Analytic Methods in Accident Research 13 (2017) 28–51

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models

Richard Amoh-Gyimah [a], Meead Saberi [a,*], Majid Sarvi [b]

[a] Institute of Transport Studies, Department of Civil Engineering, Monash University, Australia
[b] Department of Infrastructure Engineering, The University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 29 April 2016
Received in revised form 21 November 2016
Accepted 21 November 2016
Available online 21 January 2017

Keywords:
Unobserved heterogeneity
Random parameter negative binomial model
Semi-parametric geographically weighted Poisson regression
Macro-level crash prediction model

ABSTRACT

Macroscopic safety models establish a relationship between crashes and the contributing factors in a defined spatial unit. Negative binomial (NB) and Bayesian negative binomial models with conditional autoregressive prior (CAR) are techniques widely used to establish this relationship. However, these models do not account for unobserved heterogeneity and their output is global and fixed irrespective of the spatial unit of the analysis. There is a timely need to understand how variations in spatial units affect unobserved heterogeneity. This study uses two advanced modeling techniques, the random parameter negative binomial (RPNB) and the semi-parametric geographically weighted Poisson regression (S-GWPR), to investigate whether explanatory variables found to be significant and random in one spatial aggregation will remain significant and random when another spatial aggregation is used. The key finding is that variations in spatial units do have an impact on unobserved heterogeneity. We also found that variations in spatial units have a greater impact on unobserved heterogeneity in the RPNB models compared to the S-GWPR models. We found that the S-GWPR model performs better than the RPNB model with the lowest value of mean absolute deviation (MAD) and Akaiki information criterion (AIC) but the two modeling techniques produce similar results in terms of the sign of the coefficients across the selected spatial units of analysis. Overall, the study provides a methodological basis for assessing the impact of spatial units on unobserved heterogeneity.

1. Introduction

A considerable amount of research has been devoted to developing models to predict the number of crashes at the micro level and individual elements of the urban transportation network such as intersections and road segments. However, the use of planning level or macro-level crash prediction models is fairly new and becoming increasingly popular in safety research. The purpose of these models is to explain the observed cross-sectional variations in crash counts (Quddus, 2008; Xu et al., 2014) that could be integrated as part of the transportation planning process (Hadayeghi et al., 2003; Washington et al., 2006). The first step in macro-level crash prediction modeling is to aggregate the available data in a defined spatial or geographic entity. In the macro-level crash prediction models, the relationship between crash counts or crash rate and explanatory factors such as socioeconomic, land use, demographic and network characteristics are then established in a defined spatial unit.

* Corresponding author.
E-mail address: meead.saberi@monash.edu (M. Saberi).

http://dx.doi.org/10.1016/j.amar.2016.11.001

Crash modeling studies in the past have used different levels of spatial aggregation. A review of the literature shows that most studies rely on the geographical hierarchies used for planning purposes of the study location. A number of studies used traffic analysis zones (TAZ) (Abdel-Aty et al., 2011, 2013; Duddu and Pulugurtha, 2012), census tracts (Wier et al., 2009; Abdel-Aty et al., 2013) and census wards (Quddus, 2008; Dissanayake et al., 2009; Wang et al., 2009), as common units of analysis. Other units of analysis include block groups (Levine et al., 1995), statistical area levels (Amoh-Gyimah et al., 2016a, 2016b), counties (Huang et al., 2010; Li et al., 2007), enumeration districts (Priyantha Wedagama et al., 2006) and states (Noland, 2003). Other studies including Kim et al. (2006) used a uniformly sized and shaped grid cell of about 0.259-km^2 in Hawaii, whereas MacNab (2004) used local health areas in British Columbia, Canada as their unit of analysis. However, very few studies have explored how different spatial units affect modeling results when the same dataset is used. Abdel-Aty et al. (2013) investigated the effect of three spatial units (i.e., TAZs, block groups and census tracts) on the performance of three models; total crashes, severe crashes and pedestrian crashes. Using the Bayesian Poisson lognormal models, they drew two important conclusions. The first was that the signs of coefficients are consistent if the variables are significant in models with the same response variables, even if geographical units are different. The second conclusion was that both response variables and geographical units affect the number of significant variables.

Building upon the study of Abdel-Aty et al. (2013), a more recent study by Xu et al. (2014) conducted a sensitivity analysis to quantitatively investigate the effect of the modifiable area unit problem (MAUP) in the context of regional safety modeling. This was done by aggregating 738 TAZs in the county of Hillsborough to 14 zoning schemes at an incremental step-size of 50 zones based on spatial homogeneity of crash risk. They estimated the Bayesian Poisson lognormal and the Bayesian spatial models to explain the observed variations in both total and severe injury crashes at each level of aggregation. They found that zoning schemes with a larger number of zones tend to have an increasing number of significant variables, more stable coefficient estimation, and smaller standard error.

The two studies summarized above provide significant insights into the effects of zonal variation on crash models. However, the models used in these studies overlooked the issue of unobserved heterogeneity. The effects of observable variables were restricted to be the same across all observations (global) and their outputs consisted of a set of fixed parameter estimates. However, constraining parameters to be fixed when they actually vary across observations could lead to inconsistent and bias parameter estimates (Washington et al., 2010; Mannering et al., 2016). If parameters vary across observations, then the issue of unobserved heterogeneity occurs.

To account for heterogeneity across observations by allowing some or all parameters to vary (across observations), a number of heterogeneity models have been employed by researchers. One of the earlier models that account for unobserved heterogeneity is the random effects model which has been applied in studies by Kim et al. (2007) and Heydari et al. (2014). The random effects models assume fixed parameters associated with the covariates but varying intercept or error term. However, they typically necessitate panel data which makes it impossible to be extended to cross-sectional data (Mannering et al., 2016). Other researchers have adopted random parameter models to overcome the issue of unobserved heterogeneity. Random parameters models allow model covariates to vary across groups of observations to account for cross-group heterogeneity in data (Yannis et al., 2008; Islam and El-Basyouny, 2015; Heydari et al., 2016). One advantage of random parameter models is that, they can be estimated even with cross-sectional data as well as panel data. Random parameters models also constitute a more comprehensive way of overcoming unobserved heterogeneity in crash data in comparison to random effects models (Heydari et al., 2016). There are several other models proposed in the literature to address unobserved heterogeneity. Some of these models include finite mixture (latent-class) models (Eluru et al., 2012; Xiong and Mannering, 2013; Yasmin and Eluru, 2013; Yasmin et al., 2014; Cerwick et al., 2014; Shaheed and Gritza, 2014; Behnood et al., 2014), finite mixture (latent-class) random parameters model (Xiong and Mannering, 2013), Markov switching models (Malyshkina and Mannering, 2009; Xiong et al., 2014), Markov switching with random parameters (Xiong et al., 2014) and bivariate/multivariate models with random parameters (Abay et al., 2013; Russo et al., 2014).

The above-mentioned models have been applied to address the issue of unobserved heterogeneity by explicitly accounting for the variations in the effects of variables across road segments and intersections (Xu and Huang, 2015). However, the application of the above-mentioned models at the macro-level where data is aggregated at different spatial unit has been limited. To account for unobserved heterogeneity at the macro-level in a recent study by Xu and Huang (2015), two advanced approaches were used: (1) random parameter negative binomial (RPNB) and (2) semi-parametric geographically weighted Poisson regression (S-GWPR) model. The advantage of these two approaches is their ability to capture unobserved heterogeneity and to provide parameter estimates that are random across zones. However, the intention of their study was not to understand the effect of variations in spatial units on unobserved heterogeneity.

Therefore, the main aim of our study is to investigate whether variation in spatial units has any impact on unobserved heterogeneity in macroscopic crash models. The research question of interest is whether parameters that are found to be significant and random in one spatial unit will remain significant and random when a different spatial aggregation is used with the same dataset. Two advanced modelling approaches, RPNB and S-GWPR, are used to develop three different models: total crashes, serious injury crashes, and minor injury crashes consistent with Xu and Huang (2015). These models are developed in six different spatial units and the performance of the selected models in the various spatial units is investigated. Four groups of explanatory variables are considered; demographic & socio-economic, network characteristics, traffic exposure, and land use. The generalized linear model was used as the basis upon which both the RPNB and S-GWPR models were built. The mean absolute deviation (MAD) and the Akaiki information criterion (AIC) are computed to assess the goodness-of-fit of the models across the selected spatial units.