COMPARISON OF ADJACENCY AND DISTANCE-BASED APPROACHES FOR SPATIAL ANALYSIS OF MULTIMODAL TRAFFIC CRASH DATA

G. Gill [a], T. Sakrani [a], W. Cheng [a,] *, J. Zhou [a]

[a] California State Polytechnic University-Pomona, California, USA (gurdiljotg, tsakrani, wcheng, jiaozhou)@cpp.edu

KEY WORDS: traffic safety, crash, spatial, Bayesian, prediction, county, multimodal

ABSTRACT:

Many studies have utilized the spatial correlations among traffic crash data to develop crash prediction models with the aim to investigate the influential factors or predict crash counts at different sites. The spatial correlation have been observed to account for heterogeneity in different forms of weight matrices which improves the estimation performance of models. But very rarely have the weight matrices been compared for the prediction accuracy for estimation of crash counts. This study was targeted at the comparison of two different approaches for modelling the spatial correlations among crash data at macro-level (County). Multivariate Full Bayesian crash prediction models were developed using Decay-50 (distance-based) and Queen-1 (adjacency-based) weight matrices for simultaneous estimation crash counts of four different modes: vehicle, motorcycle, bike, and pedestrian. The goodness-of-fit and different criteria for accuracy at prediction of crash count reveled the superiority of Decay-50 over Queen-1. Decay-50 was essentially different from Queen-1 with the selection of neighbors and more robust spatial weight structure which rendered the flexibility to accommodate the spatially correlated crash data. The consistently better performance of Decay-50 at prediction accuracy further bolstered its superiority. Although the data collection efforts to gather centroid distance among counties for Decay-50 may appear to be a downside, but the model has a significant edge to fit the crash data without losing the simplicity of computation of estimated crash count.

* Corresponding author

1. INTRODUCTION

From the past few decades, many fields have utilized the power of spatial nature of data for understanding the influence of space on different factors (Best et al., 2001). The transportation research has also exploited the capability of spatial correlations to better understand the crashes on road. The field of roadway traffic safety primarily focuses on investigation of influential factors for different types of crashes and then providing countermeasure treatments for the hazardous sites to mitigate the crashes (Gill et al., 2017a). Many researchers noted the presence of spatial correlations among crash data and incorporated them within the regression models to derive better inferences and develop more precise crash prediction models. An extensive body of literature exists which accounts for the spatial correlations at different scales (Macnab, 2004; Song et al, 2006; Huang et al., 2010). The roadway environment may be divided into two different scales: macro and micro entities. The macro level comprises of larger spatial levels such as block group, census tract, Traffic Analysis Zone (TAZ), County (Gill et al., 2017b), and so on, while the micro level includes the smaller entities such as intersections, segments, ramps, and so on. The understanding of spatial correlations among crash data may be explained by a simple example: the intersections on a roadway corridor are exposed to similar amount of vehicle traffic and roadway geometry which lends a spatial influence to the types of crashes occurring on the intersections of that corridor. During the analysis, grouping of such intersections may be advantageous for understanding the potential significant explanatory factors for crashes. Such correlations are necessary as the other factors may not be able to include the unobserved heterogeneity within the crash locations.

The selection of spatial scale of crash analysis is governed by the motivation behind such analysis. The macro level analysis of crash data usually serves the purpose to understand the impact of demographic or socioeconomic changes within an area on the crash trends. This broader perspective is highly beneficial to the planners who design and propose policies to control different factors with the aim to reduce crashes (Abdel-Aty et al., 2013). On the other hand, the micro level approach investigates the geometric or traffic flow factors and proposes engineering solutions for mitigation of crashes.

Many studies have focused on the exploration of different spatial units for macro level modeling for analysis of certain crash types (Rhee et al, 2016). As illustrated by the previous research, the spatial correlation could be incorporated to account for heterogeneity in different forms of weight matrices. The spatial models have been developed to identify usually hidden factors or improve the estimation performance of models. These correlations help account for the spatial dependency, which often escapes from the explanatory variables. But very rarely have the weight matrices been compared for the prediction accuracy for estimation of crash counts. This study focuses on this often overlooked aspect of spatial modeling which governs the appropriate selection of matrices for different approaches. Two different spatial models are developed for the crash data of 58 counties and the results are compared to assess superiority from different perspectives.

2. DATA AND METHODOLOGY

2.1 Data Description

This study developed crash prediction models for different modes of crashes occurring in the 58 counties of California during the year 2012. It should be noted that the focus of the present study was on the comparison of alternate weight matrices, hence only the traffic exposure factor of Daily Vehicle Miles Travelled (DVMT) was considered for model development, which was collected from Highway Performance Monitoring System (HPMS). DVMT was chosen as the exposure factor for county safety performance as it usually represents the vehicular activity at the planning level. The four crash modes considered for this study were: Vehicles, Motorcycles, Bikes, and Pedestrians. The crash data were obtained from Statewide Integrated Traffic Records System (SWITRS). As this study builds spatial models for 58 counties of California, the requirement for geometric distances between the centroids of the counties was fulfilled by the Southern California Association of Governments (SCAG). The data statistics are given in Table 1 for the different crash modes as well as the weight matrices. As expected, the highest number of crashes were Vehicle crashes while the other modes had a comparable number of crashes. The centroid distance among counties had a significant variation with a minimum and maximum distance of 25 and 962 miles, respectively.

Variable (unit)	Mean	Med	Min	Max	SD
Vehicle Crashes	2,171	560	16	38,477	5,388
Motorcycle Crashes	200	54.5	4	3,349	483
Bike Crashes	241	44.5	0	4,955	685
Pedestrian Crashes	228	35	0	5,024	684
Daily vehicle miles travelled (miles)	14,768,115	4,551,148	166,923	214,482,442	31,753,245
Centroid Distance (miles)	273	227	25	962	176
Weight Matrices					
Queen-1 (Number of neighbors per county)	4.91	5	2	8	1.3
Decay-50 (Number of neighbors per county)	57	57	57	57	N/A

Notes: Med refers to Median, Min refers to Minimum, Max refers to Maximum, and SD refers to Standard Deviation

Table 1. Descriptive statistics

The two alternate weight matrices considered for spatial modeling can be associated with two different approaches for selection of neighbors: adjacency-based and distance-based. These are depicted clearly in Figure 1, where Queen-1 and Decay-50 are the adjacency and distance matrices, respectively. To understand the criterion of selection for two approaches, let us consider the county of Tehama in Figure 2. In case of Queen-1, the county of Tehama has six immediate neighboring counties, namely: Plumas, Butte, Glenn, Mendocino, Trinity, and Shasta. The prefix "1" for the Queen-1 represents the order of neighbors being considered, which in this case means the immediate neighbors of the concerned county. Table 1 shows the average number of neighbors for each county in California to be around 5. In case of Decay-50 (distance-based) weight matrix, the selection of counties is not governed by the sharing of border, but rather every county is considered to be a neighbor (with different spatial influence) for every other county. As the state of California has 58 counties, so each county has 57 neighbors. It is noteworthy that based on the two aforementioned approaches, there is a significant disparity between the number of counties considered as neighbors.

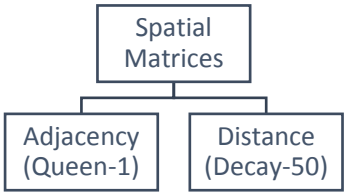

Figure 1. Types of weight matrices

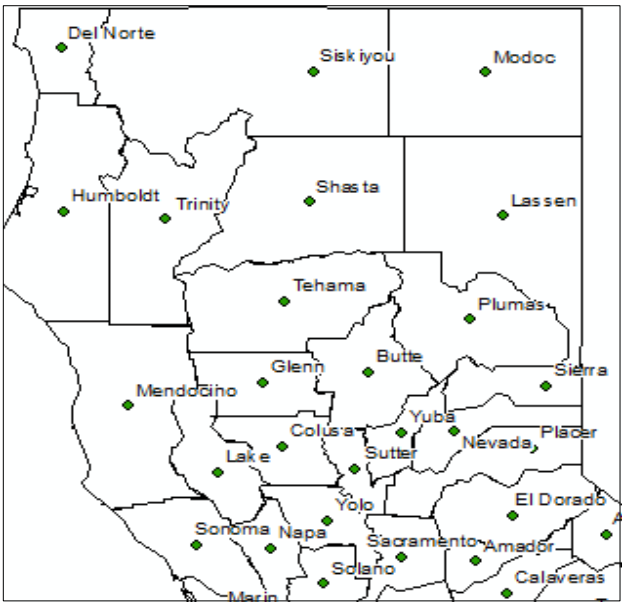

Figure 2. Section of California counties and associated centroids.

2.2 Model Development

Most of the studies model the different modes separately but some researchers observed the presence of correlations among crash modes which should be accounted for developing more robust models. To address the potential bias, this study simultaneously estimated the different crash modes by employing multivariate model. Crash data is mostly overdispersed in the field of traffic safety. To account for the overdispersion, this study developed the Poisson lognormal model under the Full Bayesian framework (Aguero-Valverde and Jovanis, 2009). To accommodate the spatial correlation, a hierarchical approach was utilized where the structural heterogeneities were incorporated. The model is of the following form:

$$y_{ij} \sim Poisson\,(e_{ij}\theta_{ij}) \quad (1)$$

Where, y_{ij} is the observed crash count at county i for the crash mode j, θ_{ij} is the Bayesian mean expected crash rate at site i for crash mode j, and e_i is the exposure in county i. In this case, the traffic exposure is accounted by DVMT of the specific county i. The crash rate is modeled as shown in the following equation:

$$Log(\theta_{ij}) = \beta_0 + \emptyset_{ij} + u_{ij} \quad (2)$$

$$u_{ij} \sim Normal\,(0,\ \sum) \quad (3)$$

$$\sum = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{14} \\ \vdots & \ddots & \vdots \\ \sigma_{41} & \cdots & \sigma_{44} \end{pmatrix} \quad (4)$$

Where, β_0 is the intercept, $\emptyset_{ij}$ is a spatially structured random effect which is fit by a zero-centered multivariate conditional auto-regressive (MCAR) model (For more details on MCAR, please refer Cheng et al., 2017), and u_{ij} is a spatially unstructured random effect which captures the extra-Poisson heterogeneity among locations. $\sum$ is called the covariance matrix. The diagonal element $\boldsymbol{\sigma_{jj}}$ in the matrix represents the variance of $\boldsymbol{u_{ij}}$, where the off-diagonal elements represent the covariance of crash counts of different modes. The inverse of the covariance matrix represent the precision matrix and has the following distribution:

$$\sum^{-1} \sim Wishart(I, J) \quad (5)$$

Where, I is the J x J identity matrix, and J is the degree of freedom, J=4 herein representing 4 crash outcomes based on mode.

The MCAR allows the incorporation of spatial structures for a specific site, where the two alternate weight matrices of this study differ on assignment of following: the identity of neighboring sites, number of neighboring sites, and the weight of spatial influence. In case of Queen-1, the distance between the county centroids was ignored and only those counties were considered as neighbors which shared an immediate border with the concerned county. This resulted in a varied number of neighbors for each county. It is noteworthy that this adjacency-based weight matrix gives a binary or dichotomous output for weight with only two responses, zero for non-neighbors and one for neighbors. In case of the distance-based matrix (Decay-50), all the counties were considered as neighbors of each other and the weight was calculated as follows:

$$w_{ik} = e^{\frac{-\text{dist}_{ik}}{\lambda_0}}$$
(6)

Where, w_{ik} is weight of the *k*th neighbor of the *i*th county, $dist_{ik}$ is the geographic centroid distance between counties *i* and *k*, and λ_0 is the decay constant. The selection of decay was done by a sensitivity analysis of total collision count and average county distance of 250 miles. It is noteworthy that the Decay-50 potentially accounts for more flexibility is weight assignment compared to the binary weights of Queen-1 matrix.

2.3 Model Comparison

The two models were assessed for goodness-of-fit with the observed crash data and the criterion employed for assessment was the Deviance Information Criterion (DIC) developed by Spiegelhalter et al., (2002). DIC provides the measure of complexity and fit of the models which are developed from same sample size. DIC is computed as the sum of the posterior mean deviance and estimated effective number of parameters:

$$DIC = \overline{D} + P_D$$
(7)

Where, $\overline{D}$ is the sum of the posterior mean deviance which measures how well the model fits the data and P_D represents the effective number of parameters utilized for model building.

As mentioned earlier, one of the primary goals for the development of crash prediction models is to achieve more precise estimate for prediction of crash count at sites. Hence, it is worthwhile to assess the prediction accuracy of the models to ensure that better fit of model estimates with crash data is transferred to similar performance at prediction accuracy. Different criteria were employed to compare the prediction accuracy of the models such as Mean Absolute Deviance (MAD), Mean Square Predictive Error (MSPE), and Total Rank Difference (TRD).

The MAD quantifies the discrepancy between the observed crash count and the ones predicted by the model for the same site. It can be calculated with the following equation:

$$\text{MAD} = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}}\left|\text{C}_\text{p} - \text{C}_\text{o}\right|$$
(8)

Where, C_p is the estimated crash count of county *i* by the model and C_o is the observed crash count of the same county *i*.

The MSPE was calculated as follows:

$$\text{MSPE} = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}}(\text{Y}_\text{i} - \text{O}_\text{i})^2$$
(9)

Where, Y_i is the predicted crash count of site *i* by a model, and O_i is the observed crash count of *i* by the same model at the same time period. The smaller value of MSPE indicates that the discrepancy between the predicted and observed crash value for a site is relatively small and hence that model has better prediction accuracy.

TRD takes into account the rank difference of a site, which is calculated based on the higher crash count. The corresponding calculation is shown as follows:

$$\text{TRD} = \sum_{k=1}^{n}|R(i_P) - R(i_O)|$$
(10)

Where, $R(i_P)$ is the rank of site *i* based on the predicted crash count while $R(i_O)$ is the rank of the same site based on observed crash count. The smaller value indicates that the particular model was able to consistently identify and rank the sites.

3. RESULTS

Model	$\overline{D}$	P_D	DIC	MAD	MSPE	TRD
Queen-1	1,819	314	2,133	282	1,762,99	3,709
Decay-50	1,742	244	1,987	199	842,669	3,561

Table 2. Model fit and prediction accuracy

DIC is a penalized criterion which is a trade-off between model complexity (P_D) and model fit ($\overline{D}$). For comparison of models with similar sample size, the difference of 5 points between DIC values of two models is considered to be comparable while higher than 10 points hints at a significant difference of model fit. As shown in Table 2, the DIC values vary substantially for the two weight matrices based on the aforementioned criterion. A significant difference of 146 points was observed, where Decay-50 had the better overall fit. As discussed earlier, DIC is a mix of model fit and complexity. To further analyze the models and understand their superiority at handling crash data, it is worthwhile to observe the values of $\overline{D}$ and P_D. Similar trend may be observed for the fit ($\overline{D}$) and complexity (P_D) as well, where the difference between two models was 77 and 70, respectively. It seems that the inclusion of continuous county-wise varying distance weights provided the flexibility to fit the crash data better compared with the rigid binary weight structure of Queen-1 matrix.

As expected, the significant performance at model fitness was observed to be correlated with the superiority at accuracy of predicting the crash counts. The prediction accuracy criteria which comprised of MAD, MSPE, and TRD, established the substantial advantage of Decay-50 at precise estimation of crashes. Since MAD and MSPE measure the deviation from observed count, so relatively low values are desirable which reflect the discrepancy between the observed and predicted crash count for the counties and hence indicate better crash prediction capability. The MAD of Queen-1 was observed to be 41% greater than Decay-50. The difference was more pronounced in case of MSPE with a 109% increase from MSPE value of Decay-50. These trends indicate that the Decay-50 model was able to capture the spatial structural heterogeneities which accounted for better prediction capability. Another measure which was calculated to compare the site ranking performance again established Decay-50 to be superior with the TRD value 148 points lower than Queen-1. This advantage of distance matrix may be accredited to the criterion which forms the basis of selection of neighbors, i.e. more counties are selected as neighbors and spatial dependency is based on mutual

distance, which potentially fits the crash data better and results in more precise estimates to rank the counties.

4. CONCLUSIONS

This study was targeted at the comparison of two different approaches for modelling the spatial correlations among crash data at macro level. Multivariate Full Bayesian crash prediction models were developed using Decay-50 and Queen-1 weight matrices for simultaneous estimation crash counts of four different modes: vehicle, motorcycle, bike, and pedestrian. The goodness-of-fit and different criteria for accuracy at prediction of crash count reveled the superiority of Decay-50 over Queen-1. Decay-50 was essentially different from Queen-1 with the selection of neighbors and more robust weights. This inclusion of extra data was expected to make the model more flexible but at the same time it was anticipated that their addition would remarkably increase the model complexity. However, the lower value of P_D reflects that the extra information may have rendered the Decay-50 model more subtle to accommodate the spatially correlated crash data. The consistently better performance of Decay-50 at prediction accuracy further bolstered its superiority. Although the data collection efforts to gather centroid distance among counties for Decay-50 may appear to be a downside, but the model has a significant edge to fit the crash data without losing the simplicity of computation of estimated crash count.

The study illustrated the remarkable superiority of a distance-based model (Decay-50) over an adjacency-based one (Queen-1). However, it is recommended not to generalize the results of this study and conclude the dominance of distance-based models as there could be many factors which may influence the performance of models such as different spatial levels; inclusion of explanatory variables; adjacency-matrices of higher order or different types (Queen-2, Queen-3, Rook-2), and so on. Hence, this study recommends the sensitivity analysis of different weight matrices to determine the suitability for development of crash prediction models.

REFERENCES

Abdel-Aty, M., Lee, J., Siddiqui, C. and Choi, K., 2013. Geographical unit based analysis in the context of transportation safety planning. *Transportation Research Part A: Policy and Practice*, *49*, pp.62-75.

Aguero-Valverde, J. and Jovanis, P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. *Transportation Research Record: Journal of the Transportation Research Board*, (2136), pp.82-91.

Best N, Cockings S, Bennett J, Wakefield J, Elliott P, 2001. Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *J R Statist Soc A*, 164:155-174.

Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X. and Zhou, J., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. *Accident Analysis & Prevention*, *99*, pp.330-341.

Gill, G. X. Wang, W. Cheng and M. Xie. "Assessment of Hit and Run Accidents-A Modification toward predicting the Contributing Factors of Hit-and-Run Accident". (2017a) WDSI Annual Conference Proceedings, Vancouver, Canada.

Gill, G. X. Wang, W. Cheng and J. Zhou "Macro-Level Annual Safety Performance Function Evaluation for Cities, Counties and State of California". (2017b) WDSI Annual Conference Proceedings, Vancouver, Canada

Huang, H., Darwiche, A.L., Abdel-Aty, M.A., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling. Transp. Res. Rec. 2148, 27–37.

MacNab, Y.C., 2004. Bayesian spatial and ecological models for small-area crash and injury analysis. Accident Analysis & Prevention 36 (6), 1019–1028.

Rhee, K. A., Kim, J. K., Lee, Y. I., & Ulfarsson, G. F. 2016. Spatial regression analysis of traffic crashes in Seoul. *Accident Analysis & Prevention*, *91*, 190-199.

Song, J. J., M. Ghosh, S. Miaou, and B. Mallick. 2006. Bayesian Multivariate Spatial Models for Roadway Traffic Crash Mapping. *Journal of Multivariate Analysis*, Vol. 97, pp. 246–273.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B 64 (4), 583–616.