Bayesian Estimation of Modified ETAS Model for Invasive Species

Binhui Deng Earvin Balderama

January 31, 2018

Abstract

The Epidemic-Type Aftershock Sequence (ETAS) model, a Hawkes (1971) self-exciting point process model, is commonly used for characterizing earthquake and aftershock activity. Red banana trees, an invasive plant in Costa Rican rainforest, have been being observed for recent few years. Adjusted ETAS model was applied onto the study of red banana tree to characterize it's spatial-temporal spreading patterns. This project proposed Bayesian Markov Chain Monte Carlo method to estimate unknown parameters in the adjusted ETAS model.

Keywords: epidemic type aftershock sequence models, bayesian, MCMC, invasive species

1 Introduction

Invasive species is a subset of introduced species. It would be defined as an invasive species if a species is artificially introduced into a region that it has not previously existed and has the capacity to grow into a certain amount in the absence of human intervention, then threaten local biodiversity and become a local hazard.

Invasive alien species of plants can also destroy the habitat of native species and disturb the natural evolution that takes place in the environment in which they spread. Many invasive plant species have been studied in the past (Higgins and Richardson 1996; Delisle et al. 2003). Too many of these studies simply report statistics based on amount of land consumed over some period of time. Of these, most do not report precise locations where these plants are spreading. There are a couple of reasons why this is the case: 1) Computational methods are either too simple or too difficult, and 2) Data on the invasive species do not contain exact locations of each individual plant. Even when trying to analyze the current and future locations of invasive plants, the models do not come from point process methodology. Instead, studies often use grid-based methods, where the surface of study is divided into an array of pixels on a grid.

The natural process by which plants can spread its seeds initiates the conversation of statistically analyzing the spatial and temporal spread of invasive plants using the theory of point processes. According to the former research done by Dr. Earvin Balderama, the estimates obtained by Newton-Rhapson numerical optimization were $\hat{\boldsymbol{\theta}} = \{\hat{\alpha} = 0.0760, \hat{\beta} = 0.0292, \hat{p} = 0.5770\}$, with corresponding asymptotic standard errors 0.0045, 0.0022 and 0.0190, respectively.

The Bayesian MCMC algorithm was of interest to retrospectively examine if more accurate estimates could be obtained and if the estimates are robust.

2 Data

The data studied in Balderama et al. (2012) contained the precise longitude and latitude coordinates, captured by Global Position System (GPS), of red banana trees that has spread through a significant part of a rainforest in Costa Rica. In addition to location, the height (in cm) and

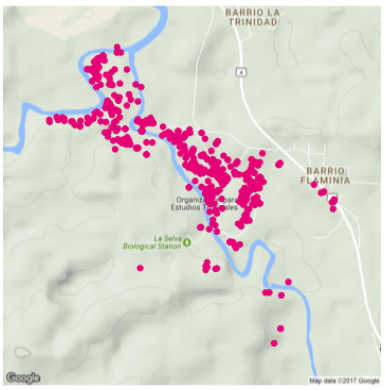

Figure 1:

other variables of each tree was also measured. With the information available, it was then easy to map the trees as a spatial point process. Note that there is no need to use the height variable to create a *marked* spatial point process, because these marks will offer no direct knowledge on how these plants are spreading. Instead, because of the partial dependency of age as it relates to the height of a tree, birth times of the red banana trees were estimated, allowing for a temporal analysis of spread.

We use the red banana data from Balderama et al. (2012). Heights (in cm) of 1008 red banana plants were observed and longitude and latitude coordinates were recorded by satellite. The heights of 318 select plants were measured weekly over a span of one year. To establish a time component, estimates of the growth rate was used to estimate any individual plant's age, and hence origin times. In the end, 788 plants had complete location and origin times data, and were used for the subsequent analyses.

3 Methods

Ogata's epidemic-type aftershock sequence models (Ogata 1988, 1998) is one of the most popular models of earthquake occurrences. The ETAS model is so aptly names because of the epidemic nature of how events are created; Earthquakes causes aftershocks, which in turn causes more aftershocks, and so on. It characterizes a sequence of earthquakes and aftershocks over time or over space and time via a conditional intensity, which represents the infinitesimal probability of an event occurrence at a single point in $\mathbb{R}^d$ given the past history of the process. This particular characterization of a sequence of earthquakes and aftershocks is a specific case of the linear, self-exciting Hawkes' point process (Hawkes 1971), which is specified by the conditional intensity

$$\lambda(t, Q \,|\, H_t) = \mu(t, Q) + \sum_{t_i < t} g(t, Q; t_i, Q_i), \tag{1}$$

where Q is additional information that may include a spatial component (x, y) and/or a mark or magnitude M, H_t is the history of the process up to time t, $\mu(\cdot)$ is the mean rate of a Poisson-distributed background process that may depend on time, space and magnitude, and $g(\cdot)$ is the "triggering function" which contributes the individual intensities of each point $\{i : t_i < t\}$ as a summation to the total conditional intensity $\lambda(\cdot)$ at time t. Thus, every past event has an additive (linear) influence on the present conditional intensity of the system.

A modified version of this Hawkes/ETAS model was empirically fitted to the red banana data (Balderama et al. 2012). First, the spatially inhomogeneous background density $\hat{\mu}(x, y)$ was estimated by a two-dimensional Guassian kernel smoother over the complete data. Next, a close examination of interevent distances and times between pairs of plants constituted an exponential decay in both time lag and squared distance. The triggering density function is

then given by

$$g(t,x,y) = \frac{\alpha\beta}{\pi} \cdot e^{-\alpha t - \beta(x^2+y^2)}, \tag{2}$$

where $\frac{\alpha\beta}{\pi}$ is a normalizing constant so that g integrates to one. The conditional intensity was then specified as

$$\lambda(t,x,y|H_t) = (1-p)\mu(x,y) + \frac{p\alpha\beta}{\pi} \sum_{\{i:t_i<t\}} e^{-\alpha(t-t_i)-\beta\{(x-x_i)^2+(y-y_i)^2\}} \tag{3}$$

where p is introduced to specify the proportion of events that were triggered, leading to a simultaneous estimation of the parameter vector $\boldsymbol{\theta} = \{\alpha, \beta, p\}$ by utilizing Bayesian Markov Chain Monte Carlo algorithm. Both α and β are parameters should be greater than zero, the prior distributions for them were then set to be Gamma(α_a, β_a) and Gamma(α_b, β_b), respectively, where α_a and α_b were set to be 1, and β_a and β_b had uninformative hyper prior distributions, Gamma(α_{hyper}, β_{hyper}), where α_{hyper}=β_{hyper}=0.5. For parameter p, which indicates an unknown proportion, should range between 0 and 1, so Uniform(0,1) was set as the prior distribution. The log-likelihood was then specified as

$$logL = \sum_{i=1}^{n} log\lambda(t_i, x_i, y_i) - \iint_A \int_0^\infty \lambda(t,x,y)dtdxdy. \tag{4}$$

The symmetric proposal distributions for α and β were set to be Normal($\alpha^{(t-1)}$,0.1) and Normal($\beta^{(t-1)}$,0.1) respectively, and the asymmetric proposal distribution for paramater p is Beta($p^{(t-1)}$,1$-p^{(t-1)}$), which has mean of $p^{(t-1)}$. The symbols $\alpha^{(t-1)}$, $\beta^{(t-1)}$ and $p^{(t-1)}$ represent the current values of parameters.

Because hyper prior distributions are conjugate priors, candidates of β_a and β_b were directly updated from their posterior distributions, $Gamma(\alpha_{hyper} + n\alpha_a, \beta_{hyper} + \sum_{i=1}^{n} \alpha)$ and $Gamma(\alpha_{hyper} + n\alpha_b, \beta_{hyper} + \sum_{i=1}^{n} \beta)$. Potential candidates of parameters would be drawn from proposal distributions and be updated by using Metropolis-Hastings ratio technique, which is specified as

$$R = \frac{p(\theta^{(c)}|y)}{p(\theta^{(t-1)}|y)}, \tag{5}$$

where $\theta^{(c)}$ represents the candidate parameter and $\theta^{(t-1)}$ indicates the current parameter. The parameter $\theta^{(t)}$ would be updated to $\theta^{(c)}$ at certain probability or remain as $\theta^{(t-1)}$.

After the Bayesian MCMC algorithm, 50,000 potential candidates for each parameter were recorded. Thining technique was applied, which discarded all but 10^{th} observations to get rid of autocorrelation to guarantee the independecy of samples, and 5,000 observations were left. Burning technique was then utilized, which threw away first 2,500 observations to minimize the effect of initial values on the posterior inference. Tuning technique was applied as well, which controlled the acceptence rate in a desired range, 25% to 60%. Bayesian inferences were then conducted based on the final 2,500 observations.

4 Results

The trace plot and posterior histogram for each parameter are shown in Figure 2.

The estimates obtained by Bayesian MCMC algorithm were $\hat{\boldsymbol{\theta}} = \{\hat{\alpha} = 0.0747, \hat{\beta} = 0.0203, \hat{p} = 0.6062\}$, with corresponding posterior standard deviations 0.0045, 0.0017 and 0.01941, respectively. These estimates are used to explain the decay rates of triggering events, where $\hat{\alpha} = 0.0747$ represents the intensity of trggering events decays at a a rate of $1 - e^{-0.0747} = 7.198\%$ for every week that passes, $\hat{\beta} = 0.0203$ suggests that the intensity of trggering events decays at a a rate of $1 - e^{-0.0203} = 2.1\%$ for every squared distance (in meters) away and $\hat{\rho} = 0.6062$ indicates that 60.62% of events were triggered by observed plants and 39.38% is due to the background rate (i.e., unobserved plants or other unknown sources). It is worth noting that this space-time

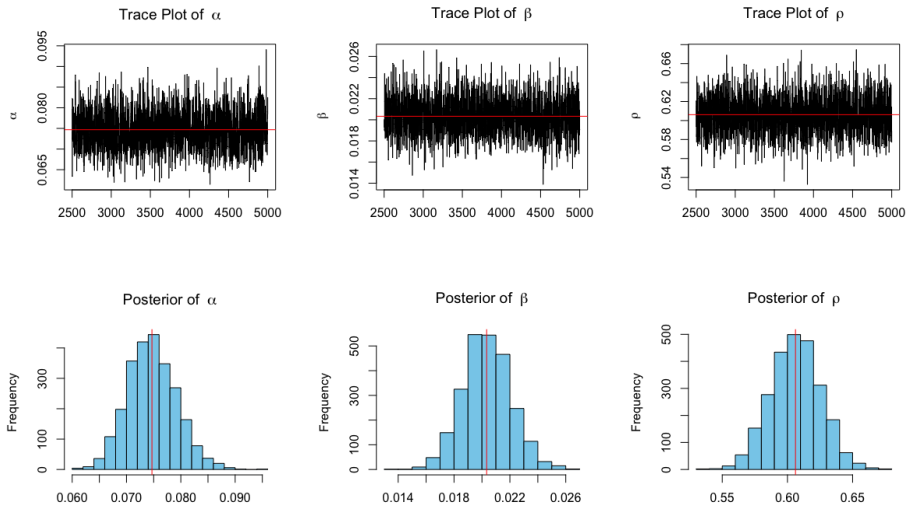

Figure 2:

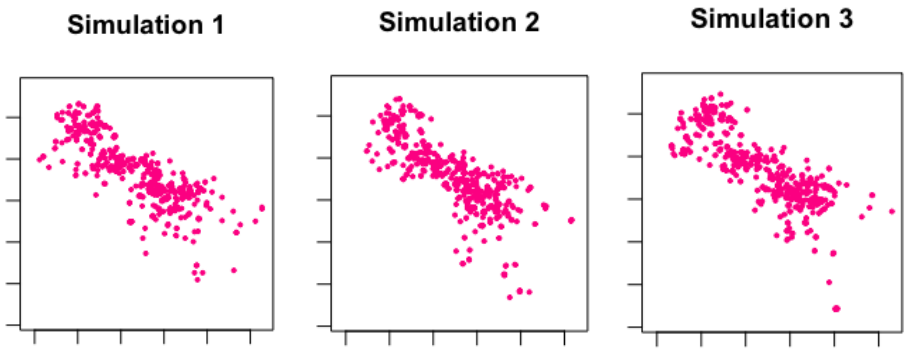

Figure 3:

modification does not contain magnitude information, as is the case for earthquakes, and that the dependence on time t is removed from the estimation of the background rate, as is also the case with Ogata's ETAS models.

The estimates from Bayesian MCMC are very close to the ones from Newton-Rhapson numerical optimization, and the posterior standard deviations are smaller than asymptotic standard errors for most of time.

Figure 3 show three simulations of the spreading pattern of red banana plants by using the modified ETAS model with Bayesian estimates.

Furthermore, this model was empirically fitted using the red banana data, and is therefore another case of a self-excited Hawkes' point process. A different species of plants may require not only different estimates of the parameters, but possibly a different set of parameters altogether, depending on the spatial and temporal distributions of interevent distances and times. However, it was the first attempt at using ETAS models to characterize the branching structure of invasive plant species.

References

Balderama, E., Schoenberg, F. P., Murray, E., and Rundel, P. W. (2012), "Application of Branching Models in the Study of Invasive Species," *Journal of the American Statistical Association*, 107, 467–476.

Delisle, F., Lavoie, C., Jean, M., and Lachance, D. (2003), "Reconstructing the spread of invasive plants: taking into account biases associated with herbarium specimens," *Journal of Biogeography*, 30, 1033–1042.

Hawkes, A. G. (1971), "Point Spectra of Some Mutually Exciting Point Processes," *Journal of the Royal Statistical Society Series B*, 33, 438–443.

Higgins, S. and Richardson, D. (1996), "A review of models of alien plant spread," *Ecological Modelling*, 87, 249–265.

Ogata, Y. (1988), "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes," *Journal of the American Statistical Association*, 83, 9–27.

— (1998), "Space-time Point-Process Models for Earthquake Occurences," *The Institute of Statistical Mathematics*, 50, 379–402.