Analysis of Road Crash Frequency with Spatial Models

Jonathan Aguero-Valverde and Paul P. Jovanis

Despite the evident spatial character of road crashes, limited research has been conducted in road safety analysis to account for spatial correlation; further, the practical consequences of this omission are largely unknown. The purpose of this research is to explore the effect of spatial correlation in models of road crash frequency at the segment level. Different segment neighboring structures are tested to establish the most appropriate one in the context of modeling crash frequency in road networks. A full Bayes hierarchical approach is used with conditional autoregressive effects for the spatial correlation terms. Analysis of crash, traffic, and roadway inventory data from a rural county in Pennsylvania indicates the importance of including spatial correlation in road crash models. The models with spatial correlation show significantly better fit to the data than the Poisson lognormal model with only heterogeneity. Parameters significantly different from zero included annual average daily traffic (AADT) and shoulder widths less than 4 ft and between 6 and 10 ft. In four models with spatial correlation, goodness of fit was improved compared with the model including only heterogeneity. More important yet is the potential of spatial correlation to reduce the bias associated with model misspecification, as shown by the change in the estimate of the AADT coefficient and other parameters.

Statistical models that use count regression to link road crash frequency to the amount of travel on a segment have existed for many years (*1*–*3*). The current development of the Highway Safety Manual (*4*) has spurred even greater interest in road crash modeling. Road crashes occur at particular points in the highway network and are traditionally viewed as evolving from the interaction of driver, vehicle, roadway, and environmental factors. Of particular interest in this study is the modeling treatment of spatial factors (e.g., roadway factors) and their correlation across road segments.

There are several reasons to consider spatial correlation in crash models. One of the most important is that by using spatial correlation, site estimates pool strength from neighboring sites and improve model estimation. This reason is especially true in circumstances with high random variability in the data, as is the case with most crash data. This issue is referred to as the "small area estimation" problem in statistics (*5*). There have been a series of papers concerning this issue in the recent literature (*6*–*8*). After conducting a simulation study, one group of researchers (*6*) concluded that excess zeros are likely under fairly common conditions and that the excess zeros arise because of low exposure, inappropriate selection of time–space scales, or both. Crash reporting thresholds also influence road crashes reported to law enforcement, thus reducing the number of crashes that appear in agency databases compared with actual crash occurrence and what might be theoretically expected from a Poisson process. Crash sample size can also be reduced because of underreporting of low-severity, property-damage-only crashes.

Spatial dependence can be a surrogate for unknown and relevant covariates and can adjust for them. By using spatial correlation as a surrogate for unmeasured covariates, model misspecification of the mean structure can be reduced by accounting for a variable that is spatially varying, improving model estimation (*9*, *10*). In highway safety analysis, potential covariates that show variability in space include weather effects, driver population, and land use; these variables are rarely measured or accounted for in road safety models.

Ignoring spatial dependence can lead to underestimation of variability (*11*). This problem is particularly important in road safety models where random variability and small sample sizes are common issues. Here a more precise estimation of the variability in the parameters of interest is clearly advantageous for model interpretation.

Finally, learning about spatial dependence of crashes is of interest in its own right. One might be interested in knowing how close intersections or segments should be to consider them correlated, or if the spatial correlation is different for different types of roads, or how spatial relationships operate between segments at intersections. Furthermore, segments identified with significant spatial correlation can be grouped in corridors for further analysis and safety treatment.

One of the first crash analyses to include any type of spatial component was published by Levine et al. in 1995 (*12*). Crashes were geocoded to the nearest intersection or ramp, and then different spatial statistics were calculated including mean center, standard distance deviation based on great circle distance, the standard deviational ellipse (first and second principal component), and the nearest-neighbor index (based on the *x*- and *y*-coordinates of the accidents). A subsequent paper by the same authors (*13*) developed a spatial model at the census-block level, which was equivalent to a time series autoregressive lag-1 (AR-1) model. The previous time was replaced by the weighted average of the neighbors. The explanatory variables included in the model were freeway crossing the block (dummy), miles of arterials or highways, miles of minor roads, miles of freeways, population, and employment. Although the model takes into account the spatial correlation of the data, its weakness is the reliance on an assumed normal distribution for crashes rather than a discrete count probability distribution such as Poisson or negative binomial.

J. Aguero-Valverde, 201 Transportation Research Building, and P. P. Jovanis, 212 Sackett Building, Department of Civil and Environmental Engineering and Pennsylvania Transportation Institute, Pennsylvania State University, University Park, PA 16802. Corresponding author: J. Aguero-Valverde, jua130@psu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2061, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 55–63.
DOI: 10.3141/2061-07

Jones et al. (*14*) conducted a classical *K*-function analysis on the residuals of a logit model in which the log-odds were fatalities compared with the seriously injured. The variables of the model were age, type of user (pedestrian, bicyclist, motor vehicle driver), and number of casualties. With this ad hoc approach, the authors found that once the trend was removed from the data, the residuals presented clustering. These findings are consistent with those of Levine et al. (*13*) although limited to illustrating qualitative association rather than an estimate of the effect of spatial correlation through a statistical parameter estimate.

Nicholson (*15*) conducted a study that sought to identify the presence of nonrandom distributions of crashes by comparing actual spatial patterns with the complete spatial randomness case. Comparisons included stationary and isotropic (accidents not clustered but arranged regularly), nonstationary and isotropic (accidents clustered at randomly distributed points), and nonstationary and anisotropic (accidents clustered along lines). Different statistical tests for spatial randomness such as quadrant methods, nearest-neighbor methods, and *K*-function were analyzed. Nicholson concluded that nearest-neighbor methods appear more powerful and robust for detecting the kind of accident patterns that can be observed in practice and that the *K*-function method enabled patterns at different spatial scales to be detected.

Spatial correlation (or network autocorrelation as named by the authors) at the segment level was first explored by Black and Thomas (*16*) using Moran's index (a standard statistic used to measure the strength of spatial association among area units; it is analogous to the lagged autocorrelation coefficient in time series). The study concluded that there was a significant level of positive spatial correlation in the data, although again the results were descriptive.

A more advanced work in terms of spatial modeling of traffic crashes was developed by Miaou et al. (*17*), who estimated a series of spatial models of crashes aggregated at the county level for data from the state of Texas. Poisson-based Bayes models of fatal (K), incapacitating (A), and nonincapacitating (B) injuries were estimated by using both frequency and rate values (with vehicle miles of travel as an offset term). A Gaussian conditional autoregressive (CAR) model was used to model spatial correlation and a Markov chain Monte Carlo (MCMC) method was used to sample the posterior probability distribution.

Another analysis of crashes with spatial Bayesian models was performed by MacNab (*18*). With hospitalization data for 83 local health areas in British Columbia, Canada, between 1990 and 1999, determinants of motor vehicle accident injury were examined. Socioeconomic variables like marriage and immigration were used along with medical variables like life expectancy, health care providers, and hospital beds. In addition, the age effects were modeled by using a spline regression. Other variables such as miles of roads and seatbelt violations were also used for the model. Random spatial effects were included in the model and assumed to have a CAR distribution. Considerable spatial correlation was found in the data.

Aguero-Valverde and Jovanis (*19*) estimated full Bayes (FB) hierarchical models with spatial and temporal effects and space–time interactions by using injury and fatality data for Pennsylvania at the county level. Covariates include sociodemographics, weather conditions, transportation infrastructure, and amount of travel. A CAR model was used for modeling spatial correlation and a time trend coefficient was included to model temporal effects. Space–time interactions were modeled by using CAR random effects for each county multiplied by the time trend. Significant spatial correlation was found in the data even after several spatially distributed covariates such as population and weather conditions were included. Given the evidence of spatial correlation at such an aggregated level, the authors speculated that spatial correlation would be more important at smaller spatial scales such as the segment and intersection level.

Recently, Wang and Abdel-Aty (*20*) performed a temporal and spatial analysis of rear-end crashes at signalized intersections by using the generalized estimating equations approach. Intersections were grouped into clusters based on their spatial location, distances, and corridor location, resulting in clusters varying in size from 1 to 13 intersections (all intersections in a cluster belong to the same corridor). Intersections within a cluster were considered correlated, whereas intersections from different clusters were considered independent. It should be noted that although the correlation between intersections in the same corridor was estimated, spatial correlation between intersections in corridors that intersect each other was ignored. For the spatial models, the authors explored three different correlation structures: independent correlation, exchangeable correlation (constant correlations between any two intersections within a cluster), and AR-1 correlation, where the correlation decreases as the gap between intersections increases. A fourth spatial correlation structure, unstructured correlation, was tried but failed to converge; in this structure different correlations are estimated for each intersection pair within a cluster. The models showed high spatial correlations between intersections for rear-end crashes. More important, when spatial correlation structures were introduced, several coefficient estimates changed noticeably; this result might, again, be a reflection of model misspecification.

On the basis of this review of the literature, it is fair to conclude that the statistical and practical consequences of omitting consideration of spatial correlation in road segment crash models are largely unknown.

NEIGHBORING STRUCTURES

The literature indicates that authors have taken a variety of approaches in addressing spatial correlation—some at the aggregate or zonal level and a few at the link level. In this section the authors' conceptual approach to link-level spatial correlation is described, specifically the approach to estimate the effect of different neighboring structure specifications on model fit. Neighboring structures are explored from the simple to the more complex. The simplest neighboring structure is the first-order neighbors as shown in Figure 1. For the purpose of this analysis, first-order neighbors are defined as all segments that connect directly with the segment in question.

The schematic road network in Figure 1 also shows the second- and third-order neighbors for segment *i*. Second-order neighbors are those connected directly to first-order neighbors, and third-order neighbors are defined as those connected to second-order neighbors. This hierarchical definition of adjacency is strictly based on network topology and it ignores distances between segments.

As will be shown in the methodology section, different neighbors can have different weights. The hypothesis behind this analysis is none other than the first law of geography according to Tobler (*21*): "Everything is related to everything else, but near things are more related than distant things." The approach followed is to create models with only first-order neighbors, with first- and second-order neighbors, and with first-, second-, and third-order neighbors. For these cases, the weights assigned are equal to the inverse of the order (i.e., 1, ½, and ⅓). This simple, clear starting point is used for the current study; other, more complex weighting approaches are discussed as future research.

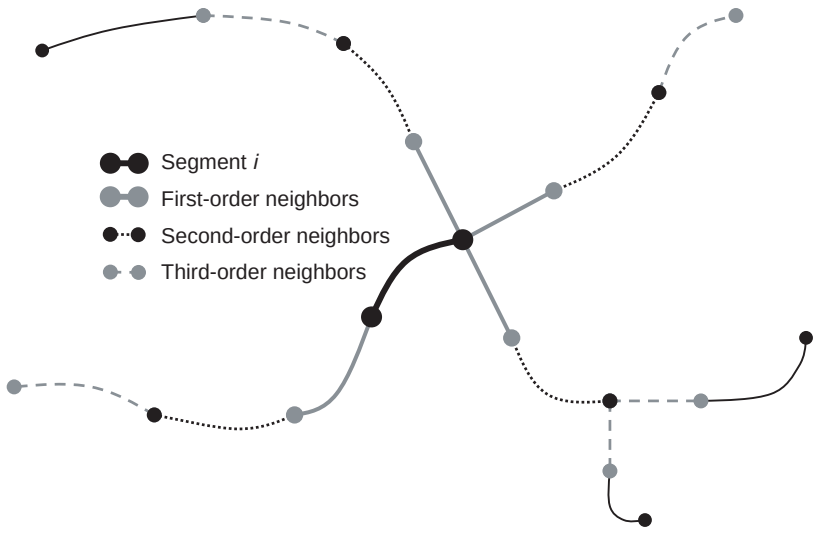

FIGURE 1 Neighbor structure definition.

MODEL IMPLEMENTATION WITH FB HIERARCHICAL MODELS

The recent enhancements in spatial modeling techniques have enabled researchers to investigate important issues related to risk estimation, unmeasured confounding variables, and spatial dependence in lattice systems such as road networks (*22*). Around 1990, the MCMC revolution took place, and methods like the Gibbs sampler and the Metropolis algorithm were coupled with faster computing to enable evaluation of complicated integrals that are usually found in Bayesian methods (*23*). Since then, Bayesian methods have been gaining popularity as the approach of choice in modeling multiple levels and incorporating random effects or complicated dependence structures.

Spatial and spatiotemporal models are better suited for Bayesian analyses in which complex correlation structures can be more easily implemented. FB hierarchical models are the more flexible approach and offer several advantages in the particular case of crash data analysis. One important characteristic of FB models is the explicit use of prior information to improve parameter estimates. The use of prior knowledge is a central part of the scientific method, and priors have a natural implementation in Bayesian analysis.

FB models also take full account of the uncertainty associated with parameter estimates and provide exact measures of uncertainty on the posterior distributions of these parameters. Frequentist and empirical Bayes (EB) methods traditionally ignore uncertainty in the correlation structures, which is translated into overestimation of the precision of parameters associated with the covariates, usually the most important quantities to estimate in the model (*24*).

Another important advantage frequently cited by Bayesians is that Bayes methods provide confidence (credible) intervals that are more in line with commonsense interpretations. For instance, a Bayesian credible interval for an unknown quantity of interest can be directly regarded as having a high probability of containing the unknown quantity, whereas a frequentist confident interval may be strictly interpreted only in relation to a sequence of similar inferences obtained by repeated sampling (*25*).

EB methods have been used in highway safety to reduce regression-to-the-mean bias (also known as selection bias) when sites are selected for engineering improvement. Typically, the prior in EB is obtained from the data (hence the empirical name), which is criticized by some as using the data twice (*26*). FB models provide correction methods for regression-to-the-mean in a different way: spatially unstructured and structured random effects can be included to smooth the estimates and reduce regression-to-the-mean bias.

Bayesian Modeling of Crashes

Bayes methods are known for their flexibility and have been recently used in analyses where spatial correlation plays an important role, such as in disease mapping. EB methods for highway safety analysis were proposed as early as 1981 (*27*). EB methods for ranking of sites by expected accident frequency as well as expected excess accident frequency have been used in several studies (*28–30*).

FB models have been used in highway safety only recently. In 1997, Schlüter et al. (*31*) proposed the use of FB models for ranking of sites based on three different criteria: posterior probability of selecting the worst site, predictive probability of future accident numbers, and expected number of future accidents. An FB approach was also used by Tanaru in 1999 (*32*) and 2002 (*33*) to model crash frequency with respect to several covariates.

Miaou and Song used an FB approach for road traffic crash mapping of sites with two different ranking criteria: ranking by probability that the site is the worst and ranking by posterior distribution of ranks (*34*). They also suggested the additional concept of a decision parameter that is site-specific and can include traffic flow, covariates, and space and time effects as well as random effects. Aguero-Valverde and Jovanis (*35*) also used FB hierarchical (Poisson lognormal) models to rank road segments for engineering improvement in Pennsylvania.

Qin et al. (*36*) used FB hierarchical models to fit zero-inflated Poisson models to four different crash types for highway segments from Michigan, California, Washington, and Illinois. Lord and Miranda-Moreno (*37*) had used the FB approach along with Monte Carlo simulation to estimate the effect of low sample mean and small sample size on the estimation of the dispersion parameter in Poisson–gamma models. Miranda-Moreno and Fu (*38*) also used FB

models in a comparison with EB approaches. Bayesian multivariate Poisson-log-normal models of crash counts were proposed by Ma et al. (*39*) as well as by Park and Lord (*40*).

Research Summary

The purpose of this research is to explore the effect of spatial correlation in models of road crash frequency at the segment level. Different segment neighboring structures are tested to establish the most appropriate one in the context of modeling crash frequency in road networks. Spatial dependency is modeled in terms of relationships such as in a directed graph.

METHODOLOGY

This section describes the general model specification and spatial correlation structures to be used in the research.

The FB hierarchical approach is used for model estimation. At the first stage the crash counts are modeled as a Poisson process, conditional on the rates:

$$Y_{it} \left| \theta_{it} \overset{\text{ind}}{\sim} \text{Pois}(\theta_{it}) \right. \tag{1}$$

where Y_{it} is the observed number of crashes in segment i at time t (in years), and θ_{it} is the expected Poisson crash rate for segment i at time t.

The Poisson rate is modeled as a function of the covariates following a lognormal distribution:

$$\ln(\theta_{it}) = \alpha + \ln(\text{length}_{it}) + \beta_V \ln(\text{AADT}_{it}) + \sum_k \beta_k X_{itk} + v_i + u_i \tag{2}$$

where

α = intercept,
β_V = coefficient for traffic volume for segment i at time t,
β_k = coefficient for indicator k,
X_{itk} = indicators for kth covariate for segment i at time t,
v_i = heterogeneity among segments, and
u_i = spatially correlated random effect for segment i.

Poisson lognormal specifications have been suggested recently in the context of highway safety analysis with FB hierarchical models as a better way to handle low sample mean, especially in comparison with the traditional Poisson–gamma or negative binomial approaches (*37*). It should be noted also that the length of the segment is included as an offset in the model, which means that crash frequency is considered proportional to segment length. Preliminary models showed that the coefficient for segment length was not significantly different from 1 (even though it was significantly different from zero); therefore, it was fixed to 1 in posterior analyses.

At the second stage, the coefficients are modeled by using normal priors, and the random effects are modeled by using a normal prior for v_i with a hyperprior to control the variance of the distribution:

$$v_i \overset{\text{iid}}{\sim} N\left(0, \frac{1}{\tau_v}\right) \tag{3}$$

where τ_v controls the amount of extra-Poisson variation due to heterogeneity among segments. The uncorrelated random effects (v_i) will basically reflect unmeasured differences among segments and are assumed to be independent and identically distributed.

The spatially correlated random effect is modeled by using a CAR prior, first proposed by Besag et al. (*41*):

$$u_i \left| u_{-i} \sim N\left(\frac{\sum_{j\sim i} w_{ij} u_j}{w_{i+}}, \frac{1}{w_{i+}\tau_u}\right)\right. \tag{4}$$

where τ_u controls extra-Poisson variation due to clustering, $j\sim i$ denotes that segment j is a neighbor of segment i, w_{ij} is the weight of jth neighbor of ith segment, and w_{i+} is the sum of weights of neighbors of segment i.

At the third stage, hyperpriors are given for τ_v and τ_u. To make these priors fair (i.e., with equal prior emphasis on heterogeneity and clustering), one might try to make them equal but that would be incorrect for two reasons: first, τ_v uses a marginal specification whereas τ_u uses a conditional specification and, second, τ_u is multiplied by the sum of the weight of neighbors w_{i+} in the prior specification. Bernardinelli et al. (*42*) suggested the following approximation as a reasonably fair specification:

$$\text{sd}(v_i) = \frac{1}{\sqrt{\tau_v}} \approx \frac{1}{0.7\sqrt{\bar{w}_+ \tau_u}} \approx \text{sd}(u_i) \tag{5}$$

where sd(·) is the standard deviation and $\bar{w}_+$ is the average sum of the weight of neighbors.

The posterior proportion of variation explained by the spatial correlation term (η) is also of interest:

$$\eta = \frac{\text{sd}(u)}{\text{sd}(u) + \text{sd}(v)} \tag{6}$$

Models are estimated with OpenBUGS 2.2 (*43*), which uses the Metropolis–Hastings algorithm to sample from the full unnormalized posterior distribution of interest, producing MCMC runs for each parameter. Quantities of interest such as means and variances are then estimated from these samples. Each model is run using two chains started from overdispersed locations. Generally, between 1,000 and 5,000 MCMC iterations are discarded as burn-in. Then 20,000 iterations for each chain are performed and summary statistics are estimated from both chains. Convergence is assessed by visual inspection of the MCMC trace plots for the model parameters as well as by using the Gelman and Rubin diagnostic (*44*). Model comparison and fit are assessed by using the deviance information criterion (DIC), which is considered the Bayesian equivalent of the Akaike information criterion (*45*).

DATA DESCRIPTION

The data for the models were obtained at the segment level for 4 years for the state-maintained rural undivided two-lane network of Centre County, located in central Pennsylvania and part of District 2-0 of the Pennsylvania Department of Transportation (PennDOT). A total of 865 rural two-lane segments were included in the analysis. A relational database was assembled with information from three different data sources: crash data, road inventory, and traffic data. All data were collected for calendar years 2003 to 2006.

Crash Data

Crash data were obtained from the PennDOT Crash Reporting System. The data include reportable crashes for road segment locations only (i.e., those that do not occur at an intersection or ramp junction). The data include state roads only and do not include Pennsylvania Turnpike crashes. A special location code was created for each crash by concatenating the county, route, and segment numbers in a single variable. This treatment created a unique location identification for each road segment. Then crashes were summarized by location code and year.

Road Inventory

Road data were obtained from the Pennsylvania road management system for the study period; the system includes data for each road segment such as county number, state route number, segment number, segment length, average daily traffic, lane width, travel lane count, posted speed limit, divisor type, functional class, and urban or rural code. These data were complemented with the state roads digital map from Pennsylvania Spatial Data Access (*46*) to be able to map crash locations. Summary statistics of the inventory data for the area of study are shown in Table 1.

RESULTS

Table 2 shows the models estimated: Poisson lognormal where the random effects reflect heterogeneity (v_i in Equation 2), Poisson lognormal where the random effects reflect clustering (first-order spatial correlation u_i in Equation 2), and a third model including heterogeneity and first-order spatial correlation. Models 4 and 5 provide estimates of heterogeneity and spatial correlation by using second- and third-order neighboring structures, respectively.

TABLE 1 Summary Statistics of Data by Segment and Year

	Mean	Std. Dev.	Min.	Max.
Variable				
Crashes	0.310	0.691	0	7
Volume (AADT)	2,636.4	3,197.7	45	18,749
Length (miles)	0.464	0.107	0.039	0.751
Indicators				
Functional class expressway and arterial	0.295	0.456		
Functional class collector and local	0.705	0.456		
Speed limit ≤ 35 mph	0.331	0.471	20	55
Speed limit > 35 mph	0.669	0.471		
Lane width < 10 ft	0.616	0.486	6	23.5
Lane width > 10 ft and < 12 ft	0.098	0.298		
Lane width 12 ft	0.017	0.128		
Lane width > 12 ft and < 14 ft	0.011	0.106		
Lane width ≥ 14 ft	0.095	0.294		
Shoulder width < 4 ft	0.608	0.488	0	14
Shoulder width > 4 ft and < 6 ft	0.161	0.367		
Shoulder width 6 ft	0.096	0.295		
Shoulder width > 6 ft and < 10 ft	0.099	0.299		
Shoulder width ≥ 10 ft	0.036	0.187		

Annual average daily traffic (AADT) is the most important covariate for explaining crash frequency in the proposed models. Other covariates such as lane width and shoulder width were included as indicator variables as shown in Table 1. For the case of functional classification, local roads and collectors were selected as the base case. For lane and shoulder width, 12 ft and 6 ft were selected as the base case, respectively, since they are the recommended designs for Pennsylvania. Speed limit is included in the data as a categorical covariate as well. Two categories are presented, with the lower-speed category (≤35 mph) being the base.

Models 1, 2, and 3 have DIC values of 4203, 4196, and 4180, respectively. Using a guideline suggested by Spiegelhalter et al. (*45*) of a difference in DIC of 7 needed for significance, Model 2 is significantly better than the model with only heterogeneity (Model 1), and the model with heterogeneity and spatial correlation (Model 3) is better than Models 1 and 2. Models 4 and 5 have DIC values of 4181; this is not an improvement with respect to the model with first-order neighbors only; as such, Model 3 is considered the model that best fits the data.

Concerning the trend in AADT first, Figure 2 shows the safety performance functions resulting from the estimates for Models 1, 2, and 3. The curves use base conditions for all the other predictors as described earlier. As shown, the model that includes only the heterogeneity effects presents the highest expected crash frequency for the range of AADT values in the data. However, Model 2, which includes only first-order adjacency spatial correlation, presents the lower expected crash frequency in the same range. Finally, the model including spatial correlation and heterogeneity is in the middle but far closer to Model 2 than to Model 1. The estimate for the AADT coefficient is affected by the inclusion of spatial correlation in Models 2 and 3, decreasing its magnitude in comparison with the value for the heterogeneity-only model. This change may be due to model misspecification, including the omission of spatial variables. As such, the parameter value for AADT in Models 2 and 3 is thought to better reflect its actual effect on the expected number of crashes. It should also be noted that the AADT parameter for Model 3 has an overlapping 95% credible set with Model 1. The overlapping 95% credible sets mean that the parameter values are indistinguishable statistically. Further research is needed with larger data sets and road segment types to verify the repeatability of this finding.

There are other changes in parameter estimates when spatial correlation and unstructured heterogeneity are added to the model. Segments with lane widths less than 10 ft are significant in the first model but not in Models 2 and 3 according to the 95% credible set. Shoulder widths less than 4 ft and 6 to 10 ft are not significant in Model 1 but rise to significance in Models 2 and 3. The signs for the coefficients for lanes widths less than 10 ft and shoulders between 6 and 10 ft are counterintuitive; one would have expected opposite results on the basis of the literature. More work needs to be conducted with additional data sets to clarify this result. Nevertheless, the fact that the parameters did change indicates the importance of considering spatial correlation in the model. Taken as a whole, these changes in parameter values indicate that spatial correlation is important at the segment level, as it was at the county level in a previous study (*19*).

Model 3 presents both the unstructured and cluster (spatial) random effects as shown in Equation 2. The priors were selected according to Equation 5. The proportion of variability in the random effects that is due to spatial correlation for Model 3 (η) is around 59% and significantly greater than 50%.

Figure 3 presents the rural two-lane segments with spatial correlation terms (u_i) significantly different from zero at 95% confidence.

TABLE 2 Models for Rural Two-Lane Center County Roads

Variable	Heterogeneity Only					Spatial Correlation Only (first-order adjacency)					Heterogeneity and Spatial Correlation (first-order adjacency)				
	Mean	Std. Dev	M.C. Error	2.5%	97.5%	Mean	Std. Dev	M.C. Error	2.5%	97.5%	Mean	Std. Dev	M.C. Error	2.5%	97.5%
Intercept	–6.007	0.510	0.011	–6.992	–5.028	–5.857	0.708	0.031	–7.234	–4.503	–6.149	0.739	0.032	–7.599	–4.672
Volume (AADT)	0.714	0.062	0.002	0.593	0.835	0.628	0.086	0.004	0.464	0.799	0.664	0.091	0.004	0.482	0.840
Functional class expressway and arterial	–0.034	0.145	0.005	–0.318	0.249	0.344	0.254	0.015	–0.153	0.830	0.228	0.251	0.015	–0.252	0.729
Speed limit > 35 mph	–0.225	0.092	0.001	–0.406	–0.045	–0.297	0.108	0.003	–0.506	–0.083	–0.301	0.116	0.004	–0.527	–0.071
Lane width < 10 ft	–0.516	0.191	0.004	–0.891	–0.143	–0.383	0.267	0.012	–0.912	0.143	–0.372	0.267	0.012	–0.906	0.160
Lane width > 10 ft and < 12 ft	–0.040	0.115	0.003	–0.264	0.186	0.097	0.184	0.009	–0.262	0.459	0.114	0.191	0.010	–0.278	0.484
Lane width > 12 ft and < 14 ft	–0.153	0.268	0.005	–0.687	0.363	–0.115	0.359	0.013	–0.834	0.582	–0.072	0.364	0.013	–0.797	0.632
Lane width ≥ 14 ft	0.422	0.273	0.004	–0.113	0.958	0.514	0.266	0.008	–0.021	1.029	0.596	0.311	0.008	–0.024	1.200
Shoulder width < 4 ft	0.240	0.146	0.003	–0.046	0.525	0.575	0.198	0.008	0.187	0.968	0.552	0.211	0.009	0.139	0.967
Shoulder width > 4 ft and < 6 ft	0.110	0.144	0.003	–0.172	0.390	0.291	0.193	0.008	–0.097	0.669	0.307	0.206	0.008	–0.097	0.708
Shoulder width > 6 ft and < 10 ft	0.082	0.142	0.003	–0.194	0.363	0.403	0.178	0.007	0.059	0.756	0.410	0.203	0.008	0.011	0.806
Shoulder width ≥ 10 ft	0.086	0.202	0.003	–0.310	0.478	0.302	0.229	0.007	–0.145	0.753	0.267	0.271	0.010	–0.265	0.791
sd(v)	0.563	0.048	0.002	0.470	0.659						0.420	0.050	0.002	0.327	0.518
sd(u)						0.629	0.055	0.003	0.532	0.749	0.616	0.056	0.003	0.519	0.743
H											0.595	0.036	0.002	0.528	0.667
Deviance	4,004	33.44	0.752	3,938	4,070	4,068	23.33	0.729	4,022	4,114	3,975	29.59	0.786	3,917	4,033
DIC	4,203					4,196					4,180				

Variable	Heterogeneity and Spatial Correlation (second-order adjacency)					Heterogeneity and Spatial Correlation (third-order adjacency)				
	Mean	Std. Dev	M.C. Error	2.5%	97.5%	Mean	Std. Dev	M.C. Error	2.5%	97.5%
Intercept	–6.195	0.699	0.028	–7.556	–4.858	–6.152	0.652	0.024	–7.450	–4.884
Volume (AADT)	0.676	0.083	0.004	0.514	0.838	0.676	0.078	0.003	0.527	0.829
Functional class expressway and arterial	0.209	0.217	0.011	–0.202	0.644	0.187	0.220	0.012	–0.246	0.610
Speed limit > 35 mph	–0.304	0.111	0.003	–0.519	–0.087	–0.296	0.108	0.002	–0.507	–0.085
Lane width < 10 ft	–0.354	0.261	0.012	–0.863	0.164	–0.374	0.259	0.011	–0.878	0.138
Lane width > 10 ft and < 12 ft	0.116	0.180	0.009	–0.242	0.466	0.101	0.178	0.009	–0.247	0.454
Lane width > 12 ft and < 14 ft	–0.043	0.345	0.011	–0.740	0.624	–0.055	0.337	0.012	–0.724	0.598
Lane width ≥ 14 ft	0.604	0.299	0.008	0.011	1.186	0.613	0.300	0.007	0.020	1.200
Shoulder width < 4 ft	0.529	0.199	0.008	0.143	0.916	0.504	0.196	0.008	0.126	0.893
Shoulder width > 4 ft and < 6 ft	0.308	0.200	0.007	–0.085	0.693	0.302	0.198	0.008	–0.087	0.689
Shoulder width > 6 ft and < 10 ft	0.408	0.193	0.006	0.038	0.785	0.394	0.185	0.007	0.029	0.754
Shoulder width ≥ 10 ft	0.264	0.259	0.008	–0.246	0.772	0.248	0.251	0.008	–0.245	0.734
sd(v)	0.423	0.050	0.002	0.329	0.526	0.423	0.050	0.002	0.329	0.526
sd(u)	0.566	0.073	0.005	0.459	0.778	0.534	0.064	0.004	0.431	0.687
H	0.572	0.041	0.002	0.496	0.659	0.558	0.042	0.002	0.474	0.642
Deviance	3,987	29.83	0.715	3,927	4,044	3,989	30.11	0.758	3,930	4,049
DIC	4,181					4,181				

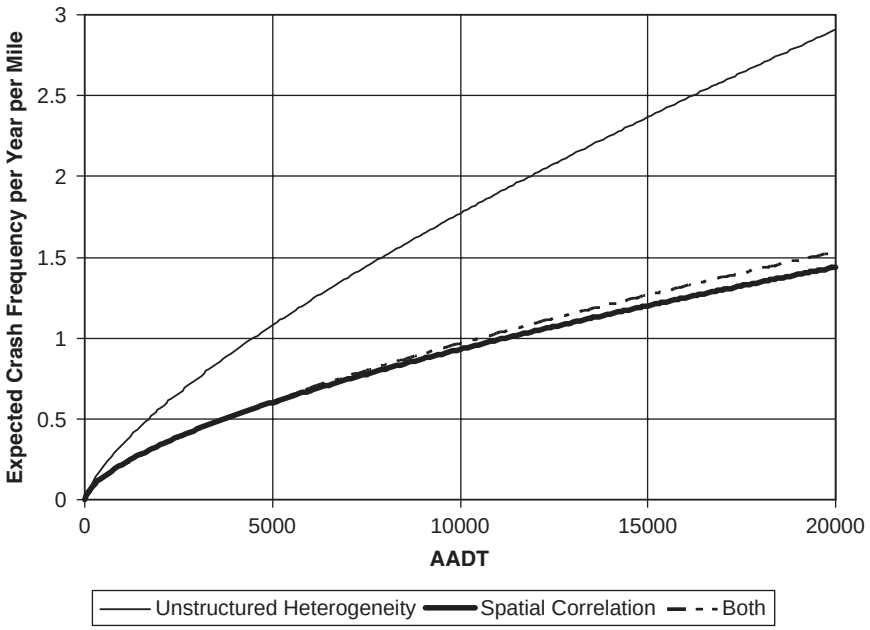

FIGURE 2 Safety performance functions for estimated models.

Clearly, segments with significant spatial correlation are clustered along well-defined corridors and in some cases those corridors even intersect other corridors with high spatial correlation. This analysis explores whether the model yielded spatially correlated segments that were contiguous. This finding would aid in the practical issue of deciding how many sections to treat as a group when highway funds are programmed. If, for example, the method yielded pairs of spatially correlated segments scattered throughout the network, the practical value would be negligible. Given the contiguous nature of the segments in Figure 3, one can conclude that, for this sample, the method yielded rational groupings of segments that can be used, along with other factors, to define projects for safety programming.

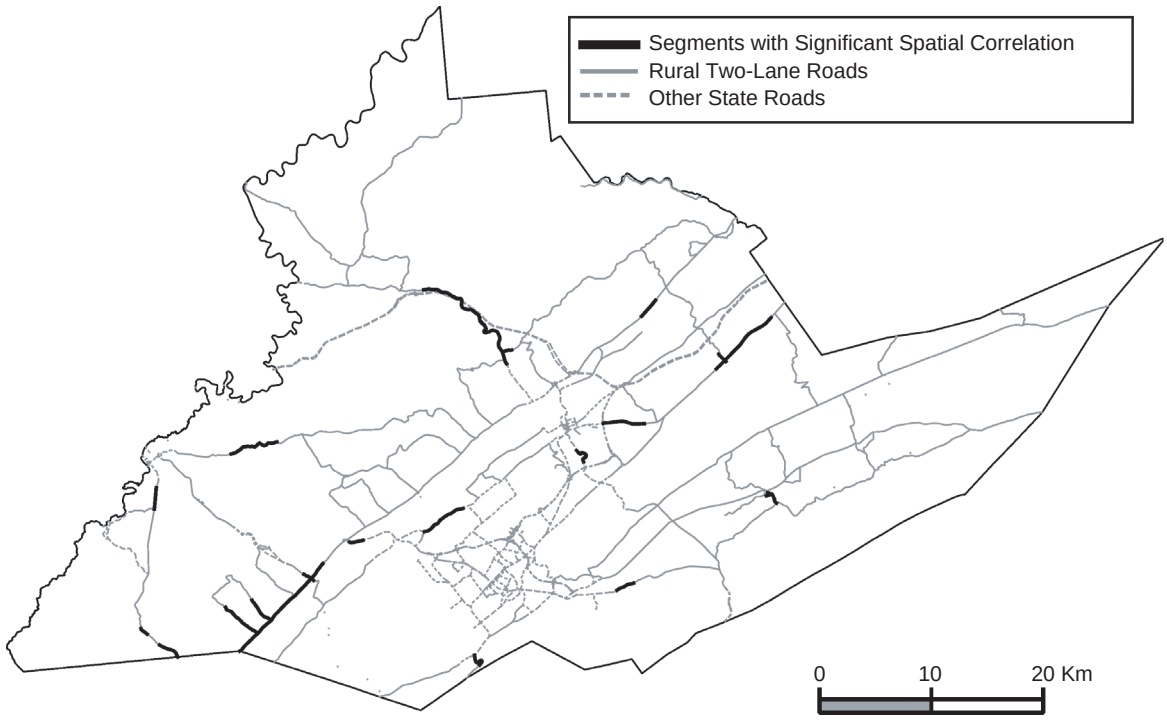

FIGURE 3 Rural two-lane segments with significant spatial correlation.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Analysis of crash, traffic, and roadway inventory data from a rural county in Pennsylvania indicate the importance of including spatial correlation in road crash models. The models with spatial correlation show significantly better fit to the data (in terms of the DIC) than the model with only heterogeneity. Even though the models with more complex spatial correlation structures (second- and third-order neighboring structures) are better than the model with only uncorrelated random effects, their goodness of fit is better than that for the model with uncorrelated heterogeneity and first-order spatial correlation. In the case of this analysis, the simplest spatial correlation structure (first-order adjacency) fits the data better.

The analysis also shows that given a fair prior for spatially unstructured and structured random effects, a higher proportion of variability is explained by the spatial correlation term. In the presence of completely uncorrelated random effects, the data still show spatial correlation, which is remarkable given the more restrictive nature of the spatial correlation structure.

More important yet is the potential of spatial correlation to reduce the bias associated with model misspecification, as shown by the change in the estimate of the AADT coefficient and other parameters. This preliminary result, if confirmed by further analysis, will have important implications in highway safety analyses since AADT is the most important variable explaining exposure in safety performance functions (SPFs). This variable is particularly critical in before-and-after studies in which the AADT changes significantly since the slopes of the SPF curves are significantly different. The result also implies that the absolute and relative deviation in the expected number of crashes from that of an average site will be underestimated if spatial correlation is not accounted for when sites are ranked for engineering improvements.

These initial findings indicate many areas of potential future research. Neighboring structures based on distance are a natural next step. The effect of spatial correlation at intersections and interchanges should also be further explored. In the current work, segments are considered first-order neighbors if they are adjacent; however, an enhanced model would consider other factors such as functional class to establish neighboring relationships. More complex neighboring structures should be considered at interchanges where the spatial array of lanes and ramps includes several levels.

Another issue with spatial correlation applied to road crash models in general is that road networks typically have different functional classes and design standards, but they work jointly as a network. This fact is ignored by traditional safety analysis methods that call for separated models for each road type. Multivariate models to include all road types in a single model for pooled spatial random effects can account for this factor. These models are said to be multivariate in the sense of estimating different coefficients for different road types but within a single model where the coefficients belong to a multivariate distribution.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of the Pennsylvania Department of Transportation, which provided the data for this analysis. The authors also thank anonymous reviewers for providing many insightful comments that improved the quality of the paper.

REFERENCES

1. Jovanis, P. P., and H. Chang. Modeling the Relationship of Accidents to Miles Traveled. In *Transportation Research Record 1068,* TRB, National Research Council, Washington, D.C., 1986, pp. 42–51.
2. Hauer, E. On the Estimation of the Expected Number of Accidents. *Accident Analysis and Prevention,* Vol. 18, No. 1, 1986, pp. 1–12.
3. Shankar, V., F. Mannering, and W. Barfield. Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention,* Vol. 27, 1995, pp. 371–389.
4. Hughes, W., K. Eccles, D. Harwood, I. Potts, and E. Hauer. *NCHRP Web Document 62: Development of a Highway Safety Manual.* Transportation Research Board of the National Academies, Washington, D.C., 2004.
5. Rao, J. N. K. *Small Area Estimation.* Wiley Interscience, New York, 2003.
6. Lord, D., S. P. Washington, and J. N. Ivan. Poisson, Poisson-Gamma, and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis and Prevention,* Vol. 37, No. 1, 2005, pp. 35–46.
7. Lord, D. Modeling Motor Vehicle Crashes Using Poisson-Gamma Models: Examining the Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter. *Accident Analysis and Prevention,* Vol. 38, No. 4, 2006, pp. 751–766.
8. Lord, D., S. P. Washington, and J. N. Ivan. Further Notes on the Application of Zero-Inflated Models in Highway Safety. *Accident Analysis and Prevention,* Vol. 39, No. 1, 2007, pp. 53–57.
9. Dubin, R. A. Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms. *Review of Economics and Statistics,* Vol. 70, No. 3, 1988, pp. 466–474.
10. Cressie, N. A. C. *Statistics for Spatial Data*. John Wiley & Sons, New York, 1993.
11. Congdon, P. *Bayesian Statistical Modeling.* John Wiley & Sons, England, 2001.
12. Levine, N., K. E. Kim, and L. H. Nitz. Spatial Analysis of Honolulu Motor Vehicle Crashes: I. Spatial Patterns. *Accident Analysis and Prevention,* Vol. 27, No. 5, 1995, pp. 663–674.
13. Levine, N., K. E. Kim, and L. H. Nitz. Spatial Analysis of Honolulu Motor Vehicle Crashes: II. Zonal Generators. *Accident Analysis and Prevention,* Vol. 27, No. 5, 1995, pp. 675–685.
14. Jones, A. P., I. H. Langford, and G. Bentham. The Application of K-Function Analysis to the Geographical Distribution of Road Traffic Accident Outcomes in Norfolk, England. *Social Science and Medicine,* Vol. 42, No. 6, 1996, pp. 879–885.
15. Nicholson, A. Analysis of Spatial Distributions of Accidents. *Safety Science,* Vol. 31, 1999, pp. 71–91.
16. Black, W. R., and I. Thomas. Accidents on Belgium's Motorways: A Network Autocorrelation Analysis. *Journal of Transport Geography,* Vol. 6, No. 1, 1998, pp. 23–31.
17. Miaou, S-P., J. J. Song, and B. K. Mallick. Roadway Traffic Crash Mapping: A Space-Time Modeling Approach. *Journal of Transportation and Statistics,* Vol. 6, No. 1, 2003, pp. 33–57.
18. MacNab, Y. C. Bayesian Spatial and Ecological Models for Small-Area Accident and Injury Analysis. *Accident Analysis and Prevention,* Vol. 36, 2004, pp. 1019–1028.
19. Aguero-Valverde, J., and P. Jovanis. Spatial Analysis of Fatal and Injury Crashes in Pennsylvania. *Accident Analysis and Prevention,* Vol. 38, No. 3, 2006, pp. 618–625.
20. Wang, X., and M. Abdel-Aty. Temporal and Spatial Analysis of Rear-End Crashes at Signalized Intersections. *Accident Analysis and Prevention,* Vol. 38, No. 6, 2006, pp. 1137–1150.
21. Tobler, W. R. A Computer Model Simulation of Urban Growth in the Detroit Region. *Economic Geography,* Vol. 46, No. 2, 1970, pp. 234–240.
22. Richardson, S. Statistical Methods for Geographical Correlation Studies. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies* (P. Elliott, J. Cuzick, D. English, and R. Stern, eds.), Oxford University Press, London, 1992, pp. 181–204.
23. Banerjee, S., B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data.* Chapman & Hall/CRC, 2004.
24. Goldstein, H. *Multilevel Statistical Models,* Second Edition. Edward Arnold, 1995.
25. Gelman, A., J. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis,* Second Edition. Chapman & Hall/CRC, 2003.
26. Carlin, B. P., and T. A. Louis. Empirical Bayes: Past, Present and Future. *Journal of the American Statistical Association,* Vol. 95, No. 452, 2000, pp. 1286–1289.

27. Abess, C., D. Jarret, and C. C. Wright. Accidents at Blackspots: Estimating the Effectiveness of Remedial Treatment, with Special Reference to the "Regression-to-Mean" Effect. *Traffic Engineering and Control*, Vol. 22, 1981, pp. 532–542.
28. Persaud, B., C. Lyon, and T. Nguyen. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1665*, TRB, National Research Council, Washington, D.C., 1999, pp. 7–12.
29. Heydecker, B. G., and J. Wu. Identification of Sites for Road Accident Remedial Work by Bayesian Statistical Methods: An Example of Uncertain Inference. *Advances in Engineering Software*, Vol. 32, 2001, pp. 859–869.
30. Miranda-Moreno, L. F., L. Fu, F. F. Saccomanno, and A. Labbe. Alternative Risk Models for Ranking Locations for Safety Improvement. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1908*, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 1–8.
31. Schlüter, P. J., J. J. Deely, and A. J. Nicholson. Ranking and Selecting Motor Vehicle Accident Sites by Using a Hierarchical Bayesian Approach. *The Statistician*, Vol. 46, No. 3, 1997, pp. 293–316.
32. Tanaru, R. Hierarchical Bayesian Models for Road Accident Data. *Traffic Engineering and Control*, Vol. 40, 1999, pp. 318–324.
33. Tanaru, R. Hierarchical Bayesian Models for Multiple Count Data. *Austrian Journal of Statistics*, Vol. 31, 2002, pp. 221–229.
34. Miaou, S-P., and J. J. Song. Bayesian Ranking of Sites for Engineering Safety Improvements: Decision Parameter, Treatability Concept, Statistical Criterion, and Spatial Dependence. *Accident Analysis and Prevention*, Vol. 37, No. 4, 2005, pp. 699–720.
35. Aguero-Valverde, J., and P. Jovanis. Identifying Road Segments with High Risk of Weather-Related Crashes Using Full Bayesian Hierarchical Models. Presented at the 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
36. Qin, X., J. N. Ivan, N. Ravishanker, and J. Liu. Hierarchical Bayesian Estimation of Safety Performance Functions for Two-Lane Highways Using Markov Chain Monte Carlo Modeling. *Journal of Transportation Engineering*, ASCE, Vol. 131, No. 5, 2005, pp. 345–351.
37. Lord, D., and L. F. Miranda-Moreno. Effects of Low Sample Mean Values and Small Sample Size on Estimation of Fixed Dispersion Parameter of Poisson-Gamma Models for Motor Vehicle Crashes: Bayesian Perspective. Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
38. Miranda-Moreno, L. F., and L. Fu. Traffic Safety Study: Empirical Bayes or Full Bayes? Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
39. Ma, J., K. M. Kockelman, and P. Damien. Bayesian Multivariate Poisson-Lognormal Regression for Crash Prediction on Rural Two-Lane Highways. Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
40. Park, E. S., and D. Lord. Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2019*, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 1–6.
41. Besag, J., J. York, and A. Mollié. Bayesian Image Restoration with Two Applications in Spatial Statistics. *Annals of the Institute of Statistical Mathematics*, Vol. 43, 1991, pp. 1–59.
42. Bernardinelli, L., D. Clayton, and C. Montomoli. Bayesian Estimates of Disease Maps: How Important Are Priors? *Statistics in Medicine*, Vol. 14, 1995, pp. 2411–2431.
43. Thomas, A., B. O'Hara, U. Ligges, and S. Sturtz. Making BUGS Open. *R News*, Vol. 6, No. 1, 2006, pp. 12–17.
44. Gelman, A., and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, Vol. 7, No. 4. 1992, pp. 457–472.
45. Spiegelhalter, D., N. Best, B. P. Carlin, and A. Linde. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society*, Vol. 64B, Part 4, 2002, pp. 583–639.
46. Pennsylvania Spatial Data Access. Pennsylvania State University. http://www.pasda.psu.edu. Accessed May 2007.

The Statistical Methodology and Statistical Computer Software in Transportation Research Committee sponsored publication of this paper.