Spatial Statistics 29 (2019) 160–176

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Handling missing data in self-exciting point process models

J. Derek Tucker *, Lyndsay Shand, John R. Lewis

Sandia National Laboratories, PO Box 5800 MS 1202, Albuquerque, NM 87185, United States

ARTICLE INFO

Article history:
Received 22 August 2018
Accepted 11 December 2018
Available online 19 December 2018

Keywords:
Bayesian inference
Hawkes process
Missing data
Point process

ABSTRACT

Self-exciting point processes have been applied to a wide variety of applications to understand event rates and clustering as a function of time and space. Typically, estimation procedures require a full temporal history of the data and do not handle cases where some of the history of the process is missing and unobserved. However, in many applications data collection is non-persistent resulting in known intervals of time where events of the process are unobserved. Motivated by these situations, a Bayesian estimation procedure for self-exciting point processes with missing histories is developed. The method naturally handles the missing data mechanism probabilistically through a specific step and is demonstrated on simulated data and a real conflict monitoring data where records over a period of time have been lost.

1. Introduction

Self-exciting point processes, also known as Hawkes processes, are an important class of point processes that can model clustered events in both time and space. Specifically, Hawkes processes allow for a *parent* event to trigger a set of subsequent *direct offspring* events. In addition to the events themselves, marks, or features associated with each event can be incorporated. Such models are known as marked point processes. Since its proposal by Hawkes and Oakes (1974), variations of these models have been applied to a wide variety of problems. Notably, they have been used to quantify the degree of seismic activity in a region (e.g., the Epidemic Type Aftershock Sequence (ETAS) model from Ogata (1988, 1998)), where the magnitude of an earthquake constituted the mark. It has also found use in modeling of violent crime activity (Mohler et al., 2011), social networks

* Corresponding author.
E-mail addresses: jdtuck@sandia.gov (J. Derek Tucker), lshand@sandia.gov (L. Shand), jrlewi@sandia.gov (J.R. Lewis).

https://doi.org/10.1016/j.spasta.2018.12.004
2211-6753/

(Zipkin et al., 2016, and Rizoiu et al., 2018), and in biology applications (Balderama et al., 2012). A comprehensive review of self-exciting point processes from Reinhart (2018) provides additional examples. Due to its wide use, a large amount of research is dedicated to estimating the parameters of a Hawkes process efficiently and accurately.

This paper addresses an important feature of self-exciting point process applications: missing data. Data collection method is often non-persistent, resulting in intervals of time where events are occurring, but unobserved. A motivating example can be seen when considering recorded terrorism events from the Global Terrorism Database (2017) (GTD), where the database is missing records for the entire year of 1993. Some of the missing records have since been recovered, but the database provides a natural situation where the missing data mechanism should clearly be accounted for in statistical inferences.

This data is revisited and describes in further detail in Section 7. Additional examples where it is critical to account for missing data include situations where sections of the process history are unobserved due to sensor errors, lack of records, or deliberate action of no observation. Accounting for missing intervals in point-processes, where the number of missing points is random and points cannot be considered independent, is minimally addressed in the literature. Zipkin et al. (2016) presents a maximum likelihood (ML) method for filling in missing data in records of communications in social networks. Our approach differs in that we take a Bayesian approach which overcomes the parameter estimation problems that result in a ML approach.

To account for the missing data mechanism, this paper leverages the Bayesian approach of Rasmussen (2013) who developed an efficient MCMC estimation procedure for the Hawkes process which takes advantage of the *branching structure* of the process that specifies relationships between *parent* and *direct offspring* points. The branching structure is considered as a set of latent variables sampled within the MCMC algorithm. Parameter updates are then conditioned on the branching structure, which results in marked efficiency gains over MCMC procedures which do not condition on the branching structure.

A Bayesian approach to handling the missing data is preferred for two reasons. First, Veen and Schoenberg (2008) demonstrate that due to the complexity of the likelihood function; a direct maximum-likelihood (ML) estimation can be fraught with problems and proposed using an expectation — maximization (EM) approach, which they observed to be more numerically stable and more accurate. In the Bayesian framework, we also more accurately represent the estimation uncertainty in the presence of missing data. Second, and more important, the Bayesian approach naturally handles the missing history by augmenting the observed data with (latent) missing data occurring in the unobserved sections of the history to form a "complete" set of data. The MCMC algorithm used to sample from the complete data posterior is augmented with an additional Gibbs step to sample missing data conditional on the observed data and parameters. The data augmentation approach taken here is the standard Bayesian approach to handle missing data (Tanner and Wong, 1987) and can be seen as a combination of the EM algorithm with multiple imputation (Little and Rubin, 2014).

A challenge in implementing the data augmentation approach is sampling from the appropriate full conditional distribution, as the missing data is dependent on the observed data. In this paper, we present a Metropolis–Hastings approach to sample from the missing data full-conditional distribution for a Hawkes process utilizing the historical data observed prior to the missing data interval.

The organization of the paper is as follows: Section 2 provides an overview of the marked Hawkes process. Section 3 outlines the MCMC step in the Bayesian framework for handling missing data in a marked point process. Section 4 outlines three different model variations considered. Section 5 describes the entire Bayesian estimation procedures using a Metropolis within-Gibbs approach. Sections 6 and 7 demonstrate the advantage of accounting for missing data on sets of simulated data and on a real data set from the Global Terrorism Database (2017) respectively. Lastly, Section 8 provides concluding remarks and a discussion on possible extension of the methods.

2. Self-exciting point process models

In the literature there are two equivalent ways to define a Hawkes process: (1) with its conditional intensity function or (2) as a Poisson cluster process. Let $X = \{(t_i, \kappa_i)\}$ be a marked point process,

where $t_i \in \mathbb{R}$ denotes the time points of the point process and $\kappa_i \in \mathbb{M}$ denotes the marks where $\mathbb{M}$ is a measurable space. Let N be the corresponding counting measure for the process on $\mathbb{R}$, counting the number of points falling within any arbitrary Borel set.

Assuming the first definition, a temporal point process $N(0, t)$ is characterized by its conditional intensity $\lambda(t)$, which is defined as the limit of the expected number of points around t given the history $\mathcal{H}_t$ of all the points up to time t (Daley and Vere-Jones, 2003),

$$\lambda^*(t) = \lim_{\Delta t \downarrow 0} (E[N\{(t, t + \Delta t)\}|\mathcal{H}_t]/(\Delta t)) \tag{1}$$

In the Hawkes process, the conditional intensity takes the form

$$\lambda^*(t) = \mu(t) + \sum_{k:t_k<t} \alpha(\kappa_k) g(t - t_k, \kappa_k), \tag{2}$$

where $\mu(t)$ is the background intensity with parameters $\mu = (\mu_1, \ldots, \mu_{\eta_\mu})$ and $\alpha(\kappa)g(t, \kappa)$ is the kernel or the offspring intensity. The function $\alpha(\kappa)$ is a non-negative function on $\mathbb{M}$ with parameter $\alpha = (\alpha_1, \ldots, \alpha_{n_1})$. $g(t, \kappa)$ is the normalized offspring intensity with parameters $\beta = (\beta_1, \ldots, \beta_{n_1})$. The mark distribution density conditional on the current time and the past points is $\gamma^*(\kappa|t) = \gamma(\kappa|t, \{t_i, \kappa_i\}_{t_i<t})$ with parameter $\gamma = (\gamma_1, \ldots, \gamma_{n_2})$. Assuming observed points $x = \{(t_1, \kappa_1), \ldots, (t_n, \kappa_n)\}$ on $[0, T)$ for some fixed time $T > 0$, then Proposition 7.3III of Daley and Vere-Jones (2003) show the likelihood of the collection of all parameters ϕ given x is

$$p(x|\phi) = \left(\prod_{i=1}^{n} \lambda^*(t_i)\gamma^*(\kappa_i|t_i)\right) \exp\left(-\int_0^t \lambda^*(s) \int_{\mathbb{M}} \gamma^*(\kappa|s) d\kappa ds\right) \tag{3}$$

$$= \left(\prod_{i=1}^{n} \lambda^*(t_i)\right) \left(\prod_{i=1}^{n} \gamma^*(\kappa_i|t_i)\right) \exp\left(-\int_0^t \lambda^*(s) ds\right) \tag{4}$$

where, the equality holds since $\int_{\mathbb{M}} \gamma^* d\kappa = 1$. The * notation as used in Rasmussen (2013) is to denote the dependence on the temporal history, i.e. $\{t_i, \kappa_i\}_{t_i<t}$.

Different specifications for μ, α, and g give rise to different models, the most popular of which are the Epidemic Type Aftershock Sequences (ETAS) models found in seismology Ogata (1998); Ogata and Zhuang (2006) which have historically been estimated using MLEs. Non-parametric estimation of μ, α, and g has also been proposed, e.g. Marsan and Lengliné (2008), Mohler et al. (2011) and Chen and Hall (2016), and is discussed later on in the paper.

Alternatively defining the Hawkes process as a Poisson cluster process implies the process is generated by a latent branching structure where we have two types of points: parents and offspring. The definition of this process is as follows:

1. The set of immigrants (I), or first generation of parents, S_0 follow a marked Poisson process with intensity $\mu(t)$.
2. Each immigrant $t_i \in S_0$ has an associated mark κ_i with mark density function $\gamma_I(\kappa_i|t_i)$.
3. Each marked point $\{t_i, \kappa_i\}$ of immigrants generates a cluster S_i, where the clusters are assumed to be independent.
4. The cluster S_i consists of marked points $\{t_j, \kappa_j\}$ of offspring generated as a marked Poisson process with intensity $\alpha(\kappa_i)g(t - t_i, \kappa_i)$ and mark density $\gamma_O(\kappa|t_j, \{t_i, \kappa_i\})$. $\alpha(\kappa_j)$ is the mean number of points with mark κ_j.
5. Each point $\{t_j, \kappa_j\} \in S_i$ becomes a potential parent to the next generation of offspring generating a cluster S_j analogously to Step 4. The process continues iteratively for each subsequent point.
6. The process, X is the union of all the clusters.

The set of parent–offspring relationships is known as the branching structure. Assuming we have a collection of arrival times $t_j \in X$, we represent the branching structure as $Y = \{y_j\}$, where $y_j = 0$ means t_j is an immigrant point and $y_j = i$ means t_j is an offspring of t_i. Conditional on Y, the arrival times can be partitioned into $n + 1$ sets $S_0, \ldots, S_n$, where

$$S_i = \{t_j; Y_j = i\}, \ i = 1, 2, \ldots, n - 1$$

so that S_0 is the set of all immigrants points, and S_i is the set of all offspring of the point at time t_i. The sets S_i are conditionally independent of each other and their union is the entire time history. The likelihood of (ϕ, Y) conditional on x (i.e., the distribution of x given the model parameters ϕ and the branching structure Y) can be written as in Rasmussen (2013):

$$p(x|\phi, Y) = p(S_0|\phi, Y)\prod_{i=1}^{n} p(S_i|\phi, Y)$$

Using the respective intensities and mark distributions for parents and offspring, we have,

$$p(x|\phi, Y) = \exp(-M(T))\prod_{t_j\in S_0}\mu(t_j)\gamma_I(\kappa_j|t_j) \times \prod_{i=1}^{n}\left(\exp(-\alpha(\kappa_i)G(T-t_i))\prod_{t_j\in S_i}\alpha(\kappa_j)g(t_i-t_j)\gamma_O(\kappa_j|t_j,\{t_i,\kappa_i\})\right), \quad (5)$$

Specifying the process as a Poisson cluster process is shown to be more computationally efficient than using the conditional intensity function (Rasmussen, 2013). Thus, all examples here will utilize this latent branching structure. Details for the specific variations of point-process models considered in this paper are given in Section 4.

Parameter estimation of models for cluster processes such as the one presented above can be challenging when the mark distribution is not independent of time, i.e. when the mark of the offspring depends on the mark of the parent. To ensure reasonable estimation of the model in this case, the model must be stable. Zhuang et al. (2013) shows that for models with i.i.d. marks, the stability is ensured when the criticality parameter (cp) is less than 1, i.e.,

$$cp = \int_K \alpha(\kappa)\gamma_O(\kappa|t)d\kappa < 1,$$

where K represents the space of all possible offspring marks. When the offspring marks depend on the parent mark, the stable condition might be quite complicated. Zhuang et al. (2013) provides some examples for various models. The criticality parameter characterizes the asymptotic population behavior after sufficiently many generations of the branching process.

3. Accounting for missing temporal histories

This section sets up the estimation procedure used to account for missing time histories. Assume a point process model given parameters ϕ with likelihood $p(x|\phi)$ generates the points $x = \{(t_1, \kappa_1), \ldots, (t_n, \kappa_n)\}$ with $t_i \in [0, T)$ for some fixed time $T > 0$. Suppose we only observe $t_i \notin M = \cup_{k=1}^{K} M_k$, where the M_k are a set of disjoint intervals within $[0, T)$. Write $x = (x_{obs}, x_{miss})$, where x_{obs} consists of observed points $(t_{obs,i}, \kappa_{obs,i})$ and x_{miss} consists of unobserved points $(t_{miss,j}, \kappa_{miss,j})$ with $t_{miss,j} \in M$. In this situation, it is desired to evaluate the posterior of ϕ given x_{obs}:

$$p(\phi|x_{obs}) \propto p(\phi)p(x_{obs}|\phi) \quad (6)$$

Note, the likelihood $p(x_{obs}|\phi)$ implicitly includes conditioning on the knowledge that no data are observed on M to avoid more cumbersome notation. In general, sampling from this posterior can be done by sampling from the joint posterior distribution of (ϕ, x_{miss}), $p(\phi, x_{miss}|x_{obs})$, and marginalizing over x_{miss}. To do this, Gibbs sampling (Geman and Geman, 1984; Gelfand and f. M. Smith, 1990) together with data augmentation (Tanner and Wong, 1987) can be used to iteratively sample from the two full conditional distributions:

1. $p(\phi|x)$
2. $p(x_{miss}|\phi, x_{obs})$

The first full conditional is the complete data posterior. Often algorithms are readily available to sample from this distribution. Utilizing the latent branching structure and the corresponding

likelihood (5), we use a Metropolis-within-Gibbs MCMC algorithm as described by Rasmussen (2013). A more detailed description of the MCMC algorithm for the complete data model is described in Section 5.1.

The second full conditional is the distribution of the missing data given the parameters and observed data. The approach taken for this step is context specific but in general, a Metropolis–Hastings approach can be used. For each known missing time interval, one proposes a set of missing data from a user-specified proposal distribution and either accepts or rejects this proposal probabilistically to satisfy the detailed-balance condition (Robert and Casella, 2004). This algorithm is then repeated iteratively for each missing interval. The method we develop to do this is detailed in Section 5.2. The data-augmentation approach is prevalent in missing data problems (Little and Rubin, 2014; Tanner and Wong, 1987; Kong et al., 1994; Gelfand et al., 1990), implementation of the EM-algorithm (Dempster et al., 1977), and fitting mixture models (Diebolt and Robert, 1994; McLachlan and Peel, 2004).

4. Models

To demonstrate the importance of accounting for missing time intervals, we consider variations of the following general model form which is similar to specific models presented by Rasmussen (2013) and commonly found throughout the literature. We assume a homogeneous parent intensity $\mu(t) = \mu$, constant total offspring intensity $\alpha(\kappa) = \alpha$, an exponential decay offspring intensity (free of κ) $g(t) = \beta \exp(-\beta(t))$, and a mark density for the immigrants $\gamma_I(\kappa|t)$ and offspring $\gamma_O(\kappa|t)$. The intensity at a given time point t:

$$\lambda(t) = \mu + \alpha \sum_{k:t_k<t} g(t - t_k) \tag{7}$$

and the likelihood in (5) becomes,

$$p(x|\phi, Y) = \exp(-\mu T)\mu^{|S_0|} \prod_{t_j \in S_0} \gamma_I(\kappa_j|t_j) \prod_{i=1}^{n} \left[\exp(-\alpha G(T - t_i))\alpha^{|S_i|} \prod_{t_j \in S_i} g(t_i - t_j) \right.$$
$$\left. \gamma_O(\kappa_j|t_j|\{t_i, \kappa_i\}) \right], \tag{8}$$

where $|S_i|$ denotes the size of cluster i. This is similar to the likelihood developed by Ross (2016), who points out that conditioning on the branching structure makes μ independent of the other model parameters and drastically weakens the dependence between the other parameters. The priors for μ, α, and β are independent gammas:

$$\mu \sim \text{Gamma}(\alpha_\mu, \text{rate} = \beta_\mu), \quad \alpha \sim \text{Gamma}(\alpha_\alpha, \text{rate} = \beta_\alpha), \quad \beta \sim \text{Gamma}(\alpha_\beta, \text{rate} = \beta_\beta) \tag{9}$$

where the hyperparameters $(\alpha_\mu, \beta_\mu, \alpha_\alpha, \beta_\alpha, \alpha_\beta, \beta_\beta)$ are fixed and chosen for the given application. Two variations of the Hawkes process model appear in this paper. Model 1 is the temporal process with no marks, i.e. $\gamma_I(\cdot) \propto 1$, $\gamma_O(\cdot) \propto 1$.

Model 2 assumes spatial marks $\kappa = (\kappa_x, \kappa_y)$ where (κ_x, κ_y) denotes the geolocation of the event, (i.e., a spatio-temporal process).

The immigrant mark density is uniform across the spatial domain $\gamma_I(\cdot) \propto 1$. The offspring mark density is assumed Gaussian centered on the location of the parents.

$$\gamma_O(\kappa) = \frac{1}{2\pi\sigma^2} \exp\left\{ \frac{-\|\kappa - \kappa_{pa}\|_2^2}{2\sigma^2} \right\}, \tag{10}$$

i.e., marks of offspring have spatial Gaussian decay. An inverse gamma prior is assumed for σ^2, i.e. $\sigma^2 \sim \text{IG}(\alpha_\sigma, \beta_\sigma)$. This model is similar, though simpler in form to the ETAS model of Ogata (1981) where the mark density is of a more complicated form. A third variation of the model modifies the immigrant mark density in model 2 and is described in the application Section 7. Finally, since the

branching structure is a latent parameter, a prior for it must be specified. Throughout, a discrete uniform prior is assumed on the feasible domain of each Y_i.

The relative simplicity of these models is chosen to concentrate on demonstrating the missing data framework for this class of point-processes, the main contribution of this work. Most parametric alternatives of $\mu(t)$, $g(t)$, $\gamma_I(\kappa|t)$, and $\gamma_O(\kappa|t, \{t_{pa}, \kappa_{pa}\})$, are straightforward to implement in our framework. Non-parametric forms, specifically for $\mu(t)$, could also be considered although not explored in this paper. We present a further discussion on this topic in Section 8.

5. MCMC algorithm

Here we describe the MCMC algorithms used to account for missing data in the models described in Section 4. Details are given for sampling from the complete data posterior $p(\phi|x)$ and full conditional of missing data, $p(x_{miss}|\phi, x_{obs})$.

5.1. MCMC for complete data models

The method of Rasmussen (2013) is adapted to sample from the complete data posterior, $p(\phi|x)$. For the models described above, the parameters μ, α, β, and the branching structure Y are sequentially sampled from their full-conditional distributions under the likelihood (8) and priors assumed. Variations to the data model and priors will result in variations to the derivations described here. The priors for μ and α and σ^2 are conditionally conjugate and sampled directly from:

$$\mu|x, \phi, Y \sim \text{Gamma}(\alpha_\mu + |S_0|,\ \text{rate} = \beta_\mu + T),$$

$$\alpha|x, \phi, Y \sim \text{Gamma}(\alpha_\alpha + \sum_{i}^{n} |S_i|,\ \text{rate} = \beta_\alpha + \sum_{i=1}^{n} G(T - t_i)),$$

and

$$\sigma^2|x, \phi, Y \sim \text{Inverse-Gamma}\left(\alpha_\sigma + \sum_{i}^{n} |S_i|,\ \beta_\sigma + \sum_{i=1}^{n}\sum_{j\in S_i} \|\kappa_i - \kappa_j\|_2^2\right).$$

The parameter β (in model 2) is sampled (approximately) from its full-conditional using a random-walk Metropolis algorithm. Specifically, the full conditional for β is of the form

$$p(\beta|x, \phi, Y) \propto \pi(\beta)\prod_{i=1}^{n} \exp(-\alpha G(T - t_i|\beta))\prod_{t_j\in S_i} g(t_i - t_j|\beta),$$

resulting in the random-walk Metropolis ratio

$$H_\beta = \frac{\pi(\tilde{\beta})\prod_{i=1}^{n} \exp(-\alpha G(T - t_i|\tilde{\beta}))\prod_{t_j\in S_i} g(t_i - t_j|\tilde{\beta})}{\pi(\beta)\prod_{i=1}^{n} \exp(-\alpha G(T - t_i|\beta))\alpha^{|S_i|}},$$

where $\pi(\beta)$ is the prior distribution, β denotes the current value, and $\tilde{\beta}$ is the proposed value sampled from a $\mathcal{N}(\beta, s_\beta^2)$. s_β^2 is a user-specified parameter tuned to have an acceptance ratio between 0.2 and 0.4. Additional parameters described in the third variation of the model presented in Section 7 are sampled in a similar fashion.

Finally, the branching structure Y can be sampled directly, and element-wise, from the full-conditional $p(Y|x, \phi)$ using the method presented in Ross (2016), which was first proposed by Zhuang et al. (2002) under the context of stochastic declustering. Note that since offspring points can only be triggered by parent points that have occurred previously in the time history, the jth element of Y

can only take integer values in the range $[0, j-1]$. With a discrete uniform prior, the full conditional probability of the jth element of Y is:

$$p(Y_i = j|x, \phi) = \begin{cases} \dfrac{\mu}{\lambda(t_i)} & \text{if } j = 0 \\ \dfrac{\alpha g(t_j - t_i)}{\lambda(t_i)} & \text{if } j \in 1, 2, \ldots, i-1. \end{cases}$$

By conditioning on Y, we reduce the amount of points to be processed for each parameter. For μ, we only use the identified background points and for α and β, we only use the identified children points. This greatly reduces the computational cost as n gets large.

5.2. Missing data model

As described in Section 3, to sample the posterior $p(\phi|x_{obs})$, we can augment the complete data MCMC algorithm with a step to sample the missing points. To explain, assume first that there is only one missing interval ($K = 1$), say $M_1 = [T_1, T_2]$ with $0 \leq T_1 < T_2 \leq T$. Sampling directly from $p(x_{miss}|\phi, x_{obs})$ is difficult because x_{obs} includes future data occurring after time T_2 is part of the conditioning. One can imagine if there is a cluster of points just after T_2 there is likely a cluster of (unobserved) points just before T_2; it is hard to derive the likelihood of such a cluster. Hence we turn to Metropolis–Hasting. For the proposal distribution, we recognize that it is easy to simulate the future of a point process starting at time T_1 conditioning only on the past. This suggests a proposal distribution that conditions only observed data up to time T_1 and ignores observed data after T_2. By conditioning on the past, we will at least capture part of the structure expected in the missing data and will be more likely to propose plausible values of x_{miss} as compared to simulating independently all of the observed data.

For further detail, let x_{miss} and $\tilde{x}_{miss}$ be the current and proposed set of missing data respectively, and set $x = (x_{miss}, x_{obs})$ and $\tilde{x} = (\tilde{x}_{miss}, x_{obs})$. Let x_{T_j} and $\tilde{x}_{T_j}$ be the subsets of x and $\tilde{x}$ including only points up to time T_j, $j = 1, 2$, respectively. The full-conditional (i.e., the target distribution of the Metropolis–Hastings algorithm) is $p(x_{miss}|\phi, x_{obs}) = \frac{p(x|\phi)}{p(x_{obs}|\phi)}$. The numerator is the likelihood (4) and the denominator will cancel in the Metropolis–Hastings ratio given below.

The proposal conditions on all the observed data up to time T_1 as well as the current values of ϕ. Data on the missing interval M_1 is proposed from $p(x_{miss}|\phi, x_{T_1}) = \frac{p(x_{miss}, x_{T_1}|\phi)}{p(x_{T_1}|\phi)}$. Recognizing that $(x_{miss}, x_{T_1}) = x_{T_2}$ we see the numerator is $p(x_{T_2}|\phi)$, the likelihood (4) up to time T_2. As with the target distribution, the denominator will cancel in the Metropolis–Hastings ratio given below.

To simulate from the proposal, we can adjust strategies to simulate from a Hawkes process to account for conditioning on the history. Algorithm 1 (adapted from Algorithm C in Zhuang et al. (2004)) describes one such method utilized in this paper. Step 1 is a general first step to generate immigrant points of a Hawkes process on the interval $[t_{start}, t_{end}]$. For the proposal, $t_{start} = T_1$ and $t_{end} = T_2$ and conditioning on the history before T_1 is handled in Step 2 by adding the history to the points generated in Step 1. If there is no history, Step 2 is ignored and Algorithm 1 generates a Hawkes process on $[t_{start}, t_{end}]$. To generate the marks, the mark distributions γ_I and γ_O are used depending on if the point generated is an immigrant or offspring. Other options for simulating Hawkes process data include, the thinning method developed by Ogata (1981), the perfect simulation algorithm of Møller and Rasmussen (2005), or a faster approximation of perfect simulation developed by Møller and Rasmussen (2006).

Cancellation of the proportionality constants in the full-conditional and the proposal allows for evaluation of the Metropolis–Hasting ratio:

$$H_t = \frac{p(\tilde{x}_{miss}|\phi, x_{obs})}{p(x_{miss}|\phi, x_{obs})} \frac{p(x_{miss}|\phi, x_{T_1})}{p(\tilde{x}_{miss}|\phi, x_{T_1})} = \frac{p(\tilde{x}|\phi)}{p(x|\phi)} \frac{p(x_{T_2}|\phi)}{p(\tilde{x}_{T_2}|\phi)}. \tag{11}$$

x_{miss} is set to $\tilde{x}_{miss}$ with probability $\min(H_t, 1)$. Otherwise, x_{miss} is unchanged.

In the case of multiple missing time intervals, $K > 1$, the method is applied iteratively to each interval in separate Gibbs steps. For example, consider the case of $K = 2$, $M_1 = [T_1, T_2]$ and

Algorithm 1 Simulate Marked Hawkes Process

1: Generate a set of immigrant points as a Poisson process in time interval $[t_{start}, t_{end}]$ over spatial region W with background intensity $\mu(t)$ and mark density $\gamma_I(\kappa|t)$ and record as Generation 0, $G^{(0)}$. The process can be either homogeneous or in-homogeneous.
2: If conditional history exists prior to t_{start}, add to $G^{(0)}$
3: Set $l = 0$
4: **while** $G^{(l)}$ is not empty **do**
5: For each event i, (t_i, κ_i), in the catalog $G^{(l)}$, simulate its $N_i^{(l)}$ offspring, $O^{(l)} = \{(t_k^{(i)}, \kappa_k^{(i)}) : i = 1, 2, \ldots, N_i^{(l)}\}$, where $N_i^{(l)}$ is a Poisson random variable with mean α. The quantities $t_k^{(i)}$ and $\kappa_k^{(i)}$ are generated from probability densities $\beta(t - t_i)$ and $\gamma_O(\kappa_i|\kappa_{pa})$, respectively.
6: Set $G^{(l+1)} = \cup_{i \in G^{(l)}} O_i^{(l)}$
7: Set $l = l + 1$
8: Remove all events whose $t_i \notin [t_{start}, t_{end}]$ from $G^{(l+1)}$
9: **end while**
10: Combine all events $X = \cup_{j=0}^{l} G^{(l+1)}$
11: Remove any events not in W, and return X

$M_2 = [T_3, T_4]$ where $0 < T_1 < T_2 < T_3 < T_4 < T$. Missing data $x_{miss_1} \in M_1$ is sampled from $p(x_{miss_1}|\phi, x_{obs}, x_{miss_2})$ the same way as above with x_{obs} augmented with $x_{miss_2} \in M_2$. x_{miss_2} is similarly sampled. This procedure is trivially extended to $K > 2$.

When simulating a Hawkes process this way, there are practical issues with spatial boundary conditions, also referred to as edge-effects. If no parent events are allowed to occur outside the spatial region W, then the offspring that would have been generated by these parents will be unobserved. We overcome this problem by simulating on a region W' that is larger than W and sufficiently large enough to capture the offspring. We then remove the events generated outside W at the end of the simulation (step 11 of Algorithm 1). Specifically for our model, we increased the region W on all sides by the length that would include 95% of the children produced by parents on the boundary. For a further discussion on how to handle edge-effects see Diggle (2014).

6. Simulated data

We first demonstrate the method on simulated data from both the temporal and spatio-temporal models, models 1 and 2 respectively, described in the previous section. For each setting, we compare parameter estimates for three different data set/model pairs over many simulation:

1. Using the complete data and fitting the complete data model, referred to as the complete/complete pair
2. Removing a significant time interval from the complete data and fitting the complete data model. The data set with the time interval removed is referred to as the 'incomplete data set'. Likewise, this pair is referred to as the incomplete/complete pair.
3. Fitting the incomplete data with the missing data model, referred to as the incomplete/missing pair.

The first data/model pair is a baseline — fitting the correct model to the complete data set. The second simply ignores that there is missing data and treats the available data as complete when fitting the complete-data model. The third correctly accounts for the missing data. The purpose of these simulated data sets is to test model performance, giving a baseline of performance that we can use when we are analyzing real data. Results for the temporal model (model 1) and the spatio-temporal model (model 2) are presented separately. To simulate the data in the following sections we use Algorithm 1.

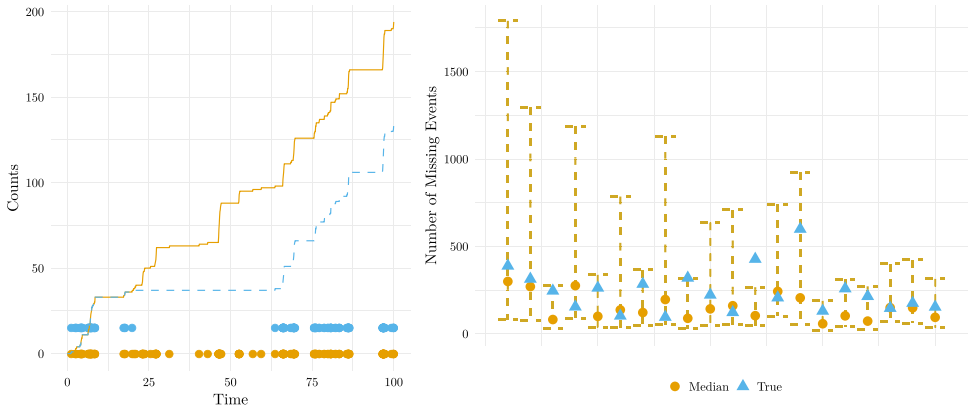

Fig. 1. Left: Observed time points for complete data (orange) and incomplete data (blue) for a single simulation. Right: 90% credible intervals of the number of events occurring in the missing interval for 20 random temporal simulations . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

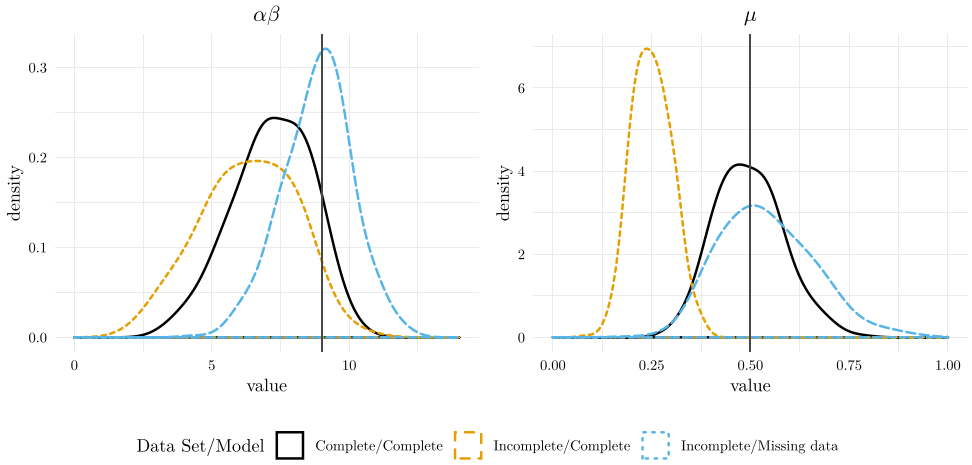

Fig. 2. Distribution of posterior means for 500 simulations of temporal data (model 1) for three data/model pairs: complete/complete (black), incomplete/complete (orange), incomplete/missing (blue).

6.1. Temporal simulation

We simulate 500 data sets from model 1 using parameters $\mu = .5, \alpha = 0.9$ and $\beta = 10$ on the time interval $[0, 100]$. For the incomplete data, the interval $[20, 60]$ is removed. The left panel of Fig. 1 displays example times and cumulative counts of the number of events for both the complete data set (green) and incomplete data set (red).

Each of the data/model pairs is fit to each simulated data set using the algorithm described in Section 5. The priors are $\mu \sim \text{Gamma}(\alpha_\mu = 1,\ \beta_\mu = 0.01)$, $\alpha \sim \text{Gamma}(\alpha_\alpha = 1,\ \beta_\alpha = 0.1)$ and $\beta \sim \text{Gamma}(\alpha_\beta = 0.1,\ \beta_\beta = 0.1)$ with proposal variance $s^2_\beta = 1$. Fig. 2 shows the distribution of posterior means for the parameters $\alpha\beta$ and μ for each of the 500 simulations and data/model pairs. The vertical gray line is the true parameter value. The estimate of $\alpha\beta$ is given because Zhang (2011) shows this to be a more statistically consistent estimator than α and β separately. As expected, using the complete/complete pair results in good estimates on average for both μ and $\alpha\beta$. The missing data has the most effect on the estimation of μ. When using the missing data model (incomplete/missing),

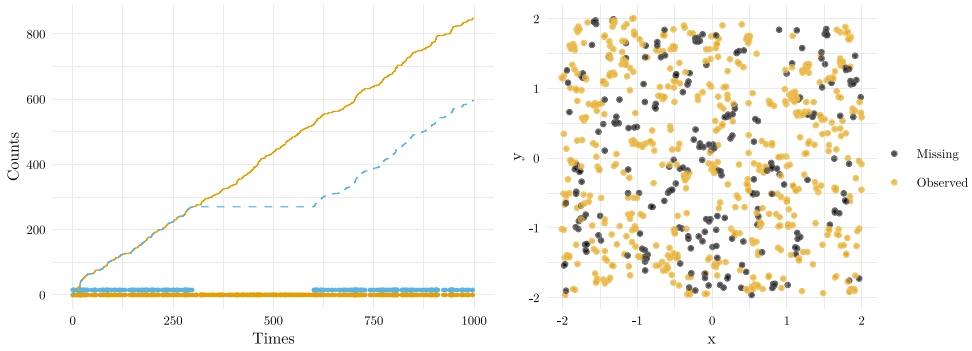

Fig. 3. Left: Observed time points for complete data (orange) and incomplete data (blue) for a single simulation from model 2. Right: Spatial locations with color indicating the point is missing (black) or observed (orange) . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the estimates of μ are good on average and more variable than when the complete data set is available (complete/complete); reflecting the additional uncertainty induced by the missing data. This parameter is drastically under-estimated when the missing data is ignored (incomplete/complete). Unlike μ, all models seem to estimate $\alpha\beta$ with similar uncertainty. Investigating α and β separately shows that the estimates of α are slightly more variable when under the missing data model, but the estimates of β are slightly less variable. This makes sense as incorporating the missing data within the MCMC as described in Section 5 has the most impact on the parent and offspring intensities, driven by μ and α respectively, and may or may not affect the estimation of β.

We can also evaluate the missing data model in terms of its ability to predict data within the missing interval. The right panel of Fig. 1 displays the number of missing events along with the medians and 90% credible intervals of the posterior distribution of number of missing events for 20 random simulations. The posteriors tend to be positively skewed and the empirical coverage over all 500 simulations is 88%. Overall, the missing data model performs well for parameter estimation and prediction.

6.2. Spatio-temporal simulation

For the spatio-temporal model (model 2), we generate 500 data sets on the time interval $[0, 1000]$ with spatial domain $W = [-2, 2]x[-2, 2]$ and fixed parameter values: $\mu = .5$, $\alpha = 0.4$, $\beta = 1$ and $\sigma^2 = 0.001$. The incomplete data is comprised by removing the interval $[300, 600]$. The left panel of Fig. 3 shows the times for a single simulated data set along with the cumulative number of events for both the complete and incomplete data. The right panel shows corresponding spatial locations colored to indicate if the point was missing (within the time interval $[300, 600]$) or observed (outside the time interval $[300, 600]$).

As with the temporal case, the three data/model pairs are fitted assuming the priors $\mu \sim \text{Gamma}(\alpha_\mu = 1,\ \beta_\mu = 0.01)$, $\alpha \sim \text{Gamma}(\alpha_\alpha = 1,\ \beta_\alpha = 0.01)$, $\beta \sim \text{Gamma}(\alpha_\beta = 0.01,\ \beta_\beta = 0.1)$ and mark prior $\sigma^2 \sim \text{IG}(\alpha_\sigma = 1,\ \beta_\sigma = 0.1)$ with proposal variances $s^2_\beta = 0.5$ and $s^2_\sigma = 0.0003$.

Fig. 4 shows the distribution of posterior means for μ, $\alpha\beta$, and σ^2 over the 500 simulations for each data/model pair. The true parameter value is indicated by the gray vertical line. Again we see that using the complete data set with the correct model (complete/complete) results in good estimates on average. The effect of accounting for the missing data is most drastic on the mean parameter μ. Without accounting for the missing data (incomplete/complete) the parameter is underestimated because the complete model incorrectly assumes no events occur in the missing interval. Further, the variance in the posterior means for each parameter is larger when accounting for the missing data, reflecting the additional uncertainty induced by the missing data.

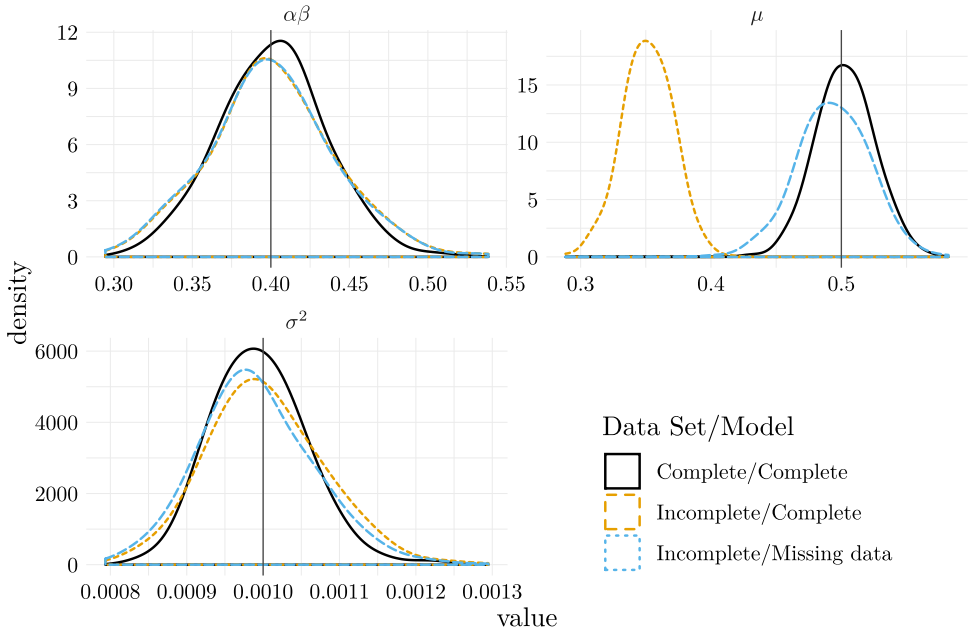

Fig. 4. Distributions of posterior parameter means of 500 random spatio-temporal simulations on the complete simulated data (black), the incomplete simulated data set (orange), and the incomplete simulated data set when accounting for missing data (blue).

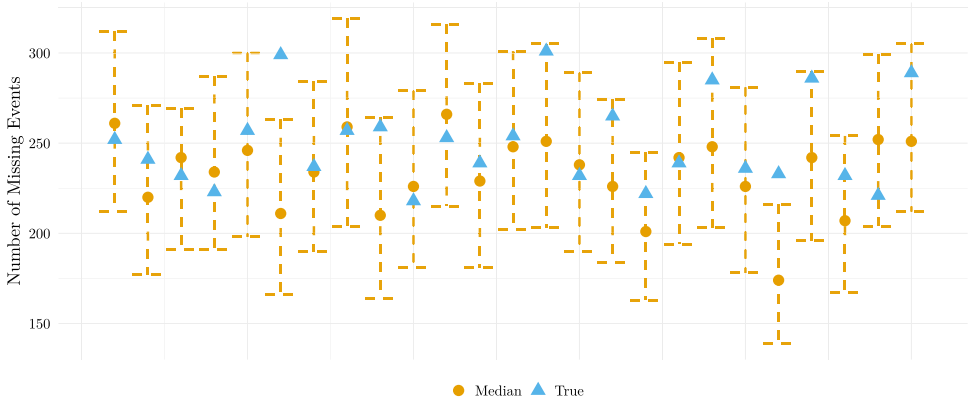

Fig. 5. 90% credible intervals of the number of events occurring in the missing interval for 25 random space–time simulations.

As with the temporal case, we can again evaluate the model based on its ability to predict the number of events in the missing time interval, which now have spatially-dependent marks. Fig. 5 displays the true number of missing events along with the medians and 90% credible intervals of the posterior distribution of number of missing events for 25 random simulations.

The empirical coverage over all 500 simulations is 84%, slightly smaller and more symmetric than the temporal simulation empirical coverage. In contrast to the temporal case, there is actually a slight tendency for the missing data model to underestimate the number of missing events when the data has spatial marks. This is likely due to the removal of points outside the spatial window when handling

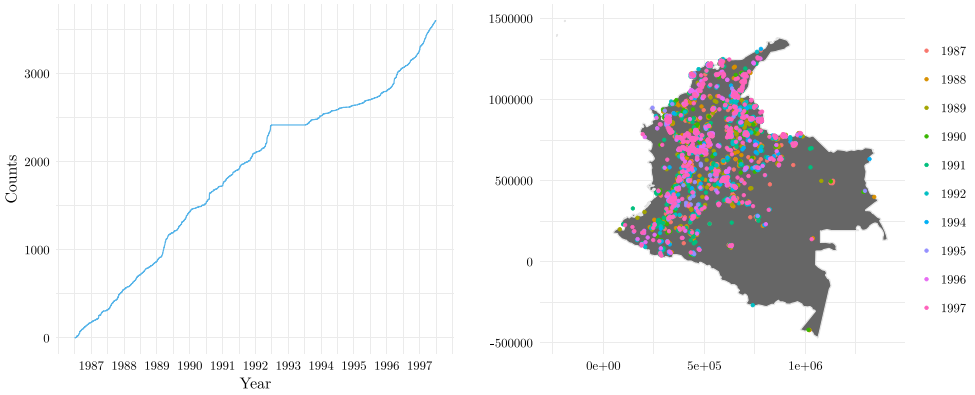

Fig. 6. Jittered arrival times (left) and spatial locations by year of observed event in Colombia between 1987 and 1997. The time is in days from the start of the record (January 1, 1987). The locations are given in terms of UTM coordinates (meters).

boundary effects in addition to removing those outside the missing time interval when proposing $\tilde{x}_{miss}$ as described in Section 5.2.

7. Global terrorism database

We apply the described missing data method on data taken from the Global Terrorism Database (GTD), an open-source database containing information on terrorism events around the world from 1970–2015. The database is systemic and has more than 150,000 cases with multiple types of event information. Additionally, the database is missing records for the entire year of 1993 due to the loss of hard-copy index cards on which the events were recorded before the data was fully digitized. For our study we consider the events in Colombia between 1987–1997. Colombia has an interesting history in terms of terrorism events during this time in which it dealt with multiple problems with guerrillas, paramilitaries, and narcotics trafficking. Fig. 6 gives the cumulative count of number of observed events as a function of time in days for the Colombia data along with the recorded spatial locations of the events. The flat line over 1993 reflects the period of known missing records. The time scale is the number of days since January 1, 1987 and the spatial resolution is also coarse due to imprecise location information, a common issue with tracking data. This results in multiple events recorded on the same day and location. To adjust for this, we minimally jitter the events in time and space prior to model fitting. This is done by specifying a uniform $U(0, 0.5)$ for the temporal displacement (in days) and a $U(-10, 10)$ for the spatial displacement (in km) of both κ_x and κ_y. The spatial jittering is expected to minimally impact the estimation of σ^2 in $\gamma_0(\kappa)$.

The data plotted in Fig. 6 indicate the spatial marks to be non-uniformly distributed across Colombia. Thus, specifying a uniform distribution for the spatial marks of the parents as was done for the simulated data in Section 6.2 does not make sense for Colombia. A modification to the spatio-temporal model (model 2) is made by specifying a bi-variate normal distribution for the event locations of the immigrants:

$$\gamma_I(\kappa|t) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right].$$

Note that since a bivariate normal distribution is specified for γ_I and $\mathbb{M}$ is restricted to the region of Colombia, $\int_{\mathbb{M}} \gamma_I < 1$ and the exact likelihood in this case takes the form of (3). Although when $\int_{\mathbb{M}} \gamma_I < 1$ is difficult to compute, as is the case here, (4) is an appropriate approximation (Schoenberg, 2013; Reinhart, 2018). The prior distributions for the additional parameters are specified to be

$$\mu_x \sim \mathrm{N}(m_x,\ s_x^2),\ \ \mu_y \sim \mathrm{N}(m_x, s_x^2),\ \sigma_x^2 \sim \mathrm{IG}(\alpha_{\sigma_x},\ \beta_{\sigma_x}),\ \sigma_y^2 \sim \mathrm{IG}(\alpha_{\sigma_y},\ \beta_{\sigma_y}).$$

Sampling steps for these parameters are easily added to the MCMC algorithm of Section 5.

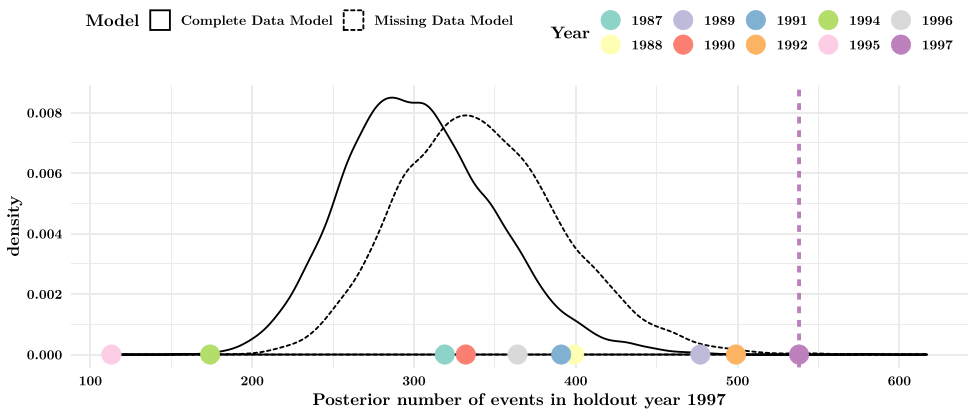

Fig. 7. Posterior predictive distribution of the number of events in 1997 under the complete data model and missing data model.

On this real dataset, we examine the effects of explicitly modeling the missing data. We fit each model to the data from 1987 to 1996, using the events in 1997 as a holdout set. The complete data model simply ignores the fact that records are missing in 1993. In principle, this is the wrong approach. The goal here is to understand how accounting for the known missing data interval over 1993 can affect the prediction of future events.

To examine the effects, we evaluate the posterior predictive distribution of 1997:

$$p(\tilde{x}|x_{obs}) = \int p(\tilde{x}|\phi, x_{obs})p(\phi|x_{obs})d\phi \tag{12}$$

where $\tilde{x}$ is predicted data (times and spatial locations) in 1997, x_{obs} is the observed data from 1987 to 1996, ϕ are the model parameters, $p(\phi|x_{obs})$ is the posterior under the model, and $p(\tilde{x}|\phi, x_{obs})$ is the distribution of 1997 data given the model, ϕ, and x_{obs}. The predictive distribution (12) is evaluated by iteratively sampling the posterior and $p(\tilde{x}|\phi, x_{obs})$. The events from the predicted year 1997 are sampled in the same fashion as the proposal for missing data in Section 5.2. Under the missing data model, the predictive distribution is also integrated over the missing data (i.e., $p(\tilde{x}|x_{obs}) = \int p(\tilde{x}|\phi, x_{obs}, x_{miss})p(\phi, x_{miss}|x_{obs})d\phi dx_{miss}$).

First, the posterior predictive distribution of the total number of events in 1997 is considered by simply counting the number of events for each posterior predictive sample. The distributions under the complete and missing data models appear in Fig. 7 along with the total number of events for each year plotted on the x-axis. The predictive distribution of the missing data model is shifted right, with an average of about 40 events higher than under the complete data model. More events are predicted because the missing data model does not incorrectly assume no events occur in 1993. Additionally, the variance in the predictive distribution is larger, reflecting the uncertainty induced by the missing data. By incorrectly assuming no events occur in 1993, the variance of the posterior predictive distribution is (unjustifiably) smaller.

Both models under predict the true number of events in 1997 of 538. This is due to the fact that the number of recorded events in 1997 is large compared to the other years making it difficult to predict. For example, the previous three years have 364, 113, and 174 total events which are all much lower than the total in 1997. For a more holistic comparison between models that does not just concentrate on one summary statistic (i.e., the number of predicted events) we seek a proper scoring rule (Daley and Vere-Jones, 2003). For this we consider the log-likelihood of the observed 1997 data. This takes the form: $\log p(\tilde{x}|\phi, x_{obs})$ for the complete data model and $\log p(\tilde{x}|\phi, x_{obs}, x_{miss})$ for the missing data model where $\tilde{x}$ is the fixed 1997 data. These are just the first factor in the integrand of the posterior predictive distribution (11). Density estimates of the posterior distributions of these log-likelihoods under each model appear in Fig. 8. Overall, the log-likelihood of the missing data model is slightly

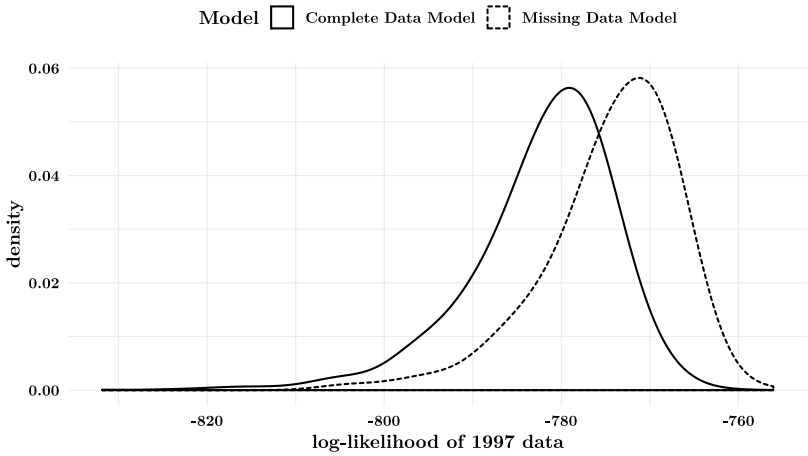

Fig. 8. Posterior distribution of log-likelihoods of the holdout data under each model. For the complete data model the posterior of $\log p(\tilde{x}|\phi, x_{obs})$ is shown where $\tilde{x}$ is the fixed 1997 holdout data. For the missing data model the posterior of $\log p(\tilde{x}|\phi, x_{obs}, x_{miss})$ is shown. The expected values for the missing data and complete data models are -774 and -782, respectively.

larger (better) than that of the complete data model with respect to predicting the holdout data. The expected log likelihood values are -774, and -782 for the missing data and complete data models, respectively.

Next, the spatial distribution of predicted events for 1997 is evaluated using nonparametric estimates of the spatial intensity. For each sample from the predictive distribution (12), the spatial intensity was estimated using the fixed-bandwidth kernel estimates of Diggle (1985) and implemented in the R function `spatstat::density.ppp` (Baddeley et al., 2015). The mean (top row) and standard deviation (bottom) of these estimates appear in Fig. 9 for both models. The black points represent the observed event locations in 1997. The intensity is scaled by the area of Columbia (A_W). Roughly speaking, the scaled intensity is the expected number of events in 1997 at a location times A_W. As seen in the plots, accounting for the missing data results in a larger mean estimate of the intensity across the spatial domain, particularly in regions where events are more dense. The standard deviation is also larger, again reflecting the additional uncertainty induced by the missing data.

While accounting for the missing data results in a spatial intensity estimate that better reflects the data for 1997, there are arguably improvements that could be made. For example, several events are observed in the southwest region of Colombia that do not appear to be properly captured by the predictions. One potential improvement to the model would be to include spatial information in the assumed intensity (7) by, for example, scaling it by population information. The idea being that recorded events are more likely in more populated areas. Additionally, a nonparametric form for γ_l could be specified to make the intensity more flexible and spatial mark distribution more reasonably fit the data. Such refinements of the model can theoretically be incorporated into the missing data model framework developed here, if deemed necessary for the application.

8. Conclusions and future work

It is clearly important to account for missing data mechanisms in statistical models. This paper concentrates accounting for missingness on Hawkes process models where the missing data is in the form of known intervals of time where data is not being observed. A natural Bayesian approach to the problem is taken by treating the missing events as latent parameters augmented the MCMC algorithm for the complete data model with a step to impute the missing data. While the general data augmentation approach is standard, the method for data-imputation required the development of an

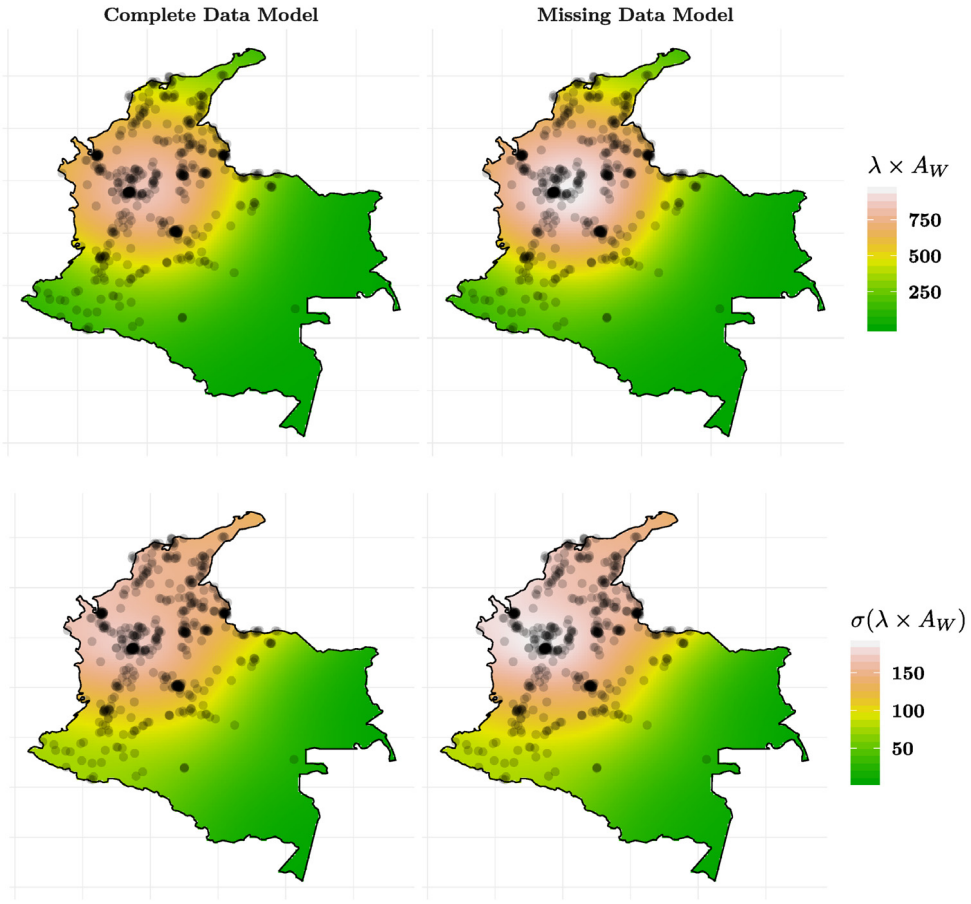

Fig. 9. Mean (top row) and standard deviation (bottom row) of intensity predictions for 1997 using observed data from 1987 to 1996 based on the complete data model (left column) and missing data model (right column). The black points represent the observed event locations in 1997. The intensity was scaled to the area of Columbia, A_W.

efficient proposal distribution to use in a Metropolis–Hasting steps. The efficiency of the proposal rests on the fact that it conditions on the data prior to the missing interval in question, rather than ignoring this information. This proposal was chosen because we could readily sample from it and because its distribution is similar to that of the target distribution. It should also be clarified that the proposed MCMC algorithm for the missing data model Section 5.2 is an approximation to an exact solution. It was suggested by a reviewer that it may be more exact to implement importance sampling instead of the proposed MCMC algorithm. This could be explored in future research.

Other future work could include extending our model to incorporate missing spatial regions in addition to the missing time intervals handled in this paper. Further, the missing data approach described is not limited to the parametric model forms presented here. For example, in many applications it is desirable to specify more flexible, nonparametric forms for the intensity function. For example, Fox et al. (2016) and Marsan and Lengliné (2008) propose novel ways to incorporate a nonparametric inhomogeneous background rate $\mu(t)$. Kernel density estimation can be used to model the decay process or triggering function $g(\cdot)$, for which estimation procedures have been explored by Zhou et al. (2013), Kirchner and Bercher (2018) and Yang et al. (2018) just to name a few. Extending the framework to incorporate such model refinements should be achievable.

Acknowledgments

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. The authors would like to thank Dr. Katherine Simonson (Sandia National Laboratories) for her technical support, John Rowe (Sandia National Laboratories) for his programmatic support during this work, Jonathon Lane for his helpful discussions early on and Stephen Rowe (Sandia National Laboratories) for helping with software development by making the code immensely more efficient.

References

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press.

Balderama, E., Schoenberg, F., Murray, E., Rundel, P., 2012. Application of branching point process models to the study of invasive red banana plants in costa rica. J. Amer. Statist. Assoc. 107 (498), 467–476.

Chen, F., Hall, P., 2016. Nonparametric estimation for self-exciting point processes – a parsimonious approach. J. Comput. Graph. Statist. 25 (1), 209–224.

Daley, D., Vere-Jones, D., 2003. An Introduction to the Theory of Point Processes, second ed. New York:Springer.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum elikelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (7), 1–38.

Diebolt, J., Robert, C.P., 1994. Estimation of finite mixture distributions through bayesian sampling. J. R. Stat. Soc. Ser. B Stat. Methodol. 56 (2), 363–375.

Diggle, P.J., 1985. A kernel method for smoothing point process data. J. R. Stat. Soc. Ser. C. 34, 138–147.

Diggle, P.J., 2014. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. CRC Press.

Fox, E.W., Schoenberg, F.P., Gordon, J.S., 2016. Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric hawkes point process models of earthquake occurrences. Ann. Appl. Stat. 10 (3), 1725–1756.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., Smith, A.F., 1990. Illustration of bayesian inference in normal data models using gibbs sampling. J. Amer. Statist. Assoc. 85 (412), 972–985.

Gelfand, A.E., f. M. Smith, A., 1990. Sampling-Based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 88, 398–409.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6 (6), 721–741.

Global Terrorism Database, 2017. National consortium for the study of terrorism and responses to terrorism (START), ULR https://www.start.umd.edu/gtd.

Hawkes, A.G., Oakes, D., 1974. A cluster representation of a self-exciting process. J. Appl. Probab. 11 (3), 493–503.

Kirchner, M., Bercher, A., 2018. A nonparametric estimation procedure for the hawkes process: comparison with maximum likelihood estimation. J. Stat. Comput. Simul. 88 (6), 1106–1116.

Kong, A., Liu, J.S., Wong, W.H., 1994. Sequential imputations and bayesian missing data problems. J. Amer. Statist. Assoc. 89 (425), 278–288.

Little, R.J., Rubin, D.B., 2014. Statistical Analysis with Missing Data. John Wiley & Sons.

Marsan, D., Lengliné, O., 2008. Extending earthquake' reach through cascading. Science 319 (2008), 1076.

McLachlan, G., Peel, D., 2004. Finite Mixture Models. John Wiley & Sons.

Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., Tita, G.E., 2011. Self-exciting point process modeling of crime. J. Amer. Statist. Assoc. 106 (493), 100–108.

Møller, J., Rasmussen, J.G., 2005. Perfect simulation of Hawkes processes. Adv. Appl. Probab. 37 (3), 629–646.

Møller, J., Rasmussen, J.G., 2006. Approximate simulation of Hawkes processes. Methodol. Comput. Appl. Probab. 8 (1), 53–64.

Ogata, Y., 1981. On Lewis' simulation method for point processes. IEEE Trans. Inform. Theory 27 (1), 23–31.

Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. J. Amer. Statist. Assoc. 83 (401), 9–27.

Ogata, Y., 1998. Space-time point-process models for earthquake occurrences. Ann. Inst. Statist. Math. 50 (2), 379–402.

Ogata, Y., Zhuang, J., 2006. Space-time ETAS models and an improved extension. Tectonophysics 413 (1–2), 13–23.

Rasmussen, J.G., 2013. Bayesian inference for hawkes processes. Methodol. Comput. Appl. Probab. 15 (3), 623–642.

Reinhart, A., 2018. A review of self-exciting spatio-temporal point processes and their applications. Statist. Sci. 33 (3), 299–318.

Rizoiu, M., Lee, Y., Mishra, S., Xie, L., 2018. Hawkes processes for events in social media. In: Frontiers of Multimedia Research. Association for Computing Machinery and Morgan & Claypool, pp. 191–218.

Robert, C., Casella, G., 2004. Monte Carlo Statistical Methods, second ed. Springer-Verlag New York.

Ross, G.J., 2016. Bayesian estimation of the etas model for earthquake occurrences, URL http://www.gordonjross.co.uk/bayesianetas.pdf.

Schoenberg, F.P., 2013. Facilitated estimation of etas. Bull. Seismol. Soc. Am. 103 (1), 601–605.
Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. J. Am. Statist. Assoc. 82 (398), 528–540.
Veen, A., Schoenberg, F.P., 2008. Estimation of space–time branching process models in seismology using an EM-type algorithm. J. Amer. Statist. Assoc. 103 (482), 614–624.
Yang, Y., Etesami, J., He, N., Kiyavash, N., 2018. Nonparametric hawkes processes: online estimation and generalization bounds, arXiv:1801.08273.
Zhang, H., 2011. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J. Amer. Statist. Assoc. 63 (465), 250–261.
Zhou, K., Zha, H., Song, L., 2013. Learning triggering kernels for multi-dimensional hawkes processes. In: Proceedings of the 30th International Conference on Machine Learning. pp. 1301–1309.
Zhuang, J., Ogata, Y., Vere-Jones, D., 2002. Stochastic declustering of space-time earthquake occurrences. J. Amer. Statist. Assoc. 97 (458), 369–380.
Zhuang, J., Ogata, Y., Vere-Jones, D., 2004. Analyzing earthquake clustering features by using stochastic reconstruction. J. Geophys. Res.: Solid Earth 109 (B05301).
Zhuang, J., Werner, M.J., Harte, D., 2013. Stability of the earthquake clustering models: criticality and branching ratios. Phys. Rev. E 88 (6), 062109.
Zipkin, J.R., Schoenberg, F.P., Coronges, K., Bertozzi, A.L., 2016. Point-process models of social network interactions: Parameter estimation and missing data recovery. European J. Appl. Math. 27 (3), 502–529.