A Doubly Stochastic Point Process Model for Modeling Crashes along a Corridor

Sami Demiroluk, M.Sc. (Corresponding Author)
Graduate Research Assistant,
Rutgers Intelligent Transportation Systems Laboratory
Department of Civil and Environmental Engineering,
Rutgers, The State University of New Jersey,
623 Bowser Road, Piscataway, NJ 08854 USA,
Tel: (732) 445-4012
E-mail: dsami@rci.rutgers.edu

Kaan Ozbay, Ph.D.
Professor, Department of Civil & Urban Engineering,
Polytechnic School of Engineering,
Center for Urban Science + Progress (CUSP), New York University (NYU)
One MetroTech Center, 19th Floor, Brooklyn, NY 11201, U.S.A.
Tel: +1 (646) 997 0552
Email: kaan.ozbay@nyu.edu

Word count:
Text: 5,721
Tables: 3 x 250
Figures: 3 x 250
Total: 7,221
Submission Date: August 1st, 2014
ABSTRACT

In this paper, a doubly stochastic model for point level modeling of the crashes is proposed. Unlike traditional approaches, the roadway is represented as a continuous entity in this model. Such model enables us to include covariates at different levels of resolution in the model. Many spatial covariates are incorporated into the model and their effects are analyzed. It is found that pavement characteristics, which are at point level, such as surface distress and rut depth significantly affect the crash risk. Moreover, it is observed that speed limit, number of lanes, lane drop and off ramp covariates are found to be significant and they might be associated with the drivers’ difficulty to adjusting to changing traffic and geometric conditions. Then, a Poisson MCMC model is developed at the link level using link level covariates for comparison purposes, since this is the common approach used for crash mapping. Poisson MCMC model was unable to capture the significant effects captured in the log-Gaussian Cox process (LGCP) model. Finally, the modeling results and Bayesian inference are used to develop risk maps for visually identifying the crash risks along the roadway. These maps showed that it is easier to pinpoint the locations with higher crash risk using the LGCP model. We believe that the maps generated from the modeling results gives better insights to transportation professionals then the maps based on simple link level crash counts since the point level maps allows zooming into the problematic locations with greater accuracy compared with the common approach of averaging out crashes over a predefined link segment. More accurate crash rate maps that can be used to pin-point local hot spots will make it easier to allocate funds for safety improvements more accurately by the state or federal governments.
INTRODUCTION

Crashes are serious events resulting in loss of property, injuries and lives. Effective ways to determine the higher risk zones on the roadways might provide invaluable information to the transportation professionals to locate higher risk zones. Mapping of crashes is a relatively new trend in safety analysis pioneered by Miaou et al. (1). Traditional models generally deal with the crash frequency (mapping) at the macroscopic level such as county and roadway levels. There are limited examples which try to mapping the crashes at the intersection and link levels. Another problem with the existing models is that, in general, they provide point estimates. Hence, these models lack the capability for providing the probabilistic output. This makes it harder to draw location specific conclusions based on those models.

The doubly stochastic model, which is also often referred as Cox process (2), might offer a solution to the some of the current issues faced in crash risk modeling. Cox process is an extension of Poisson process model where the intensity function of the Poisson process is modeled as a realization of a random field (3). One popular method used in the spatial modeling literature is the use of log-Gaussian random field and consequently this model is called log-Gaussian Cox process (LGCP) (4). There are several articles in other fields such as epidemiology (5), ecology (6) that use LGCP. For example, Diggle et al. (5) develops a model for the predictive probability that relative risk of a disease at a location exceeds a threshold set by the experts. In the model, Cox model is used to represent spatial variation in risk as a result of known risk factors and unexplained spatial variation. So the question arises, although they are very useful, why such models are not considered for mapping of crashes. First of all, the spatial point process models, so far, are used for modeling spatially continuous events such as contamination in a lake, the disease modeling, etc. In the case of crashes, we cannot directly threat them as spatially continuous events as they are bounded by the roadways. Moreover, spatial covariates are bounded by the roadways as well. Having many point locations in the data also introduces another problem, if the Bayesian statistical modeling is used in conjunction with the point models, MCMC simulations can be too complex. The simulations of such models might take hours or even days for a single run. Recently, a tool developed by Rue et al. (7) saves the researches from this struggle by using a method called R-INLA. Unlike, MCMC which gives exact results, R-INLA uses Integrated Nested Laplace Approximation (INLA) which tremendously reduces time and computing power needed for MCMC simulations. As the name suggest this approximation comes with a cost of lower accuracy than MCMC. But the research shows that in most cases the difference between MCMC and INLA results is negligible. The comparison of these methods is the out of scope of this paper and interested readers are advised to review the paper by Taylor and Diggle (8).

In this paper, we are proposing a doubly stochastic model for point level modeling of the crashes along a corridor. Then, a Poisson MCMC model, which is the general approach used in the crash mapping, is developed at the link level using link level covariates for comparison purposes. Finally, the modeling results and Bayesian inference are used to develop risk maps for visually identifying the crash risks along the roadway. To our knowledge, this is the first example of this approach for highway crash modeling. We believe that the maps generated from the modeling results gives better insights then the maps based on link level crash counts since the point level maps enable the decision maker to pinpoint the problematic locations instead of averaging out the crashes over a predefined link segment.
LITERATURE REVIEW

Previous research on the spatial analysis of crashes utilizes data aggregated at different levels. For example, Noland (9) analyzed the effects of infrastructure improvements on fatalities and injuries. Data from 50 states was used to develop a negative binomial regression model which includes infrastructure and socio-economic variables. Additionally, negative binomial models were used by various researchers to analyze crashes at the county level (10-12). While others utilized such models for modeling crashes along studied road sections (13-14).

Unfortunately, negative binomial regression models cannot handle spatial and temporal correlation (15-17). Therefore, more sophisticated methods are proposed by other researchers such as hierarchical Bayesian models (1, 15-16, 18-19).

Miaou et al. (1) developed a set of hierarchical Poisson Bayes models for estimating crash risk using crash frequencies for fatal, incapacitating and non-incapacitating injuries at the county level on low volume roads in Texas. Spatial covariates include percentage of time that the road surface is wet, number of sharp curves, and roadside hazards. However, surrogate variables are used in modeling due to limitations of the data. Conditional Auto-Regressive model (CAR) (20) is also used to model spatial correlation. Then, the samples are drawn from posterior probability distributions using Markov Chain Monte Carlo (MCMC) simulation. In the same study, it is recognized that although most of the disease mapping was done for area-based data, traffic crashes occur along the roadway network.

MacNab (15) developed a hierarchical Bayesian model to analyze variations of accident risk factors at the regional level. The study considers regional analysis of accident and injury variations, covariate effects, random spatial effects and age effects simultaneously. Hospital separation data for 83 local health areas in British Columbia (BC), Canada is used as the study area to investigate regional/contextual factors of accident injury among males between 0 to 24 ages. Different spatial covariates such as socio-economic indicators, residential environment indicators (roads and parks), availability of medical services and utilizations are included in the study. The model for injury rates assumes injury hospitalization counts to follow a Poisson distribution and the spatial random effects are modeled using a Markov random field (MRF) Gaussian distribution.

Aguero-Valverde and Jovanis (16) estimated hierarchical Bayesian models considering spatial and temporal effects and space–time interactions by using injury and fatality crash data from Pennsylvania at county level. Weather conditions, transportation infrastructure, socio-demographics and amount of travel are included as covariates in these models. Crashes are assumed to follow a Poisson distribution and spatial correlation; CAR model is used to quantify spatial correlation. Hierarchical Bayesian crash models are also compared to negative binomial regression models in the study. It is reported that while negative binomial and Bayesian models are consistent for significant variables. However, marginally significant variables in negative binomial models are not found significant in Bayesian models.

Quddus (18) developed a set of negative binomial regression and Bayesian hierarchical models for London crash data. Census wards in Greater London metropolitan area are used as the spatial unit. Variables in different categories such as traffic characteristics, road characteristics and socio-economic factors are considered as the explanatory variables. The relationships between these variables and crash casualties are analyzed in the study using both techniques. It is reported that “Bayesian hierarchical models are more appropriate in developing a relationship between area wide traffic crashes and the contributing factors associated with the road
infrastructure, socioeconomic and traffic conditions of the area”. This is an expect result since the hierarchical models can spatial dependence and heterogeneity, whereas negative binominal models fail to do so.

Huang et al. (19) analyzed the crashes at the county level in Florida using hierarchical Bayesian models. It is argued that the previous research on aggregate crash data did not explicitly differentiate exposure variables and risk factors which might have resulted in inconsistent results in similar studies. Hence, the exposure variables such as daily vehicle miles traveled and population was explicitly controlled in the study. Existence of spatial correlation between counties was checked by using Moran’s I. Moran’s I is a metric that range from -1 to 1 and a positive value indicates spatial correlation in the region under investigation whereas negative value indicates dispersion (21). It was found that the data was spatially correlated. They used the set of road and traffic related variables, and demographic and socio-economic variables in their model development.

Aguero-Valverde and Jovanis (22) examined the effect of spatial correlation in models of road crash frequency at the link level. They proposed different link neighboring schemes to determine the best one for modeling crash frequency in road networks. They used hierarchical Bayesian approach for modeling crash data and CAR model for the spatial correlation terms. Their analysis of rural roadways in Pennsylvania indicated the importance of including spatial correlation in highway crash models. In their study, the models with spatial correlation showed significantly better fit with the data than the Poisson lognormal model that incorporated only heterogeneity.

In her dissertation, Noyan (23), proposed the use of Cox process for crash modeling for the first time. A multi-level approach is used to model risk values from the severities of each individual involved in the crashes on Route 70, Ohio. A total risk value for each crash is calculated by giving different weights to different severities. Then, a hierarchical Bayesian model is developed for the risk estimation. Next, the risk values obtained from the risk estimation model are associated with crash locations as categorical marks. Finally, the intensity functions are estimated for different mark levels using Cox process. Crash risk profile map along the roadway are generated using these intensity functions to demonstrate the variation in the crash risk along the roadway. It is stated that these maps give traffic engineers and planners a powerful analysis tool to see the patterns in crashes of interest.

Ossenburgsen et al. (24) is the only example of the modeling of crash data at the point level mapping. They assumed that fatal and nonfatal crash rates are randomly distributed as Poisson processes and they developed a marked homogeneous Poisson process model. They analyzed the crashes that occurred in New Hampshire by constructing a grid of one square miles and assumed that traffic exposure (AADT) is unknown. The traffic exposure was estimated from a linear geostatistical model. They assumed a linear north-south trend (which may not always the case) and assigned continuous traffic exposure throughout the state (in reality, only the roadways are exposed to traffic). Finally, they proposed a detection scheme by calculating P value of each grid cell from their model and recommended its use as a decision making tool for the identification of hazardous locations along a roadway.
METHODOLOGY

Data Specification

New Jersey Department of Transportation (NJDOT)’s crash records database is used for developing the crash risk model (25). NJDOT keeps this yearly database of crashes occurred on all roadways in New Jersey. The database contains locations of the crashes along with many crash specific information. For this study, the spatial patterns of the crashes occurred on I-287 in 2011 are analyzed. I-287 is a 67.54 miles long interstate highway and it is one of the busiest roadways in the state of New Jersey.

For the modeling, a collection of cells of the region of interest is required. First, the roadway is divided to 0.1x0.1 mile regular grid lattice. As a result, 676 square cells are obtained. Then, the cell crash counts are calculated in ArcGIS using the locations of 2,844 crashes occurred on the roadway in 2011.

In addition to crash locations, several spatial covariates are considered. These include AADT, skid number, surface distress, rut depth, speed, number of lanes, lane drop, on-ramp, off-ramp, population variables. All covariates except population are extracted from the Roadway Centerlines database provided by NJDOT. Population data is obtained from the 2010 Census Block data (26).

Previous research shows that there is a relationship between pavement performance and crash occurrence (27). Hence, the effects of three pavement performance variables are investigated. These are skid number, surface distress, and rut depth. Skid resistance is the force developed when a tire that is prevented from rotating slides along the pavement surface (28). Skid resistance is an important pavement parameter because inadequate skid resistance might lead to higher incidences of skid related accidents. Skid resistance is generally quantified using some form of friction measurement such as a friction factor or skid number. Surface distress is another indication of poor or unfavorable pavement performance or signs of impending failure (29). Surface distress is related to roughness (the more cracks, distortion and disintegration, the rougher the pavement will be) as well as structural integrity (surface distress can be a sign of impending or current structural problems). Rutting is defined as the permanent or unrecoverable traffic-associated deformation within pavement layers that accumulates over time (30). Rut depth is a measure of permanent deformation due to rutting.

It is safe to assume traffic exposure directly affects the crash counts in each cell. Hence, AADT is considered as the exposure variable in the study. We also intend to observe the significance of number of lanes and lane drop at a location on spatial crash patterns so they are included as covariates in the model. Another pair of variables (on-ramp and off-ramp) are included to see if these locations are more prone to crashes. Last, the population variable is included to test the effect of the population of the nearest census block on the crashes patterns.

The descriptive statistics of the variables included in the study are given in Table 1.
TABLE 1 Descriptive statistics of the variables (N=676)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crashes</td>
<td>4.2071</td>
<td>4.266494</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>AADT</td>
<td>82665.25</td>
<td>19456.63</td>
<td>39118</td>
<td>117671</td>
</tr>
<tr>
<td>Skid Number</td>
<td>45.68402</td>
<td>9.590207</td>
<td>0</td>
<td>63.3</td>
</tr>
<tr>
<td>Surface Distress</td>
<td>3.447988</td>
<td>1.309394</td>
<td>1.27</td>
<td>5</td>
</tr>
<tr>
<td>Rut Depth</td>
<td>0.186095</td>
<td>0.102482</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>Link Speed</td>
<td>63.56509</td>
<td>3.508326</td>
<td>55</td>
<td>65</td>
</tr>
<tr>
<td>Number of Lanes</td>
<td>3.094675</td>
<td>0.66882</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Lane Drop (binary)</td>
<td>0.014793</td>
<td>0.120813</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Population</td>
<td>1917.778</td>
<td>812.1769</td>
<td>551</td>
<td>4392</td>
</tr>
<tr>
<td>On-ramp (binary)</td>
<td>0.136095</td>
<td>0.343143</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Off-ramp (binary)</td>
<td>0.122781</td>
<td>0.328429</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Modeling

Log-Gaussian Cox Model

A Cox process or ‘doubly stochastic’ process is an inhomogeneous Poisson process with stochastic intensity function $\tilde{\lambda}(x)$. The point process is defined by following two assumptions:

i) $\Lambda = \{\Lambda(x) : x \in \mathbb{R}^2\}$ is a non-negative valued stochastic process

ii) Conditional on the realization $\Lambda(x) = \tilde{\lambda}(x) : x \in \mathbb{R}^2$, the point process is an inhomogeneous Poisson process with intensity function $\tilde{\lambda}(x)$

In the spatial point process modeling context, intensity stands for the number of events per unit area. Hence, in the case of crashes, intensity represents the number of crashes per unit area.

Møller et al. (4) developed log-Gaussian Cox processes (LGCPs). LGCP is a special case of Cox process with $\Lambda(x) = \exp\{S(x)\}$, where $S$ is a Gaussian process. In this model, spatially varying intensity can be modeled by including one or more spatially indexed explanatory variables $z(x)$. The typical approach would be to retain the stationarity of $S(x)$ but to replace the constant intensity $\lambda$ by a regression model (31) as in the following:

$$\Lambda(x) = \exp\{\beta + S(x)\}$$ (1)

The resulting Cox process is now an intensity re-weighted stationary point process (32). This is the analogue for point process data of the linear Gaussian latent model with a spatially varying mean and a stationary residual (33). However, this approach provides a flexible and relatively tractable class of empirical models for describing spatially correlated phenomena. This makes it useful in applications where the scientific focus is on spatial prediction rather than on hypothesis testing (37).

Cox processes are very useful models for spatial point patterns (31). Cox process is also a suitable candidate for modeling the spatial distribution of crashes. The observed spatial pattern of crashes results from the spatial variation in the traffic exposure to observed and unobserved...
factors. On the other hand, these models may not be useful for modeling the spatial distribution of secondary incident sites. Since, in this case, the spatial pattern is a result of interactions between the crashes.

In this paper, we propose a LGCP model for crash locations with intensity $\Lambda(x) = E(x)R(x)$, where $E(x)$ stands for traffic exposure, and $R(x)$ represents crash risk, $R(x) = \exp\{S(x)\}$. Conditional on $R(\cdot)$, crash counts in grid cells $L_i$ of the corridor of interest are independent and Poisson distributed with means:

$$\mu_i = \int_{L_i} E(x)R(x)dx$$  \hspace{1cm} (2)

One advantage of LGCP is that it allows for the easy inclusion of spatial covariates into a model at different spatial resolutions. For the crash data, the covariates are incorporated into the model in the following form:

$$\Lambda(x) = E(x)\exp\{z(x)\beta + S(x)\}$$  \hspace{1cm} (3)

where $z(x)$ denotes the covariate surfaces and $\beta$ is the vector of parameters for covariates. Above equation applies to the continuous case. For the computation purpose, a discrete version of this equation is needed.

Assume that crash risk within each cell, $\Lambda_i, i=1,\ldots,676$, is constant or has a small spatial variation. Then we can use a GLM structure for the likelihood and estimate log crash risk (or intensity) in each cell in the following form:

$$\eta_i = \beta_0 + \log(E_i) + \sum_j \beta_j z_{ij} + S_i$$  \hspace{1cm} (4)

where $\eta_i$ is log crash risk in cell $i$, $\beta_0$ represent the intercept, $E_i$ is the exposure, $z_{ij}$ spatial covariates, $\beta_j$ the parameters for covariates and $S_i$ the spatial dependence. As the number of cells tends to go to infinity, this process behaves like its spatially continuous counterpart as shown in equation (3).

For spatial dependence, a valid spatial covariance function needs to be defined. For LGCP, a widely used family is the Matern ($\kappa$) covariance function:

$$C(u) = \sigma^2 r(u;\phi,\kappa)$$  \hspace{1cm} (5)

where

$$r(u;\phi,\kappa) = \left\{2^{\kappa-1}\Gamma(\kappa)\right\}^{-1}(u/\phi)^\kappa K_\kappa(u/\phi)$$

where $\Gamma(\cdot)$ is the complete Gamma function, $K_\kappa(\cdot)$ is a modified Bessel function of order $\kappa$, and $\phi>0$ and $\kappa>0$ are parameters. $\phi$ has units of distance and $\kappa$ is a dimensionless shape parameter that determines the differentiability of the corresponding Gaussian process. Generally $\kappa$ is selected from a set of values as it is hard to estimate it empirically. We use values of $\kappa = 0.5, 1.5, 2.5$ (31).
Considering a bounded region $\Omega \subseteq \mathbb{R}^2$, the likelihood for an LGCP has the following form (6):

$$
\pi(Y \mid \lambda) = \exp \left( \left| \Omega \right| - \int_{\Omega} \lambda(x) dx \right) \prod_{x \in Y} \lambda(x)
$$

(6)

where the integral is complicated by the stochastic nature of $\lambda(x)$. However, this integral can be numerically computed using simulation based methods such as Markov Chain Monte Carlo (MCMC). LGCP fits within the Bayesian hierarchical modeling framework. Moreover, it is a latent Gaussian model; hence, it can be easily embedded within the Integrated Nested Laplace Approximation (INLA) framework. INLA produces results faster and avoid the need to assess the convergence and mixing properties as in MCMC algorithm (7) because INLA is based on numerical integration rather than simulation (as in MCMC).

**Alternative Link Level Model: Poisson MCMC**

An alternative model to the one proposed above would be the estimation of the spatial variation using the hierarchical Poisson-Gaussian Markov random field model, due to Besag (20). This model is used in traffic safety literature for the past decade for modeling crashes at, generally, the aggregate level such as county level (1, 15-16, 18-19). However, there are also some examples of its use at intersections and at the link level (22). For comparison purposes, the model at the link level will be estimated for the same study corridor. This model would use the predefined links along the study corridor instead of the point locations. In this case, the crash counts will be assigned to the links.

The full hierarchical Bayesian modeling requires a three step approach. At the first step, conditional on mean, $\mu_i$, crash counts, $Y_i$ are assumed to follow a Poisson distribution:

$$
Y_i \sim \text{Po}(\mu_i)
$$

(7)

where $Y_i$ is the observed number of crashes in link $i$, $i=1,..,N$; and $\mu_i$ is the mean of the Poisson process for link $i$. The mean is further formulated as:

$$
\mu_i = e_i \lambda_i
$$

(8)

In the above equation, crash risk rate, $\lambda_i$, is assumed to be proportional to traffic exposure, $e_i$, hence AADT is considered as an offset for exposure to traffic. Then, the crash risk rate, $\lambda_i$, is modeled as:

$$
\log(\lambda_i) = \alpha + \sum_k \beta_k X_{ik} + \theta_i + \phi_i
$$

(9)

where $\alpha$ represents the intercept term, $\beta_k$ is the coefficient for spatial covariate $k$, $X_{ik}$ is the observed value of $k$th covariate for link $i$, $\theta_i$ stands for spatial heterogeneity, $\phi_i$ represents spatially correlated random effects for link $i$. 

TRB 2015 Annual Meeting  Paper revised from original submittal.
Next, coefficients are modeled by using non-informative priors. For the random effect terms, second level of hierarchy is defined. The heterogeneity term is modeled using a normal prior:

$$\theta_i \sim N \left(0, \frac{1}{\tau_h}\right)$$

(10)

where $\tau_h$ is a precision parameter that controls the amount of $\theta_i$ among municipalities. The heterogeneity term enables us to include extra-Poisson variability due to unobserved variables over the entire state. For modeling the spatially correlated random effects, a CAR prior is adopted as proposed by Besag (20).

$$\phi_i | \phi_{-i} \sim N \left(\sum_{j \neq i} \frac{w_{ij}}{w_{ii}} \phi_j, \frac{1}{\tau_c w_{ii}}\right)$$

(11)

where $\tau_c$ is a precision parameter that controls clustering, $ij$ is a neighbor link adjacent to link $i$, $w_{ij}$ is the weight of the neighbor $j$, and $w_{ii}$ represents the sum of the weights of the neighbors of link $i$. Note that, in this study, it is assumed that all neighbors have equal weight since this is common practice in the literature. By including a spatially correlated random effects term, extra-Poisson variability in the log-relative risk which varies from municipality to municipality can be modeled in such a way that nearby municipalities will have similar rates. In the next step, the same non-informative hyperpriors are defined for heterogeneity and clustering precision parameters.

**RESULTS AND DISCUSSION**

Based on the model specifications presented in the previous section, LGCP for point level data and Poisson MCMC for link level data are estimated. Model estimation results are presented in Table 2 and Table 3.

The spatial covariates in the data have different resolutions. LGCP model is a versatile approach for easily incorporating these covariates. Based on the 95% credible set (which is counterpart of confidence interval in Bayesian statistic) of the posterior distribution of the variables, we can make the following observations (for the variables presented in Table 1):

- The skid number has a positive relationship with respect to the crash risk. However, the credible set contains zero, hence, the skid number is insignificant. These results suggest that the skid resistance along the roadways is sufficient; hence, it does not impact crash risk significantly.

- The surface distress index is significant and negatively correlated with the crash risk. Before interpreting this value, it is necessary to give some background information on the surface distress index. Surface distress index is a value given to a roadway section on a scale of 0-100. While the index value of 100 would indicate a perfect road with no measurable distress or rough ride, an index value of 60 is considered as terminable serviceability and the road is considered failed (35). Hence, this result suggests that the rough roadway section increases the crash risk.
Rut depth is significant and has a positive relationship with the crash risk. Therefore, the higher degree of permanent traffic-related deformation of the pavement can be associated with the higher crash risk on the roadway.

Speed limit is found to be significant and negatively correlated with the crash risk. Although this sounds counter intuitive, this might point out to the fact that the drivers might not be following lower speed limits along certain roadways and consequently experience greater crash risk at the lower speed locations.

Number of lanes has a positive and significant effect. This variable might be correlated with the traffic exposure to some degree. Therefore, the positive value of this covariate is attributed to its relationship with the traffic exposure.

Lane drop is found to be significant and positively related to the crash risk. It might be due to the difficulty of drivers to adapt changing traffic conditions.

Population is found to have no impact on the crash risk. Some might expect the population is another indicator of the traffic exposure. However, the model uses fine grid of cells and there is a good chance that the population of the nearest town might not greatly affect the interchange on the roadway.

The existence of an on ramp variable is found to be insignificant in terms of its effect on the crash risk. Drivers might be more cautious while entering the roadway due to yield or stop signs. Hence, the crash risk is not significantly affected by the on ramp covariate.

On the other hand, off ramp covariate has significant and positive impact on the crash risk. This might also be another sign that the drivers need to take their time and be more careful in decelerating and merging before exiting the roadway to avoid higher risk of crashes.

Spatial dependence of grid cells is found to be a significant variable. Figure 1 shows the range of the marginal posterior distribution of the range of the spatial dependence. It can be seen that the range of the highest spatial dependence around ±0.5 miles and the spatial dependence is diminishes around ±1 miles.

### TABLE 2 The results of LGCP model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>S.D</th>
<th>2.50%</th>
<th>97.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.4897</td>
<td>1.0859</td>
<td>1.3549</td>
<td>5.6196</td>
</tr>
<tr>
<td>Skid number</td>
<td>0.008</td>
<td>0.0051</td>
<td>-0.0021</td>
<td>0.018</td>
</tr>
<tr>
<td>Surface distress</td>
<td>-0.1615</td>
<td>0.0431</td>
<td>-0.2461</td>
<td>-0.0769</td>
</tr>
<tr>
<td>Rut depth</td>
<td>1.0669</td>
<td>0.4557</td>
<td>0.1731</td>
<td>1.9626</td>
</tr>
<tr>
<td>Speed limit</td>
<td>-0.0479</td>
<td>0.0134</td>
<td>-0.0742</td>
<td>-0.0215</td>
</tr>
<tr>
<td>Number of lanes</td>
<td>0.2871</td>
<td>0.0763</td>
<td>0.1374</td>
<td>0.437</td>
</tr>
<tr>
<td>Lane drop</td>
<td>0.5651</td>
<td>0.2401</td>
<td>0.0942</td>
<td>1.037</td>
</tr>
<tr>
<td>Population</td>
<td>0</td>
<td>0.0001</td>
<td>-0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>Onramp</td>
<td>0.1321</td>
<td>0.0804</td>
<td>-0.0258</td>
<td>0.2898</td>
</tr>
<tr>
<td>Offramp</td>
<td>0.2737</td>
<td>0.0837</td>
<td>0.1093</td>
<td>0.437</td>
</tr>
<tr>
<td>Spatial dependence</td>
<td>0.7655</td>
<td>0.2663</td>
<td>0.3441</td>
<td>1.3735</td>
</tr>
</tbody>
</table>

\[ \hat{D} : 2618.56, D(\hat{\theta}) = 2255.05, p_d = 363.52, DIC = 2982.08 \]
For Poisson MCMC model, only the covariates at the link level can be included in the model as seen in Table 3. Based on the results of the Poisson MCMC model:

- Link length has a positive effect but it is statistically insignificant which highlights that the longer links do not necessarily have higher crash risk.
- Number of lanes has a positive effect but it is found to be statistically insignificant. In the previous model, we claim that number of lanes might be an indication of exposure. In this model, there is a chance that the traffic exposure is represented better by another covariate which possibly makes “number of lanes” variable insignificant.
- Lane drop is also found to have a positive effect but also statistically insignificant. This covariate is found to be significant in LGCP model. It is possible that assigning lane drop to a link might average out the effect of lane drop. On the other hand, for the point level model the effect of this covariate can be captured better.
- The population covariate is found to be positive and significant. Unlike LGCP model, the population covers all neighboring census blocks along a link and there is an interchange at the start and at the end of the link. Therefore, the population might be a sign of the demand on the roadway (traffic exposure).
- Spatial correlation and spatial heterogeneity terms are both found to be positive and statistically significant. Spatial correlation is larger than spatial heterogeneity which indicates that the clustering of crash risks along the roadway.
### TABLE 3 The results of link level MCMC model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>S.D.</th>
<th>MC Error</th>
<th>2.50%</th>
<th>97.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.225</td>
<td>0.05588</td>
<td>3.19E-04</td>
<td>4.112</td>
<td>4.334</td>
</tr>
<tr>
<td>Link length</td>
<td>0.1106</td>
<td>0.06939</td>
<td>0.002209</td>
<td>-0.0254</td>
<td>0.2494</td>
</tr>
<tr>
<td>Number of lanes</td>
<td>0.2044</td>
<td>0.1486</td>
<td>0.004937</td>
<td>-0.09799</td>
<td>0.4936</td>
</tr>
<tr>
<td>Lane drop</td>
<td>0.233</td>
<td>0.1113</td>
<td>0.0063</td>
<td>-0.1961</td>
<td>0.6433</td>
</tr>
<tr>
<td>Population</td>
<td>0.08772</td>
<td>0.03398</td>
<td>0.001125</td>
<td>0.02004</td>
<td>0.1538</td>
</tr>
<tr>
<td>SD of Spatial Correlation</td>
<td>0.3137</td>
<td>0.09144</td>
<td>0.004171</td>
<td>0.1044</td>
<td>0.4795</td>
</tr>
<tr>
<td>SD of Spatial Heterogeneity</td>
<td>0.2954</td>
<td>0.07618</td>
<td>0.00383</td>
<td>0.02174</td>
<td>0.4148</td>
</tr>
</tbody>
</table>

\[\bar{D} : 255.696, D(\theta) = 223.007, p_{D} = 32.688, DIC = 288.384\]

The two models estimated in this paper have different number of parameters and covariates. Hence, it is not possible to compare the model fit based on DIC values which is used in the analysis of model fit for Bayesian hierarchical models. Instead, their ability of accurately mapping the crash risk is investigated. For this purpose, the crash risk maps shown in Figure 2 are generated. While the map on the left side of Figure 2 represents the posterior mean of crash risk estimated from LGCP model, the one on the right side shows the posterior mean of crash risk based on Poisson MCMC model. To get the same scale in both cases, the crash risk from the link level model is normalized by its length*10, as the point level grids cells are 0.1 mile long. Figure 2 shows five categories of crash risks based on natural breaks, which is a data clustering method to determine the best arrangement of classes by minimizing each class mean from its standard deviation, in LGCP model. From Figure 2, the high crash risk locations are observed in both cases. However, Poisson MCMC model aggregates the crash risk over each link which makes it impossible to pinpoint locations with the highest risk along an individual link. On the other hand, since LGCP model uses a fine grid for estimation, the locations with higher crash risk can be determined with more accuracy. To demonstrate this difference better, a section of the roadway between mileposts 0 and 13.5 is zoomed in Figure 3.
FIGURE 2 Point level (left) and link level (right) crash risks in 2011
CONCLUSION AND FUTURE WORK

This paper presents a point level LGCP model for interstate highway I-287 in New Jersey. Unlike Bayesian models estimated at a more aggregate level, this new model specification enables us to include covariates at different resolutions in the model. Moreover, the roadway is represented as a continuous entity without predefined links in this model.

Many spatial covariates are incorporated into the model and their effects are analyzed. It is found that pavement characteristics, which are at the point level such as, surface distress and rut depth significantly affect the crash risk. Moreover, speed limit, number of lanes, lane drop and off ramp covariates are found to be statistically significant and they might be associated with the drivers’ difficulty to adjusting to changing traffic and geometric conditions. Population covariate is found to be insignificant and this result might be due to the fact that the point level model might not be correlated with the population of surrounding municipalities since the model is developed for a limited access interstate roadway that does not directly interact with local municipalities / towns at each point.

Then, a Poisson MCMC model is developed at the link level using link level covariates for comparison purposes. Poisson MCMC model was unable to capture some of the statistically significant covariates captured in the LGCP model. However, in this model, population was found to be significant. Since it is at the link level and the interchanges are located at the start and end of link, it is natural for the population at these entry and exit locations to affect the traffic exposure and the crash risk.

Finally, the crash risk maps are developed based on the modeling results. These maps show that it is easier to visually detect locations with higher crash risk using the LGCP model. Hence, LGCP model captures the continuous nature of the roadways better than any existing model that pre-discretizes the highway into links between interchanges or some other major monuments. Although this model is used in other fields, to our knowledge, this is the first attempt to employ it in the context of traffic safety. It is believed that the results of this study will help transportation professionals in identifying and ranking high risk locations at the point level. More accurate crash rate maps that can be used to pin-point local hot spots will make it easier to allocate funds for safety improvements more accurately by the state or federal governments.
For future studies, it is recommended that the methodology described in this paper is extended to the analysis of point level crashes in a spatio-temporal setting. This will allow observation of temporally varying effects of the crash risk. Sensitivity of the model to different cell sizes needs to be investigated as the cell size selection might affect the modeling results. Future efforts should also focus on incorporating the severity of crashes in the crash risk model. Different crash severities can be treated as marks and their inclusion might give better insights about the variation of the crash risks among crashes with different severities.

ACKNOWLEDGEMENTS

The contents of this paper reflect views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents of the paper do not necessarily reflect the official views or policies of the agencies.
REFERENCES


