

# Spatial Regression for Marked Point Processes

KATJA ICKSTADT and ROBERT L. WOLPERT

*University of North Carolina at Chapel Hill and Duke University, USA*

## SUMMARY

In a wide range of applications, dependence on smoothly-varying covariates leads spatial point count intensities to feature positive correlation for nearby locations. In applications where the points are “marked” with individual attributes, the distributions for points with varying attributes may also differ. We introduce a class of hierarchical spatial regression models for relating marked point process intensities to location-specific covariates and to individual-specific attributes, and for modeling the remaining intensity variation that arises from dependence on unobserved or unreported covariates. The magnitude of residual intensity variation is a measure of how completely the covariates explain the observed variations in point intensities. The models, extending those recently introduced by Wolpert and Ickstadt, treat the point patterns as doubly-stochastic Poisson random fields with a random inhomogeneous Poisson intensity given by a spatial mixture of gamma or other infinitely-divisible independent-increment random fields. Bayesian prior distributions for a third level of hierarchy are elicited to reflect beliefs about the homogeneity, continuity, and similar features of the intensity. Inference is based on posterior and predictive distributions computed using Markov chain Monte Carlo methods featuring data augmentation and an efficient method for sampling from independent-increment random fields. The models are illustrated in an application to a four-dimensional spatial regression analysis of origin/destination trip data from the 1994/95 Metro survey of Portland, Oregon.

*Keywords:* COX PROCESS; DATA AUGMENTATION; LÉVY PROCESS; TRANSPORTATION.

## 1. INTRODUCTION

In many applications point count intensities will be positively correlated for nearby locations due to the effect of smoothly-varying underlying covariates. Some of these location-specific covariates may be observed, while others may not. In addition we will often observe individual attributes (marks) for the point counts, and may expect to find positive correlations for the intensities for similarly-marked points. The goal of this paper is a spatial regression analysis for point count data that reflects the effects of both observed and latent spatial covariate data and individual marks.

Data analysis using non-spatial logistic or log-linear regression models allows one to incorporate the individual attributes, but does not allow for modeling the possible spatial correlation of either the data or the covariates. Spatial dependence can be incorporated using lattice-based Markov random field models, in which the logarithms of the Poisson intensities for specified subregions are modeled as a Gaussian Markov random field with partially unknown covariance structure (see, e.g., Bernardinelli *et al.*, 1997; Besag, York and Mollié, 1991; Clayton and Kaldor, 1987; Møller, Syversveen and Waagepetersen, 1998). These models can be viewed as a spatial version of a generalized linear model (Ghosh *et al.*, 1998) with logarithmic link. In the lattice Markov random field approach, however, individual-level covariates cannot enter the

model directly, but only after aggregation over the subregions; this will bias the analysis whenever the relationship between area-level covariates and aggregated response differs from that between individual-level covariates and individual response, a common phenomenon known as the ecological fallacy (see, e.g., Richardson, 1992).

In this paper we introduce a class of hierarchical spatial regression models for relating point process intensities to both location-specific covariates and individual-specific attributes. These doubly-stochastic (Cox) Poisson random field models have intensities that are spatial mixtures of inhomogeneous gamma or other infinitely-divisible random fields. They generalize those recently introduced by Wolpert and Ickstadt (1998a) for estimating and making inference about unobserved Poisson point intensities through the new inclusion of covariates and attributes.

At the first or lowest stage of the hierarchy we model the number of points in each region  $R$  in some set  $\mathcal{Y}$  in Euclidean space as a Poisson random variable  $N(R)$  with random mean measure  $\Lambda(R) = \int_R \Lambda(y) w(dy)$  with density  $\Lambda(y)$  with respect to some reference measure  $w(dy)$ .

At the second stage of the hierarchy we model the density  $\Lambda(y)$  as the sum  $\sum_j a_j \theta_j$  of a linear combination of individual-specific or location-specific attributes  $a_j$ , with uncertain coefficients  $\theta_j$ , and a kernel mixture  $\int_{\mathcal{S}} k^\theta(y, s) \Gamma(ds)$  for an independent-increment infinitely divisible random measure  $\Gamma(ds)$  on an auxiliary space  $\mathcal{S}$  and uncertain kernel  $k^\theta(y, s)$ , representing the effect of additional unobserved but spatially-correlated covariates, leading to

$$\Lambda(y) = \sum_j a_j \theta_j + \int_{\mathcal{S}} k^\theta(y, s) \Gamma(ds).$$

In the following we take  $\Gamma(ds)$  to have an inhomogeneous gamma process distribution  $\Gamma(ds) \sim \text{Ga}(\alpha^\theta(ds), \beta^\theta(s))$  with possibly uncertain shape measure  $\alpha^\theta(ds)$  and scale function  $\beta^\theta(s) > 0$ .

At the third stage of the hierarchy we introduce a prior distribution  $\pi(d\theta)$  on the parameter space  $\Theta$  to express uncertainty about the coefficients, the kernel, and some aspects of the distribution of  $\Gamma(ds)$ .

The models may be viewed as spatial analogues to Poisson regression models with identity link functions; covariate influences are additive under these models rather than multiplicative as in common Poisson regression models with log-link functions. The identity link scales properly under aggregation and refinement of partitions, and thus allows for a consistent random field version of the models, while the logarithmic link would lead to products rather than sums in the Poisson means for unions of neighboring sets. The identity link and the infinitely-divisible random fields at the first two levels of the hierarchy allow us to model data and covariates given on disparate spatial scales by relating all observable quantities to an underlying random field. This possibility is explored in detail in Best, Ickstadt and Wolpert (1998).

Section 2 introduces the Poisson/gamma random field models as a generalization of a spatial Poisson regression model with an identity link. Section 3 offers a sketch of the computational Markov chain Monte Carlo (MCMC) scheme. The methods are illustrated in Section 4 with a transportation example of urban commuting, using data from the 1994/95 Metro survey of Portland, Oregon. Section 5 discusses connections to other models, including conjugate Poisson/gamma models, mixture of Dirichlet process models, point source models, and process convolution models. Section 6 provides a summary discussion and an outlook on future work.

## 2. SPATIAL REGRESSION MODELS

In this section we present our general marked point process spatial regression models in several steps. We begin with a spatial formulation of Poisson regression and its random field analogue (Section 2.1) and successively add individual attributes (Section 2.2), latent spatial covariates (Section 2.3), and full hierarchical Bayesian structure (Section 2.4). The section ends summarizing some key features of the model class.

### 2.1 A spatial generalization of Poisson regression

Consider an observed point process  $N(dy)$  on some space  $\mathcal{Y}$ , along with a collection of observed covariates indexed by a set  $J'$ . For any partition  $\{R_i\}_{i \in I}$  of  $\mathcal{Y}$  we can let  $w_i$  be some measure of the size of the  $i^{\text{th}}$  partition element  $R_i$  (at-risk population for epidemiological studies, area for environmental applications, etc.),  $N_i = N(R_i)$  the number of counts in  $R_i$  and  $X_{ij}$  a summary of the  $j^{\text{th}}$  covariate there. The Poisson regression model with identity link

$$N_i \sim \text{Po}(\Lambda_i w_i) \quad \Lambda_i = \sum_j X_{ij} \theta_j \quad (1)$$

would attempt to represent point intensities  $\Lambda_i w_i$  as linear combinations of the explanatory variables for uncertain coefficients  $\theta_j$ , which in a Bayesian analysis would have a joint prior density function  $\pi(\theta)$ .

If we set  $\Lambda_{ij} \equiv X_{ij} \theta_j$  then Equation (1) accords  $N_i$  a Poisson distribution whose mean is a sum  $\Lambda_i = \sum \Lambda_{ij}$  of rates associated with the different covariates. We can write  $N_i = \sum N_{ij}$  as the sum of conditionally independent unobserved quantities  $N_{ij} \sim \text{Po}(\Lambda_{ij} w_i)$ , the number of counts in  $R_i$  associated with the  $j^{\text{th}}$  covariate. The introduction of  $N_{ij}$  may be regarded as a form of data augmentation (Tanner and Wong, 1987), and will be used to simplify the computations in Section 3 below.

One disadvantage of this partition-based approach is that the covariates  $X_{ij}$  must be associated with regions  $R_i$  rather than with individual locations or points. To summarize spatial covariates by region would lead to a form of the ecological fallacy whenever intensity depends nonlinearly upon covariates or when covariate levels and exogenous influences on intensity vary across regions.

This shortcoming may be reduced by refining the partition  $\{R_i\}$  and eliminated by the ultimate refinement, treating  $N(dy)$  as a Poisson random field with mean

$$N(dy) \sim \text{Po}(\Lambda(y)w(dy)) \quad \Lambda(y) = \sum_j X_j(y)\theta_j \quad (2)$$

for the reference measure  $w(dy)$  on  $\mathcal{Y}$ .

### 2.2 Covariates and marked point processes

In many applications the natural set of covariates under consideration will include attributes associated with individuals, rather than with locations, that are best modeled by letting  $N$  be a point process on a space  $\mathcal{X} = \mathcal{Y} \times \mathcal{A}$  of marked points  $x = (y, a)$  that include explicitly a vector  $a$  of attributes. In the example of Section 4 the case attributes include trip destination; in epidemiological applications they might include gender and age along with known or suspected risk factors; in botanical applications the attributes might include plant sizes, species, and conditions.

We can model the observation of random ordered pairs of point locations  $Y_i \in \mathcal{Y}$  and attribute vectors  $A_i \in \mathcal{A}$  as a marked point process (see, e.g., Karr, 1991), a random integer-valued measure  $N(dx)$  on the product space  $\mathcal{X} = \mathcal{Y} \times \mathcal{A}$ . The simplest extension of Equation (2) is to treat  $N(dx)$  as a Poisson random field on  $\mathcal{X}$  with mean  $\Lambda(x)w(dx)$ ,

$$N(dy da) \sim \text{Po}(\Lambda(y, a)w_Y(dy)w_A(da)) \quad \Lambda(y, a) = \sum_j a_j \theta_j. \quad (3)$$

The choice of  $a_j = X_j(y)$  would reduce this to Equation (2), but Equation (3) is more general because the covariates are now associated with marked points  $x = (y, a)$  rather than only with locations  $y$  and so could differ for distinct individuals sharing a common location (indeed this occurs in the application of Section 4). We usually use a reference weight measure of product form  $w(dx) = w_Y(dy)w_A(da)$ , normalized by  $w_A(\mathcal{A}) \equiv 1$ . This gives  $w_Y(dy)$  the role of a population reference measure (at-risk population for epidemiological applications, for example) while the probability measure  $w_A(da)$  describes the reference distribution of attributes, which we take to be stationary in  $y \in \mathcal{Y}$ . The stationarity restriction, which could be removed by replacing  $w_A(da)$  with a location-specific conditional measure  $w_A(da|y)$ , does not imply that points' locations and attributes are independent since  $\Lambda(dy da)$  need not be of product form.

### 2.3 Unobserved spatial covariates

In many applications it is unrealistic to expect all variation to be explainable by the reported covariates. If there are additional unreported covariates that vary continuously over space we should expect an additional component  $X_*(y)\theta_*$  in the expression for the Poisson mean  $\Lambda(y, a)$  in Equation (3) with unobserved but spatially correlated covariates  $X_*(y)$  and new regression coefficient  $\theta_*$ . A convenient way to model this is to introduce a set  $\{s_m\}_{m \in M}$  of locations in  $\mathcal{Y}$  and associated latent positive magnitudes  $\Gamma_m$ , and set  $X_*(y) \equiv \sum_{m \in M} k(y, s_m)\Gamma_m$  for some kernel function  $k(y, s)$  decreasing in the distance  $|y - s|$ . Upon setting  $a_* \equiv X_*(y)$  and  $J \equiv J' \cup \{*\}$  for some new index  $* \notin J'$ , we can again describe the model with Equation (3). In this context  $N(dx)$  may be viewed as the sum of a cluster process with intensity  $X_*(y)$  (see, e.g., Cox, 1955; Cox and Isham, 1980; Neyman and Scott, 1958) and an independent marked point process with intensity proportional to  $w(dx)$ .

For a Bayesian analysis we must introduce a joint prior distribution for the parameter vector  $\theta$  and the latent magnitudes  $\{\Gamma_m\}$ . In typical applications the number  $|J|$  of explanatory variables will be relatively small, while the number  $|M|$  of spatially distributed latent variables will be quite large, making the choice of independent variates from the conjugate gamma prior distributions  $\Gamma_m \sim \text{Ga}(\alpha_m, \beta_m)$  and an arbitrary prior density  $\pi(\theta)$  computationally convenient (but see Section 3 for wider choices).

The additional term  $X_*(y)\theta_* = \sum_{m \in M} k(y, s_m)\Gamma_m\theta_*$  may be viewed as the effect of unobserved point sources of magnitudes  $\{\Gamma_m\}$  at locations  $\{s_m\}$ , or simply as an approximation to any unobserved spatially varying covariate. In the latter case the approximation may be improved by enlarging the set  $\{(s_m, \Gamma_m)\}_{m \in M}$  of latent spatial variables while reducing the shape parameters  $\alpha_m$  appropriately; in the limit this leads to an inhomogeneous gamma random field  $\Gamma(ds)$  with shape measure  $\alpha(ds)$  and scale function  $\beta(s)$  on some space  $\mathcal{S}''$  containing  $\mathcal{Y}$ , with

$$N(dx) \sim \text{Po}(\Lambda(x)w(dx)) \quad \Lambda(x) = \sum_{j \in J'} a_j \theta_j + \int_{\mathcal{S}''} k(y, s)\Gamma(ds)\theta_* \quad (4)$$

where once again  $x = (y, a)$ . More general choices for the distribution of  $\Gamma(ds)$  are discussed in Sections 2.4 and 3 below.

If we allow  $k(\cdot, s)$  to depend on marked points  $x = (y, a)$  (rather than merely on the locations  $y \in \mathcal{Y}$ ) and make it  $\theta$ -dependent, and extend  $\Gamma(ds)$  to the union  $\mathcal{S} \equiv \mathcal{S}' \cup \mathcal{S}''$  of  $\mathcal{S}''$  with a disjoint set  $\mathcal{S}' = \{s_j\}_{j \in J'}$  indexed by  $J'$  by setting  $\Gamma(\{s_j\}) \equiv 1$ , we can set

$$k^\theta(x, s) \equiv \begin{cases} a_j \theta_j & \text{for } s = s_j \in \mathcal{S}' \\ k(y, s) \theta_* & \text{for } s \in \mathcal{S}'' \end{cases}$$

and simplify the model description to

$$N(dx) \sim \text{Po}(\Lambda(x)w(dx)) \quad \Lambda(x) = \int_{\mathcal{S}} k^\theta(x, s)\Gamma(ds). \quad (5)$$

An additional benefit of permitting the kernel to depend on  $x$  and  $\theta$  is added flexibility in modeling the dependence on covariates, as described below in Section 2.5.

#### 2.4 Hierarchical Poisson/gamma random field models

The gamma shape measure  $\alpha(ds)$  and scale function  $\beta(s)$  may also be treated as uncertain by making them  $\theta$ -dependent, adding additional components to  $\theta$  if necessary. This leads to a three-stage hierarchical Bayesian Poisson/gamma random field model:

Level 1:	Marked Points	$N(dx) \sim \text{Po}(\Lambda(x)w(dx)),$	$x = (y, a) \in \mathcal{X} \equiv \mathcal{Y} \times \mathcal{A}$	
	Intensity	$\Lambda(x) = \sum_{j \in J'} a_j \theta_j + \int_{\mathcal{S}''} k(y, s)\Gamma(ds) \theta_*$		
		$= \int_{\mathcal{S}} k^\theta(x, s)\Gamma(ds)$		(6)
Level 2:	Sources	$\Gamma(ds) \sim \sum_{j \in J'} \delta_{s_j}(ds) + \text{Ga}(\alpha^\theta(ds), \beta^\theta(s))$		
Level 3:	Parameter	$\theta \sim \pi(\theta)d\theta$		

where  $\delta_{s_j}(ds)$  denotes a unit point mass at the added point  $s_j \in \mathcal{S}' \subset \mathcal{S}$ . The overall influence of the latent covariates on the Poisson intensity depends on the product of the kernel  $k(y, s)$ , the gamma scale  $\beta^\theta(s)$ , and  $\theta_*$ ; we maintain identifiability by constraining  $k$  and  $\beta^\theta$ .

The structure of Equation (6) accommodates a wide variety of elaborations. Multiple varieties of latent sources are reflected by letting  $\mathcal{S}$  include a disjoint union of multiple copies of  $\mathcal{Y}$ , one for each spatially varying latent source, each with its own shape measure  $\alpha^\theta(ds)$  and scale function  $\beta^\theta(s)$ . Spatial random effects models in which we would like to treat attributes  $a_j$  as uncertain for some  $j$  may be accommodated as additional “latent” source varieties, with  $\alpha^\theta(ds)$  and  $\beta^\theta(s)$  chosen to reflect the mean and variation of the random effect. An overall non-spatial background rate may be included as an intercept term by introducing a constant attribute  $a_j \equiv 1$  for some  $j \in J'$ .

#### 2.5 Aggregation and the identity link function

Poisson regression models most commonly feature the canonical log link function (see, e.g., McCullagh and Nelder, 1989, p. 32), under which the Poisson rates are given as products of terms associated with the different covariates. The models described here feature instead the identity link, under which the covariates enter additively in determining the Poisson rates (see Equation (4)). This is more appropriate for applications with multiple causes that do not interact, and for applications in which the inferential aim is to apportion the causes of events, but less so for applications in which covariates are associated with the susceptibility to other causes where log-linear models would be better. Note that log-linear models also can be accommodated within our setting—for example, one or more covariates  $\{a_j\}_{j \in J_0}$  may be included in a

log-linear fashion with  $k^\theta(x, s) = k_0(y, s)e^{\sum_{j \in J_0} a_j \theta_j}$  for some baseline kernel  $k_0(y, s)$ , appropriate for attributes  $a_j$  that modulate Poisson rates multiplicatively by affecting the sensitivity to other causes, rather than entering additively as independent causes.

Choosing a discrete model usually begins with selecting a partition  $\{R_i\}_{i \in I}$  of the region of interest and enumerating the numbers  $\{N_i = N(R_i)\}$  of points in partition elements, with all further analysis based on this initial partition choice. Covariates must be associated with the same partition, and model inference and prediction is limited to this partition. In many problems different covariates are reported naturally at different levels of aggregation, i.e. on different partitions; for example, demographic covariates are often reported at the census block level, environmental variables are often given on a grid, while socioeconomic and employment figures are often given on political units (congressional districts, counties or states, etc.).

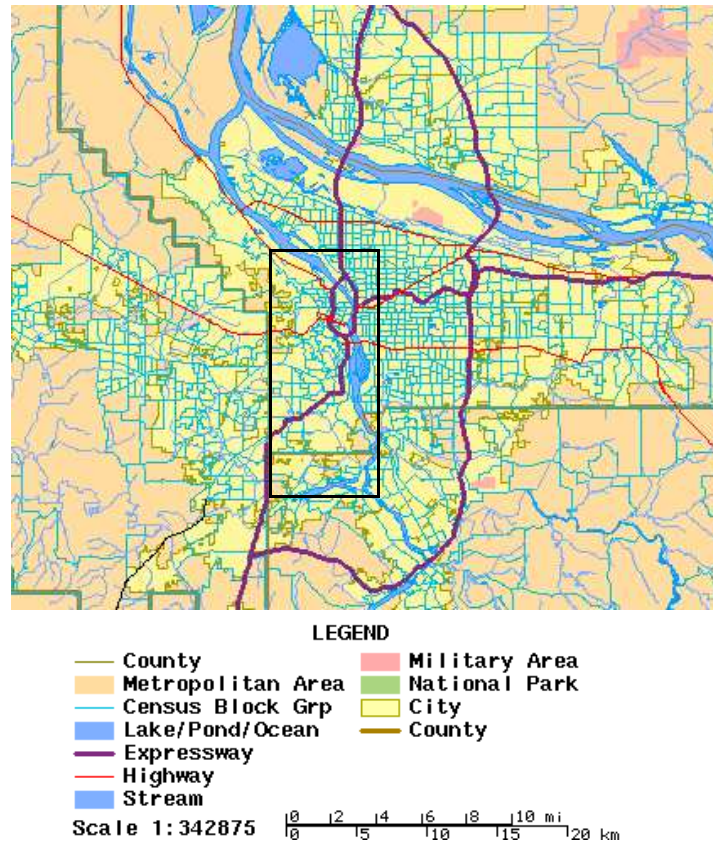
An advantage of the identity link function and of infinitely-divisible random fields in Equation (6) is that all these data may be used at their own level of aggregation, while the results can be presented at any (or several different) level(s) of aggregation, without the reanalysis that would be required to change the partition choice in the usual approach.

### 3. COMPUTATIONS

An MCMC computational scheme was introduced in Wolpert and Ickstadt (1998a) for drawing dependent samples of  $\theta$  and  $\Gamma(ds)$  from their joint posterior distribution in a version of the model with unmarked points (i.e. without attributes or covariates, so  $J' = \emptyset$  and  $\mathcal{X} = \mathcal{Y}$ ). It uses a Metropolis/Hastings step (Hastings, 1970; Tierney, 1994) for drawing  $\theta^t$  at each time-step  $t$ , allowing the use of a completely arbitrary prior density  $\pi(\theta)$ , and employs data augmentation to permit a Gibbs draw (Gelfand and Smith, 1990) for the then-conjugate gamma random field  $\Gamma^t(ds)$ . The method has a straightforward extension to accommodate marked points and covariates, i.e.  $J' \neq \emptyset$  and  $\mathcal{X} = \mathcal{Y} \times \mathcal{A}$ , based on a further data augmentation scheme writing  $N(dx) = \sum_{j \in J} N_j^t(dx)$  as a sum of unobserved Poisson random fields associated with each covariate including the latent spatial covariate  $X_*^t(y)$ .

Sampling from the gamma random field is not trivial. Even if we begin with uniform  $\alpha^\theta(ds) \equiv \alpha ds$  and constant  $\beta^\theta(s) \equiv \beta$  the complete conditional distributions of the gamma random field  $\Gamma(ds)$  will have a  $\text{Ga}(\alpha_+^\theta(ds), \beta_+^\theta(s))$  conditional distribution (given the data and  $\theta$ ), needed in the MCMC scheme, with  $s$ -dependent  $\alpha_+^\theta(ds)$  and  $\beta_+^\theta(s)$ . An efficient method for generating samples from such inhomogeneous gamma random fields, called the Inverse Lévy Measure or ILM algorithm, is introduced in Wolpert and Ickstadt (1998a) and extended to more general infinitely-divisible random fields such as the fully-skewed stable in Wolpert and Ickstadt (1998b), allowing modelers concerned about robustness to replace the gamma random field in Equation (6) by fields with tails heavier than those of the likelihood.

Sample realizations of all infinitely divisible random fields are discrete, with countably many point masses of random magnitudes  $v_m$  at locations  $s_m \in \mathcal{Y}$ . A novel feature of the ILM algorithm is that it allows us to draw the simulated  $\{s_m\}$  from any sampling distribution  $\Pi(ds)$  whatsoever, allowing us to achieve efficiency by drawing heavily from areas in  $\mathcal{Y}$  where we expect large numbers of points associated with the latent variables. The most efficient choice for constant  $\beta^\theta(s)$  would be to take  $\Pi_\alpha(ds) \equiv \alpha^\theta(s)/\alpha^\theta(\mathcal{S})$ , proportional to the gamma shape measure (as is done for the gamma process in one dimension in a related sampling method by Laud, Smith and Damien (1995)), if it were possible to draw efficiently from that distribution. In spatial applications the measure  $\alpha^\theta(ds)$  is often taken to be uniform on complex geometrical figures representing geographical or political regions (states, counties, census blocks, etc.), making it impractical to employ  $\Pi_\alpha(ds)$  as a sampling distribution; instead we draw from a finite mixture of uniform distributions on rectangles.



**Figure 1.** Portland Metropolitan area, with Study Region indicated by box.

#### 4. A TRANSPORTATION EXAMPLE

In 1994/95 the Portland Metropolitan regional government (Metro) conducted a household activity survey including 4,400 households, selected by random dialing, with about 10,000 people altogether. Each respondent's activity and travel behavior were recorded for a 48 hour period, either in fall 1994 or spring 1995, with the periods staggered across the seven days of a week. For each change in location the destination coordinates and activity were recorded. Household characteristics (e.g. household size, income, type of dwelling, number of vehicles) and personal socio-demographic variables (e.g. age, gender, employment status, occupation) were also recorded. The data are available at

`ftp://ftp.metro.dst.or.us/sys/ftp/planning/tf/pub/.`

For our analysis we take only data entries of employed individuals at least sixteen years old, who commuted from home to work on the survey day. We restrict ourselves to the first survey day to avoid correlation; we pool observations from Monday to Friday; we use only the spring sample; and we limit ourselves to the 290 records with source and destination in a  $6\text{km} \times 15\text{km}$  rectangular study region including downtown Portland (see Figure 1). We follow the United States Geological Survey (USGS) in using the Lambert conic projection to convert the original data from GPS coordinates (latitude and longitude in radians with respect to the equator and prime meridian) to state plane coordinates  $(x, y)$  in kilometers east and north, respectively, of an arbitrary reference point. The same data were used by Ickstadt, Wolpert, and Lu (1998) with a different model and inferential aim.

We treat the origin (home) coordinates as the “location” of a commuting trip segment and treat the destination coordinates as individual covariates, or “marks”; for the present illustrative

analysis we ignore other recorded covariates such as the driver's age and sex, though these too would be included as marks in a complete analysis.

With both the spatial location space  $\mathcal{Y}$  and the attribute space  $\mathcal{A}$  subsets of  $\mathbb{R}^2$  the Poisson measure  $N(dx)$  is defined on subsets of the space  $\mathcal{X} \equiv \mathcal{Y} \times \mathcal{A} \subset \mathbb{R}^4$ . We take the mean measure  $\Lambda(x)w(dx)$  to have a density function

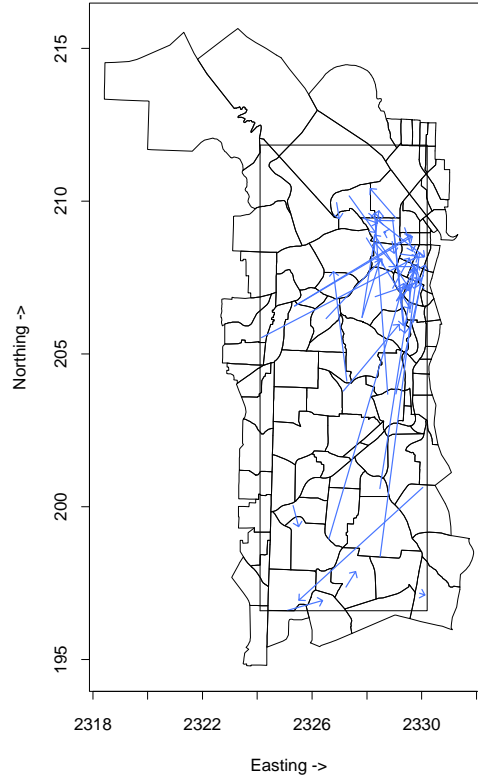
$$\Lambda(x) = \int_{\mathcal{S}} k(x, s) \Gamma(ds) \quad (7)$$

with respect to a baseline measure  $w(dx)$ , which we take to be proportional to the product of household density (in the origin space) and employment density (in the destination space), so that completely random assignment of workplaces to workers would lead to a constant  $\Lambda(x)$ .

We use a four-dimensional Gaussian kernel function  $k(x, s)$  parameterized by  $\theta \in \Theta \subset \mathbb{R}^3$  given, for origin/destination (o/d) pairs  $x = (x_o, x_d)$  and  $s = (s_o, s_d)$  in  $\mathbb{R}^4$  by

$$k^\theta(x, s) = \frac{2e^{2\theta_1} + e^{2\theta_2}}{4\pi^2 e^{4\theta_1 + 2\theta_2}} \exp \left[ \theta_0 - \frac{(x_o - s_o)^2 + (x_d - s_d)^2}{2e^{2\theta_1}} - \frac{(x_o - x_d)^2}{2e^{2\theta_2}} \right].$$

Evidently  $e^{\theta_0} = \int_{\mathbb{R}^4} k^\theta(x, s) ds$  is an overall scale factor,  $e^{\theta_1}$  is a measure of how far an o/d four-tuple  $x$  might be from an unobserved prototype  $s$ , and  $e^{\theta_2}$  is a measure of typical trip length  $|x_o - x_d|$ . The gamma random field  $\Gamma(ds) \sim \text{Ga}(\alpha(ds), \beta(s))$  on  $\mathbb{R}^4$  has constant scale  $\beta(s) \equiv 500$  and uniform  $\alpha(ds) \equiv 0.002 ds$ , giving (four-dimensional) Lebesgue measure for the prior mean  $E[\Gamma(ds)] = ds$  and constant induced prior mean  $E[\Lambda(x)] \equiv e^{\theta_0}$  for the Poisson mean density.



**Figure 2.** Fifty randomly selected origin/destination pairs and Travel Analysis Zone boundaries that intersect Study Region in Portland, Oregon.

Independent normal prior distributions were used for the three components of  $\theta$  with means

5.60,  $-1.00$ , and  $2.00$ , all with variance  $0.50$  to put half the prior mass in an interval of  $\pm 0.48$  centered at the mean.

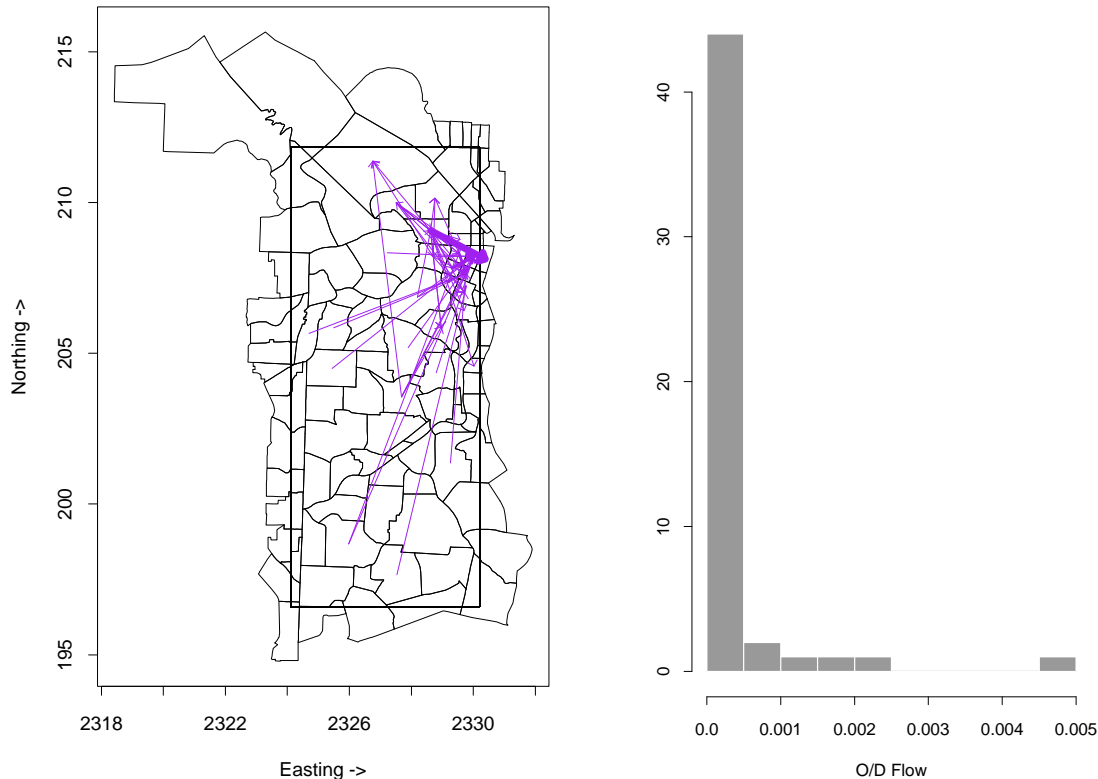
Log scale parameter  $\theta_0$  had a posterior median of  $6.53$  with posterior interquartile range (IQR) of  $[6.29, 6.73]$ . Log distance parameter  $\theta_1$  had a posterior median of  $-0.30$  with posterior IQR of  $[-0.42, -0.14]$ , so trip origins and destinations were about  $\exp(-0.30) = 0.74$  km from their prototype's origin and destination, with posterior IQR  $[0.66, 0.87]$  km; the log commute distance  $\theta_2$  had a posterior median of  $2.38$  with IQR of  $[2.08, 2.54]$ , so commuting distances have posterior median of  $10.8$  km with IQR of about  $[8.0, 12.6]$ .

These four-dimensional quantities can be presented as arrows, with the origin as base and the destination as head (for morning commuting trips these are the homes and workplaces, respectively). Figure 2 presents fifty randomly selected data in arrow form; note that multiple trips may share the same origin.

The posterior distributions of o/d flows  $E[\Lambda(x)w(x)|\text{data}]$  or their densities  $E[\Lambda(x)|\text{data}]$  are harder to present. Trip data are traditionally studied by partitioning the study region into a number  $N$  of Traffic Analysis Zones (TAZ's)  $\{A_i\}$ , then counting and modeling the numbers of trips for each of the  $N^2$  o/d pairs. Although our analysis does not begin with such a partition, we can present our output in a comparable manner by evaluating the flow

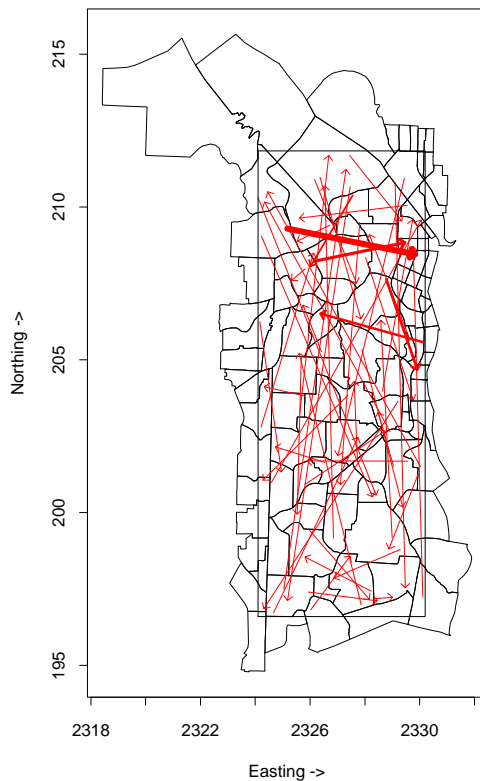
$$F_{ij} = \int_{A_i} \int_{A_j} E[\Lambda(x_o, x_d)w(x_o, x_d)|\text{data}] dx_o dx_d$$

from the  $i$ 'th to the  $j$ 'th TAZ. There are 114 TAZ's that intersect the Study Region, making an unwieldy 12,996 o/d pairs; we summarize the output in Figure 3 by presenting the fifty largest o/d flows as arrows connecting TAZ centers, with width proportional to the flow, along with a histogram of these o/d flows. Evidently a few flows are much larger than the others.



**Figure 3.** The fifty largest posterior mean TAZ O/D flows and their histogram.

In our model the flow densities are given by an integral  $\Lambda(x_o, x_d) = \int_{\mathcal{S}} k(x, s) \Gamma(ds)$  (see Equation (7)) with respect to a random measure  $\Gamma(ds)$  whose realizations are almost-surely discrete, so that  $\Lambda(x_o, x_d) = \sum k(x, s_m) v_m$  for a countably infinite random set of prototype trips  $s_m$  with magnitudes  $v_m$ . Figure 4 shows the fifty largest mass points  $\{s_1, \dots, s_{50}\}$  of one posterior realization of the gamma random field  $\Gamma(ds)$  as arrows with width proportional to the magnitudes  $\{v_1, \dots, v_{50}\}$ .



**Figure 4.** Fifty largest mass points of one realization of the gamma random field.

## 5. CONNECTIONS TO OTHER MODELS

The Bayesian models introduced in Section 2 are related to a variety of others including conjugate Poisson/gamma models, mixture of Dirichlet process models, point source models, and convolution models.

Although motivated and introduced as continuous, these models also include discrete versions simply by using discrete sets for  $\mathcal{X}$  and  $\mathcal{S}$  or discrete measures for  $w(dx)$  and  $\alpha(ds)$ . The discrete form of Equation (4), which extends Equation (1) with the addition of a latent source term

$$N_i \sim \text{Po}(\Lambda_i w_i) \quad \Lambda_i = \sum_j X_{ij} \theta_j + \sum_m k_{im} \Gamma_m \theta_*$$

reduces to the conjugate Poisson/gamma models of Clayton and Kaldor (1987) when  $k_{im}$  is diagonal; the continuous form of Equation (4) with a singular kernel also reduces to a similar conjugate random field model.

In the special case of a Gaussian kernel  $k^\theta(x, s)$  normalized to satisfy  $\int_{\mathcal{X}} k^\theta(x, s) w(dx) \equiv 1$ , empty covariate set  $J' \equiv \emptyset$ , and constant  $\beta^\theta(s) \equiv \beta$ , the conditional density of points  $x \in \mathcal{X}$

(given their total number) is given by the normalized intensity field

$$\Lambda(x)/\Lambda(\mathcal{X}) = \int_{\mathcal{S}} k^{\theta}(x, s)\Gamma(ds)/\Gamma(\mathcal{S}),$$

a mixture of Dirichlet process model (Antoniak, 1974). Allowing non-constant  $\beta^{\theta}(s)$  would extend this class of models, allowing the modeler to express more informed prior opinion in some parts of the space  $\mathcal{S}$  than in others.

Diggle (1990) studied the problem of quantifying disease risk close to a pre-specified point source by using Poisson random field models with intensity of the form  $\Lambda(y) = \theta_0 + k^{\theta}(y, s)\theta_*$ ; the model of Equation (4) extends this to multiple sources of uncertain strength and location with location-specific covariates. Recently Bayesian methods have been applied to related problems (Wakefield and Morris, 1998).

Higdon (1998) uses a process convolution approach to produce a smooth spatial surface of temperatures in the North Atlantic ocean. His model is similar to that of Equation (4) without covariates ( $J' = \emptyset$ ) and with a discrete lattice for  $\mathcal{S}$ , so  $\Gamma(ds)$  reduces to a Gaussian lattice field; indeed our approach may also be viewed as a process convolution model.

Nonspatial Bayesian Poisson models using the identity link function and individual covariates, differing from those of Equations (2) and (3) in their use of truncated normal prior distributions, are studied by Clyde (1998).

## 6. SUMMARY AND FUTURE WORK

In this paper we extend the class of spatial hierarchical mixture models introduced by Wolpert and Ickstadt (1998a) to allow for individual level attributes and for spatially varying covariates, by treating event locations as a marked point process in a generalized linear spatial regression model with identity link function. In our transportation example we treat trip origins as locations, and destinations as marks, leading to a four-dimensional spatial model for the marked process describing trips as origin/destination pairs. This example could be elaborated by removing the restriction to commuting trips and to trips that begin and end in the study region, and by including such individual attributes as time of day, trip purpose and duration, driver's age and gender, etc.

A similar four-dimensional approach would apply to other network analysis problems, including packet transport in communications networks. A two-dimensional epidemiological case study based on these models will appear in Best, Ickstadt and Wolpert (1998), relating disease incidence to both individual attributes (such as gender and parental smoking) and spatially varying environmental and socio-demographic covariates reported at different levels of aggregation. One-dimensional versions of the models can be used to relate survival to individual attributes and time-varying covariates.

One remaining issue for future work is to gain more experience in the modeling of individual attributes. In which cases and applications should these enter the model additively, as possible "causes" of events? In which should they enter multiplicatively (see Section 2.5), as modifiers of susceptibility? Should covariate attribute interaction terms be included? How should we make variable selection decisions? How can more general kernel functional forms be tailored to particular applications, reflecting the ways observed and latent covariates are expected to relate to intensities in specific cases? Many questions remain about model selection and model validation for this class of models.

Another remaining issue concerns model predictions under both prior and posterior distributions. What prior predictions would be helpful for facilitating the elicitation of informed prior distributions? What posterior predictions would be most useful for model validation?

The authors are working in collaboration with N. Best, D. Spiegelhalter and A. Thomas to explore the extension of the BUGS Bayesian MCMC software package (Spiegelhalter *et al.*, 1995) to include these spatial point-process regression models, to make them more widely available and easy to use.

### ACKNOWLEDGMENTS

This work was supported by U.S. E.P.A. grant CR822047-01-0, N.S.F. grant DMS-9626829, and the Deutsche Forschungsgemeinschaft. The work would have been impossible without the ideas, advice and encouragement of Eric Pas, who passed away while the work was in progress. We miss him deeply. We would also like to thank Xuedong Lu for data preparation and conversion, Keith Lawton and his staff at Metro Portland for making the data available, and the U.S. Census Bureau for offering the TIGER Mapping Service used to produce Figure 1.

### REFERENCES

- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.
- Bernardinelli, L., Pascutto, C., Best, N. G., Gilks, W. (1997). Disease mapping with errors in covariates. *Statistics in Medicine* **16**, 741–752.
- Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43**, 1–59 (with discussion).
- Best, N. G., Ickstadt, K. and Wolpert, R. L. (1998). Ecological modeling of health and exposure data measured at disparate spatial scales. (Work in progress).
- Clayton, D. G. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671–681.
- Clyde, M. A. (1998). Bayesian model averaging and model search strategies. (this volume).
- Cox, D. R. (1955). Some statistical methods connected with series of events. *J. Roy. Statist. Soc. B* **17**, 129–64 (with discussion).
- Cox, D. R. and Isham, V. (1980). *Point Processes*. New York: Chapman & Hall.
- Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *J. Roy. Statist. Soc. A* **153**, 349–362.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.
- Ghosh, M., Natarajan, K., Stroud, T. W. F., Carlin, B. P. (1998). Generalized linear models for small-area estimation. *J. Amer. Statist. Assoc.* **93**, 273–282.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
- Higdon, D. (1998). A process-convolution approach to modeling temperatures in the North Atlantic Ocean. *J. Environmental and Ecological Statist.* **5**, (to appear).
- Ickstadt, K., Wolpert, R. L. and Lu, X. (1998). Modeling travel demand in Portland, Oregon. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Müller and D. Sinha, eds.). New York: Springer-Verlag, 305–322.
- Karr, A. F. (1991). *Point Processes and their Statistical Inference*. New York: Marcel Dekker.
- Laud, P. W., Smith, A. F. M. and Damien, P. (1995). Monte Carlo methods for approximating a posterior hazard rate process. *Statistics and Computing* **6**, 77–84.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.
- Møller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian J. Statist.* (to appear).
- Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *J. Roy. Statist. Soc. B* **20**, 1–43 (with discussion).

- Richardson, S. (1992). 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, 181–204.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R. (1995). BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50. *Tech. Rep.*, MRC Biostatistics Unit, Cambridge University.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**, 528–550.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–1762 (with discussion).
- Wakefield, J. C. and Morris, S. E. (1998). The Bayesian modelling of disease risk in relation to a point source. (Submitted).
- Wolpert, R. L. and Ickstadt, K. (1998a). Poisson/gamma random field models for spatial statistics. *Biometrika* **85**, (to appear).
- Wolpert, R. L. and Ickstadt, K. (1998b). Simulation of Lévy random fields. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Müller and D. Sinha, eds.). New York: Springer-Verlag, 227–242.