Accident Analysis and Prevention 119 (2018) 290–306

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

An empirical evaluation of multivariate spatial crash frequency models

Wen Cheng[a,*], Gurdiljot Singh Gill[a], Mohan Dasu[b], Xudong Jia[a]

[a] Department of Civil Engineering, California State Polytechnic University, Pomona 3801 W. Temple Ave., Pomona, CA 91768, United States
[b] California Department of Public Health, Sacramento, CA, United States

ARTICLE INFO

Keywords:
Multimodal
ANOVA
Spatiotemporal
Interaction
Spatial weight matrices
Multivariate

ABSTRACT

Many studies have employed spatial, temporal, or a combination of both specifications for analysis of roadway crashes at different spatial levels. However, there is lack of a comprehensive study which compares the crash estimation performance of different spatial weight matrices and their combination with various temporal treatments. The current study fills the research gap by comparing different Full Bayesian (FB) multivariate spatiotemporal crash models. The pedestrian and bicyclist crash data across an eight-year period for 58 counties in California were used as a case study. Three groups of models were developed based on temporal treatment, where each group comprised of 17 models differing on the basis of different adjacency- and distance-based spatial weight matrices. The first group of multivariate models incorporated only unstructured random error term and spatially structured conditional autoregressive (CAR) term. The second group built upon the former and introduced a linear time trend to develop a spatiotemporal model, while the third group allowed the interaction of space and time. The predictive performance of the alternate models across and within groups was assessed by employing several evaluation criteria.

The modeling results demonstrated the robustness of models based on the similar signs and closeness of coefficients for the posterior estimates of parameters. For overall model comparison, the pure-distance model D0.5 demonstrated the best performance for different evaluation criteria based on training and test errors across three groups. The variability in performance of other distance models suggested that caution must be exercised for the choice of exponents. The correlation analysis revealed the presence of positive correlations among the criteria based on training errors, as well as with cross-validation. However, a very strong positive correlation was observed between the criteria based on effective number of parameters and posterior deviance, indicating that an increased number of parameters may not lead to improved model fit. This finding reinforced the importance of selecting the optimum weight matrix for spatial correlation as a more complex structure may not lead to expected advantages at model performance. For comparison among three groups of different temporal treatments, the third group demonstrated the best performance and conveyed the benefits of incorporating the spatial and temporal interaction. The results from ANOVA (analysis of variance) and HSD (Honest Significant Differences) tests also established the existence of statistical differences for the superiority of space-time interactions models. However, the box and whisker plots demonstrated high variability among the models of the third group, suggesting that some models may not benefit from interaction term. For comparison among adjacency- and distance-based models, the distance-based models were mostly observed to be superior. However, the greater variability of model performance associated with distance-based models suggested for careful consideration during their selection. Additionally, it is important to note that the results observed in this study are specific to the county-level crash data of California. As such, the study does not recommend generalization of the results for extension to other spatial levels of roadway network, and readers and future research studies are advised to exercise caution before implementing the models.

1. Introduction

Majority of safety research studies have focused on the analysis of crash counts for a single transportation mode (such as motorist, pedestrian or cyclist crash count) aggregated at a certain spatial level of roadway entities (such as segments, intersections, cities, and so on). Such a conventional approach has demonstrated adequate performance for investigation of influential factors pertaining to only the interested

* Corresponding author.
E-mail addresses: wcheng@cpp.edu (W. Cheng), gurdiljotg@cpp.edu (G.S. Gill), rkmdasu@gmail.com (M. Dasu), xjia@cpp.edu (X. Jia).

https://doi.org/10.1016/j.aap.2018.07.001
Received 28 March 2018; Received in revised form 11 May 2018; Accepted 1 July 2018
0001-4575/

crash mode. However, some studies observed that different crash modes (or crash types) share some influential features, which may not be captured by the conventional method. Hence, a multivariate model structure is recommended to improve the model performance by accounting for the unobserved heterogeneity by jointly estimating multiple dependent variables (crash modes or types). Some studies also stated that aggregation of crash counts at a spatial level requires the accommodation of spatial correlation structures to capture shared features across different roadway entities and generate more precise estimates (Song et al., 2006). Similarly, many studies have employed different temporal treatments to address the serial correlations within crash data across a time period (Andrey and Yagar, 1993). Moreover, some studies observed that crash data tend to be clustered both spatially and temporally (Aguero-Valverde and Jovanis, 2006) and hence adopted the spatiotemporal models which allowed the interaction of space and time.

Albeit the safety literature has extensively employed multiple approaches to accommodate the spatial and temporal correlation structures, a study focused on the comparison of different spatial approaches and their combination with temporal treatments is lacking. To this end, the primary objective of the current study is to conduct the comprehensive comparison of different Full Bayesian (FB) multivariate spatiotemporal crash models at the macro-level (County). The crash data for active modes of transportation (bicyclist and pedestrian) across an eight-year period for 58 counties in California are used. Three groups of models are developed based on temporal treatment, where each group comprises of 17 spatial models of different adjacency- and distance-based spatial weight matrices. The first group of multivariate models incorporates only unstructured random error term and spatially structured conditional autoregressive (CAR) term. The second group builds upon the former and introduces a linear temporal trend while the third group allows the interaction of space and time. The crash estimation performance of the alternative models across and within groups is assessed by employing several evaluation criteria. The results of this study are expected to provide new insights into the performance of different spatial weight matrices and their association with temporal treatments. The contributions of this study to the literature are further discussed in the following sections.

2. Literature review

The traditional statistical analyses may be prone to serious specification problems due to ignorance of the unobserved heterogeneity which may lead to inaccuracies in predicted estimates and inferences (Mannering and Bhat, 2014; Mannering et al., 2016). The safety literature has employed different approaches to account for the unobserved heterogeneity and minimize the erroneous estimation of models (Aguero-Valverde et al., 2016). This section discusses these efforts and then highlights the contributions of the present study.

2.1. Spatial

In the past decade, it has been clearly established that spatial correlation should be incorporated in the development of crash frequency models since the roadway entities under investigation share common features due to close spatial proximity (Quddus, 2008). Such correlation structures “pool strength” from neighboring sites to improve the accuracy of posterior estimates (Lord and Mannering, 2010). Many studies have been undertaken to address the unobserved heterogeneity by accommodating the spatial correlations at different level of roadway network such as intersections (Wang and Abdel-Aty, 2006), segments (Aguero-Valverde and Jovanis, 2008), corridors (Wang and Abdel-Aty, 2006; Guo et al., 2010), census tracts (MacNab, 2004), traffic analysis zones (Washington et al., 2010), counties (Miaou et al., 2003), and so on. All such studies revealed the significance of incorporating spatial correlations for better model fit and crash estimation performance.

Among a plethora of studies employing spatial correlations under various approaches, very few studies have attempted to focus on the comparison of different spatial weight matrices and evaluate their advantages for model fit and crash estimation performance. At the micro-level, Aguero-Valverde and Jovanis (2008) compared three adjacency-based spatial weight matrices to develop crash frequency models based on segment data. The three models differed on the number of neighbors considered: first-order, first-second-order, and first-second-third-order, where the weights were assigned as inverse of the specific order. Subsequently, Aguero-Valverde and Jovanis (2010) explored the effects of different neighboring structures of varying complexity including exponential decay; adjacency-based, adjacency-route information; and distance order structures. The results demonstrated superior model fit for distance order models. At the macro-level of TAZ, Dong et al. (2014) evaluated four alternative spatial-proximity structures: 0–1 first order, adjacency-based; common boundary-length-based; geometric centroid-distance-based; and crash-weighted centroid-distance-based, and observed superior performance of boundary length model. Another TAZ study by Wang et al. (2016) adopted a few models from Dong et al. (2014) and compared them with three different spatial features: land use type, adjacency-based; land use intensity, adjacency-based; and geometric centroid-distance-order (GCO), where the GCO approach was designated as the best. Recently, a county-level study by Gill et al. (2017b) proposed two distance-based gravity models which accounted for the population and DVMT between neighboring counties. The model evaluation results for comparison between different weight matrices revealed the superiority of the exponential decay model and observed the presence of a strong correlation between model fit and site ranking performance.

The above studies employed and compared different spatial structures, however, the crash frequency models were developed for analysis of one crash type or crash mode (univariate) and focused on the spatial correlations while ignoring temporal correlation, whose advantages are discussed in the subsequent Sections 2.3 and 2.4.

2.2. Multivariate spatial

Similar to the spatial specification which allows incorporation of structured correlations among crashes at neighboring sites to address the unobserved heterogeneity, many studies observed the presence of unstructured correlations among different crash types (or severity levels, or crash modes). It was noted that different crash types share unaccounted factors and the ignorance of such interdependence may lead to erroneous inferences from model estimates (Congdon, 2001; Bijleveld, 2005; Park and Lord. 2007). Following the recommendation, many studies addressed the unobserved heterogeneity by adopting the multivariate specification, which allowed for simultaneous estimation of different crash types, and observed superior model performance (Miaou and Song, 2005; Barua et al., 2014; Mothafer et al., 2016; Anastasopoulos, 2016; Serhiyenko et al., 2016). Given the benefits associated with joint modeling and spatial random effects, some studies combined the two specifications and obtained more precise model estimates than the models which only adopted univariate, multivariate, or spatial specification (Aguero-Valverde, 2013; Aguero-Valverde et al., 2016; Barua et al., 2016; Gill et al., 2017a).

2.3. Spatiotemporal

Apart from the spatial random effects, accommodation of serial correlations has also been found to enhance the model fitness and precision by numerous research studies focused on vehicle crashes (Andrey and Yagar, 1993; Hay and Pettitt, 2001; Wang et al., 2013; Cheng et al., 2017a, b). Building on the advantages of spatial and temporal correlation structures to address the issue of unobserved heterogeneity, some studies further extended the benefits by developing spatiotemporal models which accounted for the spatial as well as