Accident Analysis and Prevention 98 (2017) 330–337

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Revisiting crash spatial heterogeneity: A Bayesian spatially varying coefficients approach

Pengpeng Xu [a], Helai Huang [b,*], Ni Dong [c], S.C. Wong [a]

[a] Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, China
[b] School of Traffic & Transportation Engineering, Central South University, Changsha, Hunan, China
[c] School of Transportation & Logistics, Southwest Jiaotong University, Chengdu, Sichuan, China

ARTICLE INFO

Article history:
Received 19 June 2016
Received in revised form 8 September 2016
Accepted 11 October 2016

Keywords:
Crash frequency
Spatial heterogeneity
Unobserved heterogeneity
Conditional autoregressive prior
Bayesian inference

ABSTRACT

This study was performed to investigate the spatially varying relationships between crash frequency and related risk factors. A Bayesian spatially varying coefficients model was elaborately introduced as a methodological alternative to simultaneously account for the unstructured and spatially structured heterogeneity of the regression coefficients in predicting crash frequencies. The proposed method was appealing in that the parameters were modeled via a conditional autoregressive prior distribution, which involved a single set of random effects and a spatial correlation parameter with extreme values corresponding to pure unstructured or pure spatially correlated random effects.

A case study using a three-year crash dataset from the Hillsborough County, Florida, was conducted to illustrate the proposed model. Empirical analysis confirmed the presence of both unstructured and spatially correlated variations in the effects of contributory factors on severe crash occurrences. The findings also suggested that ignoring spatially structured heterogeneity may result in biased parameter estimates and incorrect inferences, while assuming the regression coefficients to be spatially clustered only is probably subject to the issue of over-smoothness.

1. Introduction

Modeling crash data involving contiguous spatial units, such as road networks and traffic analysis zones (TAZs), has gained growing research interests in the road traffic safety domain. This allows safety analysts to identify the clustering pattern of crashes, to better understand the factors that contribute to crash occurrences, and to recommend targeted countermeasures. Conventional crash prediction models, including the commonly used negative binomial and Poisson lognormal models, have an underlying assumption that their observations should be mutually independent. This fundamental requirement is almost always violated, because crash data collected in close proximity usually display spatial dependence (Quddus, 2008). The inclusion of spatially correlated effects typically has two main benefits. First, considering spatial correlation helps site estimates to pool strength from their neighbors, thereby improving model estimations (Aguero-Valverde and Jovanis, 2008). Second, spatial dependence can serve as a surrogate for unobserved covariates that vary smoothly over the region of interest (Cressie, 1993).

A range of spatial statistical techniques have been used to incorporate this spatial dependence into crash frequency modeling. The Bayesian hierarchical models are primarily used in these analyses, in which the spatial correlation is modeled via a set of random effects at the second level of hierarchy (Miaou et al., 2003; MacNab, 2004; Aguero-Valverde and Jovanis, 2006, 2008, 2010; Aguero-Valverde, 2014; Quddus, 2008; El-Basyouny and Sayed, 2009a; Mitra, 2009; Guo et al., 2010; Huang and Abdel-Aty, 2010; Siddiqui and Abdel-Aty, 2012; Flask and Schneider, 2013; Wang et al., 2013a, 2016; Xie et al., 2013; Dong et al., 2014, 2016; Xu et al., 2014; Zeng and Huang, 2014; Lee et al., 2015; Huang et al., 2016; Wang and Huang, 2016). This effect is mostly derived from the intrinsic conditional autoregressive (CAR) prior distribution proposed by Besag et al. (1991), which is a special case of Gaussian Markov random fields (Rue and Held, 2005). Alternative CAR specifications were also introduced by Richardson et al. (1992), Cressie (1993), and Leroux et al. (1999). Lee (2011) made a comprehensive comparison and concluded that the model of Leroux et al. (1999) was most appealing, as it performed consistently well in the presence of independence and strong spatial correlation.

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

http://dx.doi.org/10.1016/j.aap.2016.10.015
0001-4575/

Although most safety analysts have made an effort to handle the spatially correlated effects in model residuals, a limited number of studies have specifically focused on another issue related to the location dimension of crash data, i.e., spatial heterogeneity or spatial non-stationarity (Xu and Huang, 2015). Variables do not usually vary identically across space, and the relationship between crashes and related risk factors may not necessarily be constant or fixed across the study area. The possibility of accounting for this spatial heterogeneity by allowing some or all parameters to vary spatially holds considerable promise.

One possible method is the random parameters count-data models. Some of the many factors that influence crash occurrences are not observed or are nearly impossible to collect. If these unobserved factors were correlated with observed ones, biased parameters would be estimated and incorrect inference could be drawn (Mannering and Bhat, 2014). The random parameters approach has therefore been used to account for the unobserved heterogeneity in crash frequency (Anastasopoulos and Mannering, 2009; El-Basyouny and Sayed, 2009b, 2011; Dinu and Veeraragavan, 2011; Ukkusuri et al., 2011; Venkataraman et al., 2013; Barua et al., 2015, 2016). The regression coefficients in these random parameters models typically arise independently from some univariate distributions, and no attention is paid to the locations to which the parameters refer. This hypothesis may be inappropriate, particularly in cases where the unobserved factors are correlated over space (Xu and Huang, 2015). To capture this spatially structured variability in the effects of contributory factors, Xu and Huang (2015) advocated the development of a model based on the principle that the estimated parameters on a geographical surface are related to each other with closer values more similar than distant ones.

To address this potential spatial correlation in varying coefficients, two competing approaches are promising, i.e., the geographically weighted Poisson regression (GWPR; Fotheringham et al., 2002; Nakaya et al., 2005) and the Bayesian spatially varying coefficients (BSVC) models (Congdon, 1997, 2003; Assuncao et al., 2002; Gelfand et al., 2003). The geographically weighted approach is one of the most innovative techniques in geography and has become increasingly prevalent in spatial econometrics, ecology analysis and disease mapping (Yao et al., 2015a). The method is similar in spirit to local linear models, relying on the calibration of multiple regression models for different geographical entities. Recently published studies have empirically demonstrated the superiority of the GWPR model with a substantial improvement in model goodness-of-fit and the ability to explore the spatially varying relationships between crash counts and predicting factors (Hadayeghi et al., 2010; Li et al., 2013; Pirdavani et al., 2014a, 2014b; Shariat-Mohaymany et al., 2015; Xu and Huang, 2015; Yao et al., 2015b).

Another potential method is the BSVC. The BSVC model has long been emerging in statistics as a methodological alternative to examine the non-constant linear relationships between variables (Congdon, 1997). The varying coefficients in the BSVC model can be selectively modeled as the geostatistical (Gelfand et al., 2003), intrinsic CAR (Congdon, 1997; Assuncao et al., 2002), or multiple membership processes (Congdon, 2003). Such an approach fits naturally into the Bayesian paradigm, where all parameters are treated as unknown random quantities. Obviously, the BSVC model differs from the GWPR in that the former is a single statistical model specified in a hierarchical manner, whereas the latter is an assembly of local spatial regression models, each fits separately. Wheeler and Calder (2007) conducted a series of simulation studies to evaluate the accuracy of regression coefficients in these two types of models under the presence of collinearity. Their evidence suggested that the BSVC model produced more accurate and more easily interpreted inferences, thus providing more flexibility (Wheeler and Calder, 2007). However, to assume the regression coefficients to be spatially clustered only is a strong prior belief. In reality, spatial pooling with smoothly varying coefficients over contiguous areas may be implausible, especially when clear discontinuities exist (Congdon, 2014; p. 340). In this vein, a robust model with a mechanism to accommodate the global and local smoothing collectively would be preferable.

This study intends to investigate the spatially varying relationships between crash frequency and relevant risk factors using a fully Bayesian approach. To simultaneously determine the strength of the unstructured and spatially structured variations in model regression coefficients, the CAR prior distribution derived from Leroux et al. (1999) is elaborately extended to the spatially varying coefficients framework. The proposed method is illustrated based on a case study with a comprehensive dataset from Hillsborough County, Florida.

2. Methodology

We begin this section with a quick review of the fixed coefficients model commonly used for modeling spatially correlated errors in crash prediction. We then move on to detail how this basic model can be readily generalized to estimate the varying regression coefficients within a fully Bayesian context.

Let Y_i denote the observed number of crashes in location $i(i = 1, 2, \ldots, n)$, $\mathrm{EV_i}$ the exposure, and X_{ik} the kth$(k = 1, 2, \ldots, p)$ explanatory variable. On the basis of Huang and Abdel-Aty (2010), we have:

$$Y_i \sim \mathrm{Poisson}(\lambda_i)$$

$$\ln(\lambda_i) = \beta_1 + \beta_2 \ln(\mathrm{EV}_i) + \sum_{k=3}^{p} \beta_k X_{ik} + u_i + s_i \tag{1}$$

where λ_i is the parameter of the Poisson model (i.e., the expected number of crashes in site i); β_1 is the intercept; $\beta_k(k = 2, \ldots, p)$ refers to the kth regression coefficient to be estimated; u_i denotes the pure unstructured effect, which could be specified via an exchangeable normal prior, i.e., $u_i \sim \mathrm{N}(0, \sigma_u^2)$; and s_i is the spatially structured or spatially correlated error.

One widely used joint density for the spatial effects $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ is in terms of pairwise differences in errors and a variance term σ_s^2 (Besag et al., 1991):

$$P(s_1, s_2, \ldots, s_n) \propto \exp[-0.5(\sigma_s^2)^{-1} \sum_{i \sim j} c_{ij}(s_i - s_j)^2] \tag{2}$$

This joint density implies a normal conditional prior for s_i conditioning on the effect of s_j in the remaining observations:

$$s_i | s_{j \neq i} \sim \mathrm{N}\left(\frac{\sum_j c_{ij} s_j}{\sum_j c_{ij}}, \frac{\sigma_s^2}{\sum_j c_{ij}}\right) \tag{3}$$

where c_{ij} represents the non-normalized weight, e.g., $c_{ij} = 1$ if i directly connects with j, otherwise $c_{ij} = 0$ (with $c_{ii} = 0$); and σ_s^2 is the variance parameter, which controls the amount of extra variations due to spatial correlation. It is worth noting that this intrinsic CAR specification permits contiguity and distance-based weight matrices, but precludes the kth-nearest neighbor weighting scheme as such weights violate the symmetry condition.

Although the univariate conditional prior distribution in Eq. (3) is well defined, the corresponding joint prior distribution for $\mathbf{s}$ is now improper (i.e., undefined mean and infinite variance; Sun et al., 1999). This fact probably leads to problems in convergence and identifiability in Bayesian estimation (Eberly and Carlin, 2000).