Investigating the Existence of Second Order Spatial Autocorrelation in Crash Frequency across Adjacent Freeway Segments

Eneliko Mulokozi, Hualiang (Harry) Teng

Department of Civil and Environmental Engineering and Construction, Howard R. Hughes College of Engineering, University of Nevada, Las Vegas, NV, USA
Email: mulokozi@unlv.nevada.edu, Hualiang.teng@unlv.edu

Abstract

This study developed spatial Poisson model to incorporate spatial autocorrelation in crash frequency across contiguous freeway segments. Spatial autocorrelation is the presence of spatial pattern in crash frequency over space due to geographic proximity. Usually crash caused congestion on a freeway segment propagates upstream and creates chance of occurring secondary crashes. This phenomenon makes the crash frequency on the contiguous freeway segments correlated. This correlation makes the distributional assumption of independence of crash frequency invalid. The existence of spatial autocorrelation is investigated by using Conditional autoregressive models (CAR models). The models are set up in a Bayesian modeling framework, to include terms which help to identify and quantify residual spatial autocorrelation for neighboring observation units. Models which recognize the presence of spatial dependence help to obtain unbiased estimates of parameters quantifying safety levels since the effects of spatial autocorrelation are accounted for in the modeling process. Based on CAR models, approximately 51% of crash frequencies across contiguous freeway segments are spatially auto-correlated. The incident rate ratios revealed that wider shoulder and weaving segments decreased crash frequency by factors of 0.84 and 0.75 respectively. The marginal impacts graphs showed that an increase in longitudinal space for segments with two lanes decreased crash frequency. However, an increase of facility width above three lanes results in more crashes, which indicates an increase in traffic flows and driving behavior leading to crashes. These results call an important step of analyzing contagious freeway segments simultaneously to account for the existence of spatial autocorrelation.

Keywords

Freeway Segments, Spatial Autocorrelation, Conditional Autoregressive Model,
1. Introduction

Crash frequency for contiguous roadway segments tends to be correlated because congestion caused by a crash on a segment propagates to upstream. The propagation could reach to upstream segment quickly if the crash is severe, the segment where the crash occurs is short and traffic entering the downstream is high. The congestion reached to upstream may cause secondary crashes, which makes the crash frequencies on the contiguous segments correlated. Usually crash frequency is modeled by a Poisson model (Noland et al., 2004 [1]; Quddus, 2008 [2] and Li et al., 2007 [3]) which includes a set of influencing factors such as geometric elements of the freeways, traffic characteristics, environmental factors, and human factors. The correlation of crash frequency on contiguous freeway segments cannot be taken account in such a Poisson regression model in previous study.

To deal with the correlation of crash frequency, the Poisson regression model is integrated with a Conditional Autoregressive Model (CAR) (Aguero-Valverde, 2013 [4] and 2014 [5]) where the spatial effects leading to residual autocorrelation in the crash frequency can be specified. CAR models contain a precision matrix to control the spatial autocorrelation structure of the random effects based on the weight matrix. Contiguity of freeway segments can be specified in the model by a binary coding where a code equals to 1 if the freeway segments share a common border and is zero otherwise. A spatial autocorrelation parameter with variance is used to indicate the amount of autocorrelation in the crash frequency. The Poisson model integrated with CAR model can be estimated in a Bayesian framework. Under this framework, the unknown parameters are set to reflect prior knowledge. Posterior means can be derived based on Markov Chain Monte Carlo (MCMC) simulation using Gibbs and Metropolis steps as sampling techniques. This study developed a statistical model that can incorporate spatial autocorrelation on contiguous freeway segments. Crash data and related influencing factor data were collected from the freeways in the Las Vegas area of Nevada. The spatial models are calibrated and interpreted, which shows significant different observations than those from using the traditional methods.

The remaining part of the paper is organized as follows. The first section presents review of previous work on addressing spatial effects of transportation measures. In the second section, the spatial Poisson model is described with the presentation of the estimation method. The third section provides the description of data collection, which is followed by calibrating the spatial Poisson models in the fourth section. Comparison between the non-spatial and spatial Poisson regression models is also included in the fourth section. The last section includes the conclusions and recommendations for future study needs.

2. Review of Previous Work

Crash events occur spatially along the highway network and including spatial effects in
crash prediction models help to explain variability observed in crash frequency and avoid making inference on biased estimates. Black et al., 1998 [6] employed a network autocorrelation analysis to examine accidents distributed along the segments of a highway system and found a significant level of positive spatial autocorrelation. El-Basyouny et al., 2009 [7] used Gaussian conditional autoregressive and multiple membership models on 281 urban segments in Vancouver, Canada and found that spatial autocorrelation across urban segments explained approximately 87.6% of the variability in crash rates for CAR model while it was approximately 98.5% for multiple membership models. In addition to these findings, it was also revealed that AADT, business land use, number of lanes between signals and density of unsignalized intersections had significant positive impact on the number of crashes.

Guo et al. (2010) [8] employed Bayesian count and Gaussian models to incorporate corridor-level and intersection proximity spatial autocorrelations in predicting crash rate and crash frequency. It was revealed that the size of an intersection, traffic conditions for both through and turning movements and the coordination of signal phase have significant impacts on intersection safety. This implies that closeness of coordinated intersections is likely to stimulate differential driving behavior compared to isolated intersections.

Arthur R. M. [9] identified the existence of spatial autocorrelation based on Moran’s I statistics applied on neighboring network intersections. To be able to apply the concepts of spatial autocorrelation on intersections as opposed to network roadway segments, the analysis considered the roadways as links and the intersections and the adjusted frequencies of collisions as areas. The Moran’s I statistics values indicated the existence of clusters of collision frequencies while graphing these values identified a temporal fluctuation that follows a diurnal pattern which indicates clustering patterns. It was also revealed that daytime pattern suggests a high frequency of collisions on major arteries during the day, especially over rush hour where it would be reasonable to assume a more clustered pattern.

Wang et al. (2013) [10] used Poisson-based multivariate conditional autoregressive (CAR) models estimated by Bayesian Markov Chain Monte Carlo methods to examine the relationship between pedestrian crash counts across tracts areas and various attributes characterizing the network, land use and demography. The results indicated the existence of positive spatial autocorrelation across neighborhoods as a result of the existence of latent heterogeneity or missing variables that trend in space which are likely to generate spatial clustering of crash counts. In addition to spatial autocorrelation identification, their results also showed that there is a greater association of residences and commercial land uses with pedestrian crash risk across different severity levels due to high potential conflicts between pedestrian and vehicle movements.

Miaou et al. (2005) [11] used a multivariate spatial Generalized Linear Mixed Model (GLMM) to model crashes by injury severity type simultaneously and to rank sites by crash cost rate as decision parameter in ranking. Ranking results was based on relative standards which imply that rank and select among a predetermined group of sites based
on their relative risk levels. The results showed that including spatial effects components in modeling processes improved the overall goodness-of-fit performance of the model and affected the ranking results for site improvements. The results further revealed that including CAR model in modeling process accounts for the degree of over dispersion.

3. Model Specification and Estimation

Bayesian Hierarchical Model for Crash Frequency

Let \( y_i \) denote the number of crashes observed in a freeway segment for the \( i \)th covariate pattern. Let the expected value of \( y_i \) depend on the explanatory variables \( x_i \). The Poisson generalized linear model (Guo et al., 2010 [8]) with the natural link function in the logarithmic function can be specified as:

\[
\ln(\mu_i) = x_i'\beta, \quad i = 1, \cdots, n . \tag{1}
\]

\( \mu_i \) denotes the expected value of the crash frequency for segment \( i \), \( x_i' \) is the matrix of observed influencing factors including an intercept, and \( \beta \) is the matrix of regression coefficients which quantifies the impact of covariates on the expected crash frequency. Equation (1) can be used to model spatial pattern in the crash frequency across freeway segments via a matrix of the covariates which in this case are the geometric and traffic characteristics observed on the freeways. However, the observed crash frequency for Poisson model exhibit over-dispersion and to capture this effect we extend equation 1 to include random effects, \( \varnothing_i \) to account for the possible effects of over-dispersion:

\[
\ln(\mu_i) = x_i'\beta + \varnothing_i, \quad i = 1, \cdots, n . \tag{2}
\]

Under Bayesian modeling frame work, prior distributions for the unknown parameters are set to reflect prior knowledge about the parameters of interest (El-Basyouny et al., 2009 [7]; Guo et al., 2010 [8]; Aguero-Valverde and Jovanis, 2006 [12]; Bailey et al., 1995 [13]). In this case we assume an independent Gaussian prior (diffuse normal priors) for each regression coefficient, \( \beta_j \sim N(m_j, \nu_j) \) with mean, \( m_j \) equals 0 and variance, \( \nu_j \) necessarily large. We also assume uniform priors for random effects, \( \varnothing_i \sim U\left(0, \sigma^2\right) \) with large variance, \( \sigma^2 \).

We further assume the existence of second order spatial effects (Bolstad, 2005 [14] and Gelman et al., 2014 [15]) unaccounted for by the covariates and specify a Conditional Autoregressive Priors (Lee, 2011 [16] and Kery, 2010 [17]):

\[
\varnothing_i | \varnothing_{-i} \sim N\left(\frac{\rho \sum_{i=1}^{n} w_{ii} \varnothing_i}{\rho \sum_{i=1}^{n} w_{ii} + 1 - \rho} \left(1 - \frac{\nu^2}{\rho \sum_{i=1}^{n} w_{ii} + 1 - \rho}\right) \right). \tag{3}
\]

Equation (3) is a special case of the Gaussian Markov random field which contains a precision matrix to control the spatial autocorrelation structure of the random effects based on the weight matrix \( W \). Contiguity of freeway segments can be specified in the model by a binary coding where \( w_{ii} = 1 \) if the freeway segments share a common bor-
der and is zero otherwise. $\rho$ is a spatial autocorrelation parameter with variance equals $\tau^2$. A statistically significant value of $\rho$ away from 0 implies the existence of spatial autocorrelation of crash frequency for contiguous freeway segments. Both $\rho$ and its variance parameter, $\tau^2$ have an independent prior specified as follows:

$$\text{Spatial autocorrelation: } \rho \sim U(0,1); \quad \tau^2 \sim U(0,M) .$$  \hspace{1cm} (4)$$

Our decision to adopt Equation (3) is based on the appealing fact that (Kery, 2010 [17]) conducted a comparative research and identified that the random effects modeled by a conditional autoregressive (CAR) prior distribution specified by Equation (3) is the best because it produces consistently good results across the range of spatial correlation scenarios considered. It also represents a range of strong and weak spatial correlation structures with a single set of random effects which is beyond the models proposed.

Inference for the above models is based on Markov Chain Monte Carlo (MCMC) simulation (Lee, 2011 [16]; Kery, 2010 [17] and Dobson, 2010 [18]) using a combination of sampling techniques. The variance parameters are Gibbs sampled from their full conditional truncated inverse gamma distributions, while the remaining parameters are updated using Metropolis steps. An important key part of the analysis based on sampling techniques is to be able to make valid inferences. This is possible by monitoring Markov chain convergence to the target densities. To ensure the Markov chain lies within the stable area of high likelihood we apply a burn-in of 20,000 samples to ensure that the samples drawn from the chains approximate the posterior distribution. We also apply a thinning equal to 10, to reduce autocorrelation of neighboring samples (Dobson, 2010 [18]). We monitor our results of convergence and stable posterior distributions based on trace plots and posterior densities of covariates. To estimate our models, we apply CARBayes package in an R software environment (Lee, 2014 [19]) and WINBUGS version 1.4.3 (Medical Research Council, 2015 [20]). Choosing the most parsimonious model is based on the Deviance information criterion (DIC), which is a generalization of Akaike Information Criterion (AIC) for Bayesian models. We also evaluate the significance of estimated parameters based on 95\% credible intervals.

4. Data and Descriptive Statistics

This section requires dataset with a structure focusing on the investigation of a special case of unobserved heterogeneity: spatial dependence of crash frequency for contiguous freeway segments. The model inputs are traffic and geometric characteristics from contiguous freeway segments extracted from loop detectors managed by FAST. Freeway segments which shares a common border identified as natural delineation between entrance and exit were considered. Since the purpose is to identify the existence of spatial dependence, contiguous freeway segments with missing traffic characteristics were removed from the study and retained only segments with all information required.

Based on the aforementioned criteria, a total of 36 Segments were selected for study. Using ArcMap, a polygon shapefile was created for all segments under study with visual
aid from a base map as a tracing tool. Furthermore, sensor codes with their locations were observed from Google maps and matched with the created GIS shapefile of freeway segments and traffic characteristics which included vehicular speed and traffic volumes were extracted for each sensor located on those segments.

Geometric characteristics were obtained by changing a GIS shapefile to KMZ and overlay the resulting KMZ file on Google earth map for visual aid. Number of lanes, median shoulder and right shoulder were observed and measured from the overlaid KMZ file as shown on Table 1 which shows summarized data.

From Table 1, it is evident that there is a high variability of crash counts and an indication of the presence of over-dispersion since the standard deviation of the crash counts is greater compared to its mean. Geometric characteristics are also diverse as shown by the minimum and maximum values in addition to their mean and standard deviations. However, less can be learnt concerning the existence of spatial effects and our detailed discussion is focused on the results section which follows.

5. Model Results and Discussion

CARBayes package version 4.0 and WINBUGS version 1.4.3 were used in estimating our models as shown on Table 2. To reduce autocorrelation of samples from the posterior distribution, we thinned the sequence by keeping every 10th simulation draw from each sequence. We also discarded the first 20,000 samples and concentrate on the last 80,000 samples to be able to diminish the influences of early iterations and achieve the target distribution. This implies that our final results are summarized from 8,000 drawn samples. To ensure that the chain’s stationary distribution approximates the target distribution, we monitored our chain based on trace plots, historical plots of chain process as well as density plots of posterior means of the covariates and autocorrelation

Table 1. Estimated posterior means of covariates.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Estimates (95% credible interval)</th>
<th>Non-spatial GLM Poisson</th>
<th>Spatial GLM Poisson model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.45 (6.78, 8.06)</td>
<td>4.46 (−3.69, 7.91)</td>
<td></td>
</tr>
<tr>
<td>Log (right shoulder)</td>
<td>−1.08 (−1.25, −0.92)</td>
<td>−0.17 (−1.63, 1.21)</td>
<td></td>
</tr>
<tr>
<td>Segment length in mile</td>
<td>−8.03 (−8.10, −7.15)</td>
<td>−4.21 (−8.16, 6.00)</td>
<td></td>
</tr>
<tr>
<td>No. of lanes</td>
<td>−0.57 (−0.68, −0.45)</td>
<td>−0.19 (−0.91, 1.25)</td>
<td></td>
</tr>
<tr>
<td>No. of lanes * seg. length</td>
<td>1.99 (1.80, 2.19)</td>
<td>1.18 (−1.20, 2.22)</td>
<td></td>
</tr>
<tr>
<td>Weaving segment</td>
<td>−0.03 (−0.14, 0.07)</td>
<td>−0.28 (−0.97, 0.48)</td>
<td></td>
</tr>
<tr>
<td>$\tau^2$</td>
<td></td>
<td>1.22 (0.65, 2.89)</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td></td>
<td>0.51 (0.09, 0.86)</td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td>1567.6</td>
<td>243.8</td>
<td></td>
</tr>
</tbody>
</table>
term. Our final results include the trace plots for only the autocorrelation term as shown on Figure 1.

Based on the model fit criteria, the Spatial GLM Poisson model had a deviance Information criterion (DIC) equal to 243.8 which is small compared to a Non-spatial GLM model. This implies that the spatial model exhibited better fit to our data and therefore we interpret our results using the spatial GLM Poisson model. The final results of a Spatial GLM Poisson model contain posterior means of covariates and autocorrelation term. The significance of these terms is based on 95% credible intervals. When the 95% credible intervals include zero, the corresponding factor is not significant at the 95% level and vice versa. We interpret our results based on the incident rate ratios which are exponentiated estimates and credible intervals as well as marginal impact for the interaction terms involving number of lanes and segment length.

Based on the results on Table 2, approximately 51% of crash frequencies across contiguous freeway segments are autocorrelated with a variance equals to 1.22. This results supports our aforementioned hypothesized situation that, there are spatial correlations of underlying processes generating crashes and these are likely to propagate across the adjacent segments. Most of the research activities analyze crash events on freeways based on the assumption that crash frequency observed on freeway segments are independent. This results lead to biased estimates if spatial effects are not included in our modeling processes.

In addition to the aforementioned findings, we also investigated the impact of geometric elements on the crash frequency. The results are interpreted based on the exponentiated coefficients and credible intervals. As the results showed, right shoulder was found insignificant on individual bases. However, overall all variables were found

<table>
<thead>
<tr>
<th>Variable</th>
<th>Scale</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crashes</td>
<td>Number</td>
<td>57.7778</td>
<td>86.5801</td>
<td>1</td>
<td>411</td>
</tr>
<tr>
<td>Centred width of log of right shoulder</td>
<td>Natural logarithm</td>
<td>0.9969</td>
<td>0.3323</td>
<td>−0.0163</td>
<td>1.6832</td>
</tr>
<tr>
<td>Log of base length of segment</td>
<td>Mile</td>
<td>0.6766</td>
<td>0.4409</td>
<td>0.2019</td>
<td>2.5727</td>
</tr>
<tr>
<td>Number of lanes</td>
<td>Number</td>
<td>4.3056</td>
<td>1.2833</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Number of lanes * Log of base length of seg.</td>
<td>Mile</td>
<td>2.7146</td>
<td>1.5166</td>
<td>0.8077</td>
<td>7.7180</td>
</tr>
<tr>
<td>Weaving segment</td>
<td>Categorical</td>
<td>0.4167</td>
<td>0.5000</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Figure 1.** Posterior distribution of spatial correlation parameter.
to have an impact in the model. With these results it decreased crash frequency by a factor of 0.84 for segments with the same length and number of lanes and of the same type. This is reasonably true because wider shoulder provides enough space for drivers to maneuver to avoid crashes.

Weaving segments reduced crash frequency by a factor of 0.75 compared to non-weaving segments for segments with the same level of other influencing factors included in the model. This implies that drivers across these segments have time to make decisions to execute accelerations and decelerations activities to and from the freeway while still on the speed change lanes. In this case risk hazards to be encountered are likely to be avoided.

The impact of the number of lanes and segment lengths are better interpreted based on their marginal impacts as shown on Figure 2 including its corresponding marginal effect function and in the context of longitudinal space and transverse space of the freeway facility. Including an interaction term is based on the fact that the influence of the extent longitudinal space depends on the transverse space available to accommodate the number of vehicles available. As shown on Figure 2, as the longitudinal space increases for segments with two lanes, crash frequency decreased. This was also true for segments with three lanes. It should be understood that we interpret number of lanes as representing width of freeway in a transverse dimension.

\[
\frac{\partial (\text{marginal eff.})}{\partial (\text{length})} = (1.18 \times \text{lanes} - 4.21) \exp \{1.09(1.18 \times \text{length} - 0.19) \times \text{lanes}\}.
\]

However, when the width of freeway increases above three lanes, the facility experienced
an increase in crash frequency. This can be explained by the fact that there are more traffic flows and driving activity associated with change lanes which is likely to stimulate driver behavior as more vehicles occupy these segments. This is counter-intuitive when compared to the normal intuition indicated by two and three lanes as well as the previous research. The marginal effect graph also provides a clear picture of the interaction which could not be obvious based on the estimates alone.

6. Conclusions and Recommendation

Second order spatial effects are one of the components leading to unobserved heterogeneity. If these effects are not included in quantifying the impacts of influencing factors, the results obtained are likely to be biased. This study investigated the existence of second order spatial effects (spatial autocorrelation) for contiguous freeway segments. This is motivated by the fact that there are spatial correlations of underlying processes generating crashes and these are likely to propagate across the adjacent segments. Furthermore, there are interdependences across freeway characteristics in influencing crash occurrence which means the effect of one of the geometric elements depends on other geometric elements. Including interaction terms in modeling reduces the impact of unobserved heterogeneity because it accounts for effect modification resulting from modification variables.

Based on the aforementioned findings, it was found that wider shoulders and weaving segments reduced the level of crash frequency. Longitudinal space which was reflected in the segment length exhibited differential impact across long freeway segments. This differential impact can be explained by the existence of differential transverse space reflected in the number of travel lanes for long segments. This implies that interacting influencing factors helps to reduce the impact of unobserved modifiers leading to biasing results. It is imperative that safety modeling includes terms which explain any general or specific forms of unobserved heterogeneity. This helps to come up with actual impacts of the influencing factors retained in the model.

6.1. Model Application

The developed model can be applied as a discriminant model. This is based on the fact that spatial effects terms are included in the modeling process. Theories on the estimation process require integrating out these effects and summarize them in terms of variance, a method which leaves out the actual influence of the remaining factors. Based on Figure 2, a researcher can point out locations on the freeway network from which its factors exhibited more impacts on the crash frequency. For instance, with these results segments with four and five lanes increased crash frequency compared to those with fewer number of lanes. These lanes require further investigation to be able to understand causes of crash frequency occurrence.

Another important application is based on the natural interpretation of most of the regression coefficients. Negative coefficients in most of the cases means the corresponding factors had a negative impact and therefore by increasing those factors help to
reduce more crash frequency on freeways. For instance, our results for long segments with few numbers of lanes indicated that by increasing the longitudinal space we are able to reduce the number of crash frequency experienced. This is counter-intuiting with segments of the same length but have more lanes. This means we cannot adopt an alternative of increasing segment length on these locations. This helps to narrow down countermeasures alternatives and focus on Intelligent Transportation Systems designed for safety improvements on the freeway systems.

6.2. Future Research

The results from this study have indicated the existence of spatial autocorrelation across contiguous freeway segments. This implies that segments with spatial proximity constitute traffic and geometric characteristics which influence crash occurrence in a similar trend. This phenomenon violates the distribution assumption of Poisson process under which crash events occur. Future research may involve simultaneity treatment of freeway segments in analyzing safety effects of factors which are believed to influence crash occurrence. Simultaneity behavior of observation unit can be incorporated in safety analysis by using simultaneous equation models which are the special case of the general structural equation models. In addition to solving the aforementioned problem of distributional assumption, general unobserved heterogeneity terms can further be incorporated to account for random effects.

References


Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: http://papersubmission.scirp.org/
Or contact jtts@scirp.org