	158.00
Total crash1	Total number of crashes by M1 in 250 ft boundary size	11.64	15.48	0.00	155.43
Total crash2	Total number of crashes by M2 in 250 ft boundary size	11.65	13.41	0.00	141.40
Total crash3	Total number of crashes by M3 in 250 ft boundary size	21.39	25.10	0.00	223.00
Total crash4	Total number of crashes by M4 in 250 ft boundary size	11.67	15.21	0.00	137.20

Total crash5	Total number of crashes by M1 in 300 ft boundary size	11.67	15.23	0.00	152.07
Total crash6	Total number of crashes by M2 in 300 ft boundary size	11.63	13.38	0.00	141.00
Total crash7	Total number of crashes by M3 in 300 ft boundary size	21.62	25.36	0.00	224.00
Total crash8	Total number of crashes by M4 in 300 ft boundary size	11.66	15.22	0.00	137.67
Total crash9	Total number of crashes by M1 in 350 ft boundary size	11.62	15.55	0.00	153.00
Total crash10	Total number of crashes by M2 in 350 ft boundary size	11.62	13.38	0.00	142.00
Total crash11	Total number of crashes by M3 in 350 ft boundary size	21.74	25.44	0.00	225.00
Total crash12	Total number of crashes by M4 in 350 ft boundary size	11.65	15.24	0.00	137.14
DVMT	Daily vehicle-miles traveled (Vehicle*miles/ day) in thousands)	95.07	110.24	0.06	788.77
Intersection	Number of intersections	13.2	10.98	1	83
seglen15	Road segment length with 15 mph speed limit (mile)	0.2	0.41	0	4.02
seglen25	Road segment length with 25 mph speed limit (mile)	6.51	6.4	0	43.46
seglen35	Road segment length with 35 mph speed limit (mile)	1.22	1.35	0	14.47
seglen45	Road segment length with 45 mph speed limit (mile)	0.14	0.39	0	3.76
seglen65	Road segment length with 65 mph speed limit (mile)	0.27	0.66	0	6.1
TOTALP	Total trip productions	5016.01	4033.4	0	28638
TOTALA	Total attractions	5393.28	6464.14	0	79717
MHINC	Median household income (in thousands dollars)	40.14	20.24	0	115.66

Note: M_1: iterative boundary crash allocation method; M_2: half-and-half ratio method; M_3: one-to-one ratio method; M_4: ratio of exposure variables method. The bold numbers mean statistical significance at 95% significance level

4. Results and Discussion

4.1 Model comparison

To assess the performance of proposed method, three boundary sizes, i.e. 250ft, 300ft, and 350ft, are considered when allocating the boundary crashes. Therefore, altogether twelve

models with respect to four boundary crash aggregation methods, i.e. the proposed one and three other traditional approaches were estimated using the statistical package WinBUGS (Lunn et al., 2000). Table 2 and Table 3 present the results of model assessment and parameter estimates, respectively. Concerning the effect of spatial autocorrelation, the variations contributed by spatial correlations in all the models were statistically significant, which justified the usage of spatial models.

Table 2. Results of Model Assessment

Boundary Size	250ft				**300ft**				350ft			
Method	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$	$\mathbf{M_4}$	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$	$\mathbf{M_4}$	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$	$\mathbf{M_4}$
MAD	2.24	2.40	3.28	2.33	**2.16**	2.32	2.83	2.32	2.24	2.40	3.25	2.33
MSPE	11.20	11.56	21.82	11.55	**10.74**	11.62	16.26	11.65	11.21	11.55	21.51	11.51
DIC	3283	3662	4123	3639	3135	3302	3715	3281	**3121**	3658	4002	3568

Note: M_1: iterative boundary crash allocation method; M_2: half-and-half ratio method; M_3: one-to-one ratio method; M_4: ratio of exposure variables method.

Roughly, differences in the DIC of more than 10 points can rule out a model with a higher DIC. Model with lower value of MAD and MSPE implies a better prediction capability. As shown in Table 2, Method 1 (M1, the iterative boundary crash allocation method) resulted in the best models with the lowest DIC, MAD and MSPE scores, and Method 4 (M4, the ratio of exposure variables method) performed better than the Method 2 (M2, half-and-half ratio method) and Method 3 (M3, one-to-one ratio method) for all three boundary size. Such results indicated that allocating the boundary crashes based on the attributes of the neighboring zones would greatly improve the model performance than equally assigning the boundary crashes into adjacent zones, and the use of more attributes is correlated to better model performance. The M1 models outperformed other models, for considering the complete prior information including spatial proximity structure, exposure and explanatory variables in

boundary crash allocation process. The consistent results revealed for different boundary sizes justify the validity of the proposed method at TAZ-level. It could possibly be attributed to the high proportion of boundary crashes.

The model results indicated that traditional boundary crash allocation methods had some drawbacks. As shown in Table 3, the intercepts in M3 based models were larger than other three methods, which conforms to previous research (Cui et al. 2015). This could be explained by that the M3 repeatedly aggregated boundary crashes, or in other words, one boundary crash was assigned into more than one TAZ. As such, the total number of crashes increased by using M3, which was impractical and gave it the worst fitting offset. The M4 based model resulted in the fewest significant factors. The variables of the total trip productions, roadway length with a speed limit of 45 and 65mph were not significant in M4, which did not maintain consistency with other models. It may be due to the fact that M4 assigned the boundary crashes only according to the ratio of the exposure variables, and too much prior information had been ignored in allocating the boundary crashes, which might confuse the responsibility that the crash risk factors undertake for crash counts.

4.2 Interpretation for Parameter Estimation

Given that the model performance was better, the parameter estimation results for boundary size of 300ft were presented in the Table 3 and the effects of risk factors are discussed as follows.

Intersections have long been recognized as hazardous locations on roads (Wang et al., 2009). As shown in Table 3, the number of intersections was negatively associated with the

crash occurrence in all the four models, which implied that crash rates decreased with the increase of intersection numbers. Although intersections are notorious as crash hotspots in roadway systems, recently some safety researchers found that intersection density was related to a decrease in fatal crashes (Guevara et al., 2004). Intersections are usually associated with slow speed and high levels of congestion. In such cases, drivers tended to be more cautious while driving, so as to decrease the crash risk.

Appropriate speed limit should be set in accordance to the land use, landscape, road class, traffic pattern and their interactions (Xu and Huang, 2015). The road lengths with a speed limit equal to or greater than 35 mph was found negatively associated with the crash occurrence. This finding was consistent with the well-known fact that, when exposure and other risk factors were equal, higher speed led to a substantial increase in crash risk, especially for severe crashes (Aarts and Van, 2006). Thus, more attention should be paid to the road segments with higher speed limit for road safety improvement.

We included the total trip production as the explanatory variable in the models. Several previous studies have examined trip data as a surrogate exposure factor to possibly account for the planning-level safety conditions in zonal crash prediction models (Huang et al., 2013). The coefficients of total trip production were significantly positive in M1 and M2 based models, which indicated more trip productions led to more crashes in general. Special attention should be paid to the places with more trip productions, such as the commercial zones and the residential places.

The coefficients of the median household income were significantly negative as shown in

Table 3. The results implied that TAZs with a lower median household income were associated with a worse safety level, which is consistent with the previous research (Xu and Huang, 2015; Xu et al., 2014). Wealthier areas could afford to develop and implement more effective risk mitigation and avoidance measures, and thus resulted in favorable safety effect. Besides, higher income households could afford to buy more expensive vehicles with more advanced safety features.

Table 3. Results of Parameter Estimation

Method	$\mathbf{M}_1$		$\mathbf{M}_2$		$\mathbf{M}_3$		$\mathbf{M}_4$	
Variable	Mean	95% BCI	Mean	95% BCI	Mean	95% BCI	Mean	95% BCI
Intercept	-1.47	(-1.96, -1.00)	-1.54	(-1.90, -1.06)	-0.75	(-1.04, -0.47)	-1.30	(-1.60, -0.96)
Intersection	**-2.80**	**(-4.19, -1.23)**	**-2.64**	**(-3.61, -1.59)**	**-2.26**	**(-3.15, -1.34)**	**-1.71**	**(-2.78, -0.72)**
seglen35	**-1.97**	**(-3.98, -0.42)**	**-3.00**	**(-4.24, -1.81)**	**-2.38**	**(-3.90, -1.06)**	**-2.22**	**(-3.87, -0.82)**
seglen45	**-1.95**	**(-3.29, -0.29)**	**-1.88**	**(-2.56, -1.09)**	**-1.98**	**(-3.05, -0.98)**	-1.55	(-3.26, 0.08)
seglen65	**-1.02**	**(-2.03, -0.15)**	**-0.86**	**(-1.63, -0.02)**	**-1.99**	**(-3.20, -0.83)**	-0.96	(-1.89, 0.23)
TOTALP	**0.73**	**(0.22, 1.39)**	**0.81**	**(0.18, 1.56)**	0.57	(-0.25, 1.32)	0.24	(-1.48, 1.36)
MHINC	**-1.31**	**(-2.30, -0.32)**	-0.50	(-1.49, 0.36)	**-1.16**	**(-1.73, -0.63)**	**-1.52**	**(-2.15, -0.88)**
tau.f	1.11	(0.32, 2.27)	1.04	(0.57, 2.27)	1.60	(0.80, 2.98)	0.64	(0.32, 1.12)
tau.h	0.33	(0.27, 0.43)	0.43	(0.36, 0.52)	0.48	(0.41, 0.56)	0.53	(0.43, 0.65)

Note: M_1: iterative boundary crash allocation method; M_2: half-and-half ratio method; M_3: one-to-one ratio method; M_4: ratio of exposure variables method. The bold numbers mean statistical significance at 95% significance level

Across the results based on different allocation methods, the sign of the parameters kept the same, but the value and significance of the parameter estimates were varied. Only the roadway length with a speed limit of 35 mph and the number of intersections were significant in all the four models. To determine whether the variations depicted above were statistically significant, the standard difference-in-means tests were performed (Fotheringham and Wong, 1911). Figure 3 illustrates the ranges of parameter estimates (mean +/-1.5 standard errors) for individual risk factors with respect to different boundary crash allocation methods. It was

apparent that the results of parameter estimates fluctuated across different boundary crash allocation methods for the variables including number of intersections, road segment length with 35 mph speed limit, road segment length with 65 mph speed limit and median household income. It implies that those variables were greatly sensitive to the boundary effect and should be taken into more consideration in zonal safety analysis.

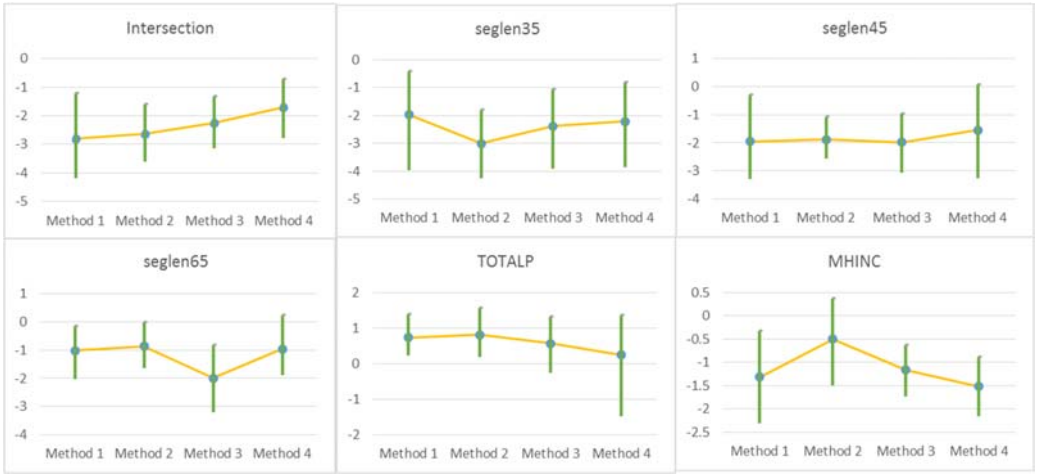

Figure 3. Results of Standard Difference in Means Test for Parameter Estimates

5. Conclusion

For proactive road safety management framework, zonal-level crash prediction models are increasingly popular. It becomes a vital component in transportation safety planning for estimating the crash potentials (measures) for alternative policy schemes. Whenever the aggregation of crash across spatial unit exists, the boundary effect must be explicitly considered. The ambiguous allocation of boundary crashes would lead to bias estimation for zonal safety analysis. This study focused on the compensation of the boundary effect and the improvement of model prediction performance. An iterative boundary crash allocation method was proposed to assign the boundary data according to the ratio of expected crash number, using Bayesian

spatial approach. Based on the results of case study, the proposed method was found superior to the traditional boundary crash aggregation method in terms of model prediction performance and number of contributory variables revealed. As expected, the iterative boundary crash allocation method using crash predisposing agents should be a promising solution for boundary crash data aggregation. Factors including number of intersections, road segment length with 35 mph speed limit, road segment length with 65 mph speed limit and median household income were found to be sensitive to the boundary effect in the case study.

A suitable boundary size is critical for boundary crash identification. However, it might vary with the neighborhood condition, road condition, and spatial units. Future research is recommended to study the determination of suitable boundary zone size under different conditions.

Acknowledgement

This work was jointly supported by: 1) the Joint Research Scheme of National Natural Science Foundation of China/Research Grants Council of Hong Kong (Project No. 71561167001 & N_HKU707/15), 2) National Key Research and Development Plan (No. SQ2018YFB120181), 3) the Research Grants Council of Hong Kong (Project No. 25203717), and 4) the Research Committee of the Hong Kong Polytechnic University (Project No. 1-ZE5V). We would like to thank Dr. Mohamed Abdel-Aty at the University of Central Florida and the Florida Department of Transportation for providing the data.

Reference

1 Aarts, L., Van Schagen, I. 2006. Driving speed and the risk of road crashes: a review. Accid.
2 Anal. Prev. 38, 215-224.
3 Abdel-Aty, M., Siddiqui, C., Huang, H., Wang, X., 2011. Integrating trip and roadway
4 characteristics to manage safety in traffic analysis zones. Transport. Res. Rec. 2213, 20-28.
5 .
6 Besag, J., York, J., Molli, E.A., 1991. Bayesian image restoration with two applications in
7 spatial statistics. Annu. Inst. Stat. Math. 43(1), 1-59.
8 Bureau, U.S.C., 2007. Summary file 3, 2000 census of population and housing, technical
9 documentation. In: Bureau, U.S.C. ed. U.S. Department of Commerce, Economics and
10 Statistics Administration, and U.S. Census Bureau.
11 Cui, G., Wang, X., Kwon, D.-W., 2015. A framework of boundary collision data aggregation
12 into neighbourhoods. Accid. Anal. Prev. 83, 1-17.
13 Dong, N., Huang, H., Xu, P., Ding, Z., Wang, D., 2014. Evaluating spatial-proximity structures
14 in crash prediction models at the level of traffic analysis zones. Transport. Res. Rec. 2432, 46-
15 52.
16 Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level
17 of traffic analysis zones: Assessing the spatial proximity effects. Accid. Anal. Prev. 82, 192-
18 198.
19 Fotheringham, A.S., Wong, D.W., 1991. The modifiable areal unit problem in multivariate
20 statistical analysis. Environ. Plan. A 23, 1025-1044.
21 Fotheringham, A.S., Wegener, M., 2000. Spatial Models and GIS: New Potential and Models.

Taylor & Francis, London.

Guevara, F.L.D., Washington, S.P., Oh, J., 2004. Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. Transport. Res. Rec. 1897, 191-199.

Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prev. 42, 1556-1565.

Huang, H., Abdel-Aty, M.A., Darwiche, A.L., 2010. County-level crash risk analysis in Florida Bayesian spatial modeling. Transport. Res. Rec. 2148, 27-37.

Huang, H., Chin, H.C., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis. Accid. Anal. Prev. 40 (1), 45-54.

Huang, H., Xu, P., Abdel-Aty, M., 2013. Transportation safety planning: a spatial analysis approach. Transportation Research Board 92nd Annual Meeting Compendium of Papers, 13-1855.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., & Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. J. Tran. Geogr. (54), 248-256.

Ivan, J.N., Deng, Z., Jonsson, T., 2006. Procedure for allocating zonal attributes to link network in GIS environment. Transportation Research Board 85th Annual Meeting (No. 06-2561) .

Khondakar, B., Sayed, T., Lovegrove, G.R., 2010. Transferability of community-based

collision prediction models for use in road safety planning applications. J.Transp Eng-Asce, 136 (10), 871-880.

Lee, J., Abdel-Aty, M., & Jiang, X., 2014. Development of zone system for macro-level traffic safety analysis. J. Tran. Geogr. 38, 13-21.

Lovegrove, G.R., 2007. Road safety planning: new tools for sustainable road safety and community development VDM-Verlag, Müller, Germany.

Lovegrove, G.R., Sun, J., 2010. Using community-based macrolevel collision prediction models to evaluate safety level of neighborhood road network patterns. Transportation Research Board 89th Annual Meeting Compendium of Papers DVD, 10-0535.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. 10(4), 325-337.

Peng, Y., Abdel-Aty, M., Shi, Q., & Yu, R., 2017. Assessing the impact of reduced visibility on traffic crash risk using microscopic data and surrogate safety measures. Transp. Res. C 74, 295-305.

Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of london crash data. Accid. Anal. Prev. 40, 1486-1497.

Schneider, R.J., Grembek, O., Braughton, M., 2013. Pedestrian crash risk on boundary roadways university campus case study. Transport. Res. Rec. 2393, 164-173.

Siddiqui, C., Abdel-Aty, M., 2012. Nature of modeling boundary pedestrian crashes at zones. Transport. Res. Rec. 2299, 31-40.

Siddiqui, C., Abdel-Aty, M., Choi, K., 2012. Macroscopic spatial analysis of pedestrian and

1 bicycle crashes. Accid. Anal. Prev. 45, 382–391.

2 Spiegelhalter, D., Best, N.G., Carlin, B.P., Linde, A.V.D., 2001. Bayesian measures of model

3 complexity and fit. J. R. Stat. Soc. 64 (4), 583-639.

4 Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2007. WinBUGS version 1.4.1 user manual.

5 MRC Biostatistics Unit, Cambridge University, United Kingdom, 2007.

6 Wakefield, J.C., Best, N.G., Waller, L., 2000. Bayesian Approaches to Disease Mapping, on

7 Spatial Epidemiology: Methods and Applications. Oxford University Press.

8 Waldo, T., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr.

9 46 (2), 234-240.

10 Wang, C., Quddus, M.A., Ison, S.G., 2009. Impact of traffic congestion on road accidents: a

11 spatial analysis of the M25 motorway in England. Accid. Anal. Prev. 41, 798-808.

12 Wang, X., Jin, Y., Abdel-Aty, M., Tremont, P., Chen, X., 2012. Macrolevel model development

13 for safety assessment of road network structures. Transport. Res. Rec. 2280, 100-109.

14 Washington, S.P., Van Schalkwyk, I., Mitra, S., Meyer, M., Dumbaugh, E., Zoll, M., 2006.

15 Incorporating safety into long-range transportation planning. In: NCHRP Report 546,

16 Transportation Research Board, Washington, DC.

17 Wei, F., 2010. Boundary Effects in Developing Macro-level CPMs: A Case Study of City of

18 Ottawa. University of British Columbia.

19 http://www. cite7org/scholarships_awards/documents/2010_StudentPaper_FengWei.pdf.

20 Xu, P., Huang, H., Dong, N., & Abdel-Aty, M. 2014. Sensitivity analysis in the context of

21 regional safety modeling: identifying and assessing the modifiable areal unit problem. Accid.

Anal. Prev. 70, 110-120.

Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: random parameter versus geographically weighting. Accid. Anal. Prev. 75, 16-25.

Xu, P., Huang, H., Dong, N., Wong, S.C., 2017. Revisiting crash spatial heterogeneity: a Bayesian spatially varying coefficients approach. Accid. Anal. Prev. 98, 330-337.

Xu, C., Wang, W., Liu, P., Guo, R., & Li, Z., 2014. Using the Bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models. Transp. Res. C 38(1), 167-176.

Yu, R., Abdel-Aty, M., 2013. Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions. Accid. Anal. Prev. 56, 51-58.

Yu, R., Abdel-Aty, M., 2014. An optimal variable speed limits system to ameliorate traffic safety risk. Transp. Res. C 46, 235-246.

Yu, R., Xiong, Y., Abdel-Aty, M., 2015. A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway. Transp. Res. C 50, 68-77.