Statistica Sinica **26** (2016), 861-879
doi:http://dx.doi.org/10.5705/ss.2014.150

A NOTE ON THE CONSISTENT ESTIMATION OF SPATIAL-TEMPORAL POINT PROCESS PARAMETERS

Frederic Paik Schoenberg

University of California, Los Angeles

Abstract: For models used to describe spatial-temporal marked point processes with covariates, the high number of parameters typically involved can make model evaluation, construction, and estimation using maximum likelihood quite difficult. A further complication is that some relevant covariates may be missing from the fitted model, and the impact of these missing variables is typically unclear. Conditions are explored here under which parameters governing a space-time marked point process may be estimated simply and consistently by maximizing a partial likelihood, essentially ignoring other terms in the model and any missing covariates. Under the given conditions, the resulting estimates may have the desirable properties of maximum likelihood estimates for the full model. An application to southern California earthquake forecasting using weather data is provided.

Key words and phrases: Conditional intensity, consistency, maximum likelihood estimation, Poisson process, spatial-temporal point process, weighted least squares estimation.

1. Introduction

Recent increases in spatial-temporal marked point process data with multiple covariates have led to the development of point process models with relatively large numbers of parameters. The high dimensionality of the models can pose problems when it comes to the improvement, evaluation, and estimation of these models. In such situations, one may initially choose to focus on an individual portion of the process, e.g. by modeling first the purely temporal aspects of the process, and then adding spatial, mark, and covariate components. In seismology, for instance, a decade elapsed between the introduction of temporal marked Epidemic-Type Aftershock Sequence (ETAS) models (Ogata (1988)) and the development of spatial-temporal marked versions (Ogata (1988)) that are now commonly used. When examining just a portion of the process, however, one may inquire whether the modeled process may be accurately estimated in the absence of the components being ignored. In modeling the times and marks of a spatial-temporal marked point process, for instance, it is important to determine under what conditions the spatial components of the process may be ancillary to

the parameters governing the temporal and mark components, or when ignoring the spatial coordinates of the observations would substantially bias estimates of temporal and mark parameters.

In addition, there is a sense among point process practitioners that, when certain possibly confounding variables have been omitted from the analysis, the resulting estimates of parameters should not be too substantially biased provided the omitted variables have a small influence on the conditional rate. Theoretical justification for this idea is hard to find, however. This paper attempts to provide some mathematical support for the notion that omitted variables with small effects on the conditional rate should have small effects on parameter estimates, and conditions are described under which conventional maximum likelihood estimates are consistent despite the omission of certain variables.

The current paper explores conditions under which such *partial* models may be consistently estimated despite missing information. Certain special cases are well known. In point process models for earthquake occurrences, for instance, the distribution of earthquake magnitudes is typically modeled as constant over time. Thus, while earthquake magnitudes can affect the times and locations of future earthquakes, the magnitude *distribution* of an event at a particular location and time, given that an event occurs, is typically thought to be constant. Under this assumption, the estimation of the earthquake size distribution is especially straightforward. When one or more dimensions of a point process have coordinates whose entries are i.i.d. draws from a fixed distribution, the process is called *separable*; see e.g. Rathbun (1996) or Schoenberg (2004) for examples. This paper investigates more general conditions under which components of a point process, when estimated separably using maximum likelihood methods, yield consistent parameter estimates.

The separability of a component in a point process model is very important in that if a parameter or collection of parameters may be estimated individually, this greatly facilitates model building, fitting, and assessment. For multi-dimensional models, the rate at which points occur versus each separable coordinate may be plotted individually to suggest functional forms for the model, and also the fit of the model is much more readily inspected from such a plot due to the reduction in the number of dimensions. Further, while maximum likelihood estimates have such well-understood properties as consistency and asymptotic efficiency under rather general conditions, in practice maximum likelihood typically requires an iterative optimization procedure which, when many parameters are being estimated, can fail to converge to a global maximum and which often relies heavily on starting values, the choice of which can be very problematic. Estimation is greatly facilitated when only a few parameters are estimated at a time. Hence it is worth exploring situations in which point process models can be decomposed

so that certain parameters can be consistently estimated separably, without optimizing over all values of the other parameters.

Tests for separability of point process models have been proposed by Schoenberg (2004) and Chang and Schoenberg (2010). Here, we focus on the estimation of separable point process models, including processes with covariates, and address the question of what types of models have components that may be estimated consistently. Rathbun (1996) noted that models that are multiplicative in all dimensions may be estimated separably, and methods for estimating such models are detailed by Baddeley and Turner (2000). The present paper treats a much wider class of models. Our main results may roughly be summarized as follows: for models that are multiplicative in the dimensions of the point process, and either multiplicative or additive in the covariates, the individual components of the model may, under general conditions, be consistently estimated separately. The resulting estimates will be equivalent, or, in the case of Theorem 2 below, will converge in probability to the ordinary maximum likelihood estimates. An application is given involving earthquake weather and point process modeling of southern California seismicity.

2. Preliminaries

Suppose N is a point process whose domain D is a measurable product space, $D = D_0 \times D_1 \times \cdots \times D_k$, equipped with measure μ. For instance, in the case of earthquake occurrences, D might be the product of a portion of space-time and a mark space. Suppose that each of the domains D_i is measurable and is equipped with measure μ_i, and that in particular $D_0 = [0, T]$ is a portion of the real (time) line.

For any point $x = (t, m_1, m_2, \ldots, m_k)$ in D, let $\lambda(x)$ denote the conditional intensity of the point process. $\lambda(x)$ reflects the infinitesimal expected rate of accumulation of points around location x, given the entire history of the process over all previous times. More precisely, following the notation in Brown, Ivanoff, and Weber (1986) or Merzbach and Nualart (1986), beginning with a partial ordering on D where $(t_1, m_1, \ldots, m_k) \leq (t_2, m'_1, \ldots, m'_k)$ iff. $t_1 \leq t_2$, let $\mathcal{F}_x$ be a filtration on D, and define $\mathcal{F}_x^1$ as the filtration generated by the $\mathcal{F}_x$-adapted, left-continuous processes. We say a process is *predictable* if it is $\mathcal{F}_x^1$-adapted. Then the *conditional intensity* (or 1-*intensity*) λ is any non-negative, $\mathcal{F}_x$-predictable process such that for any measurable subset S of $D_1 \times D_2 \times \ldots \times D_k$, $N([0, t] \times S) - \int_0^t \int_S \lambda(u, m_1, m_2, \ldots, m_k) d\mu_1 \ldots d\mu_k dt$ is an $\mathcal{F}_x$-martingale. Intuitively, a conditional intensity at a particular location of space-time is the rate at which one expects points to occur, given everything that has happened *previously*, i.e., to the left on the time line of the location in question. It thus makes sense

to require the process λ to have left-continuous sample paths, which is what is meant by the imposition of predictability.

Suppose that λ is governed by a parameter vector θ from some compact parameter space Θ, and that Θ is a product of compact parameter spaces Θ_0, Θ_1,..., Θ_k, Θ_{k+1}. We assume in what follows that Θ_{k+1} is a compact subset of $\mathbf{R}^+$, but each of the other spaces Θ_i may be multi-dimensional.

To ease notation, it will be useful to introduce the following conventions. For any integer i in $\{0, 1, \ldots, k\}$, let D_{-i} represent the product space $D_0 \times D_1 \ldots \times D_{i-1} \times D_{i+1} \times D_{i+2} \times \ldots \times D_k$, and let μ_{-i} be a measure on D_{-i}. Similarly, let $m_{-i} = (m_1, m_2, \ldots, m_{i-1}, m_{i+1}, m_{i+2}, \ldots, m_k)$, and let θ_{-i} denote the parameter vector $\{\theta_0, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \theta_{i+2}, \ldots, \theta_k\}$.

We say λ is *completely separable* if

$$\lambda(t, m_1, \ldots, m_k; \theta) = \theta_{k+1}\lambda_0(t; \theta_0)\lambda_1(t, m_1; \theta_1) \ldots \lambda_k(t, m_k; \theta_k), \tag{2.1}$$

where $\theta_i \in \Theta_i$, and each λ_i is $\mathcal{F}^1$-predictable. θ_{k+1} represents a multiplicative constant; if this is not desired, Θ_{k+1} may simply be taken to be the constant 1. In some applications, it may be unreasonable to suppose that the process is completely separable. However, more generally one might suppose that a given component is separable, as in the following definition.

We say λ (or equivalently, the point process N) is *separable in mark* m_i if the 1-intensity may be written as

$$\lambda(t, m_1, \ldots, m_k; \theta) = \theta_{k+1}\lambda_i(t, m_i; \theta_i)\lambda_{-i}(t, m_{-i}; \theta_{-i}). \tag{2.2}$$

Mark m_i may be multiplicative and yet may influence the conditional rates λ_i and λ_{-i} at future times and the distribution of mark m_i may vary with t and may depend on any facets of the history of the process. The key feature in (2.2) is that the parameter θ_i only influences the process λ_i. The idea is that the rate λ may vary in time and might depend on mark m_i in potentially complicated ways, but the component related to mark m_i and the components related to other marks m_{-i} influence the rate in multiplicative fashion.

For point processes in general, the loglikelihood for the full parameter vector θ may be written (eq. 7.1.2 of Daley and Vere-Jones (2003)) as

$$L(\theta) = \int_D \log \lambda(x; \theta) dN - \int_D \lambda(x; \theta) d\mu. \tag{2.3}$$

The parameter vector $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k, \hat{\theta}_{k+1})$ is called the maximum likelihood estimate (MLE) of θ.

For a point process thought to be separable in mark m_i, one might choose to estimate only the parameters θ_i and θ_{k+1}, ignoring the other parameters and

other components of the process N. That is, one may consider maximizing the *partial loglikelihood*

$$\tilde{L}_i(\theta_i, \theta_{k+1}) = \int_D \log[\theta_{k+1}\lambda_i(t, m_i; \theta_i)]dN - \theta_{k+1}\int_{D_0}\int_{D_i} \lambda_i(t, m_i; \theta_i)d\mu_i dt. \quad (2.4)$$

The parameters $\tilde{\theta}_i, \tilde{\theta}_{k+1}$ maximizing $\tilde{L}$ may be called *partial maximum likelihood estimates* (PMLEs).

Under quite general conditions, maximum likelihood estimates of point process parameters can be shown to be asymptotically normal, with standard errors obtainable via the Fisher information, i.e., the diagonal of the inverse of the Hessian of the log-likelihood, or its estimate when this Hessian is unknown (see e.g., Ogata (1978) for the purely temporal case, or Rathbun (1996) for spatial-temporal point processes). Of course, since the PMLE is essentially a maximum likelihood estimate of a reduced model, such results generally apply to the PMLE as well.

3. PMLEs for Multiplicative Models

For processes that are separable in a certain mark, partial maximum likelihood estimates of the parameters governing the component of the rate related to this mark are often quite similar to maximum likelihood estimates. Some conditions under which the two estimates are exactly the same are given in the following elementary result.

Lemma 1. *Let N be a point process whose 1-intensity λ is separable in mark m_i as in (2.2). Suppose that both L and $\tilde{L}_i$ are differentiable with respect to θ_i, and that $\tilde{\theta}_i$ is the unique value of θ_i satisfying $\partial\tilde{L}_i/\partial\theta_i = 0$ and $\partial\tilde{L}_i/\partial\theta_{k+1} = 0$. Suppose also that (at least) one of the following three conditions holds, for some scalar γ, for all $t \in D_0$:*

$$\int_{D_{-i}} \lambda_{-i}(t, m_{-i}; \theta_{-i})d\mu_{-i} = \gamma, \forall\theta_{-i}, \quad (3.1)$$

$$\int_{D_i} \lambda_i(t, m_i; \theta_i)d\mu_i = \gamma, \forall\theta_i, \quad (3.2)$$

$$\int_D \lambda(t, m_1, \ldots, m_k; \theta)d\mu = \theta_{k+1}\int_{D_0}\int_{D_i} \lambda_i(t, m_i; \theta_i)d\mu_i dt = \gamma, \forall\theta, \quad (3.3)$$

then

$$\tilde{\theta}_i = \hat{\theta}_i. \quad (3.4)$$

The proof of Lemma 1 is given in the online supplement. Equations (3.1)–(3.3) are not impossibly restrictive. The following examples illustrate conditions under which these assumptions are met.

Example 1. The Epidemic-Type Aftershock Sequence (ETAS) model of Ogata (1988, 1998) is a type of branching model that is widely used in seismology. According to the ETAS model, the conditional rate λ is separable with respect to magnitude, and can be written $\lambda(t, m, \mathbf{x}) = \lambda_1(t, \mathbf{x}; \theta_{-i})\lambda_2(t, m; \theta_i)$, where $\lambda_2(t, m) = f(m)$ is the magnitude density, which is posited not to change over time. Thus the LHS of (3.2) becomes $\int f(m; \theta_i)dm = 1$, since f is a density. As a result, the parameters governing the magnitude density can be estimated separately, using only the observed magnitudes and not the times and spatial locations of the events. As noted in Schoenberg (2004), the magnitudes of prior events can influence the conditional intensity subsequently, but the process can nevertheless be separable in magnitude provided (2.2) holds, i.e., the parameters governing the magnitude distribution do not influence the other marginal distributions of the process.

Example 2. In the analysis of wildfires, one important mark is the amount of area burned, and it has often been noted that the density of area burned can change from year to year. This density (assuming it exists) may depend on the fuel age distribution and other dynamic conditions. It nevertheless must always integrate to unity, and models have been proposed that posit that the parameters governing this density do not interact with the other parameters governing the other distributions of the process. For instance, Schoenberg, Pompa, and Chang (2009) consider the model $\lambda(t, x, y, m) = f(m)\mu(x, y)\beta_1 \exp\{\beta_2 W(t) + \beta_3 R(t) + \beta_4 P(t) + \beta_5 L(t) + \beta_6 Temp(t) - \beta_7(\beta_8 - D(t))^2\}$, where $f(m)$ is the tapered Pareto burn area density, $\mu(x, y)$ is the spatial background fire rate at location (x, y) obtained by kernel smoothing the estimated origin locations of wildfires recorded prior to the beginning of the catalog used in the remainder of the model fitting, and $W(t)$, $R(t)$, $P(t)$, $L(t)$, $Temp(t)$, and $D(t)$ represent the wind speed, maximum relative humidity over the previous 24 hours, precipitation, lagged precipitation over the previous 60 days, temperature for day t, and the index of the day of the year corresponding to time t, respectively. For such models, (3.2) is satisfied with m_i the burn area of a fire (or equivalently (3.1) is satisfied where m_{-i} is the burn area, and m_i contains information on all other marks).

Example 3. When implementing maximum likelihood estimation algorithms in practice, one must verify that the optimization routine converges to the global maximum rather than some other local maximum. A common way of checking whether the routine's output is reasonable is by ensuring that the integral term in (2.3) is approximately equal to the number $N(D)$ of observed points, since $E\int_D \lambda(x; \theta)d\mu = E\int_D dN = EN(D)$. Similarly, in maximizing the partial likelihood, one would typically ensure that $\theta_{k+1}\int_{D_0}\int_{D_i} \lambda_i(t, m_i; \theta_i)d\mu_i dt$ is approximately equal to $N(D)$. If one imposes the constraint that each of these integrals must equal $N(D)$, then (3.3) is satisfied with $\gamma = N(D)$.

Example 4. In some models, spatial background rates are fitted by kernel smoothing of a certain fixed subset of n points (e.g., Ogata (1988), Schoenberg (2003)), and the bandwidth of the kernel density may be estimated by maximizing the partial likelihood governing only the spatial coordinates. In such situations, if the spatial domain has no boundary, or if boundary effects are negligible, or if a correction is used in the fitting so that each of the n points identically contributes a value of one to the total background rate, then as in the previous example, (3.3) holds with $\gamma = n$.

Recall that in the parameterization of each component $\lambda_i(t, m_i; \theta_i)$, the parameter θ_i need not be a scalar, but may rather be a vector in $\mathbf{R}^d$. (Similarly, m_i may also be vector-valued.) Although $\lambda_i(t, m_i)$ must be $\mathcal{F}$-predictable, it may depend on covariates, including external observations and/or functionals of the history of the point process. We turn now to the estimation of the parameters governing the effect of these covariates on λ.

Suppose that the parameterization of one particular component $\lambda_i(t, m_i; \theta_i)$ of the 1-intensity can be decomposed into a product of terms

$$\lambda_i(t, m_i; \theta_i) = f_1(X(t, m_i); \beta_1) f_2(Y(t, m_i); \beta_2), \tag{3.5}$$

where $\theta_i = (\beta_1, \beta_2)$, and X and Y are predictable processes. Such a model can arise for example when f_1 represents the effect on the rate caused by one collection of covariates, and f_2 represents the effect of another group of covariates. Here X and Y need not be scalars, but can be vector-valued or take values in an arbitrary measurable space.

Let $H_1(t, x, m_1), H_2(t, y, m_2)$, and $H(t, x, y, m_1, m_2)$ denote the cumulative distribution functions of $X(t, m_1)$, $Y(t, m_2)$, and of the pair (X, Y), respectively. Of particular interest is the special case where X and Y are independent, for then H has the multiplicative form

$$H(x, y) = H_1(x) H_2(y). \tag{3.6}$$

Let $\check{\beta}_1$ denote the maximum likelihood estimate when the parameter (vector) β_1 is estimated separately, i.e., the value of β_1 maximizing

$$\begin{aligned} \check{L}(\beta_1) := &\int_{D_0} \int_{D_i} \log[\theta_{k+1} f_1(X(t, m_i); \beta_1)] dN(t, m_i) \\ &- \theta_{k+1} \int_{D_0} \int_{D_i} f_1(X(t, m_i); \beta_1) d\mu_i dt. \end{aligned} \tag{3.7}$$

Theorem 1. *Suppose that the conditions of Lemma* 1 *hold and that* λ_i *is multiplicative as in* (3.5). *Suppose that* $\check{L}$ *is differentiable with respect to* β_1, *and that there exists a unique solution* $(\check{\theta}_{k+1}, \check{\beta}_1)$ *satisfying* $\frac{d\check{L}}{d\beta_1} = 0$. *If* H *has the multiplicative form* (3.6), *then* $\check{\beta}_1$ *is the MLE of* β_1.

Example 5. Baddeley and Turner (2000, 2005) consider a *log-linear* or *exponential* family of spatial point process models, which readily extend to the spatial-temporal case with random covariates via $\lambda(t,x) = \exp\{\theta^T S(t,x)\}$, where $S(t,x)$ is a vector of covariates observed at spatial-temporal location (t,x). The conditional rate thus is purely multiplicative with respect to all covariates, satisfying conditions (2.2) and (3.5). According to Theorem 3.2, if two of the covariates X and Y satisfy (3.6), then the parameters governing their components in the conditional rate λ may equivalently be estimated separately.

4. Additive Models

The result in Theorem 1 may seem intuitively obvious given (3.6), but note that this condition does not necessarily imply that the effects of X and Y may be estimated separately. For additive models, for instance, the result in Theorem 1 does not generally hold. For a simple example, suppose that N is a 1-dimensional point process whose conditional intensity has the form $\lambda(t) = \alpha X(t) + \beta Y(t)$, and suppose that $X(t) = 1$ and $Y(t) = t$, for all t. Then (3.6) holds, but the estimate $\check{\beta}$ obtained by separately estimating the coordinate $f_2(Y(t)) = \beta Y(t)$ is simply the MLE of β for the model $\lambda(t) = \beta t$, which is obviously different from the MLE of β for the model $\lambda(t) = \alpha + \beta t$.

This section explores conditions under which parameters can be estimated separately for the case of components of λ that are additive rather than multiplicative. As an alternative to the product form in (3.5), suppose instead that λ_i is parameterized as a sum of functions of the covariates X and Y,

$$\lambda_i(t, m_i; \theta_i) = f_1(X(t, m_i); \beta_1) + f_2(Y(t, m_i); \beta_2), \tag{4.1}$$

where $\theta_i = (\beta_1, \beta_2)$, and X, Y are predictable processes.

Consider the maximum likelihood estimate $\acute{\beta}_1(T)$ when the parameter (vector) β_1 is estimated individually, using observations on $[0,T] \times D_1 \times \ldots \times D_k$. That is, $\acute{\beta}_1(T)$ is the value of β_1 maximizing

$$\begin{aligned} &\acute{L}^{(T)}(\beta_1) \\ &:= \int_0^T \int_{D_i} \log[f_1(X(t, m_i); \beta_1)] dN(t, m_i) - \int_0^T \int_{D_i} f_1(X(t, m_i); \beta_1) d\mu_i dt. \end{aligned} \tag{4.2}$$

General conditions for the convergence in probability of the MLE $\hat{\theta}$ to the true parameter vector θ^* have been given by a variety of authors; see for instance Theorem 2 of Ogata (1978) for stationary one-dimensional processes, conditions for which are in the online supplement, or Theorem 1 of Rathbun (1996) for more general multi-dimensional point processes.

Theorem 2. *Suppose N satisfies the conditions for Theorem 2 of Ogata (1978). Suppose also that N satisfies the conditions of Lemma 1, and that λ_i has the additive form (4.1), where f_1 and f_2 are continuous in β_1 and β_2, respectively. Suppose also that $E\int\int|\lambda(t,m_i;\theta_i^*)\log\lambda(t,m_i;\theta_i)|d\mu_i dt<\infty$ and $E\int\int|\lambda(t,m_i;\theta_i^*)\log f_1(X(t,m_i);\beta_1)|d\mu_i dt<\infty$, and that there exists an open neighborhood U of the true parameter vector θ^*, such that for θ in U, $(1/T)\int_0^T\int_{D_i}f_2(Y(t,m_i);\beta_2)\,d\mu_i dt$ and $(1/T)\int_0^T\int_{D_i}\{[\lambda(t,m_i;\theta_i^*)f_2(Y(t,m_i);\beta_2)]\,/\,[f_1(X(t,m_i);\beta_1)]\}\,d\mu_i dt$ converge to zero in probability as $T\to\infty$. Then $\acute{\beta}_1(T)$ is a consistent estimate of β_1.*

The basic idea here is the following. By Lemma 1, $\tilde{\beta}_1=\hat{\beta}_1$, and this MLE $\hat{\beta}_1$ is known to be consistent under standard conditions. $\acute{\beta}_1$ is the estimator one would obtain by fitting the incorrectly specified likelihood in (4.2) by maximum likelihood, ignoring the effect of the covariate Y. Under the assumptions of Theorem 2, the effect of Y is so small that the true likelihood is sufficiently similar to (4.2) that $\acute{\beta}_1$ is consistent in this situation as well. For a formal proof, see the supplemental materials.

Example 6. The conditions on f_1 and f_2 in Theorem 2 may be satisfied when f_2 is small, both in absolute terms and relative to f_1. Let f_1 and f_2 be shorthand for $f_1(X(t,m_i);\beta_1)$ and $f_2(Y(t,m_i);\beta_2)$, respectively. Suppose that, for θ in a neighborhood U of θ^*, $|\lambda|$ is bounded in absolute value by some value b with probability going to one, and $\int_{D_i}f_2 d\mu_i$ and $\int_{D_i}f_2/f_1 d\mu_i$ converge to zero in probability as $t\to\infty$. Then so do $(1/T)\int_0^T\int_{D_i}f_2 d\mu_i dt$ and $(1/T)\int_0^T\int_{D_i}[\lambda(t,m_i;\theta_i^*)f_2/f_1]d\mu_i dt$, thus satisfying the last conditions in Theorem 2.

Example 7. If f_1 is bounded away from zero and $|\lambda|$ is bounded above, then the conditions on f_1,f_2 in Theorem 2 simply amount to the convergence to zero in probability of $(1/T)\int_0^T\int_{D_i}f_2 d\mu_i dt$ as $T\to\infty$. In particular, if $\int_{D_i}f_2 d\mu_i\to_p 0$ as $T\to\infty$, then these conditions are trivially satisfied.

5. Simulations

The accuracy of PMLEs can be demonstrated under various conditions using simulations. Specifically, we consider the case where a point process N is repeatedly simulated in the presence of some covariate or noise process Z, and then for each simulation, the model is estimated by maximum likelihood as though the covariate were completely ignored. The question is whether the parameters in various point process models can be accurately estimated by maximum likelihood even though this covariate Z is ignored in the parameter estimation.

For example, consider an inhomogeneous Poisson process N with intensity

$$\lambda(t,x,y)=\mu+\alpha x+\beta y+Z(t), \tag{5.1}$$

where N is observed in the time span $(0, T)$ and in the spatial domain $[0, 1]\times[0, 1]$, and where Z is a uniform white noise process contained in $\mathcal{F}_0$ such that for any t,

$$Z(t) \sim U(0, \frac{1}{t}). \tag{5.2}$$

Thus, N and the covariate Z are correlated, but letting $Y(t) = t$ and $Z(t) = f_2(t)$, both f_2 and $f_2\lambda(t)$ converge to zero as $t \to \infty$ so that the conditions of Theorem 2 are satisfied.

Such a process N can easily be simulated using the simulation technique of Lewis and Shedler (1979). Specifically, for given T and positive α, β, γ, and μ, one can set $b = \mu + \alpha + \beta + 1.0$, generate a homogeneous Poisson process with rate b on the spatial-temporal domain $[0, 1] \times [0, 1] \times [0, T]$ and, for each point (x_i, y_i, t_i), draw a uniform random variable u on $(0, 1/t_i)$, keeping the point independently of the others with probability $(\mu + \alpha x_i + \beta y_i + u)/b$.

For each of the estimates reported in this Section, the parameters were estimated by minimizing the negative partial log-likelihood, with the minimization done in R using the Broyden-Fletcher-Goldfarb-Shanno quasi-Newton gradient method (see e.g., Nocedal and Wright (1999)), and using starting values of double the actual parameter values.

The means and root-mean-squared errors (RMSE) of the parameters in model (5.1) are shown in Figure 1. Specifically, for each value of T, 100 simulations of model (5.1) were constructed, and for each such simulation, the parameters (μ, α, β) in the model

$$\lambda(t, x, y) = \mu + \alpha x + \beta y \tag{5.3}$$

were estimated, where $\mu = \alpha = \beta = 1$. One sees in Figure 1 the rapid convergence toward one in the means of all of the parameter estimates, despite the presence of a confounding factor Z in the simulations that is completely ignored in the fitting of model (5.3). The decrease to zero of the RMSE of the estimates $\hat{\mu}, \hat{\alpha}$, and $\hat{\beta}$ as T increases is also evident.

In contrast to the purely additive models (5.1) and (5.3), a multiplicative inhomogeneous Poisson model with intensity

$$\lambda(t, x, y) = \exp\{\mu + \alpha x + \beta y + Z(t)\} \tag{5.4}$$

was also simulated 100 times, with Z as in (5.2); the results, when the model

$$\lambda(t, x, y) = \exp\{\mu + \alpha x + \beta y\} \tag{5.5}$$

was estimated using data from the simulations of model (5.4), are shown in Figure 1 of the online supplement. As with the linear intensity case, the very rapid

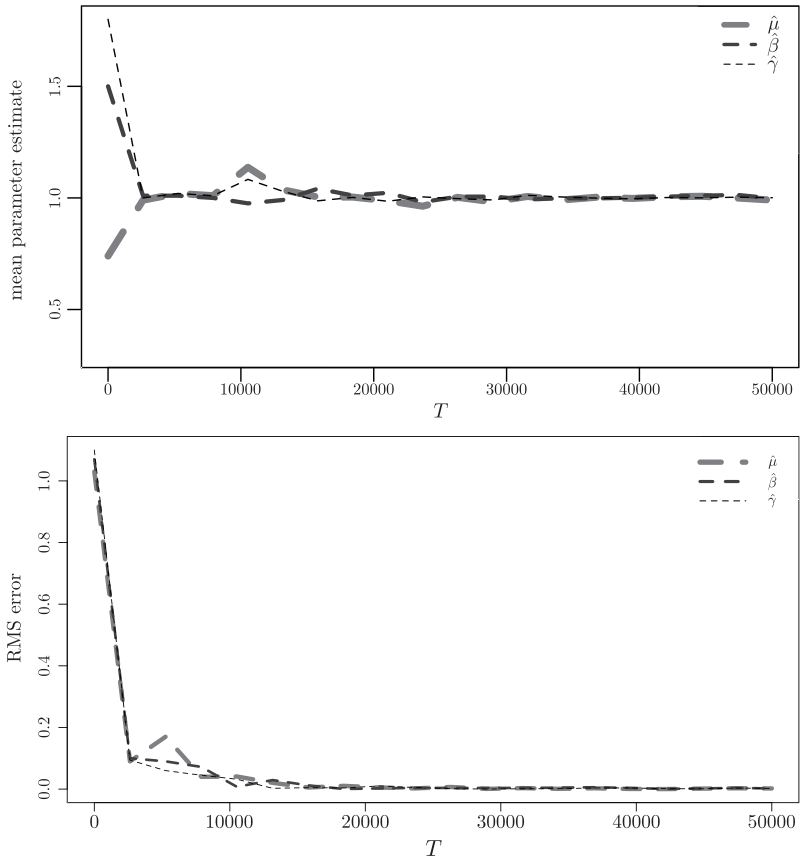

Figure 1. Means and RMSE in partial maximum likelihood estimates of parameters (μ, α, β) in model (5.3), using simulations of model (5.1) with $(\mu, \alpha, \beta) = (1, 1, 1)$. 100 simulations were performed for each T, for 50 equally spaced values of T between 1 and 50,000.

convergence to one of the means of the partial maximum likelihood estimates and the rapid convergence to zero of the RMSE of these estimates is evident. The confounding factor Z appears to have a rapidly diminishing impact on the PMLEs.

Simulations were run using alternative error models and alternative parameters and the results were similar. For instance, Figure 2 of the online supplement shows the means and RMSE of parameter estimates when model (5.3) is fit by PMLE, but the model being simulated is a Poisson process with intensity

$$\lambda(t, x, y) = \mu + \alpha x + \beta y + Z(t)W(x), \tag{5.6}$$

where N is observed in the time span $(0, T)$ and in the spatial domain $[0, 1] \times [0, 1]$, $Z(t)$ is again a uniform $\mathcal{F}_0$-measurable white noise process on $(0, 1/t)$ as described

above, and $W(x)$ is a uniform $\mathcal{F}_0$-measurable white noise process, independent of Z, such that $W(x)$ is uniform on $(0, x)$. Thus the error term in model (5.6) now varies with x and can thus be anticipated to interfere adversely with the estimation of α when model (5.6) is simulated and model (5.3) is estimated. However, as shown in Figure 2 of the supplemental materials, the bias in the parameter estimates appears to converge to zero and the RMSE of the PMLEs, when model (5.6) is simulated and model (5.3) is estimated, appear to converge rapidly to zero.

A variety of models, choices of parameters, and time spans T were selected, and a sampling of the results is summarized in Table 1 in the online supplemental materials. For simplicity, the focus in Table 1 is on models with similar parameter values. For instance, the means and RMSEs of the parameter estimates for when model (5.6) is simulated and model (5.3) is estimated are shown in the Table, for various time spans T, and similar results for self-exciting and self-correcting point process models are summarized in Table 1 of the online supplement as well. Specifically, we consider the case when the model

$$\lambda(t, x, y) = \mu + \alpha x + \beta y + Z(t) + K\nu \sum_{i:t_i<t} \exp(-\nu(t - ti)) \tag{5.7}$$

was simulated, with $Z(t)$ in (5.2), and the results were used to estimate the model

$$\lambda(t, x, y) = \mu + \alpha x + \beta y + K\nu \sum_{i:t_i<t} \exp(-\nu(t - ti)) \tag{5.8}$$

with the covariate $Z(t)$ removed. The parameter $\nu > 0$ governs the temporal scale of the clustering or inhibiiton. K is the branching ratio of the process, which is highly clustered, or self-exciting, when $K > 0$ and is inhibitory, or self-correcting, when $K < 0$. Models with different noise processes,

$$\lambda(t, x, y) = \mu + \alpha x + \beta y + K\nu \sum_{i:t_i<t} \exp(-\nu(t - ti)) + \tilde{Z}(t) \tag{5.9}$$

and

$$\lambda(t, x, y) = \mu + \alpha x + \beta y + \tilde{Z}(t) \tag{5.10}$$

were also simulated, where $\tilde{Z}(t) = (1 - (-1)^{N(0,t)})/t$.

Table 1 of the online supplement shows the root-mean-square errors (RMSE) of the parameters (μ, α, β) in some of the models described, but with different missing covariates. The RMSE of the parameter estimates appears to converge to zero as T gets large for both the self-exciting and self-correcting model, even though the potential confounders Z or $\tilde{Z}$ in the simulations are completely ignored in the fitting of the models. For comparison, the case Table. The confounding factors Z and $\tilde{Z}$ apparently have minimal effect on the estimates of

the parameters in model (5.7), for both self-exciting and self-correcting point processes.

The estimation reported in Table 1 of the online supplement is non-standard, as in each case other than rows 2 and 5, the model governing the simulations is different from the model being estimated. For each of the given models, the model with the covariate Z is simulated, and the model without this covariate is estimated. In rows 2 and 5, however, the results represent standard maximum likelihood estimates and may be useful for comparison. One sees from the table that the missing covariate seems to have little effect quite generally on the estimates of the parameters for the inhomogeneous Poisson, self-exciting, and self-correcting models, and one sees that the RMSEs of these parameter estimates decrease as the time span T increases.

6. Application to Earthquake Weather

Theorems 3.2 and 4.1 suggest that as long as a given covariate's effect on the conditional rate is small, estimates of the parameters governing the rest of the process will scarcely be influenced by the omission of the given covariate. The simulations in Section 5 support this. In this Section, we explore how the parameters governing models for the rate of earthquakes in California are affected by the inclusion or omission of weather data.

The idea that weather can influence earthquakes has been the subject of much debate and controversy and has been largely dispelled since the work of Humphreys, who stated that earthquake weather was merely a misleading, psychological impression (Abbe (1919)). Recently, the topic was raised again by Enescu, Hainzl, and Ben-Zion (2009) and Hainzl et al. (2013), who claimed that in zones of near critical stress, small changes in stress due to temperature can substantially influence seismicity. A question we address here is whether the addition of weather data into existing models for seismicity would lead to a substantial change in the parameter estimates governing earthquake triggering and other aspects of seismicity.

Southern California weather data are publicly available from the National Weather Service Forecast Office of the National Oceanic and Atmospheric Administration (NOAA). We follow Hainzl et al. (2013) by considering the daily average of the daily high temperature recorded at three local stations, in this case the National Weather Service's Ontario Station, Santa Ana Fire Station, and Anaheim Station. The corresponding geographical coordinates of the three stations are reported as (Lon: 117.57583^{o}W, Lat. 34.05333^{o}N, Elevation 287.1m), (Lon: 117.868^{o}W, Lat. 33.743^{o}N, Elevation 33.5m), and (Lon: 117.89374^{o}W, Lat. 33.77635^{o}N, Elevation 35.7m), respectively. Daily weather information has been compiled for each of the three stations since May 23, 1998, and we consider here

the daily high temperature records from the start date of May 23, 1998 to March 11, 2014. The records at the stations were averaged each day, and in the event of missing data, records at all available stations were averaged. Curiously, on 16 days during this period, each of the stations had duplicate records; NOAA officials were contacted about this and indicated that the cause of this problem is unclear but in such circumstances the first record should be ignored and only the second record should be used, so this is what was done here. Data are missing on only 4.9% of days during this time period, and on no days were all 3 records missing.

Data on earthquakes in southern California have been compiled and are publicly available from the Southern California Earthquake Data Center. The catalog contains estimates of the origin time, hypocenter, moment tensor, and various measures of magnitude for earthquakes dating back to 1932, and is believed to be complete in recent years down to magnitude 1.8 (Hutton, Woessner, and Hauksson (2010)). We restrict our attention here to the subset of shallow (estimated depth $\leq 75km$) recorded earthquakes in southern California with moment magnitude at least 3.0, as these are of greater practical interest and are fraught with fewer missing data issues. Epicentral locations of these earthquakes are shown in Figure 2a, with larger circles corresponding to circles of greater magnitude. To match the temperature data, we focused on the time span of May 23, 1998 to March 11, 2014, and the spatial region from longitude 117.0 to 118.0 and latitude 33.0 to 35.0.

The mean daily high temperature on days with at least one recorded $M \geq 3.0$ earthquake is 24.56^oC, compared to a mean daily high temperature of 25.06^oC across all days. Over the observed spatial-temporal region, the correlation between the daily high temperature and the daily number of recorded $M \geq 3.0$ earthquakes is -0.0124. The permutation test standard error is 0.0135 and the corresponding p-value is 0.360. Because earthquakes are clustered, the daily seismicity totals are highly correlated as are daily high temperatures, so the permutation standard error is likely a considerable underestimate, but even according to the permutation standard error, the observed sample correlation is not statistically significant.

The question is whether, despite its minimal correlation with seismicity, the daily temperature data may nevertheless induce substantial changes in the parameter estimates for such standard models of seismicity as ETAS. The results of the preceding sections suggest that the influence on ETAS parameters caused by the introduction of temperature would be small. On the other hand, practitioners of maximum likelihood estimation are familiar with how even slight changes to a model and the introduction of a single extra parameter can sometimes result in sharp changes to maximum likelihood estimates.

The ETAS models of Ogata (1998) referred to in Example 1 are commonly used to model earthquake occurrences. One version of the model has the form

$$\lambda(t,x,y) = \mu\rho(x,y) + \sum_{i=1}^{n} g(t-t_i, x-x_i, y-y_i; M_i), \tag{6.1}$$

where $\rho(x,y)$ is a spatial density, M_i is the magnitude of earthquake i, and g is a triggering function indicating how the expected rate of seismicity increases following an earthquake of magnitude M_i. We consider here the normalized version of the triggering function proposed in Ogata (1998) with normalization suggested in Schoenberg (2013),

$$\begin{aligned} &g(t-t_i, x-x_i, y-y_i; M_i) \\ &\quad = \{\frac{K(p-1)c^{p-1}(q-1)d^{q-1}}{\pi}\}(t_i - t + c)^{-p} \exp\{a(M_i - M_0)\}(r_i^2 + d)^{-q}, \end{aligned}$$

where $r_i = ||(x,y) - (x_i, y_i)||$ and M_0 is the lower magnitude cutoff for the catalog, which here is 3.0.

Figure 3 of the online supplement shows the fitted parameters of model (6.1) over time, using the progressive approximate MLE technique of Schoenberg (2013) and using a constant for ρ. For starting values of the parameters, the values from the 2nd row of Table 3 of Ogata (1998) were used, as in Schoenberg (2013). Standard errors for the parameters were estimated using the inverse of the Hessian of the loglikelihood. One sees that, after about 6-7 years of data, most of the parameters have nearly converged, though parameter d seems somewhat unstably estimated using this limited dataset.

Using temperature data, one may modify the ETAS model in (6.1) by replacing the term μ in (6.1) with a term exponential or linear in temperature, for example, resulting in

$$\lambda(t,x,y) = \mu\rho(x,y)\exp\{\nu\ Temp(t)\} + \sum_{i=1}^{n} g(t-t_i, x-x_i, y-y_i; M_i), \tag{6.2}$$

or

$$\lambda(t,x,y) = \{\mu + \nu\ Temp(t)\}\rho(x,y) + \sum_{i=1}^{n} g(t-t_i, x-x_i, y-y_i; M_i), \tag{6.3}$$

respectively, where $Temp(t)$ denotes the daily high temperature on day t.

Figure 4 of the online supplement shows the fitted ETAS parameters using model (6.2) compared to those of model (6.1). The differences in parameter estimates induced by the addition of temperature in the model are modest, as expected. The introduction of temperature does noticeably change the estimates of parameter d, however. This is not surprising, given the high volatility in the estimate of d. Results for model (6.3) were similar.

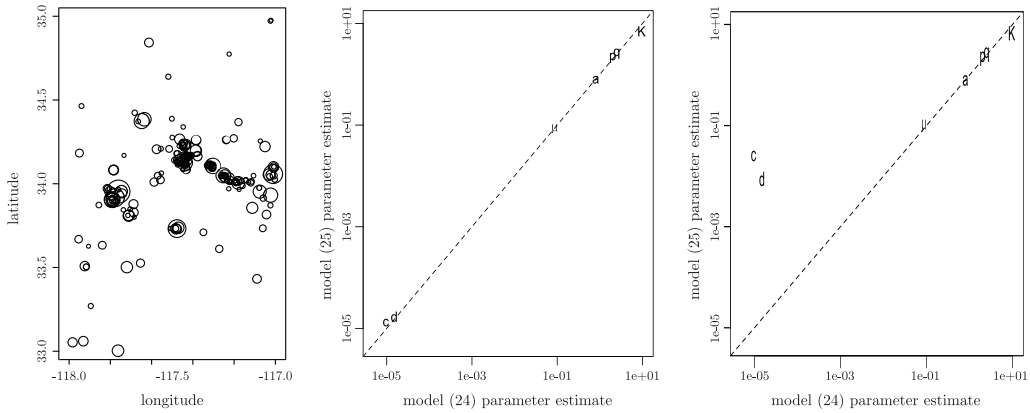

Figure 2. (a) Estimated epicenters of shallow southern California earthquakes of magnitude $M \geq 3.0$ with longitude -117.0 to -118.0 and latitude 33.0 to 35.0, from May 23, 1998 to March 11, 2014. Circle radii are proportional to recorded earthquake magnitude. (b) Comparison of estimated ETAS parameters for model (6.1) on the x-axis vs. model (6.2) on the y-axis. (c) Comparison of estimated ETAS parameters for model (6.1) on the x-axis vs. model (6.3) on the y-axis. For parameter a, the estimate of $a/2$ is shown instead of a so that the point would not overlap on the plot with the estimate of p.

One can inspect how the estimated effect of daily high temperature affects the background seismicity rate, according to the fitted versions of models (6.2) and (6.3). The estimate of the parameter ν governing the effect of temperature on seismicity is 0.00421 (0.000286) for model (6.2) and 0.00393 (0.0000512) for model (6.3). These coefficients are positive, in agreement with the conclusions of Hainzl et al. (2013) that warmer weather may lead to increased seismicity. Although there is a negative correlation between overall seismicity and temperature in our dataset, according to the fitted model, when triggered earthquakes are accounted for, higher temperatures have a slight positive association with background seismicity, or mainshocks.

Figures 2b and 2c compare final maximum likelihood ETAS parameters estimates from model (6.1) with those from models (6.2) and (6.3). One sees very close general agreement among most of the parameters. Parameters c and d for model (6.3) are an exception, and these parameters are substantially changed by the introduction of temperature into the model. Note that the changes in the parameter estimates in absolute terms are actually quite small, however.

7. Discussion

Theorems 1 and 2 imply that parameters governing individual covariates in multi-dimensional point process models may often be estimated separately. Indeed, as supported by the simulations in Section 5 and the application in Section 6, estimates of the parameters governing a given covariate's effect on the conditional intensity will hardly be influenced by the omission of other covariates, even if these missing covariates can influence the conditional intensity overall and may even be confounded with the given covariate in an additive or multiplicative way. Both results require the point process to be stationary in time; maximum likelihood estimates for non-stationary point processes are frequently inconsistent even for very simple models. The conditions in Theorems 1 and 2 essentially mandate that the impacts of the missing covariates on the conditional intensity are not too large. Thus these results are not surprising, and the general conclusion that parameters may be well estimated in the absence of data on covariates that do not greatly affect the conditional intensity of the process seems to be widely acknowledged among practitioners, though this is perhaps the first attempt at mathematical support for this notion.

These results may have implications for point process estimation. It is typically far easier (and faster) to obtain a PMLE $\check{\beta}$ or $\acute{\beta}$ than to search over values of all parameters in order to find the value $\hat{\beta}$ maximizing the full likelihood. It is important to note, however, that ignoring relevant covariates is not advocated here. In the application to weather and earthquake modeling, the point here is not that weather or climate should be ignored in modeling or forecasting earthquakes. Indeed, if such information can lead to more accurate forecasts then of course it should be included. The argument here is that a model's omission of weather and other variables with little influence on the conditional rate of seismicity need not cause one to doubt the accuracy of all parameter estimates and inferences based on the model.

The results in Sections 3 and 4 may have implications for model *building* as well. It is typically extremely difficult to construct realistic models for multi-dimensional point processes with many covariates. Ideally such models should be based on well-understood physical principles and subject-matter expertise. However, in some situations empirically-based models may be sought, and one method for constructing such a model would be to individually investigate the distribution of the coordinates, and the individual contribution to the conditional intensity of each (or perhaps small collections of) covariates. These marginal distributions of the process could then be estimated separately, and the parametric forms for each could readily be inspected for goodness-of-fit. The results above suggest circumstances under which a model may be thus constructed and estimated.

Supplementary Materials

The supplementary materials contain a proof of Lemma 1, a proof of Theorem 1, a counterexample to Theorem 1, the conditions for Theorem 2, a detailed proof of Theorem 2, figures showing the means and RMSE in PMLEs of parameters (μ, α, β) in models (5.5) and (5.3), using simulations of models (5.4) and (5.6), respectively, a table indicating the RMSE for parameter estimates for various models simulated in Section 5, and figures showing how ETAS parameters in (6.1) and (6.2) vary with catalog length, when fit to the data described in Section 6, both in absolute terms and relative to their estimated values using the entire earthquake catalog.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 9978318. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Thanks to the NOAA and SCEDC for their wonderfully collected and catalogued datasets.

References

Abbe, C. (1919). Earthquake weather. *Monthly Weather Review.* War Department, Office of the Chief Signal Officer, 180-181.

Abramowitz, M. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* (Edited by M. Abramowitz and I. Stegun), U.S. Government Printing Office, Washington D.C.

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Austral. N. Z. J. Statist.* **42**, 283-322.

Baddeley, A. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *J. Statist. Software* **12**, 1-42.

Brown, T., Ivanoff, B. G. and Weber, N. C. (1986). Poisson convergence in two dimensions with application to row and column exchangeable arrays. *Stochastic Process. Appl.* **23**, 307-318.

Chang, C. and Schoenberg, F. P. (2010). Testing separability in multi-dimensional point processes with covariates. *Ann. Inst. Statist. Math.*, to appear.

Daley, D. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes.* Springer, New York.

Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes.* 2nd ed., Vol. 1: Elementary Theory and Methods. Springer, New York.

Enescu, B., Hainzl, S. and Ben-Zion, Y. (2009). Correlations of seismicity patterns in Southern California with surface heat flow data. *Bull. Seismol. Soc. Amer.* **99**, 3114-3123.

Hainzl, S., Ben-Zion, Y., Cattania, C. and Wassermann, J. (2013). Testing atmospheric and tidal earthquake triggering at Mt. Hochstaufen, Germany. *J. Geophys. Res.* **118**, 5442-5452.

Hutton, K., Woessner, J. and Hauksson, E. (2010). Earthquake monitoring in southern California for seventy-seven years (1932-2008). *Bull. Seismol. Soc. Amer.* **100**, 423-446.

Lewis, P. A. W. and Shedler, G. S. (1979). Simulation of non-homogeneous Poisson processes by thinning. *Naval Res. Logistics Quart.* **26**, 403-413.

Lipster, R. S. and Shiryaev, A. N. (1977). *Statistics of Random Processes, I: General Theory.* Springer-Verlag, New York.

Merzbach, E. and Nualart, D. (1986). A characterization of the spatial Poisson process and changing time. *Ann. Probab.* **14**, 1380-1390.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization.* Springer, New York.

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Ann. Inst. Statist. Math.* **30**, 243-261.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83**, 9-27.

Ogata, Y. (1998). Space-time point process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50**, 379-402.

Rathbun, S. L. (1996). Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *J. Statist. Plann. Inference* **51**, 55-74.

Schoenberg, F. P. (2003). Multi-dimensional residual analysis of point process models for earthquake occurrences. *J. Amer. Statist. Assoc.* **98**, 789-795.

Schoenberg, F. P. (2004). Testing separability in multi-dimensional point processes. *Biometrics* **60**, 471-481.

Schoenberg, F. P. (2013). Facilitated estimation of ETAS. *Bull. Seismol. Soc. Amer.* **103**, 601-605.

Schoenberg, F. P., Pompa, J. L. and Chang, C. (2009). A note on the non-parametric and semi-parametric modeling of wildfire hazard in Los Angeles County, California. *Environmental and Ecological Statist.* **16**, 251-269.

Department of Statistics, University of California, Los Angeles, CA 90095–1554, USA.

E-mail: frederic@stat.ucla.edu

(Received May 2014; accepted July 2015)