Integrating Machine Learning and Bayesian Nonparametrics for Flexible Modeling of Point Pattern Data

Matthew J. Heaton[a,*], Benjamin K. Dahl[a], Caleb Dayley[a], Richard L. Warr[a] and Philip White[a]

[a]Department of Statistics, Brigham Young University, 2152 WVB, Provo, UT, 84602, USA

ARTICLE INFO

Keywords:
Mixture models
Log-Gaussian Cox process
Dirichlet process
Penalized least squares

ABSTRACT

Two common approaches to analyze point pattern (location-only) data are mixture models and log-Gaussian Cox processes. The former provides a flexible model for the intensity surface at the expense of no covariate effect estimates while the latter estimates covariate effects at the expense of computation. A bridge is built between these two methods that leverages the strengths of both approaches. Namely, Bayesian nonparametrics are first used to flexibly model the intensity surface. The posterior draws of the fitted intensity surface are then transformed into the equivalent representation under the log-Gaussian Cox process approach. Using principles of machine learning, estimates of covariate effects are obtained. The proposed two-step approach results in accurate estimates of parameters, with proper uncertainty quantification, which is illustrated with real and simulated examples.

1. Introduction and Problem Background

1.1. Preliminaries

Let $\boldsymbol{s}_i \in \mathcal{D} \subset \mathbb{R}^D$ denote the location of an event of interest (e.g. the location of a car crash or the location of a disease incidence) and $\mathcal{D}$ the domain on which events can occur. A set of N events $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$ is referred to as a "point pattern" on the domain $\mathcal{D}$. Depending on the domain, events can occur across time ($D = 1$), space ($D = 2$), or space-time ($D = 3$).

Point pattern data are becoming increasingly common due to the ease at which locations can be geocoded, with recent books such as Daley, Vere-Jones et al. (2003); Moller and Waagepetersen (2003); Snyder and Miller (2012) giving many examples. The motivation for this research comes from the two point patterns displayed in Figure 1. The left panel of Figure 1 displays locations of crashes along Interstate-5 (I-5) in Washington, USA along with an associated kernel density estimate of the crashes. In this example, the "event" is a crash at mile point s_i in the domain corresponding to I-5 (a one dimensional spatial point pattern). The right panel of Figure 1 displays crash locations along I-15 in Utah, USA and the associated day the crash occurred along with a two-dimensional kernel density estimate in the background (a spatio-temporal point pattern). In this example, $\mathbf{s}_i = (m_i, t_i)'$ is the milepoint along I-15 (m_i) and the time of the crash (t_i). In both of these examples and also in most analyses of point patterns, the main statistical goal of an analysis is to (i) estimate the rate of event occurrences at all locations in the domain (i.e. identify locations where automobile crashes occur at a higher than expected rate given the traffic level) and (ii) estimate the effect of covariates, if any, on the rate of event occurrence (i.e. link the rate of a crashes to various roadway characteristics such as speed limit).

In analyzing point pattern data, the number of events N and the event locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ are the random variables to be modeled. One such model common in the literature is the Poisson process. In the Poisson process framework, both the location and number of events are governed by an intensity surface, which we will denote by $\Lambda(\boldsymbol{s})$. First, the number of events $N \sim \mathcal{P}\left(\int_D \Lambda(\boldsymbol{u})d\boldsymbol{u}\right)$ where $\mathcal{P}(\cdot)$ denotes the Poisson distribution. Second, conditional on N, the location of the N events are assumed to be independent with distribution given by the normalized intensity $\lambda(\boldsymbol{s}) = \Lambda(\boldsymbol{s})/\int_D \Lambda(\boldsymbol{u})d\boldsymbol{u}$ (notationally we write $\boldsymbol{s}_i \overset{iid}{\sim} \lambda(\boldsymbol{s})$). The kernel density estimates displayed in Figure 1 are estimates of these normalized intensity surfaces.

*Corresponding author

ORCID(s): 0000-0003-4654-9827 (Matthew J. Heaton); 0009-0008-2502-2210 (Benjamin K. Dahl); 0000-0001-8508-3105 (Richard L. Warr); 0000-0003-0907-9221 (Philip White)

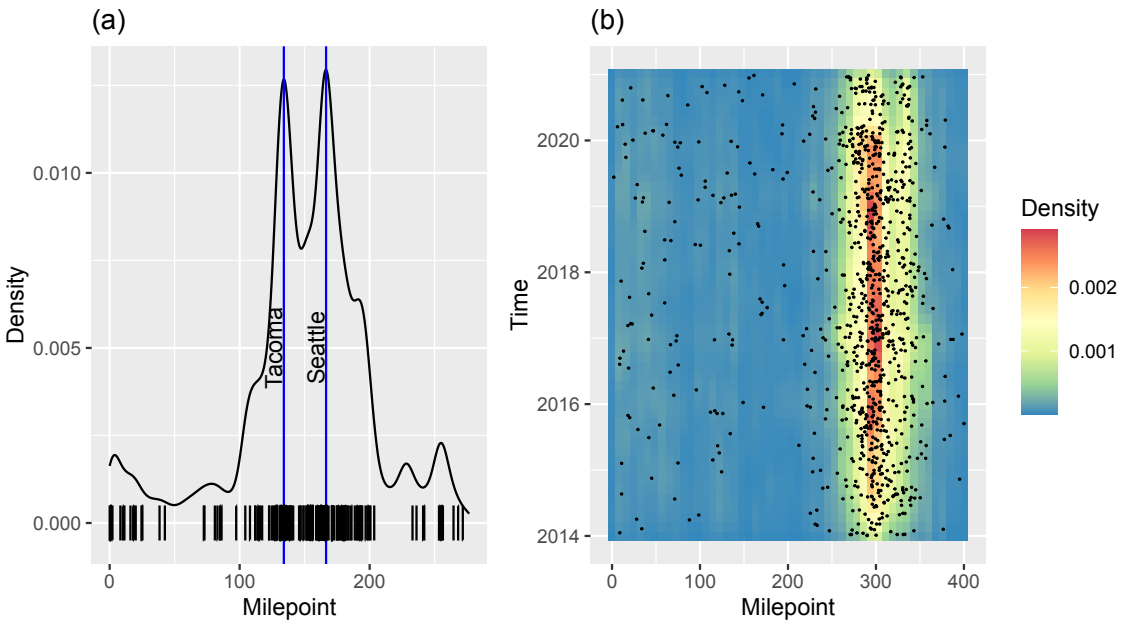

Figure 1: Motivating point pattern datasets. (a) Locations (given by the vertical bars) of automobile crashes along I-5 in Washington State, USA in the year 2017 along with an associated density estimate. (b) Locations and day of the year of automobile crashes along I-15 in Utah, USA along with associated density estimate.

The popularity of the Poisson process as a model for point pattern data is due, in part, to its ability to address both statistical goals mentioned above. First, locations where $\Lambda(\boldsymbol{s})$ is high are locations where the rate of events is high. And, second, covariate effects can be estimated by hierarchically modeling $\Lambda(\boldsymbol{s})$ as a function of the covariates. For example, researchers can set $\mathbb{E}(\log(\Lambda(\boldsymbol{s}))) = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta}$ where $\boldsymbol{x}'(\boldsymbol{s}) = (x_1(\boldsymbol{s}), \dots, x_P(\boldsymbol{s}))'$ is a set of covariates (or bases) associated with location $\boldsymbol{s}$. Hence, under the Poisson process framework, if $\Lambda(\boldsymbol{s})$ can be estimated from the data, both analysis goals mentioned above can be achieved.

1.2. Common Approaches

While, in theory, the Poisson process is useful for analyzing point pattern data, practically there are several challenges in doing so. First, the continuous nature of $\Lambda(\boldsymbol{s})$ can be problematic computationally. To avoid this issue altogether, many analyses will discretize (partition) the domain $\mathcal{D}$ then count and model the number of events per partition (see Aguero-Valverde, 2013; Barua, El-Basyouny and Islam, 2016; Gomes, Cunto and da Silva, 2017; Zeng, Gu, Zhang, Wen, Lee and Hao, 2019, for examples). This leads to straightforward generalized linear modeling using a Poisson or negative binomial distribution as the likelihood. While discretizing $\mathcal{D}$ has been used successfully to model point pattern data, the main issue with this approach is that the choice of partition is arbitrary, with no clear guidance on the best partition for the data. Further, by modeling the data as counts in regions rather than continuously, important structure in $\Lambda(\boldsymbol{s})$ within a region (partition) is lost.

Given the issues with discretizing the domain, an alternative approach to capture the continuous intensity surface are mixture models. In the mixture model approach, $\Lambda(\boldsymbol{s}) = \delta\lambda(\boldsymbol{s})$ where $\lambda(\boldsymbol{s})$ is the normalized intensity (i.e. a probability density function on $\mathcal{D}$) and $\delta = \int_{\mathcal{D}} \Lambda(\boldsymbol{u})d\boldsymbol{u}$ is a scalar representing the expected number of events. The normalized intensity $\lambda(\boldsymbol{s})$ is represented as a mixture where $\lambda(\boldsymbol{s}) = \sum_{k=1}^{K} \omega_k f_k(\boldsymbol{s})$ where $\{\omega_k\}$ are mixture weights and $f_k(\boldsymbol{s})$ is a valid density on $\mathcal{D}$ (e.g. truncated Gaussian). This approach, as well as Bayesian nonparametric extensions, have been successfully adopted by, among others, Hougaard, Lee and Whitmore (1997); Kottas and Sansó (2007); Taddy and Kottas (2012); Zhou, Matteson, Woodard, Henderson and Micheas (2015); Jiao, Hu and Yan (2021); Zhao and Kottas (2021); Geng, Shi and Hu (2021); Yin, Jiao, Yan and Hu (2022). The appeal of this approach is that mixtures are highly flexible and capture complex, nonlinear, continuous structure in $\lambda(\boldsymbol{s})$ and are often computationally reasonable to fit. However, the downside of the mixture approach is that there is not a clear way to include covariates in the mixture

representation that control the associated height of the intensity. Hence, these mixture approaches often do not include the use of covariate information.

When estimating the effect of covariates is important to the analysis, using log-Gaussian Cox processes (LGCP) may be more appropriate. In the LGCP setup, $\log(\Lambda(\boldsymbol{s}))$ follows a Gaussian process with mean $\boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta}$ and covariance function $\sigma^2\rho(\cdot)$ where $\rho(\cdot)$ is a positive definite correlation function (e.g. Matern). Examples of analyses adopting the LGCP approach include Møller, Syversveen and Waagepetersen (1998); Diggle, Moraga, Rowlingson and Taylor (2013); Serra, Saez, Mateu, Varga, Juan, Díaz-Ávalos and Rue (2014); Shirota and Gelfand (2017). The LGCP approach is able to estimate covariate effects but is computationally slow to implement. Specifically, the log-likelihood under the LGCP approach is given by

$$\log\left(\mathcal{L}(\Lambda)\right) = -\int_D \Lambda(\boldsymbol{u})d\boldsymbol{u} + \sum_{i=1}^{N}\log(\Lambda(\boldsymbol{s}_i)). \tag{1}$$

The primary issue in using this likelihood is the integral over the random intensity surface $\Lambda(\boldsymbol{s})$ is unknown and computationally expensive to calculate. Solutions to this issue of an intractable likelihood include Taylor and Diggle (2014); Simpson, Illian, Lindgren, Sørbye and Rue (2016); Johnson, Diggle and Giorgi (2019); Adams, Murray and MacKay (2009), but these can be computationally expensive in their own right.

1.3. Research Goal and Paper Outline

As discussed above, both the mixture model approach and the LGCP approach have their advantages and disadvantages. Yet, both are inherently tied to the Poisson process framework. In this paper, we seek to link the two methods and thereby leverage the strengths of both approaches. That is, we seek to exploit the flexibility and computational simplicity of the mixture model approach while utilizing the strength of LGCP to estimate associated covariate effects from the mixture model fit. Specifically, we propose a Dirichlet process mixture (DPM) model for the intensity surface as a flexible tool to capture nonlinearity in the continuous intensity surface. The DPM fit is then transformed via principles of machine learning into the corresponding fit from the LGCP approach. The effectiveness of this approach is inherent in the effective sample size of parameters of interest.

The remainder of this paper is as follows: Section 2 describes the connection between the two methods in detail along with our machine learning approach to transform parameters from the mixture model fit to the LGCP parameterization. Section 3 demonstrates a proof of concept via two-dimensional simulation studies. Section 4 then applies this methodology to the two point patterns presented in this section. Finally, Section 5 draws conclusions and highlights areas of future research.

2. Methodology

2.1. Linking Mixture Models and LGCPs

As above, let $\boldsymbol{s}_i = (s_{i1}, \ldots, s_{iD})'$ be the location of an event on the bounded domain $\mathcal{D} \subset \mathbb{R}^D$. For ease of implementation, we scale each dimension so that $s_{id} \in (0,1)$ for all $d = 1, \ldots, D$ so that $\mathcal{D} = (0,1)^D$. Further, let $\Lambda(\boldsymbol{s}) = \delta\lambda(\boldsymbol{s})$ where $\lambda(\boldsymbol{s})$ is the normalized intensity. This specification allows us to model the overall expected number of counts $\mathbb{E}(N) = \delta$ independently of the normalized intensity $\lambda(\boldsymbol{s})$.

Because $\lambda(\boldsymbol{s})$ is the density for the event locations, we use a Dirichlet process (DP) mixture model for $\lambda(\boldsymbol{s})$ so that for $j = 1, \ldots, D$,

$$s_{ij} \mid \mu_{ij}, \tau_{ij} \overset{ind}{\sim} \mathcal{B}(\mu_{ij}, \tau_{ij}), \tag{2}$$

$$(\mu_{ij}, \tau_{ij}) \mid \mathcal{G} \overset{iid}{\sim} \mathcal{G}, \tag{3}$$

$$\mathcal{G} \sim \mathcal{DP}(\mathcal{G}_0, \alpha). \tag{4}$$

Here $\mathcal{B}(\mu, \tau)$ is the beta distribution parameterized by the mean μ and precision parameter τ so that $\mathbb{E}(s_{id}) = \mu_{id}$ and $\mathbb{V}(s_{id}) = \mu_{id}(1-\mu_{id})/(1+\tau_{id})$, and $\mathcal{G}$ is a random measure from the DP with centering measure $\mathcal{G}_0$ and precision parameter α. This roughly follows the model presented in Kottas and Sansó (2007). *A priori*, we choose the centering measure for μ and τ to have an independent structure as follows:

$$\mathcal{G}_0 = \mathcal{U}(0,1) \times \mathcal{IG}(2, \kappa), \tag{5}$$

where $\mathcal{U}(a, b)$ represents the uniform distribution and $\mathcal{IG}(a, b)$ represents the inverse-gamma distribution with shape a and rate b.

Because any realization of a DP is almost surely discrete, for each iteration of an MCMC sampling algorithm (see Algorithm 8 of Neal, 2000), the DP mixture model results in

$$\lambda(\boldsymbol{s}_i) = \sum_{k=1}^{K} \left[\omega_k \prod_{d=1}^{D} \mathcal{B}(s_{id} \mid \mu_{kd}, \tau_{kd}) \right]. \tag{6}$$

The inherent advantages of the DPM model are twofold. First, the model simultaneously estimates the number of components (K) in addition to the mixture component parameters $\{\omega_k, \mu_{kd}, \tau_{kd}\}$ resulting in a flexible, continuous model specification for the intensity. And, second, each of the parameters in this model with the exception of $\{\mu_{kd}, \tau_{kd}\}$ are known to be conjugate resulting in a straightforward sampling algorithm. The parameters $\{\mu_{kd}, \tau_{kd}\}$ can be updated using simple Metropolis-Hastings type algorithms.

While the DPM model for $\lambda(\boldsymbol{s})$ excels in terms of flexibility and straightforward computation, the inherent disadvantage to this DPM model is that it does not easily lend itself to including covariates that explain the height of $\lambda(\boldsymbol{s})$. Nevertheless, estimating the effect of covariates is a key analysis goal for the motivating point pattern data shown above in Figure 1. To solve this issue, consider the LGCP approach mentioned above which sets,

$$\log\left(\Lambda(\boldsymbol{s})\right) = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta} + w(\boldsymbol{s}), \tag{7}$$

where $w(\boldsymbol{s})$ is a spatial random effect that is modeled by a Gaussian process (GP) with mean 0 and covariance function $\sigma^2\rho(\boldsymbol{s}_i, \boldsymbol{s}_j)$. Under this LGCP approach, the coefficients $\boldsymbol{\beta}$ directly correspond to the effects of covariates on the intensity surface. However, under the LGCP approach, the likelihood in (1) is computationally intractable for large datasets (see Teng, Nathoo and Johnson, 2017, for discussion).

From above, the strength of the LGCP approach is the weakness of the mixture model approach and vice versa. Hence, we propose jointly using the two methods so as to utilize the strength of both approaches. First, due to the computational issues associated with GPs, we define the spatial random effect to be $w(\boldsymbol{s}) = \boldsymbol{b}'(\boldsymbol{s})\boldsymbol{\theta}$ where $\boldsymbol{b}'(\boldsymbol{s})$ is a set of spatial bases (Heaton, Datta, Finley, Furrer, Guinness, Guhaniyogi, Gerber, Gramacy, Hammerling, Katzfuss et al., 2019) with associated coefficients $\boldsymbol{\theta}$. This basis function approach is used to primarily help with the computational bottleneck of GPs. Next, we link the two approaches via

$$\log\left[\delta \sum_{k=1}^{K} \left[\omega_k \prod_{d=1}^{D} \mathcal{B}(s_{id} \mid \mu_{kd}, \tau_{kd}) \right]\right] \approx \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{b}'(\boldsymbol{s})\boldsymbol{\theta}, \tag{8}$$

where the left-hand-side of (8) is $\log(\Lambda(\boldsymbol{s}))$ under the DPM specification and the right-hand-side is $\log(\Lambda(\boldsymbol{s}))$ under the LGCP specification. Under this link, the goal is to estimate the appropriate $(\boldsymbol{\beta}', \boldsymbol{\theta}')$ from the LGCP specification that matches the resulting DPM model fit. We do so in the following manner.

Let $\mathbf{s}_1^\star, \ldots, \mathbf{s}_L^\star$ denote a set of L well-dispersed locations on $\mathcal{D}$. Further, let $\boldsymbol{m}(\boldsymbol{s}) = (m_1(\boldsymbol{s}), \ldots, m_K(\boldsymbol{s}))'$ where

$$m_k(\boldsymbol{s}) = \delta \prod_{d=1}^{D} \mathcal{B}(s_{id} \mid \mu_{kd}, \tau_{kd}),$$

and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)'$ so that $\log(\Lambda(\boldsymbol{s}))$ under the DPM approach can be represented as $\log(\Lambda(\boldsymbol{s})) = \log(\boldsymbol{m}'(\boldsymbol{s})\boldsymbol{\omega})$. Further, let $\boldsymbol{\psi} = (\boldsymbol{\beta}', \boldsymbol{\theta}')'$ and $\mathbf{h}(\mathbf{s}) = (\mathbf{x}'(\mathbf{s}), \mathbf{b}'(\mathbf{s}))'$ so that $\log(\Lambda(\boldsymbol{s}))$ under the LGCP approach can be represented as $\log(\Lambda(\boldsymbol{s})) = \boldsymbol{h}'(\boldsymbol{s})\boldsymbol{\psi}$. Each iteration of the fitting algorithm for the DP approach gives $\{\mathbf{m}'(\mathbf{s}_\ell^\star)\boldsymbol{\omega}\}_{\ell=1}^{L}$ from which we can define

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}} \left(\left[\sum_{\ell=1}^{L} \mathcal{E}\left(\log\left(\mathbf{m}'(\mathbf{s}_\ell^\star)\boldsymbol{\omega}\right), \mathbf{h}'(\mathbf{s}_\ell^\star)\boldsymbol{\psi} \right) \right] + \xi\left(\zeta\|\boldsymbol{\psi}\|_1 + (1-\zeta)\|\boldsymbol{\psi}\|_2^2\right) \right), \tag{9}$$

where $\mathcal{E}(\cdot)$ is an error (loss) function while $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the $\mathcal{L}_1$ and $\mathcal{L}_2$ norms, respectively, $\xi \geq 0$ is a penalization term and $\zeta \in [0, 1]$ is a mixing parameter. The link in Equation (9) is an elastic net penalty wherein $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be estimated from the mixture model fit $\{\mathbf{m}'(\mathbf{s}_\ell^\star)\boldsymbol{\omega}\}_{\ell=1}^{L}$.

The above estimation procedure for $\hat{\boldsymbol{\psi}}$ is essentially a machine learning fit of the covariates and bases $\boldsymbol{h}(\boldsymbol{s})$ to the fitted intensity $\log(\Lambda(\boldsymbol{s}))$. Implementing this estimation procedure requires a choice of error function $\mathcal{E}(\cdot)$, along with an appropriate choice of penalization parameter ξ and mixing parameter ζ. In terms of the error function, the most common choice is to let $\mathcal{E}(a, b) = (a - b)^2$ but certainly absolute loss could be considered. Generally, however, the researcher will choose $\mathcal{E}(\cdot)$ based on prior knowledge. To choose $(\xi, \zeta)'$, we advocate a cross-validation approach wherein $\{\mathbf{m}'(\mathbf{s}_\ell)\boldsymbol{\omega}\}_{\ell=1}^{L}$ is split into test- and training-sets to determine an appropriate choice of (ξ, ζ).

2.2. Parameter Estimation

Unknown parameters in the above approach are those associated with the DPM model. In this work, we take a Bayesian approach as this naturally leads to some uncertainty quantification for the estimated parameters $\boldsymbol{\psi}$ rather than a simple point estimate. Under this Bayesian paradigm, we assume priors for unknown parameters and estimate them via Markov chain Monte Carlo (MCMC) sampling in the following manner.

The unknown parameters include the mixture parameters $\{\mu_{kd}, \tau_{kd}\}$, the DP precision parameter α, the hyperparameter for the inverse gamma prior for τ_{kd} (denoted as κ in (5)) and the expected number of events δ. First, we *a priori* assume that δ follows an (improper) gamma distribution with shape parameter 0 and rate parameter 0. While this is not a proper prior, it does correspond to a "uniform" prior on the positive reals. Under this prior specification, the full posterior conditional distribution for δ is simply a gamma distribution with shape n and rate 1.

We rely on Algorithm 8 of Neal (2000) to estimate the parameters of the DPM model and the reader can find details in this reference. Broadly speaking, our MCMC algorithm first iterates through each observation by removing it from its current cluster allocation and reassigning it either to an existing cluster (i.e., an existing mixture component) or to a new mixture component according to their associated Polya urn probabilities. Having removed and reassigned each observation to a mixture component, the mixture parameters $\{\mu_{kd}, \tau_{kd}\}$ are updated via a standard Metropolis algorithm, although the adaptive Metropolis algorithm of Haario, Saksman and Tamminen (2001) could also be used. Finally, we place a Gamma$(1, 1/1600)$ prior on κ and a Gamma$(3, 1/10)$ prior on α. Posterior sampling of κ is straightforward because its posterior is conditionally conjugate. We obtain posterior draws from α using the augmentation method found in Escobar and West (1995). These (conjugate) priors were chosen because they are very diffuse, but it bears noting that the particular choice of prior hyperparameters is minimally influential because there are many "layers" of parameters above them in the hierarchical model specification.

At the end of each iteration of the MCMC algorithm, we have associated draws of the intensity surface $\log(\Lambda(\boldsymbol{s}))$ via Equation (6) where the mixture weights ω_k are the proportion of observations assigned to mixture component k. To then estimate $\boldsymbol{\psi} = (\boldsymbol{\beta}', \boldsymbol{\theta}')'$ we use an equally spaced grid of L points across $\mathcal{D}$ as the target variable $\log(\boldsymbol{m}'(\boldsymbol{s}_\ell)\,\boldsymbol{\omega})$ in the minimization in (9). The tuning parameters ζ and ξ are chosen by cross-validation; we used and recommend the `train` function in R's `caret` package.

An outline of the full model fitting algorithm is given as Algorithm (1). Importantly, Algorithm (1) results in draws of $\boldsymbol{\psi}$ giving a measure of uncertainty associated with these parameters. However, the draws of $\boldsymbol{\psi}$ are draws from the posterior distribution of the elastic net estimate in (9). This is not equivalent to the posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ from the LGCP approach. That is, our resulting draws of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ from Algorithm (1) are draws of the elastic net estimates. Hence, we do not expect that the uncertainty reflected in these draws matches the uncertainty that would be reflected if the LGCP approach were used. Rather, the uncertainty in these draws more directly corresponds to the uncertainty in the maximum *a posteriori* estimates of these parameters.

3. Proof of Concept via Simulation

In this section we conduct a proof of concept that the above two step approach can result in viable estimates of model parameters. Specifically, our goal in this section is to ensure that the parameter estimates obtained via the two-step approach above can recover the associated parameters. As such, for this simulation study, we let $\boldsymbol{s} = (s_1, s_2)$ where $s_d \in (0, 1)$ for $d = 1, 2$. To simulate event locations, we define the true intensity surface as

$$\log(\Lambda(\boldsymbol{s}))) = \boldsymbol{h}'(\boldsymbol{s})\boldsymbol{\psi}, \tag{10}$$

as in the LGCP approach. To ensure that this simulation study is realistic and we can recover model parameters similar to what we do in our application, we use the posterior mean of the covariate effects (i.e. $\hat{\boldsymbol{\psi}}$) in the two dimensional example of Section 4 to define the true intensity surface. However, we decreased the intercept term in order to decrease the overall intensity surface, reducing the size of the simulated datasets and allowing for faster computation while

Algorithm 1 Fitting algorithm.

Initialize δ, $\{\mu_{kd}, \tau_{kd}\}_{k,d}$, α and κ along with an initial clustering of the events $c_1, \ldots, c_N$ where $c_i \in \{1, \ldots, K\}$ such that $c_i = k$ implies that $\mu_{id} = \mu_{c_i d}$ and $\tau_{id} = \tau_{c_i d}$ in Equation (2).
Choose $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_L^\star$
Determine spatial bases $\boldsymbol{b}'(\boldsymbol{s})$
for t in $1, \ldots, T$ **do**
    Update δ by drawing from its complete conditional distribution
    for i in $1, \ldots, N$ **do**
        Reallocate observation i to cluster $k \in \{1, \ldots, K+1\}$ with probability proportional to

$$\mathbb{P}\text{rob}(c_i = k) \propto \begin{cases} n_k \prod_{d=1}^{D} \mathcal{B}(\boldsymbol{s}_i \mid \mu_{kd}, \tau_{kd}) & \text{if } k < K+1 \\ \alpha \prod_{d=1}^{D} \mathcal{B}(\boldsymbol{s}_i \mid \mu_{(K+1)d}, \tau_{(K+1)d}) & \text{if } k = K+1 \end{cases}$$

    where n_k is the number of observations currently allocated to cluster k and $\mu_{(K+1)d}, \tau_{(K+1)d}$ are drawn from the base measure $\mathcal{G}_0$. As described in Neal (2000), if observation i is the only observation in its cluster, cluster $K+1$ to which it may be assigned should have the μ and τ parameters of the cluster it came from.
        Set K as the number of unique clusters
    end for
    for k in $1, \ldots, K$ **do**
        for d in $1, \ldots, D$ **do**
            Update (μ_{kd}, τ_{kd}) via adaptive Metropolis
        end for
    end for
    Update α and κ by drawing from their complete conditional distributions
    For each $\boldsymbol{s}_\ell^\star$, evaluate $\mathbf{m}'(\mathbf{s}_\ell^\star)\,\boldsymbol{\omega}$.
    Using cross-validation, determine ξ and ζ
    Estimate and retain $\hat{\boldsymbol{\psi}}$ as solution to (9)
end for

preserving the essential structure in the two-dimensional example of Section 4. We sampled points according to the assumed true intensity surface by a double application of the inverse CDF method, for every point first sampling one coordinate according to its marginal CDF, and then the second coordinate according to the relevant conditional CDF. The true intensity surface is given as the top left panel in Figure 2.

Using the above methods, we simulated 46 point pattern datasets and fit each of these datasets using a nonhomogenous Poisson process as well as the above two-step approach in Section 2 and Algorithm 1. For the nonhomogenous Poisson process approach, we follow Mortensen, Heaton and Wilhelmi (2018) and assume a piecewise constant intensity surface where

$$\Lambda(\boldsymbol{s}) = \sum_{r=1}^{R} \lambda_r \mathbb{1}\{\boldsymbol{s} \in \mathcal{R}_r\}, \tag{11}$$

$$\log(\lambda_r) = \boldsymbol{h}'(\boldsymbol{s})\boldsymbol{\psi}, \tag{12}$$

where $\mathcal{R}_1, \ldots, \mathcal{R}_R$ is a partition of the domain $\mathcal{D}$ into disjoint regions $\mathcal{R}_r$. This piecewise constant approach is, essentially, a discretized form of a LGCP and allows us to easily deal with the likelihood in (1) because the integral is constant over each region $\mathcal{R}_r$. In this nonhomogenous Poisson process model, the model parameters consist only of $\boldsymbol{\psi}$ so we rely on the adaptive Metropolis algorithm of Haario et al. (2001) to draw these parameters from the posterior distribution.

As shown by Figure 2, all methods seem to reliably recapture the true intensity surface, which itself reflects the crash patterns observed on I-15, with most "crashes" concentrated in the Salt Lake City area (approximately scaled milepoint 0.75). It is difficult to distinguish between the model fits based solely on their intensity surfaces, but there are some important differences in diagnostics of model fit. Notably, from the top right panel of Figure 2, the DP intensity estimate smooths the time effect. This is likely due to the mixture model representation under the DP (which

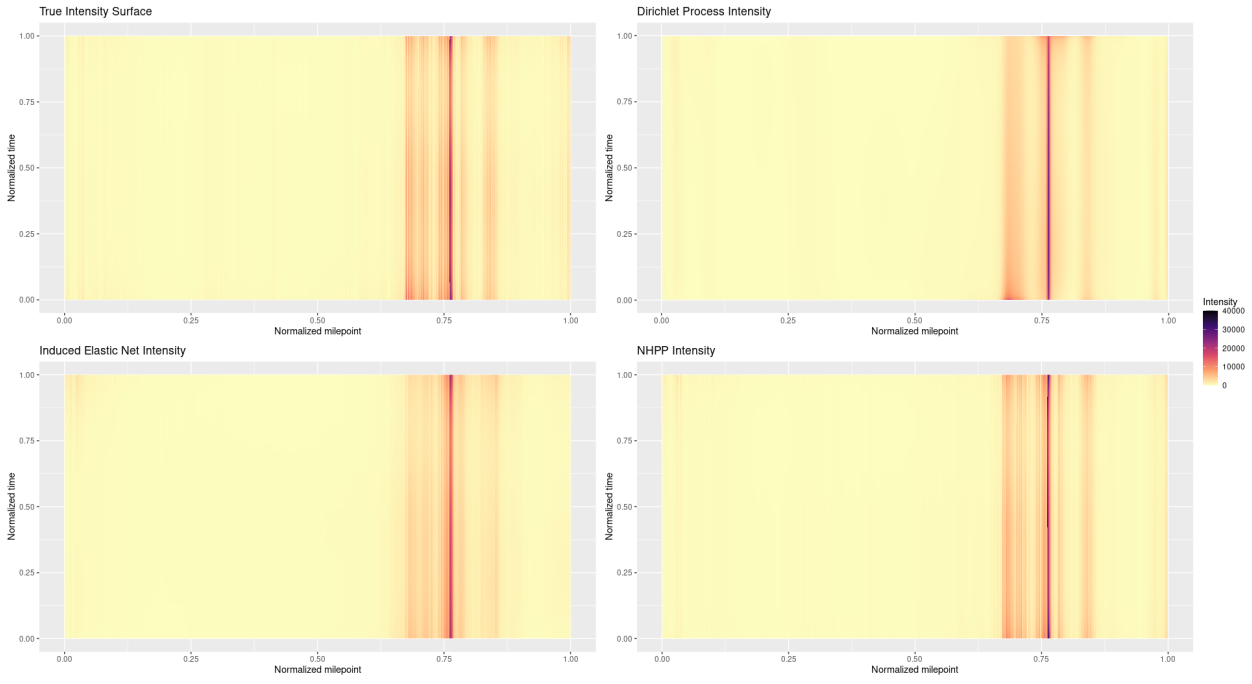

Figure 2: Assumed true intensity surface, with Dirichlet process and elastic net intensity surface estimates computed from the 36^{th} dataset generated from the true intensity surface, compared to NHPP fit generated from same data

is inherently smooth) while the NHPP approach relies on basis functions to smooth out the estimate of the intensity surface.

Beyond the qualitative comparison in Figure 2, we also compare the elastic net estimates with the NHPP estimates of the coefficients $\boldsymbol{\beta}$ in Table 1 in terms of root mean square error (RMSE), continuous rank probability score (CRPS; Gneiting and Raftery 2007) and effective sample size. Broadly speaking, the RMSE of the elastic net method is slightly better than that of the NHPP. This is to be expected as the elastic net essentially trades bias for variance to attain a better RMSE. On the other hand, the CRPS scores of the elastic net estimators are slightly worse than the NHPP estimators. This is also to be expected in that under the two-step approach we obtain, via invariance of posterior draws, the posterior of (9) (which is a minimum). This is different than the NHPP model which results the posterior distribution of $\boldsymbol{\beta}$. Hence, we expect the CRPS score for the NHPP model to be slightly better because it accurately quantifies the uncertainty in $\boldsymbol{\beta}$ rather than the uncertainty in the solution to (9).

Finally, it is worth noting the substantially better effective sample size for the two-step method over the NHPP model. Notably, this is because the two-step approach is effectively conjugate while the NHPP approach relies on adaptive tuning of a proposal distribution within a Metropolis accept-reject step. In our analysis, both approaches took roughly the same amount of time to sample an intensity surface. While the NHPP gives coefficient estimates immediately, extracting coefficient estimates from the DPM approach is fairly computationally intensive. However, the task can be easily parallelized, and on a per effective sample size basis, the method presented in the paper is faster than the NHPP.

4. Applications

In this section, we analyze the two vehicle crash examples discussed in Section 1 and displayed in Figure 1. Specifically, Section 4.1 analyzes a one-dimensional point pattern data of crashes along I-5 in Washington State while Section 4.2 considers a two-dimensional, space-time point pattern of crashes along I-15 in Utah.

Table 1
Diagnostics of model fit for NHPP vs proposed model, averaged over all 46 datasets

	Non-homogenous Poisson process			Proposed method		
Variable	Avg. RMSE	Avg. CRPS	Eff. Samp. Size	Avg. RMSE	Avg. CRPS	Eff. Samp. Size
β_0	4.44	0.77	721.48	4.18	0.82	5267.60
β_1	6.61	1.00	447.78	3.42	0.65	7955.03
β_2	9.40	1.20	567.97	5.24	0.96	6358.61
β_3	7.61	0.83	165.36	7.30	0.87	7332.17
β_4	8.98	0.87	444.32	2.93	0.63	7079.99
β_5	9.83	0.90	491.56	7.53	0.88	6482.20
β_6	8.29	0.79	193.06	7.13	0.82	7296.99
β_7	9.07	0.88	603.09	7.85	0.94	4219.95
β_8	8.95	0.92	683.90	8.71	0.84	4012.20
β_9	8.70	0.90	1229.88	17.16	1.18	6407.70
β_{10}	11.57	1.00	1489.09	3.76	0.64	7010.27
β_{11}	8.84	0.88	1419.60	6.18	0.95	4739.74
β_{12}	7.88	0.82	1094.71	3.95	0.69	5686.59
β_{13}	9.68	0.96	1242.88	7.13	0.87	7072.46
β_{14}	8.98	0.92	649.32	3.72	0.69	322.35
β_{15}	10.71	0.94	453.07	15.46	0.95	2327.51

Effective sample size is expressed per 10,000 draws.

4.1. Crashes along I-5 in Washington State

The primary goal of this analysis is to identify road characteristics that are associated with elevated crash intensity. For this analysis, $\boldsymbol{s}_i = s_i \in (0, 1)$ corresponds to the (scaled) milepoint of crash i along I-5 in Washington State. The road characteristics of interest are area classification (urban vs. rural), surface type (asphalt vs. pavement), terrain (level vs. rolling), shoulder width/type (asphalt, concrete, pavement, wall, unspecified), median width (distance between opposing traffic directions), number of lanes, and average annual daily traffic (AADT). For ease of interpretation, we scaled AADT to be measured per 10000 cars. For spatial bases ($\boldsymbol{b}'(\boldsymbol{s})$), we use 40 Gaussian bases where the knot locations are spaced evenly along the $(0, 1)$ domain. To avoid spatial confounding (Reich, Hodges and Zadnik, 2006), we orthogonalize the spatial bases with respect to the road characteristics.

We ran the two-step method in Section 2 for 10,000 iterations using the first 5,000 as burn-in. We use $L = 10,000$ locations along $(0, 1)$ to obtain $\boldsymbol{\psi}$ via the elastic net penalty and square error loss. At each iteration we used the `train` function in R's `caret` package to choose optimal ξ and ζ penalty values.

The resulting posterior means for the Dirichlet process intensity surface and elastic net coefficient estimates give us the intensity estimates for a fine grid of 10,000 points along our domain shown in Figure 3. Both methods appropriately identify large spikes in crash intensity around Seattle and Tacoma as well as smaller spikes around other populated areas. However, areas of especially high intensity are emphasized more by the Dirichlet process model while the elastic net seems to smooth over the larger peaks. This is to be expected because the elastic net intensity includes spatial bases for smoothing the surface.

Posterior summaries of $\boldsymbol{\psi}$ are shown in Table 2. Importantly, because the elastic net is used for each iteration, we obtain a measure of significance for each variable as the percent of negative, zero or positive coefficients. For reference, the baseline road segment is a rural setting, asphalt surface, level terrain, and no specified shoulder types. Road characteristics such as AADT, urban settings, and rolling terrain increase crash intensity while number of lanes and shoulder width decrease crash intensity. Further, structured left shoulder types (i.e. curb, wall) seem to decrease crash intensity which could stem from added protection from cross traffic collisions. Interestingly, structured right shoulder types seem to have the opposite effect and increase crash intensity. We hypothesize that open right shoulder types lead to added space to maneuver or park for emergency situations.

4.2. Crashes along I-15 in Utah across time

The method presented in 2 is sufficiently general that it can be used to model point patterns on any bounded rectangular domain. In the case of I-15 in Utah, we have data on crashes across milepoint and also across time. Hence, for this analysis, each crash location $\boldsymbol{s}_i \in (0, 1) \times (0, 1)$ consists of a (scaled) milepoint and a (scaled) crash time.

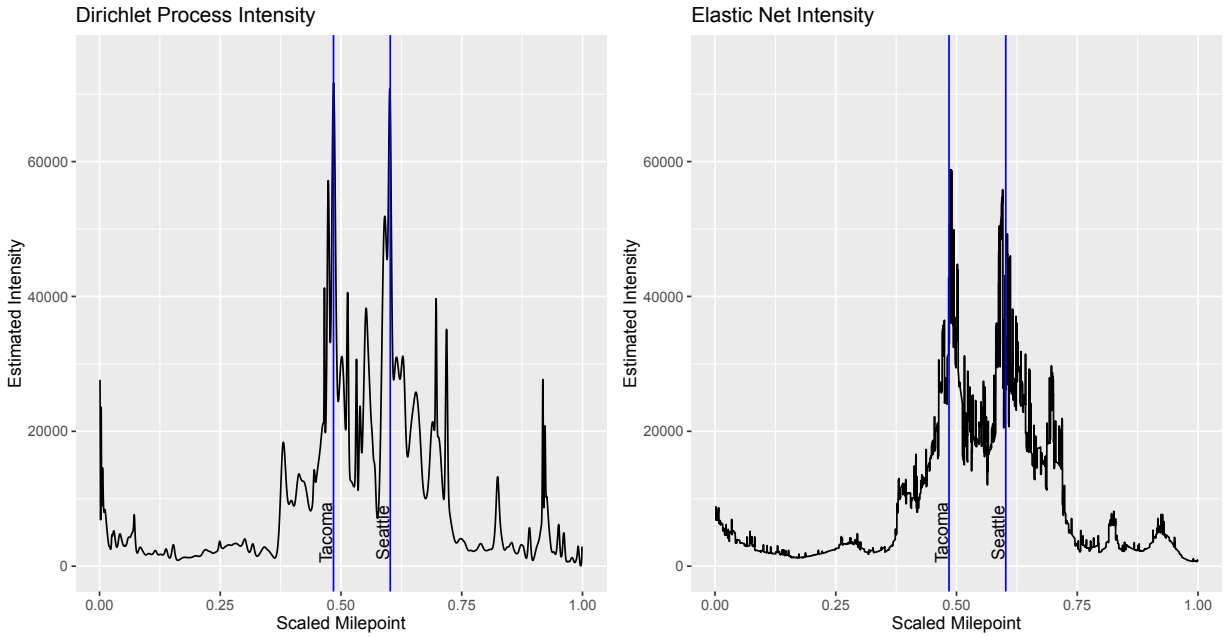

Figure 3: Dirichlet process and elastic net intensity surface estimates for I-5 crash data.

Table 2
Posterior summary of $\boldsymbol{\psi}$ for the I-5 example. The baseline category is a rural road with asphalt surface, level terrain, and no specified shoulder type. The interval endpoints reported are for a central 95% credible interval.

Variable	Mean	Lower Cred.	Upper Cred.	Prop. Negative	Prop. Zero	Prop. Positive
Intercept	7.2	7	7.3	0	0	1
AADT	0.17	0.16	0.17	0	0	1
Urban	0.29	0.22	0.34	0	0	1
Portland Concrete Surface	0.19	0.13	0.23	0	0	1
Rolling Terrain	0.16	0.09	0.24	0	0	1
Left Shoulder Width	−0.02	−0.03	−0.01	> 0.99	< 0.01	0
Median Width	0	0	0	0.89	0.05	0.06
# of Lanes	−0.09	−0.11	−0.08	1	0	0
Left Shoulder Type: Asphalt	−0.06	−0.15	0	0.92	0.06	0.02
Left Shoulder Type: Portland Concrete	0	−0.04	0	0.23	0.77	0
Left Shoulder Type: Curb	−0.37	−0.45	−0.28	1	0	0
Left Shoulder Type: Wall	−0.26	−0.32	−0.19	1	0	0
Right Shoulder Type: Asphalt	0.01	−0.03	0.12	0.07	0.73	0.2
Right Shoulder Type: Portland Concrete	0.32	0.26	0.43	0	0	1
Right Shoulder Type: Curb	0.21	0.13	0.34	0	0	1
Right Shoulder Type: Wall	0.22	0.15	0.33	0	0	1

Proportions may not add to 1 due to rounding.

The available road characteristics include information on average annual daily traffic (AADT), number of single-unit trucks (SUTRK), combo-unit trucks (CUTRK), speed limit, number and width of lanes, shoulder width, and median type (concrete barrier, depressed median, separate grades, or unprotected). We included 121 spatial bases, with the knots spaced evenly on an 11×11 grid in $(0, 1)^2$. As in the previous analysis, we orthogonalized the spatial bases with respect to the matrix of road characteristics to avoid confounding. Figure 4 displays a posterior mean intensity surface from the Dirichlet process (left), a posterior mean of the intensity surface from the elastic net fit shown (center) and the fit of a competing non-homogeneous Poisson process (right).

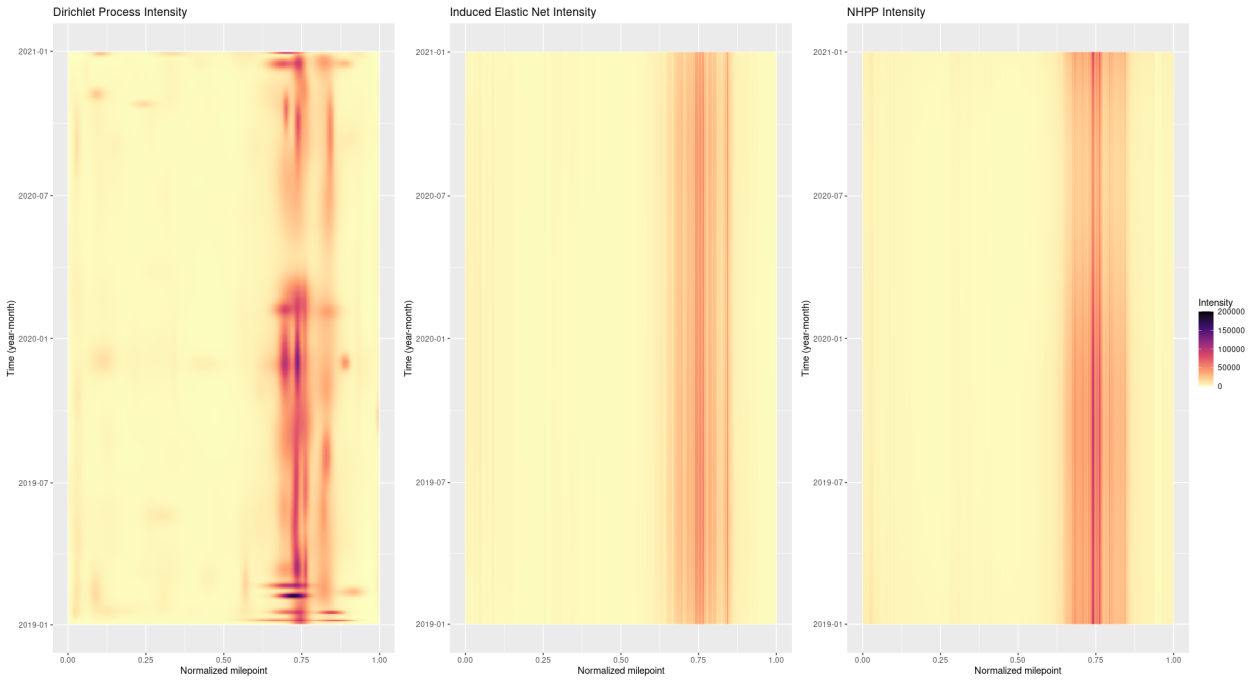

Figure 4: Dirichlet process and elastic net intensity surface estimates for I-15 data, compared to NHPP fit

The flexibility of the Dirichlet process intensity surface allows it to capture many compelling phenomena contributing to crash patterns. Most basically, the model reflects that car crashes are concentrated around the Salt Lake City area (approximately scaled milepoint 0.75) but there are narrower regions of particularly elevated risk. Further, the DP identifies that the crash intensity surface falls precipitously around March 2020 (onset of the COVID-19 pandemic) and then gradually recovers to pre-pandemic levels. Most interestingly, the DP estimate also identifies irregular “bumps" running across space but not time. These dates correspond to heavy snowstorms in Utah, and that the Dirichlet process intensity surface is able to capture their effect illustrates its flexibility (and hence the advantage of its use).

The intensity surface induced by the elastic net is not nearly so flexible as the Dirichlet process intensity surface, since it is restricted to the use of space-time basis functions. The elastic net intensity surface preserved that most crashes occur in the greater Salt Lake City area, but there shrunk the drop in intensity near March 2020. In general, even though 121 spatial bases were included, time appears to have very little impact on the intensity surface, and heavy snowstorms have no influence at all. Even so, the R^2 of estimating the Dirichlet process intensity by the elastic net intensity heights (on the log scale, where they were both computed and where inference is made) is roughly 0.72, suggesting that the elastic net fit is capturing the most important features in the Dirichlet process intensity.

In this analysis, we had substantially more information on traffic levels and roadway characteristics than in the I-5 example and our estimates of these effects are given as Table 3. From Table 3, the intensity surface is affected by AADT and the relative abundance of single- and combo-unit trucks. We note that the other effects correspond well with intuition and the results of the I-5 analysis which are, broadly speaking, that characteristics associated with heavily trafficked urban areas are associated with greater crash intensities and vice versa. The Gelman-Rubin diagnostic (Gelman and Rubin, 1992) applied to our draws of $\boldsymbol{\psi}$ gives potential scale reduction factor upper bounds close to 1 for most vector elements. Only 4 (out of 136) had an upper bound greater than 1.2, which indicates that the chains are converging properly.

Table 3
Posterior summary of $\boldsymbol{\psi}$ for the two-dimensional I-15 example. The baseline category is a painted median. The interval endpoints reported are for a central 95% credible interval.

Variable	Mean	Lower Cred.	Upper Cred.	Prop. Negative	Prop. Zero	Prop. Positive
Intercept	11.24	9.66	12.78	0	0	1
AADT in 2020	0.05	−0.02	0.12	0.06	0.07	0.86
Single-unit truck pct. in 2020	−2.22	−3.74	−0.61	> 0.99	< 0.01	< 0.01
Combo-unit truck pct. in 2020	−15.73	−21.48	−10.06	1	0	0
AADT in 2019	0.02	−0.03	0.08	0.13	0.13	0.74
Single-unit truck pct. in 2019	0.15	−1.76	1.46	0.57	< 0.01	0.42
Combo-unit truck pct. in 2019	9.53	3.80	15.24	< 0.01	< 0.01	> 0.99
Speed limit	−0.03	−0.05	−0.01	> 0.99	0	< 0.01
Number of lanes	0.18	0.12	0.23	0	0	1
Lane width	−0.01	−0.09	0.07	0.64	< 0.01	0.36
Right shoulder width	0.03	0.01	0.05	< 0.01	0	> 0.99
Left shoulder width	−0.03	−0.05	−0.01	> 0.99	< 0.01	< 0.01
Median type: Depressed median	0.28	0.16	0.40	0	0	1
Median type: Separate grades	0.96	0.54	1.52	0	0	1
Median type: Unprotected	−0.43	−0.56	−0.30	1	0	0

Proportions may not add to 1 due to rounding.

5. Conclusions

In this analysis we have described a method for linking a flexible mixture model approach to the analysis of point patterns with a log-Gaussian Cox process approach to obtain estimates of covariate effects using machine learning. We have demonstrated the viability of this approach via simulation studies and analyses of real automobile crash datasets. Overall, the approach allows accurate estimation of covariate effects at the cost of a slightly decreased ability to assess uncertainty in these parameters. However, by drawing on both a mixture and a LGCP approach, we are able to leverage the strengths of both.

While the DP was used here due to its flexibility, it came at a high computational cost. We found this flexibility to be useful, in that it detected features in the intensity surface of the I-15 example that were shrunk out using either the elastic net or NHPP approaches. However, for very large datasets, the DP may be computationally infeasible and users may need to resort to a finite mixture model, which may lack flexibility. Where it can be applied, though, the intensity surfaces generated by the DP are sufficiently flexible to model nearly any spatial process faithfully. The performance of the entire modeling framework presented here is constrained by the second step rather than the first.

Given our sampled intensity surfaces, we opted to use the elastic net to obtain interpretable estimates of model parameters. This approach works well in the absence of any high-order covariate interactions or nonlinearities in covariate effects. However, the procedure proposed here consists of two entirely separable steps, and covariates only come into play in the second. More complex machine learning algorithms such as the random forest or neural networks could also be used to estimate non-linear effects in the place of the elastic net used here. If these other algorithms are used, more complex relationships between the covariates and the intensity estimates could be modeled, but at the cost of interpretablity. The choice of which machine learning algorithm to use, then, depends on the needs of the analysis. We plan to explore the use of other machine learning algorithms in similar applications in future research.

As mentioned, one of the downsides of this approach is that the resulting uncertainty is the uncertainty in the machine learning estimates of model parameters which, in this case, correspond to a minimizer of a penalized square error loss. As we saw in the simulations, this is not equivalent to the uncertainty associated with the parameter but rather is more closely related to the uncertainty associated with its maximum *a posteriori* estimate. An interesting line of research would seek to adjust this uncertainty to more closely resemble the uncertainty in the parameter itself.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2053188. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Adams, R.P., Murray, I., MacKay, D.J., 2009. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities, in: Proceedings of the 26th annual international conference on machine learning, pp. 9–16.

Aguero-Valverde, J., 2013. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. Accident Analysis & Prevention 50, 289–297.

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research 9, 1–15.

Daley, D.J., Vere-Jones, D., et al., 2003. An introduction to the theory of point processes: volume I: elementary theory and methods. Springer.

Diggle, P.J., Moraga, P., Rowlingson, B., Taylor, B.M., 2013. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. Statistical Science 28, 542–563.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical science 7, 457–472.

Geng, J., Shi, W., Hu, G., 2021. Bayesian nonparametric nonhomogeneous poisson process with applications to USGS earthquake data. Spatial Statistics 41, 100495.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378.

Gomes, M.J.T.L., Cunto, F., da Silva, A.R., 2017. Geographically weighted negative binomial regression applied to zonal level safety performance models. Accident Analysis & Prevention 106, 254–261.

Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive Metropolis algorithm. Bernoulli , 223–242.

Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M., et al., 2019. A case study competition among methods for analyzing large spatial data. Journal of Agricultural, Biological and Environmental Statistics 24, 398–425.

Hougaard, P., Lee, M.L.T., Whitmore, G., 1997. Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes. Biometrics , 1225–1238.

Jiao, J., Hu, G., Yan, J., 2021. Heterogeneity pursuit for spatial point pattern with application to tree locations: A Bayesian semiparametric recourse. Environmetrics 32, e2694.

Johnson, O., Diggle, P., Giorgi, E., 2019. A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data. Statistics in medicine 38, 4871–4887.

Kottas, A., Sansó, B., 2007. Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. Journal of Statistical Planning and Inference 137, 3151–3163.

Møller, J., Syversveen, A.R., Waagepetersen, R.P., 1998. Log Gaussian Cox processes. Scandinavian journal of statistics 25, 451–482.

Moller, J., Waagepetersen, R.P., 2003. Statistical inference and simulation for spatial point processes. CRC press.

Mortensen, J.W., Heaton, M.J., Wilhelmi, O.V., 2018. Urban heat risk mapping using multiple point patterns in Houston, Texas. Journal of the Royal Statistical Society Series C: Applied Statistics 67, 83–102.

Neal, R.M., 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9, 249–265.

Reich, B.J., Hodges, J.S., Zadnik, V., 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics 62, 1197–1206.

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Díaz-Ávalos, C., Rue, H., 2014. Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: the case of Catalonia, 1994–2008. Environmental and ecological statistics 21, 531–563.

Shirota, S., Gelfand, A.E., 2017. Space and circular time log Gaussian Cox processes with application to crime event data. The Annals of Applied Statistics , 481–503.

Simpson, D., Illian, J.B., Lindgren, F., Sørbye, S.H., Rue, H., 2016. Going off grid: Computationally efficient inference for log-Gaussian Cox processes. Biometrika 103, 49–70.

Snyder, D.L., Miller, M.I., 2012. Random point processes in time and space. Springer Science & Business Media.

Taddy, M.A., Kottas, A., 2012. Mixture Modeling for Marked Poisson Processes. Bayesian Analysis 7, 335 – 362. URL: https://doi.org/10.1214/12-BA711, doi:10.1214/12-BA711.

Taylor, B.M., Diggle, P.J., 2014. Inla or mcmc? a tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. Journal of Statistical Computation and Simulation 84, 2266–2284.

Teng, M., Nathoo, F., Johnson, T.D., 2017. Bayesian computation for log-Gaussian Cox processes: A comparative analysis of methods. Journal of statistical computation and simulation 87, 2227–2252.

Yin, F., Jiao, J., Yan, J., Hu, G., 2022. Bayesian nonparametric learning for point processes with spatial homogeneity: A spatial analysis of nba shot locations, in: International Conference on Machine Learning, PMLR. pp. 25523–25551.

Zeng, Q., Gu, W., Zhang, X., Wen, H., Lee, J., Hao, W., 2019. Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors. Accident Analysis & Prevention 127, 87–95.

Zhao, C., Kottas, A., 2021. Modelling for Poisson process intensities over irregular spatial domains. arXiv preprint arXiv:2106.04654 .

Zhou, Z., Matteson, D.S., Woodard, D.B., Henderson, S.G., Micheas, A.C., 2015. A spatio-temporal point process model for ambulance demand. Journal of the American Statistical Association 110, 6–15.