Accident Analysis and Prevention 112 (2018) 84–93

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Bayesian spatiotemporal crash frequency models with mixture components for space-time interactions

Wen Cheng[a,*], Gurdiljot Singh Gill[a], Yongping Zhang[a], Zhong Cao[b]

[a] Department of Civil Engineering, California State Polytechnic University, Pomona 3801, W. Temple Ave., Pomona, CA 91768, United States
[b] Engineering Technology Research Center for Mathematical Educational Software, Guangzhou University, Waihuanxi Road No.230, Guangzhou Higher Education Mega Center, China

ARTICLE INFO

Keywords:
Mixture
Spatiotemporal
Time adjacency
Interaction
Predictive accuracy
Cross validation

ABSTRACT

The traffic safety research has developed spatiotemporal models to explore the variations in the spatial pattern of crash risk over time. Many studies observed notable benefits associated with the inclusion of spatial and temporal correlation and their interactions. However, the safety literature lacks sufficient research for the comparison of different temporal treatments and their interaction with spatial component. This study developed four spatiotemporal models with varying complexity due to the different temporal treatments such as (I) linear time trend; (II) quadratic time trend; (III) Autoregressive-1 (AR-1); and (IV) time adjacency. Moreover, the study introduced a flexible two-component mixture for the space-time interaction which allows greater flexibility compared to the traditional linear space-time interaction. The mixture component allows the accommodation of global space-time interaction as well as the departures from the overall spatial and temporal risk patterns. This study performed a comprehensive assessment of mixture models based on the diverse criteria pertaining to goodness-of-fit, cross-validation and evaluation based on in-sample data for predictive accuracy of crash estimates.

The assessment of model performance in terms of goodness-of-fit clearly established the superiority of the time-adjacency specification which was evidently more complex due to the addition of information borrowed from neighboring years, but this addition of parameters allowed significant advantage at posterior deviance which subsequently benefited overall fit to crash data. The Base models were also developed to study the comparison between the proposed mixture and traditional space-time components for each temporal model. The mixture models consistently outperformed the corresponding Base models due to the advantages of much lower deviance.

For cross-validation comparison of predictive accuracy, linear time trend model was adjudged the best as it recorded the highest value of log pseudo marginal likelihood (LPML). Four other evaluation criteria were considered for typical validation using the same data for model development. Under each criterion, observed crash counts were compared with three types of data containing Bayesian estimated, normal predicted, and model replicated ones. The linear model again performed the best in most scenarios except one case of using model replicated data and two cases involving prediction without including random effects. These phenomena indicated the mediocre performance of linear trend when random effects were excluded for evaluation. This might be due to the flexible mixture space-time interaction which can efficiently absorb the residual variability escaping from the predictable part of the model. The comparison of Base and mixture models in terms of prediction accuracy further bolstered the superiority of the mixture models as the mixture ones generated more precise estimated crash counts across all four models, suggesting that the advantages associated with mixture component at model fit were transferable to prediction accuracy. Finally, the residual analysis demonstrated the consistently superior performance of random effect models which validates the importance of incorporating the correlation structures to account for unobserved heterogeneity.

* Corresponding author.
E-mail addresses: wcheng@cpp.edu (W. Cheng), gurdiljotg@cpp.edu (G.S. Gill), yongpingz@cpp.edu (Y. Zhang), zhongc@gzhu.edu.cn (Z. Cao).

https://doi.org/10.1016/j.aap.2017.12.020
Received 18 July 2017; Received in revised form 24 November 2017; Accepted 28 December 2017

1. Introduction

Roadway crashes have caused an immense burden on society with respect to emotional and financial losses. Researchers are entrusted to develop analytic approaches to gain a better understanding of the causal factors for crash occurrence and develop more accurate crash prediction models to formulate road-safety policies and engineering solutions for mitigation of crashes. However, the accuracy of inferences drawn from the statistical analysis of crashes is highly dependent on the robustness of crash data and an array of potential influential factors such as roadway geometric (number of lanes, lane width, radius of horizontal curve, etc.), traffic flow (vehicle density, volume, real-time speed, speed deviations, etc.), environment (lighting, weather), driver characteristics and mental state (gender, response time, age, etc.), among others. Unfortunately, the crash related data collected by safety agencies may be inadequate or unavailable for detailed investigation (Lord and Mannering, 2010). Hence, the researchers managed to handle this issue to study the significant factors by virtually enhancing the quantity of dataset by disaggregation over some geographical space (micro or macro level) and some specified time period (e.g., division of five year accumulated crash data into five individual subsets). The crash-frequency data are obtained in the form of non-negative integers allowing the application of count-based regression models.

These regression models (or, crash prediction models) have been used in research and practice for determination of influential factors, planning purposes, or site ranking. Models of varying complexity have been employed, ranging from very basic to sophisticated. The traditional approach to analyzing roadway crashes employed generalized linear models (McCullagh and Nelder, 1989;Zeger and Karim, 1992) to establish a linear relationship between explanatory variables and log-transformed outcomes such as crash frequencies of different severities or vehicle modes. This allowed for clear interpretation of inferences drawn from model estimates. To handle over-dispersion commonly associated with crash data, over-dispersed generalized linear models such as Poisson mixtures (e.g., negative binomial or Poisson-gamma, Poisson-lognormal, etc.) were introduced (Persaud, 1994; Hauer, 1997; Milton and Mannering, 1998; Karlaftis and Tarko, 1998). These models may not fully incorporate the unobserved heterogeneity as in case of count-data models, the overdispersion may be attributed to various factors, such as the grouping of data over space (segments, neighborhood, cities, regions, etc.), unaccounted temporal correlation, and model miss-specification (Gourieroux and Visser, 1997;Poortema, 1999;Colin Cameron and Trivedi, 1998).

Many studies have been proposed to address the unobserved heterogeneity shared by roadway entities in close proximity. These studies have focused on various spatial units including micro-levels such as intersections (Wang and Abdel-Aty, 2006; Cheng et al., 2017a), segments (Aguero-Valverde and Jovanis 2008), and macro-levels like corridors (Abdel-Aty and Wang, 2006; Guo et al. 2010), census tracts (MacNab, 2004), traffic analysis zones (Washington et al. 2010), counties (Gill et al., 2017), and so on. The significance of incorporating spatial correlations was highlighted by some studies (Guo et al., 2010; Abdel-Aty & Wang, 2006) with the consistently better performance of the spatial models over those accounting for heterogeneity random effect only.

In addition to spatial correlations, the disaggregation of crash data over specified time periods also leads to temporal correlation as those datasets may share unobserved effects which remain constant over time. To remove the potential bias of estimated model parameters, some researchers addressed the serial correlation in crash data by employing different temporal treatments such as linear and/or quadratic trend (Andrey and Yagar, 1993; Hay and Pettitt, 2001), autoregressive correlation structure with a time step of one year (lag-1) (Huang et al., 2009; Wang et al., 2013; Cheng et al., 2018), fixed-over-time and independent-over-time random effects (Aguero-Valverde, 2013; Jiang et al., 2014), and time-varying model coefficients and intercepts (Cheng et al., 2017b), and so on.; These models revealed that accounting for temporal correlations significantly improved the capability of the models to fit the crash data.

Building on the advantages of spatial and temporal correlation structures to address the issue of unobserved heterogeneity, some studies incorporated temporal dimension for spatial models as the crash analysis is not curbed to a single time period. Wang and Abdel-Aty (2006) investigated the rear-end crashes at intersections while employing the generalized estimating equations with the negative binomial link function to account for the temporal and spatial correlation among the 3-year longitudinal crash data. The study by Blazquez and Celis (2013) did a spatial and temporal analysis of child pedestrian crashes occurring during a period of nine years. The spatial autocorrelation analysis indicated that the responsibility of pedestrians is the major contributing factor for the generation of child pedestrian crashes with a tendency to cluster in space and time. Also, spatial clustering distribution of crashes in terms of time of the day was also observed. While these studies treated spatial and temporal correlations independently, some studies noted that vehicle crashes tend to cluster both spatially and temporally, hence space-time interaction specification was employed at different spatial scales. Aguero-Valverde and Jovanis (2006) proposed a spatiotemporal model with a linear time-space interaction term to study the fatal and injury crashes in Pennsylvania. The authors noticed that spatial correlation, time trend, and space–time interactions were significant in the proposed county-level Bayesian crash models. They also recommended such model should be extended to road segment and intersection-level crash models, where spatial correlation is likely to be even more pronounced. Subsequently, Plug et al. (2011) explored the variation of spatial distribution of single vehicle crashes (SVCs) according to different time periods (time of day and day of the week) by employing visualization technologies. The results showed significant differences in spatiotemporal patterns of SVCs for various crash causes.

The literature review illustrated notable benefits at various fronts associated with inclusion of spatial and temporal correlations and their interactions. However, in comparison with other types of models, a very limited body of research dedicated to spatiotemporal models exists in the field of traffic safety. Moreover, most of the current limited spatial-temporal models in the field assume a linear temporal trend and linear space-time interaction which may be seen as a restrictive assumption (Lawson and Clark, 2002; Lawson et al., 2003). For example, as discussed previously, the temporal random effects may take on nonlinear shape or have autocorrelation with previous crash counts. In addition, the temporal trend might have a non-linear change across the spatial units. To add to the current literature with more spatiotemporal models, the authors aimed to develop four alternative spatial-temporal models which employ different temporal treatments with the varying complexity of random effects: (I) linear time trend; (II) quadratic time trend; (III) Autoregressive-1 (AR-1); and (IV) time adjacency. Furthermore, instead of using linear space-time interaction, the authors borrowed a flexible two-component-mixture interaction from one previous disease-mapping study (Abellan et al., 2008). Such mixture can easily capture the space-time trends that depart from the predictable patterns of overall temporal and spatial risk surface as it allows the smoothness as well as discontinuities in the space-time variations within the roadway entities. The interested readers can be referred to this study for more details of the mixture components. The study results demonstrated a number of benefits associated with the proposed mixture model. However, the performance of the mixture model in traffic safety area is unknown and is therefore worth studying. In addition, in order for a comprehensive comparison of the predictive accuracy of the four models, five evaluation criteria were utilized which include log pseudo marginal likelihoods (LPML), mean square predictive error (MSPE), mean absolute deviation (MAD), residual sum of squares (RSS), and total rank difference (TRD). LPML assesses the predictive capability using cross-validation while the rest utilize the same dataset for model