ELSEVIER

Accident Analysis and Prevention 38 (2006) 618–625

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Spatial analysis of fatal and injury crashes in Pennsylvania

Jonathan Aguero-Valverde, Paul P. Jovanis *

Department of Civil and Environmental Engineering, Pennsylvania Transportation Institute, University Park, PA 16802-1408, USA

Received 15 July 2005; received in revised form 1 December 2005; accepted 12 December 2005

Abstract

Using injury and fatal crash data for Pennsylvania for 1996–2000, full Bayes (FB) hierarchical models (with spatial and temporal effects and space–time interactions) are compared to traditional negative binomial (NB) estimates of annual county-level crash frequency. Covariates include socio-demographics, weather conditions, transportation infrastructure and amount of travel. FB hierarchical models are generally consistent with the NB estimates.

Counties with a higher percentage of the population under poverty level, higher percentage of their population in age groups 0–14, 15–24, and over 64 and those with increased road mileage and road density have significantly increased crash risk. Total precipitation is significant and positive in the NB models, but not significant with FB. Spatial correlation, time trend, and space–time interactions are significant in the FB injury crash models.

County-level FB models reveal the existence of spatial correlation in crash data and provide a mechanism to quantify, and reduce the effect of, this correlation. Addressing spatial correlation is likely to be even more important in road segment and intersection-level crash models, where spatial correlation is likely to be even more pronounced.

Keywords: Full Bayes hierarchical model; Spatial correlation; Negative binomial model; Crash risk; Weather conditions and crash risk

1. Introduction

Many factors affecting crashes operate at a spatial scale (e.g. land-use policy, demographic characteristics and highway infrastructure functional class). It is therefore reasonable to explore the use of spatial models of crash occurrence to better understand the implications of these policies.

In most roadway accident studies, crashes are grouped in spatial units that range from intersection or road section level to zip code or county level (e.g. Amoros et al., 2003; Miaou et al., 2003; Noland and Oh, 2004; Noland and Quddus, 2004; MacNab, 2004). One concern with these studies is the effect of spatial correlation (i.e. the spatial dependence among observations), which produces higher variance of the estimates and therefore, underestimated standard errors.

Recent developments in spatial modeling techniques have enabled researchers to investigate important issues related to risk estimation, unmeasured confounding variables, and spatial dependence (Richardson, 1992). An important advantage of spatial models is that spatial effects may reflect unmeasured confounding variables. This is particularly useful for unmeasured confounders that vary in space like weather, population, and others. More important yet, "the methods also facilitate spatial smoothing and data pooling when regions under investigation involve small-population areas", MacNab (2004). Here the term 'small-population areas' refers to areas that present very few events, given a rare-event phenomenon, for example roadway crashes.

Previous research has dealt with the spatial component of road crashes in different ways. Crashes have been modeled as point events (e.g. Levine et al., 1995; Jones et al., 1996), while others have modeled road crashes at different area levels, ranging from road sections to local census tracks or counties (e.g. Shankar et al., 1995; Amoros et al., 2003; Miaou et al., 2003; Noland and Oh, 2004; MacNab, 2004).

Honolulu census tract data have been used (Levine et al., 1995b) in a continuous model for predicting crashes. Analysis at the "ward" (census track) level has been conducted (Noland and Quddus, 2004) for fatalities, serious injuries, and slight injuries using four different categories of predictor variables: land-use indicator variables (employment and population density), road characteristics, demographic characteristics (age cohorts), and

* Corresponding author. Tel. +1 814 865 9431.
E-mail address: ppj2@engr.psu.edu (P.P. Jovanis).

0001-4575/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2005.12.006

traffic flow proxies (proximate and total employment). Country-level data for Illinois (Noland and Oh, 2004) were used to estimate the expected number of crashes using infrastructure characteristics and demographic indicators as independent variables in a negative binomial (NB) model. Limitations of these studies are the use of proxy variables for traffic flow estimation and the lack of spatial correlation analysis. An additional paper (Amoros et al., 2003) developed NB models at county level in France that included interactions between road type and county.

Poisson-based full Bayes (FB) hierarchical models of county-level fatal (K), incapacitating (A), and non-incapacitating (B) injuries were estimated using both frequency and rate for the state of Texas (Miaou et al., 2003). Conditional auto-regressive model (CAR) was used to model spatial correlation and Markov Chain Monte Carlo (MCMC) was used to sample the posterior probability distribution. The main limitation of this paper is the use of the surrogate variables: percent of time that the road is wet, sharp horizontal curves, and roadside hazards. These predictor variables were estimated by proportions of crashes. For example, for percent of time that the road is wet, the variable was estimated by dividing the number of crashes that occurred under wet pavement by the total number of crashes. These estimators are clearly biased in the direction of the effect. Given the poor definition of contributing factors in the model, it is likely that the spatial correlation is overestimated. In a recent paper, Miaou and Song (2005) used the same approach and data in the ranking of sites for engineering safety improvements.

The adoption of the FB hierarchical approach by Miaou is an important advance in model estimation and is a departure point for this paper. The purpose of this research is to develop spatial models of road crash frequency for the State of Pennsylvania at the county level while controlling for socioeconomic, transportation-related, and environmental factors. The results from FB hierarchical spatial models are compared with the more traditional approach using an NB distribution to model crash frequency. Particular attention is paid to the inclusion of weather as a predictor and the search for spatial correlation among neighboring counties.

2. Methodology

2.1. The Poisson and negative binomial distributions

When data arise as counts, the Poisson distribution is typically used to model them. Traffic crashes are a clear example of count data, therefore, a Poisson distribution is a useful stating point (see for example Jovanis and Chang, 1986; Shankar et al., 1995). An important characteristic of the Poisson distribution is that its variance is equal to its mean. Several authors (e.g. Shankar et al., 1995; Noland and Quddus, 2004) have argued that vehicle crashes are better represented by an NB distribution, which is a count distribution generated by a Poisson process with variance greater that the mean (see for example, Hamed et al., 1998; Hamed,1999).

Several goodness-of-fit measures have been proposed for this kind of model including the Poisson R^2, R^2_{P}, and the Freeman–Tukey R^2, R^2_{PFT} (Fristrøm et al., 1995). Another measure of goodness-of-fit uses the overdispersion parameter α of the NB model (Miaou, 1996). NB models are estimated using R statistical Software (R Development Core Team, 2004).

2.2. Spatial modeling using full Bayes hierarchical approach

Many spatial modeling techniques may be developed within a Bayesian approach because of its flexibility in structuring complicated models, inferential goals, and analysis (Miaou et al., 2003). Bayesian inference has been used in the past in disease mapping and ecological analysis and just recently, it has been applied to crash modeling (e.g. Miaou et al., 2003; MacNab, 2004; Miaou and Song, 2005). For a detail description of Bayesian inference, see Gelman et al. (2003).

The problem of group estimation, namely estimating the parameters of a common distribution thought to underlay a collection of outcomes for similar types of units, has motivated much research in Bayesian statistics. One seeks to make conditional estimates of the true outcome rate in each unit of observation (e.g. fatal crashes rate by county), given the parameters of the common density. Such estimation for sets of similar units is known as 'hierarchical modeling' because of its conditioning on higher stage densities (Congdon, 2003). At the first stage, the observed counts are modeled as a function of area-level summaries such as risk or relative risk. At the second stage, a joint distribution is specified for the collection of these risks as a function of explanatory variables. The second stage distribution depends on unknown parameters and these are assigned a (hyper) prior distribution at the third stage (Wakefield et al., 2000).

In the case of this study, the model developed by Besag et al. (1991) is the base of the formulation used, as shown in Eqs. (1) and (2):

$$y_i \sim \mathrm{Poisson}(e_i\theta_i) \tag{1}$$

where y_i is the number of fatal crashes in county i, θ_i the risk in county i, and e_i the exposure in county i; in this case the exposure is the total daily vehicle-miles traveled (DVMT) by county. DVMT was also included as explanatory variable to account for possible non-linearity between crash frequency and DVMT. This is the first stage in the model. The log risk is modeled as:

$$\log(\theta_i) = \boldsymbol{\alpha} + \boldsymbol{x_i'\beta} + v_i + u_i \tag{2}$$

where x_i represents a vector of explanatory variables, or covariates, $\boldsymbol{\beta}$ a vector of fixed effect parameters, v_i the uncorrelated heterogeneity or unstructured error and u_i the correlated heterogeneity or spatial correlation. The last two variables are known as random effects, therefore, this kind of model is commonly known as a mixture model where fixed and random effects are combined.

In the third stage, a uniform prior distribution is assigned to α and a highly non-informative normal distribution is assigned to the $\boldsymbol{\beta}$'s with mean 0 and variance 1000 corresponding to vague prior beliefs, given the scale of the covariates. On the other hand