Accident Analysis and Prevention 121 (2018) 231–237

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Boundary crash data assignment in zonal safety analysis: An iterative approach based on data augmentation and Bayesian spatial model

Xiaoqi Zhai[a], Helai Huang[a,*], Mingyun Gao[b], Ni Dong[c], N.N. Sze[d]

[a] School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan, China
[b] Department of Industrial and Business Management, Hunan University, Changsha, Hunan, China
[c] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China
[d] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

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

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.

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, 2000; 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

* Corresponding author.
E-mail address: huanghelai@csu.edu.cn (H. Huang).

https://doi.org/10.1016/j.aap.2018.09.010
Received 28 February 2018; Received in revised form 11 September 2018; Accepted 13 September 2018

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 (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

2.1.1. Step 1: divide zones into boundary and interior

Fig. 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.

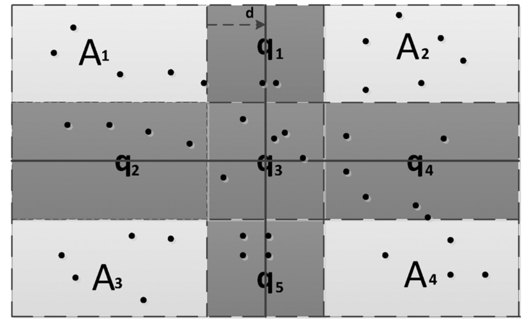

Fig. 1. The Structure of a Neighborhood: Boundary Zone and Interior Zone.

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

$$Y_{\backslash b} \cap Y_b = \varnothing \tag{1}$$

and

$$Y = Y_{\backslash b} \cup Y_b. \tag{2}$$

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}.

2.1.2. 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:

$$Y_i \sim Poisson(\lambda_i) \tag{3}$$

$$\log(\lambda_i) = \boldsymbol{\alpha} + \log(e_i) + \boldsymbol{x_i\beta} + \delta_i + \vartheta_i \tag{4}$$

where for zone i ($i = 1, 2, \cdots, N$), 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

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

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