Accident Analysis and Prevention 40 (2008) 1486–1497

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data

Mohammed A. Quddus*

Transport Studies Group, Department of Civil and Building Engineering, Loughborough University, Leicestershire LE11 3TU, United Kingdom

ARTICLE INFO

Article history:
Received 30 October 2007
Received in revised form 20 March 2008
Accepted 24 March 2008

Keywords:
Spatial correlation
Uncorrelated heterogeneity
Negative binomial models
Classical spatial models
Bayesian hierarchical models

ABSTRACT

Count models such as negative binomial (NB) regression models are normally employed to establish a relationship between area-wide traffic crashes and the contributing factors. Since crash data are collected with reference to location measured as points in space, spatial dependence exists among the area-level crash observations. Although NB models can take account of the effect of unobserved heterogeneity (due to omitted variables in the model) among neighbourhoods, such models may not account for spatial correlation areas. It is then essential to adopt an econometric model that takes account of both spatial dependence and uncorrelated heterogeneity simultaneously among neighbouring units. In studying the spatial pattern of traffic crashes, two types of spatial models may be employed: (i) classical spatial models for higher levels of spatial aggregation such as states, counties, etc. and (ii) Bayesian hierarchical models for all spatial units, especially for smaller scale area-aggregations. Therefore, the primary objectives of this paper is to develop a series of relationships between area-wide different traffic casualties and the contributing factors associated with ward characteristics using both non-spatial models (such as NB models) and spatial models and to identify the similarities and differences among these relationships. The spatial units of the analysis are the 633 census *wards* from the Greater London metropolitan area. Ward-level casualty data are disaggregated by severity of the casualty (such as fatalities, serious injuries, and slight injuries) and by severity of the casualty related to various road users.

The analysis implies that different ward-level factors affect traffic casualties differently. The results also suggest that Bayesian hierarchical models are more appropriate in developing a relationship between area-wide traffic crashes and the contributing factors associated with the road infrastructure, socioeconomic and traffic conditions of the area. This is because Bayesian models accurately take account of both spatial dependence and uncorrelated heterogeneity.

1. Introduction

The primary objective of this study is to develop a series of relationships (i.e., crash prediction models) between area-level traffic casualties and their contributing factors using both non-spatial (such as negative binomial models) and spatial models (such as traditional spatial models and Bayesian hierarchical models) and to compare the results obtained from these models. The spatial units of this analysis are the 633 census *wards* from London. Ward-level crash data are disaggregated by severity of the casualty such as fatalities, serious injuries and slight injuries and by severity of the casualty associated with various road users such as motorised transport (MT), non-motorised transport (NMT) and vulnerable road user (VRU). A range of potential contributing factors associated with ward-level road infrastructure, traffic, socioeconomic characteristics including traffic speed, flow and road curvature are considered in this study.

Crash prediction models to explain observed cross-sectional variations in crash counts using macro-structural covariates at various levels of area-aggregation (e.g., states, counties, other census tracts, etc.) are becoming a fairly routine component in crash research. Researchers usually seek to establish links between the road infrastructure, environmental, traffic, and socioeconomic conditions in spatial units with the counts or rates of traffic crashes observed at various spatial units. To isolate and identify the macro-processes leading to different types of crashes, researchers sometimes estimate crash models with disaggregated crash rates with varying bases for the disaggregation such as by severity of the casualty (such as fatalities, serious injuries, and slight injuries) or by severity of the casualty related to various road users (e.g., motorised transport, vulnerable road users, etc.).

For instance, Levine et al. (1995a) derived a series of statistics that provide explicit measurements of a spatial pattern of crashes and also provide insights into how certain relationships

* Tel.: +44 1509 22 8545; fax: +44 1509 22 3981.
E-mail address: m.a.quddus@lboro.ac.uk.

0001-4575/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2008.03.009

(for instance, between alcohol consumption and injury severity) have a spatial dimension. Honolulu census tract data have been utilised to explain spatial variation in motor vehicle crashes (Levine et al., 1995b). Kim and Yamashita (2002) conducted an empirical analysis of motor vehicle crashes and land use variables with the aid of geographic information systems (GIS). Hadayeghi et al. (2003) developed a series of macro-level crash prediction models that estimate the number of crashes in planning zones in the city of Toronto as a function of zonal characteristics. Graham and Glaister (2003) analysed ward-level (a census tract) traffic casualty data in England to see how urban scale, density and land use mix affect pedestrian casualties. Noland and Quddus (2004a) also conducted a spatially disaggregated ward-level analysis of England to identify various factors affecting road casualties and were based on cross-sectional traffic crash data associated with different levels of spatial aggregation, most of these above studies were employed a negative binomial (NB) count model. NB models have also been used to develop crash models for cross-sectional time-series data (e.g., Amoros et al., 2003; Noland and Quddus, 2004b[1], and Noland and Oh, 2004[2]). Integrating NB count models with geographic information systems (GIS), Kim et al. (2006) established the nature and magnitude of relationships between land use, population, economic development and crashes using a uniform 0.1 square mile grid structure from Hawaii. Their study confirmed the finding of Ladron de Guevara et al. (2004) that population-based metrics by spatial units are the most statistically significant predictors of crash occurrences. Kim et al. (2006) recommended the use of a spatial statistical analysis when developing relationships between area-wide land use variables and traffic crashes.

Crash data are collected with reference to location measured as points (x- and y-coordinates) in space. According to LeSage (1998), two problems arise when sample data has a locational dimension: (1) spatial correlation exists between the observations, and (2) spatial heterogeneity occurs in the relationships that are modelled. Traditional econometrics (including NB models used in crash research) has largely ignored the issue of spatial correlation that violates the traditional Gauss–Markov assumptions used in regression modelling.

An alternative approach is to employ spatial econometric models. Anselin (1988) provides a complete treatment of many aspects of spatial econometrics including the application of Bayesian methods in spatial econometrics. There are generally two methods in spatial econometrics: (1) traditional econometric methods suitable for continuous data, and (2) Bayesian hierarchical methods suitable for non-negative random count data. At higher levels of spatial aggregation (e.g., districts, counties, states), when the number of counts (e.g., crashes) is sufficiently large and non-zero counts are observed in most of the sampled spatial units, the count outcomes may be considered continuous, and traditional spatial analytical methods have been utilised (Messner et al., 1999; Baller et al., 2001).

However, Bhati (2005) indicated that inferences derived from traditional spatial models could be misleading as this does not reflect the true underlying data generating processes. Moreover, as the spatial unit of analysis becomes smaller (such as wards, zip-code, post-code, etc.), the number of count outcomes observed in each sampled unit decreases and the distribution of such counts becomes a highly skewed (to the right) distribution as the number of spatial units with zero counts increases. In order to overcome these issues, researchers used a more flexible Bayesian method in spatial econometrics (Besag et al., 1991; Mollie, 1996; Wolpert and Ickstadt, 1998; Best et al., 2000) and the application of such methods to crash modelling can be found in Miaou et al. (2003), MacNab (2004), Aguero-Valverde and Jovanis (2006) and Li et al. (2007). Miaou et al. (2003) provides a good overview on the appropriateness of employing a Bayesian hierarchical model in area-wide crash modelling.

Different area-wide characteristics were considered in previous research while developing a crash prediction model using either a non-spatial model (such as an NB model) or a spatial model. These include factors associated with land use (e.g., Graham and Glaister, 2003; Kim et al., 2006), road characteristics such as road length, junctions and roundabouts (e.g., Noland and Quddus, 2004a), environmental conditions such as total precipitation, number of rainy days per year and total snowfall (e.g., Aguero-Valverde and Jovanis, 2006), and various socioeconomic factors such as population, poverty, and employment (e.g., Kim et al., 2006). Some other important area-wide characteristics that affect area-wide traffic crashes are less considered in the literature. These include traffic speed, traffic flow, and road curvature measured at spatial-level.

The rest of the paper is structured as follows. In the next section, a brief discussion of the data used in the analysis is presented. This is followed by a description of the models considered in this research. Then, the results obtained from the developed models are presented with the similarities and differences among them. Finally, conclusions are drawn and further research suggestions are discussed.

2. Data

The spatial units of the analysis are the census wards from the Greater London metropolitan area. According to the UK Census 2001, there are 633 *wards* in London and each ward consists, on average, of about 11,350 resident population. The electronic ward boundary data were obtained from UK Ordnance Survey (OS) data via EDINA services. Data on traffic casualties were extracted from the STATS19 UK National Road Crash Database that has information on the recorded location of each crash. Ward-level casualty data were extracted from the STATS19 data using a GIS technique. Since previous research suggests that factors affecting traffic casualties vary by severity of the casualty (e.g., Noland and Quddus, 2004a; Aguero-Valverde and Jovanis, 2006), ward-level traffic casualty data were disaggregated into fatalities, serious injuries and slight injuries. Ward-level casualty data were also disaggregated by severity of the casualty associated with motorised transport (MT[3]), non-motorised transport (NMT[4]), and vulnerable road users (VRU[5]) to identify any differences in influential factors. Since a large number of wards have a zero fatality count, fatalities are combined with serious injuries resulting in a killed or serious injury (KSI) category. Casualty data were aggregated for 3 years of data, 2000–2002. Fig. 1 shows the spatial distribution of total serious casualties (for years 2000–2002). It is noticeable from this figure that traffic casualties are spatially correlated among neighbouring wards.

There are three major categories of explanatory variables: (1) traffic characteristics, (2) road characteristics and (3) socio-demographic factors. Environmental factors are not considered in this study as weather conditions such as snowfall and rainfall tend to be similar across different wards in London.

[1] They used a random effects NB model.

[2] They used a fixed effects NB model.

[3] Cars, taxi, bus, goods vehicles, and other motor vehicles (motorcycles are not included).

[4] Pedestrians, cyclists, horse riders.

[5] Motorcyclists, pedestrians, cyclists, horse riders.