Title
Spatial-Temporal Branching Point Process Models in the Study of Invasive Species

Author
Balderama, Earvin

Publication Date
2012

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Spatial-Temporal Branching Point Process Models in the Study of Invasive Species

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Earvin Balderama

2012

Abstract of the Dissertation

Spatial-Temporal Branching Point Process Models in the Study of Invasive Species

by

Earvin Balderama

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2012

Professor Frederic R. Paik Schoenberg, Chair

Earthquake occurrences are often described using a class of branching models called Epidemic-Type Aftershock Sequence (ETAS) models. The name derives from the fact that the model allows earthquakes to cause aftershocks, and then those aftershocks may induce subsequent aftershocks, and so on. Despite their value in seismology, such models have not previously been used in studying the incidence of invasive plant and animal species. Here, we apply a modified version of the space-time ETAS model to study the spread of an invasive species red banana trees (Musa velutina) in a Costa Rican rainforest. One challenge in this ecological application is that fitting the model requires the originations of the plants, which are not observed but may be estimated using filed data on the heights of the plants on a given date and their empirical growth rates. The formulation of the triggering density function, which describes the way events cause future occurrences of events is based on plots of inter-event times and distances for the red banana plants. We characterize the estimated spatial-temporal rate of spread of red banana plants using a space-time ETAS model. We then assess the triggering density more carefully using a non-parametric stochastic declustering method based on Marsan and Lengliné (2008). When the algorithm is applied to the red banana data, the results indicate similar temporal and spatial structure, compared to previous

estimates, as well as triggering of offspring running primarily to the northwest and the southeast from each parent. Non-parametric results are also used to obtain estimates of the most likely targets where immigration of red banana plants may be occurring.

The dissertation of Earvin Balderama is approved.

Philip W. Rundel

Hongquan Xu

Qing Zhou

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2012

To my mom . . .

for all your love and support

I could never have done any of this without you

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor, Rick Schoenberg, for being the best advisor one could hope for. Thanks for being so patient and understanding through the years. I have really learned a lot about being a good teacher and researcher, and also a good poker player!

Thank you to my committee members Hongquan Xu, Qing Zhou, and Phil Rundel for all your helpful suggestions and comments. Phil, I would still really love to visit Costa Rica someday!

I also want to thank Nicolas Christou for being a wonderful professor and friend. We should definitely plan another dinner at some ethnic restaurant! Also, thank you Glenda Jones for making all the paperwork and logistics easy for me, and thanks for all those funny conversations.

To my officemates at 8105C, thank you for being there when I needed help or just when I needed to talk. I am grateful for being part of such an awesome office! To my classmates during my first couple of years, thanks for all the homework and studying help, and thanks for all those fun nights out!

Finally, to all my closest friends and family, in particular my sisters Shella and Karen, and my sister and brother from other mothers, May and Brian, thank you for constantly asking me when I will be done with grad school. Here you go!!!

Vita

2004	B.S. (Applied Mathematics), San José State University, San José, CA.
2005	Teaching Assistant, Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles.
2005–2011	Teaching Assistant, Department of Statistics, University of California, Los Angeles.
2009	M.S. (Statistics), University of California, Los Angeles.
2010–2011	Quality of Graduate Education University Fellowship.
2011–present	Instructor, Mathematics Department, Chaffey College, Rancho Cucamonga, CA.
2012	Graduate Student Researcher, Department of Statistics, University of California, Los Angeles.

Publications and Presentations

Balderama, E., Applying a Modified ETAS Model to Invasive Plant Spread Data, Presented at Joint Statistical Meetings, Washington, D.C., 2009.

Balderama, E., Schoenberg, F. P., Murray, E., Rundel, P. W. (2011), Application of Branching Models in the Study of Invasive Species, *Journal of the American Statistical Association*, to appear.

CHAPTER 1

Introduction

The establishment of alien (invasive) plant and animal species outside of their natural range of distribution present a major challenge to the structure and function of natural ecosystems as well as a major economic cost through impacts on ecosystem services. Pimentel et al. (2005, 2007) estimates the financial impact of invasive species in the United States at over 120 billion dollars per year, and Colautti et al. (2006) estimates the cost of eleven invasive species in Canada at up to 34 billion canadian dollars per year. Understanding the nature and dynamics of such invasions represents a major environmental challenge (Pejchar and Mooney 2009).

There have been very few attempts at modeling alien plant spread in the literature (Higgins and Richardson 1996), despite the devastating effects alien plants have on natural ecosystems. For ecologists and statisticians alike, the motivation comes from the unique opportunity to develop new methods offer insight into ecological theory from studying the expansion of alien organisms in a new range.

In Costa Rica, an alien *Musa velutina* (red banana, Musaceae), has recently spread from areas where it is cultivated into secondary forests and to a lesser degree into primary tropical rainforests at La Selva Biological Station near Puerto Viejo in Sarapiquì Province. The establishment of these plants has been monitored by researchers, and exact locations for each observed plant have been recorded using a global positioning satellite (GPS).

The study of the detailed spatial-temporal pattern of the spread of invasive

species such as red bananas is typically fraught with difficulty, for two major reasons. First, data on such species are rarely recorded with sufficient precision, even locally, to warrant the precise estimation of parameters governing the spatial-temporal dispersal of the species. Second, the detailed estimation of spatial-temporal spread has proven to be a difficult and somewhat non-standard statistical problem. Ecological researchers have typically found it more convenient to organize the data into counts of observations within predefined grid cells and to use standard grid-based spatial statistics methods, including the application of simple birth-and-death models, though such analyses obviously result in a loss of crucial information for the study of spread on small scales (Keeling et al. 2001; Peters 2004).

Fortunately, both of these hurdles may be overcome in the present application. In the case of red bananas in Costa Rica, the precise locations of the plants have been recorded by researchers, and because the plants have only recently been identified in Costa Rica, their numbers are sufficiently small and the records sufficiently careful to warrant a detailed analysis of their spread. Regarding the problem of statistical analysis, one can appeal to modern innovations in the statistical analysis of earthquakes, which have led to substantial gains in the characterization of aftershock activity. As with lists of red banana plant sightings, catalogs of earthquakes include detailed estimates of the spatial location associated with the origin of each event, and point process methods have been used to characterize the spread, or aftershock activity, associated with each event. In fact, prevailing methods in seismology involve the use of so-called *epidemic* branching point process models, where one earthquake may trigger aftershocks, those aftershocks trigger further aftershocks, and so on. Space-time versions of these models and methods for their estimation have recently become sufficiently robust to describe seismicity on various different space-time scales, and thus may also be used to describe other phenomena such as invasive species. The potential advantage of

the use of such methods, as opposed to grid-based methods, is improved precision in the characterization of the spatial-temporal spread of the events being studied.

The spatial distribution of the plants can be thought of as a point process, where each individual plant is represented by a point on the surface of the Earth. The invasive nature of the plant leads to the assumption (as well as observation) that each plant can spread its seeds and initiate the growth of more plants around it, and each of those plants will spread its own seeds at smaller or larger distances, creating multiple nodes of invasion (Pysek and Hulme 2005). The spread of seeds can be attributed to a variety of different mechanisms, from animals consuming the fleshy fruits and dispersing the seeds (Begon et al. 2006; Gosper et al. 2005), as well as by fruits floating in frequent flood events.

Studies of invasive plants are usually concerned with the environmental impact and harmful effect of the particular species of plants on local biodiversity and ecosystem processes, and much less on the detailed spatial-temporal patterns of the plant occurrences (Higgins and Richardson 1996; Peters 2004). Simple demographic models have been proposed for plant spread, including exponential, logistic, and logistic-difference equation models, which have no spatial component and predict the total number of individuals in a population over time. As noted in Higgins and Richardson (1996), early spatial models for plant spread only modeled the total area of coverage, rather than the detailed locations of the observations: regression models, for instance, attempt to quantify the amount of area invaded over time (Thompson 1991; Lonsdale 1993; Perrins et al. 1993; Pysek and Prach 1993; Delisle et al. 2003; Peters 2004). Related models include reaction-diffusion models, which use the spatial locations and the contagious behavior of the species in modeling invasion (Peters 2004). Reaction-diffusion models use partial differential equations to characterize the rate of spread and expansion range but require good estimates of the rate of population growth and diffusivity, and like regression models, are typically applied over spatial grids (Higgins and Richardson 1996);

as a result, detailed spatial information is typically lost, relative to point process methods. Studies that do treat such epidemic-type data as a point process are usually of diseases in animal populations, and typically use grid-based birth-and-death rather than branching point process models (Keeling et al. 2001; Keeling 2005; Diggle and Zheng 2005; Diggle 2006; Riley 2007), essentially ignoring detailed information on the spatial locations.

More recent studies have looked into distributions of seed dispersal (Marco et al. 2011; Wang et al. 2011a) which may give more insight into the spatial distribution of the spread of plants after invading new territory. However, these studies generally analyze spread through either a purely spatial or a purely temporal model, essentially ignoring spatial-temporal interdependence, rather than modeling the entire spatial-temporal evolution of the process.

The process by which plants spread seeds naturally lends itself to spatial-temporal self-exciting point process analysis. As noted by Law et al. (2009), unlike grid-based studies on area occupation, where the surface of study is divided into an array of pixels on a grid, point process methodology over time and space allows for greater precision when taking exact times and locations into account, and can offer plant ecologists a more detailed and precise account of spatial heterogeneity and clustering.

In this disseration, we propose the use of a class of branching point process models called epidemic-type aftershock sequence (ETAS) models to characterize the spatial-temporal spreading patterns of the red banana plants. We then explore a completely non-parametric method of estimating the components in such a model. Such a non-parametric estimate is useful for two reasons: first, the components when estimated non-parametrically will offer superior fit to the data when the exponential model does not fit perfectly; second, such non-parametric estimates can be used to verify whether the exponential model seems to be an appropriate approximation. In order to estimate the spatial-temporal spread of

plants non-parametrically, we focus mainly on the branching structure which indicates the triggering process of events.

The remainder of this dissertation is organized as follows. The red banana data invasive to Costa Rica are described in Chapter 2. Previous models of invasive species are summarized in Chapter 3. Branching point processes, including ETAS models, are reviewed in Chapter 4. We characterize the spatial distribution and estimate the birth time of each plant to characterize the temporal structure in Chapter 5. Maximum likelihood is used to estimate the parameters of our adjusted ETAS model for red banana plants in Chapter 6.1. The model is assessed in Chapter 6.3 using an extension of the thinned residual analysis technique of Schoenberg (2003). Chapter 6.4 shows how simulations of the model can be used to forecast future spread. Chapter 7 describes non-parametric point process methods that may be used to study invasive plant spread, and the application of a non-parametric algorithm to characterize the spread of the red banana plants. A discussion is provided in Chapter 8.

CHAPTER 2

Red Banana Data

2.1 Data collection

Data on the red bananas were collected at La Selva Biological Station in Costa Rica over a 15-month period beginning in late 2006. Although highly diverse rainforest communities are often considered to be strongly resistant to the establishment of invasive species, red banana has become established in large numbers in secondary forest stands less than 2 km from rural farms where red banana has recently been planted widely as an ornament. Red banana is native to Southeast Asia but is an attractive plant in tropical gardens because of its foliage and clusters of small pink bananas. Because of their rapid range extension into La Selva forests and the potential for red banana to compete with and possibly replace native species in disturbed habitats, there is an active management program to remove plants as they are found.

Field data on the establishment of red banana included 1008 individual plants observed and mapped with globing positioning satellite (GPS) between December 2006 and January 2008. The coordinates of observed plants ranged between 83.996° and 84.017° West longitude, and 10.421° and 10.445° North latitude. Figure 2.1 shows a map of the observed plants. The general NorthWest-SouthEast trend observed in Figure 2.1 is mainly due to a small tributary stream running diagonally just south of the points.

To provide data on the relative growth rates of individual plants at various

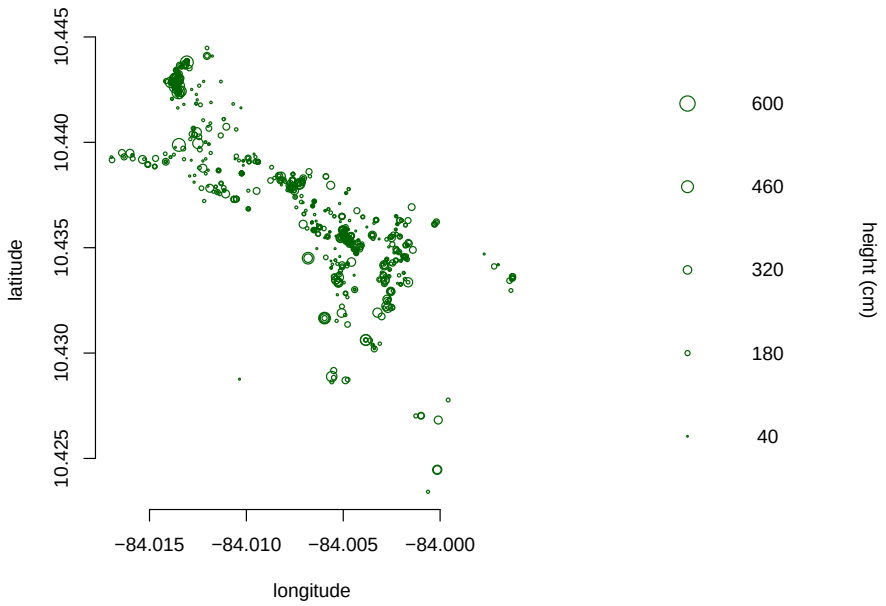

Figure 2.1: Top: Map of observed red bananas with larger circles indicating taller plants. Bottom: Latitude and Longitude each plotted against height.

life history stages, 318 plants were allowed to continue to grow and measured weekly over a span of 49 weeks for height, leaf number, and reproductive condition (flowers or fruits). In practice, plants were individually monitored for an average of 20 weeks (median of 15 weeks) because of their death or rapid growth rate that led to a need to remove plants before mature fruits were dispersed. Nevertheless, 286 of these plants provided useful data to establish growth rates in relation to plant size. Height measurements to the uppermost leaf height provided the most useful indication of continuing growth, although height measurement in bananas is somewhat subjective depending on leaf condition.

2.2 Included variables

The variables measured on each plant include the height in centimeters, the number of leaves, the number of fruits, the number of seedlings surrounding the base of the plant, and the day it was observed. Figure 2.2 offers information on the distribution of observed heights. Not all plants had complete measurements. Some were missing GPS coordinates, while others had missing height values, simply because they were too short to be considered a full grown herb. The 788 plants with complete GPS and height data were used in the subsequent estimation and model fitting analyses.

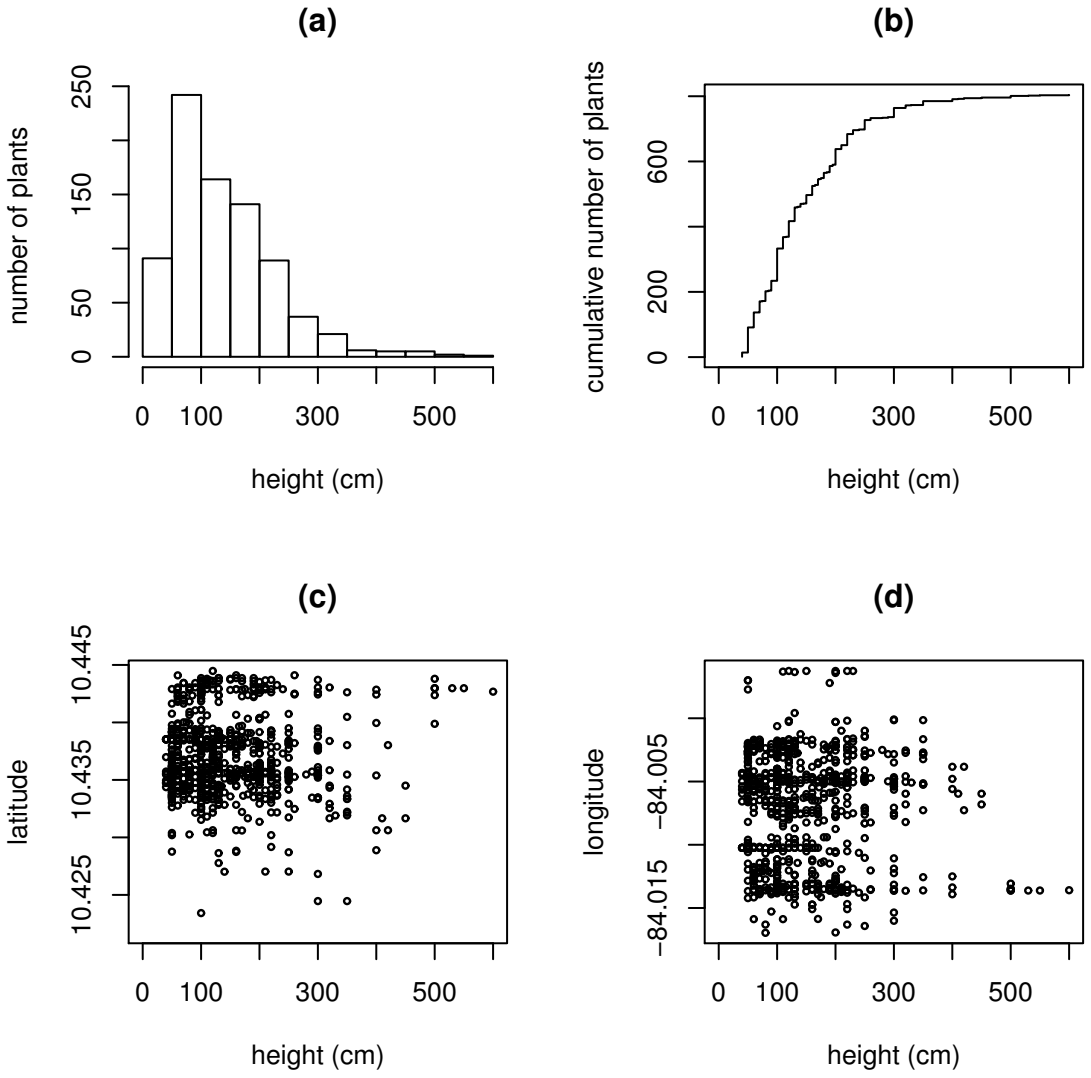

Figure 2.2: (a) Distribution of heights of observed plants. (b) A plot of the cumulative number of plants against height. (c) A plot of latitude against height. (d) A plot of longitude against height. We notice that the taller plants occurs in the area of high latitude and low longitude. This is apparent in the map in Figure 2.1.

CHAPTER 3

Models for Invasive Species

Models previously applied to invasive plants fall into one of several categories, discussed here. This categorization is based on the model's input requirements, its data sources and its output variables (Higgins and Richardson 1996). Although there are many models that fall into each category, only a few of the more popular models are used as examples. In general, model complexity increases as we move from non-spatial models to more spatially explicit models, where the occurrence of events depends on spatially dependent random variables. But as Peters (2004) notes, a parameter-heavy model may not necessarily provide better results.

3.1 Non-spatial models

Models that do not contain spatial information are sometimes referred to as demographic models because they tend to estimate population growth over time. They assume that population density is simply related to area invaded. Typically, these models are characterized by the differential $\frac{dN}{dt}$, where N is the population size at time t. The functional form of the differential equation depends on the nature of population growth. The exponential model is a simple, yet popular, model for population growth, although it assumes an infinite population density potential. Because a finite amount environmental resources may be available that limits the size of the population, a logistic model may be more appropriate.

Other models include the logistic-difference model, which discretizes the pop-

ulation size N and time t, and stochastic models, which allow the invasion rate to vary with respect to a random variable.

This class of models is not usually considered "invasion" models due to the absence of a spatial component. Non-spatial models are useful for population density estimation and serves as a theoretical foundation for actual invasion models (Higgins and Richardson 1996). Although invasion rates of spread can be estimated from non-spatial models, these estimates are usually inaccurate in plant populations. As noted in Higgins and Richardson (1996), population growth does not easily translate into a rate of spread. For example, empirical rates of spatial spread, as in Perrins et al. (1993); Pysek and Prach (1993), is an order of magnitude lower than population increase. Thus, most studies of invasion rates and patterns require spatial information.

3.2 Spatially implicit models

To investigate and forecast the amount of area being invaded over time, spatially implicit models are used. These models combine spatial information with non-spatial models and attempt to estimate the invasion rate of spread.

For example, regression models are commonly used to find the relationship between the amount of area invaded and time, and predict potential range distributions of invasive trees (Peters 2004). The slope of regression equations is used as an estimate of the rate of spread (see e.g. Lonsdale 1993; Perrins et al. 1993; Pysek and Prach 1993). Since regression, as well as multiple regression, models are fitted to empirical data, they can be used as a comparative tool and help develop further models and theory.

Other spatially implicit models include the geometric model and the Markov model, both of which have theoretical modeling potential, but have not been applied to invasive plants (Higgins and Richardson 1996). The geometric model uses

a regression approach to investigate invasions with multiple focus points. Markov models incorporates birth and death state transition probabilities in discretized time and space to estimate population densities.

3.3 Spatially explicit models

When spatial information is considered in the formulation of the model, spatially explicit models are used. This is a large category of models where the invasion rate of spread is affected by explicit spatial information, such as seed dispersal distribution, spatial environmental heterogeneity, spatial arrangement of invasion focus points, and local neighborhood interactions (Peters 2004; Perry et al. 2006; Marco et al. 2011; Wang et al. 2011a). For example, simulation models may use seed dispersal information and other spatial covariates to show environmental heterogeneity and variability of population density.

Reaction-diffusion models assume that the rate of spread is a function of the rate of population increase and the rate of movement of individuals in the population (Higgins and Richardson 1996). These models are characterized by a partial differential equation that gives the population density at some spatial location (x, y) with respect to time t. Typically used to predict range expansion of invasive animals and disease, reaction-diffusion models have rarely been applied to plant invasions (Skellam 1951; Higgins and Richardson 1996). This may be due to the underlying model assumption that seed dispersal distances are normally distributed, although many wind- and animal-dispersed plants have more of a leptokurtic dispersal distribution (Howe and Westley 2009). The reaction-diffusion model further predicts that plotting the graph of the square root of invaded area against time should be a straight line, and that the slope should correspond to the mean rates of expansion. Lonsdale (1993) concluded that the model was inadequate for plant spread data, after showing that these predictions failed to be

true.

Metapopulation models and individual-based cellular automata models consider space and time as discrete variables, and typically adopt a simulation approach. A metapopulation model of plant spread divides the space into discrete local population sites of varying degrees of invasion susceptibility, and each local site experiences some population rate of growth, i.e. exponential. As the simulation progresses, populations eventually emigrate to other sites. Cellular automata models also divides the space into local sites, or grid of cells, where each cell takes on one of a number of states. Transition probabilities are used to define the state of the system, where the current state of a cell depends on the previous state of the cell and the states of nearby cells.

3.4 Spatial point processes

The models described above do not use the exact locations of individual observations. When exact locations are given, point process methods may provide more detailed spatial analyses of the data (Law et al. 2009).

Each point is sometimes referred to as an *event*, with spatial coordinates (x, y). A point processes is typically characterized by its intensity, $\lambda(x, y)$, which represents the expected number of events per unit area at the location (x, y). A spatial point process with constant intensity λ is termed spatially homogeneous or stationary in space, and is said to exhibit complete spatial randomness (CSR). If the process is CSR, then the number of events inside a region of area A follows a Poisson distribution of rate λ, and the events inside that region are positioned independently and uniformly (Cressie 1993; Moller and Waagepetersen 2004). Testing the null hypothesis of complete spatial randomness (CSR) is usually one of the main goals (Wiegand and Moloney 2004; Perry et al. 2006; Law et al. 2009). If the null hypothesis is rejected, the process is spatially inhomogeneous.

Point process methods became popular in the fields of ecology, geography and archaeology beginning in the 1950s and 1960s (Gatrell et al. 1996). Cluster point processes is still widely used in Forestry and other applications in plant ecology (Stoyan and Penttinen 2000).

However, these point pattern analyses are purely spatial. Such models only analyze a single temporal snapshot of an underlying process. In a way, these models simply replace the location of points in time with locations in space. However, ecological situations naturally come with spatial components, so analysis is incomplete without a spatial dimension (Law et al. 2009). Without a temporal component, one should proceed with caution when inferring *process* from *pattern*, since multiple processes may be able to explain the same observed pattern (Perry et al. 2006). Hence, stronger inferences about ecological processes require the combination of temporal and spatial analyses.

3.5 Spatial point pattern analysis over time

Incorporating time in the analysis of spatial pattern poses challenges in the ecological setting, as many factors may be involved in the estimation of plant ages (Rice et al. 2012) and the collection of temporal data.

One way to assess the spatio-temporal relationships among plants, without the need for age estimation or extra data collection, is to consider snapshots of the data at several different times, and treat each snapshot as a different point pattern (Wiegand and Moloney 2004). An analysis using a bivariate K-function or O-ring statistic can investigate any changes in the behavior of the process. (Halpern et al. 2010) also used a series of stem maps to quantify the spatio-temporal pattern of a conifer invasion.

Rice et al. (2012) attempts to show the nature and strength of tree interactions over time, by analyzing snapshots of a tree species at different time lags. Nearest-

neighbor distances between trees were analyzed, beginning with a 20-year gap from 1935-1955, then gradually decreasing the time interval to 2-year gaps ending in 1997.

In a sense, these models are still purely spatial. Each snapshot is analyzed independently as its own spatial point process. Furthermore, the quantification of a time component is given to the pattern as a whole, at specific time intervals, but the occurrence of events may be continuous in time. Better spatio-temporal models may incorporate more specific time information, such as individual plant ages, into a model that makes estimates based on space and time simultaneously.

CHAPTER 4

Branching Point Process Models

4.1 Spatial-temporal point processes

A spatial-temporal point process is a random collection of points, where each point represents the time and location of an event. Typically the spatial locations are recorded in three spatial coordinates, though sometimes only one or two spatial coordinates are available or of interest. For example, earthquake occurrence locations may be specified by longitude, latitude, and depth. In the case of the births of invasive plant species, such as the red bananas, only longitude and latitude dimensions are of interest. Much of the theory of spatial-temporal point processes carries over from that of spatial point processes. However, the temporal aspect enables a natural ordering of the points that does not generally exist for spatial processes. Indeed, it may often be convenient to view a spatial-temporal point process as a purely temporal point process, with spatial marks associated with each point. Mathematically, the spatial-temporal point process N is defined as a σ-finite random measure on a region $S \subseteq \mathbb{R} \times \mathbb{R}^3$ of space-time. Denote $N(A)$ as the number of points inside a subset A of S, taking values in the non-negative integers $\mathbb{Z}^+$ or infinity. The spatial region of interest is often rectangular, but can sometimes have an irregular boundary due to geographical constraints such as city limits or shorelines. Attention is typically restricted to to points inside some time interval $[T_0, T_1]$, and to processes with only a finite number of points in any compact subset of S.

Point processes with temporal components are typically characterized by their conditional intensity functions, which represent the infinitesimal rate at which events are expected to occur around time t, based on the prior history H_t of events prior to time t (see e.g. Daley and Vere-Jones 2003). In general, the conditional intensity $\lambda(t, x, y, z)$, at any point (t, x, y, z) associated with the spatial-temporal process N may be defined as the limiting conditional expectation

$$\lambda(t, x, y, z) = \lim_{\Delta \to 0} \frac{E[N(B_\Delta)|H_t]}{|\Delta|} \tag{4.1}$$

where B_Δ is the set $(t, t + \Delta t) \times (x, x + \Delta x) \times (y, y + \Delta y) \times (z, z + \Delta z)$, and Δ is the vector $(\Delta t, \Delta x, \Delta y, \Delta z)$. For a more comprehensive overview, please see e.g. Cressie and Wikle (2011); Ripley (1981); Vere-Jones (2009).

4.2 Self-exciting point processes

In the study of branching point processes, Hawkes' linear self-exciting point process model (Hawkes 1971) is a birth process with immigration and triggering, such that each event may trigger subsequent events, which may in turn trigger future subsequent events, and so on. The model characterizes a sequence of earthquakes and aftershocks over space and time via a conditional intensity, specified by

$$\lambda(t, s|H_t) = \mu(t, s) + \sum_{t_i < t} g(t, s; t_i, s_i), \tag{4.2}$$

where s 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 indicates how previous occurrences contribute, depending on their spatial and temporal distances and marks, to the intensity $\lambda(\cdot)$ at the location and time of interest. The individual intensities of each prior point i, $\{i : t_i < t\}$ is summed and contributes 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.

4.3 Epidemic-type aftershock sequence models

Ogata (1988) introduced the Epidemic-Type Aftershock Sequence (ETAS) model, based on Hawkes' mutually exciting model (a multivariate extension of the linear, self-exciting model (Hawkes 1971)) to study the temporal pattern of earthquake and aftershock activity. The ETAS model of Ogata (1988, 1998) is aptly named because of the epidemic nature of how events are created, and is widely used to describe earthquake occurrences (see e.g. Helmstetter and Sornette 2003; Ogata et al. 2003; Sornette and Werner 2005; Vere-Jones and Zhuang 2008; Console et al. 2010; Chu et al. 2011; Wang et al. 2011b; Werner et al. 2011; Zhuang 2011; Harte 2012; Tiampo and Shcherbakov 2012).

For a temporal ETAS process, the conditional intensity is given by

$$\lambda_0(t|H_t) = \mu + \sum_{\{i:t_i<t\}} g_i(t; M_i), \tag{4.3}$$

where

$$g_i(t) = \frac{K_0}{(t - t_i + c)^p} \cdot e^{\alpha(M_i - M_0)} \tag{4.4}$$

is the triggering function describing the contribution to the rate at time t attributed to an event at a prior time t_i of size M_i. K_0, α, c and p are parameters, and M_0 is some cutoff magnitude typically based on the earthquake catalog's completeness threshold. Ogata (1998) extended (4.3) and (4.4) to accommodate the spatial pattern of earthquake occurrences. The ETAS model naturally extends to

$$\lambda(t, x, y|H_t) = \mu(x, y) + \sum_{\{i:t_i<t\}} g(t - t_i, x - x_i, y - y_i; M_i), \tag{4.5}$$

where $\mu(x, y)$ is a non-homogeneous background rate and $g(\cdot)$ is the triggering

function defined in the earthquake context, for instance, by

$$g(t, x, y; M) = \frac{K_0}{(t + c)^p} \cdot \frac{e^{\alpha(M - M_0)}}{(x^2 + y^2 + d)^q}, \tag{4.6}$$

where K_0, α, c, p, d and q are parameters to be estimated. Other formulations of the triggering function were considered in Ogata (1998), but all are written typically as a product of a temporal density, a spatial density and a cluster size factor (Ogata 1998). The power law decay of aftershock activity over time is known as the modified Omori law (Utsu 1961). The spatial component, typically a power law decay function as in (4.6), is based mostly on empirical studies of aftershock clustering with some speculative hypotheses, and takes the remote triggering phenomena into account (Ogata 1998). While each of these terms should be adjusted to fit red banana data rather than earthquake occurrences, the overall structure of the ETAS model seems to be a sensible approach for the description of data on invasive species.

Note that the background rate $\mu(x, y)$ accommodates the spatial inhomogeneity obvious in seismology, since earthquakes occur primarily along existing faults, and similar spatial inhomogeneity may be present in the case of red banana plants in Costa Rica due to external factors such as land use, topography and meteorological conditions which vary spatially. In the case where the model must accommodate temporal non-stationarity, the background rate μ can be extended to a function of space and time, $\mu(t, x, y)$, and if a vector $\mathbf{Z}(t, x, y)$ of spatial-temporal covariates are available for each location and time, then one may incorporate them by extending the background rate to a function $\mu(t, x, y|\mathbf{Z})$.

CHAPTER 5

Adjusting the Triggering Function

5.1 Temporal clustering

While earthquake catalogs typically contain precise estimates of earthquake occurrence times, similar estimates of the origin times of red banana plants are generally unavailable. Thus, we propose estimating these birth times using the heights of the plants. The data on the 318 red banana plants that were measured weekly over a span of 49 weeks provides information on the rate of height increase over time.

The rate of growth $r_i(h)$ for each red banana plant $i, \{i = 1, 2, ..., 318\}$, is the change in height with respect to time, and may be calculated as follows:

$$r_i(h) = \frac{(maximum\ height)_i - (initial\ height)_i}{\#\ weeks}. \tag{5.1}$$

The denominator in equation (5.1) is the number of weeks between the initial height measurement and the maximum height measurement of plant i. To estimate a rate function

$$r(h) = \frac{dh}{dt}, \tag{5.2}$$

that generalizes the rate of growth of a red banana plant at any stage of its life, we plot (5.1) against the mean heights over all observed weeks of each plant, and smooth the resulting graph by taking running medians of length 21 plants (see Figure 5.1).

From the rate function (5.2), we can extract the current age of any plant by taking the integral of the inverse rate function from 0 (height at birth) to h_i

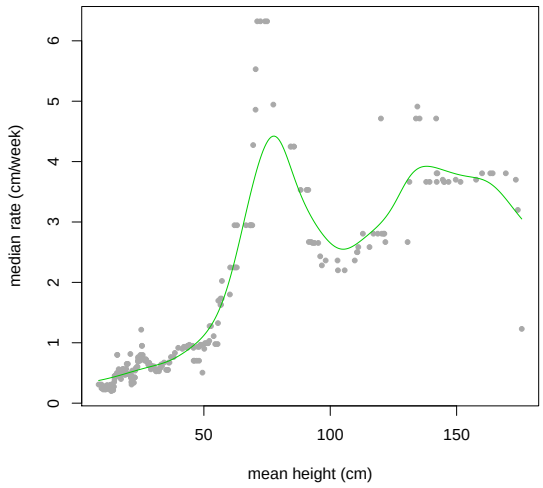

Figure 5.1: Running medians of length 21 plants of the rate of growth of a plant is plotted (as gray dots) against the mean height of the plant. The function $r(h)$ (solid green curve) is estimated by a kernel smoother.

(current observed height):

$$a_i = \int_0^{h_i} \frac{1}{r(h)} dh, \tag{5.3}$$

where a_i is the (estimated) age of plant i and $r(h)$ is the (estimated) rate of growth function of the plant with respect to height, using the fitted kernel regression curve in Figure 5.1. The age of the plant can then be converted to a birth time of the plant, where the oldest plant is given a birth time of 0. From these estimated ages, the youngest plant is given a birth time of 194.6 weeks. A histogram of estimated birth times, shown in Figure 5.2, reveals an increasing number of plants up to 160 weeks, with a slower rate of younger plants thereafter. The median birth time is 152.2 weeks and the inter-quartile range is between 130.1 and 164.0 weeks.

One may examine the inter-event birth times in order to model the temporal component of the triggering function g in (4.5). Figure 5.3 shows a histogram of the differences in estimated birth times between pairs of plants that are within 100 meters of each other. These inter-event times appear to be approximately

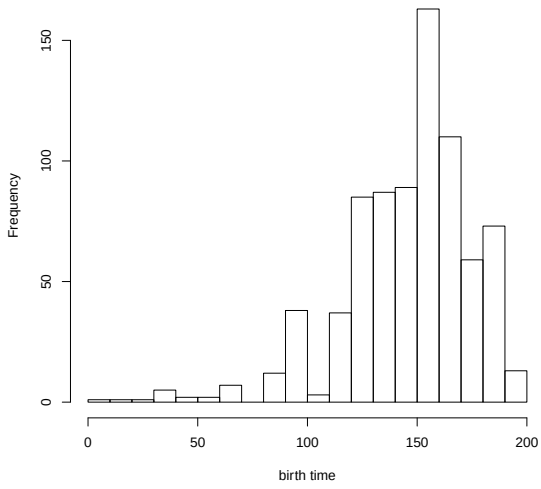

Figure 5.2: Histogram of estimated birth times.

exponentially distributed.

5.2 Spatial clustering

To assess the structure of the spatial component of the triggering function, one may examine the inter-event distances between pairs of plants. For computational ease and interpretability, we convert plant locations from the geographical longitude-latitude coordinate system to the Cartesian plane. We construct a rectangular spatial window surrounding the observed data, setting an arbitrary origin with x-coordinate equal to the minimum of all observed longitude values and y-coordinates equal to the minimum of all observed latitude values. Using the fixed length of one latitudinal degree, and the variable length of one longitudinal degree conditioned on latitude, we systematically transformed the geographical plant locations to Cartesian coordinates. Figure 5.5 shows the plot with the new transformed coordinate system.

Distances between pairs of plants were calculated by utilizing the haversine

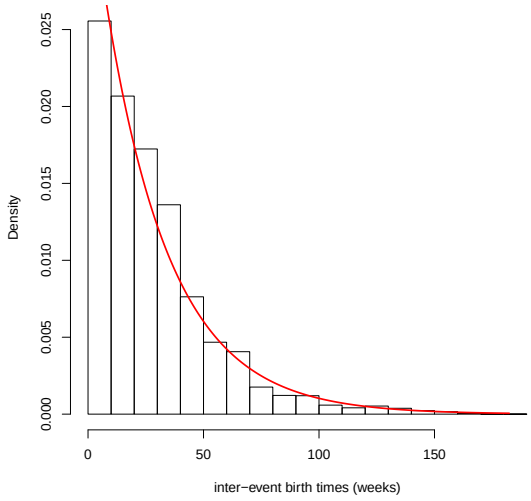

Figure 5.3: Histogram of estimated inter-event times (black), obtained by taking the differences in estimated birth times between pairs of plants that are less than 100 meters apart. The exponential density curve (red) has rate $\lambda = 0.0354$, fitted by maximum likelihood.

formula for the central angle between two points with geographical longitude and latitude coordinates (Sinnott 1984). Given two points, i and j, with geographical coordinates (φ_i, ψ_i) and (φ_j, ψ_j), respectively, the central angle between them is

$$\phi_{ij} = 2 \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos\varphi_i \cos\varphi_j \sin^2\left(\frac{\Delta\psi}{2}\right)}\right) \tag{5.4}$$

where $\Delta\varphi$ and $\Delta\psi$ are the differences between the longitude and latitude coordinates, respectively. Assuming a spherical Earth, the arc length (distance) between i and j is then given by $d_{ij} = r\phi_{ij}$, where r is the radius of the Earth. Variation in Earth radii make estimates correct to within 0.5%, but the error is less if applied to a limited area (Sinnott 1984). Nevertheless, since measurements from GPS were only accurate to within approximately 6 meters, we jitter the point locations (by moving each point a distance of mean 0 and standard deviation 3) to maintain a simple point process (Daley and Vere-Jones 2003).

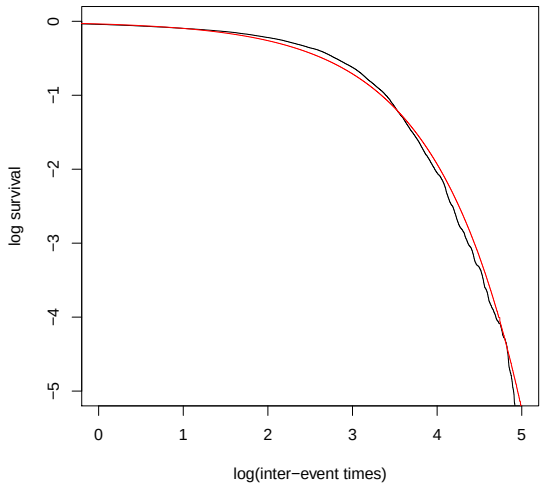

Figure 5.4: Survival function of estimated inter-event times (black). The survival curve of the fitted exponential density curve (red) has rate $\lambda = 0.0354$.

In the ETAS model of Ogata (1998), the spatial component of the triggering function has a power law form. However, the inter-event distances of plants that are estimated to have originated within six weeks of each other suggests an approximately exponential spatial decay, as shown in Figure 5.6 and Figure 5.7.

Indeed, Figures 5.6 and 5.7 suggests replacing the spatial component of the triggering function in (4.6) with an exponential function of the squared distance, i.e. $g(t, x, y) \propto \exp\{-\beta(x^2 + y^2)\}$.

5.3 Magnitude and productivity

In the earthquake context, the frequency of events of a certain magnitude is observed to depend on the magnitude in question, with a magnitude frequency distribution posited to follow an exponential distribution (Gutenberg and Richter 1944), and furthermore events of large size are thought to induce more after-

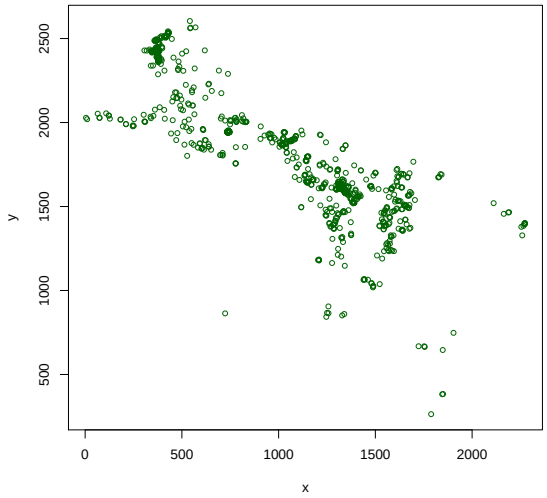

Figure 5.5: Map of observed red banana plants in the transformed coordinate plane.

shocks, also according to an exponential law (Ogata 1988). The magnitude of an earthquake is a measure of the amplitude of ground motion recorded by a seismograph which affects the speed and overall coverage of aftershock sequences.

There does not seem to be one obvious corresponding analogue for plants, however. Some studies analyze the reproduction and spread of invasive plants using only general traits of the species, such as habitat preference, climate reliance or viability of seeds (Delisle et al. 2003; Pysek and Prach 1993; Thompson 1991). In some purely spatial point pattern analyses, categorical marks (m_i) are often used (such as species, sex, age class, or health) but may also be continuous (such as plant size), especially where representing temporal phenomena (e.g. time of establishment) (Law et al. 2009; Perry et al. 2006). In our analysis, we were already able to estimate the temporal structure of the data by using height to estimate occurrence times. A mark characteristic in our model should have a profound effect on the triggering probabilities of future occurrences. The number of leaves, the number of flowers, or the height at maturity may have potential for

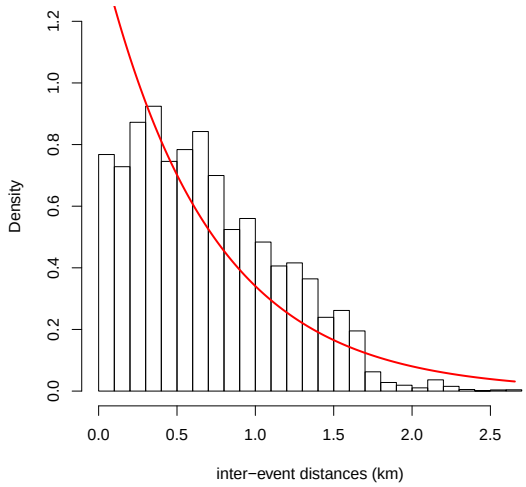

Figure 5.6: Histogram of the inter-event distances of plants estimated to have originated within six weeks of each other (black). The exponential density curve (red) has rate $\lambda = 1.44$, fitted by maximum likelihood.

being a useful measure of productivity for plants, and this is a subject of ongoing research. For this analysis, we assume uniform productivity of the plants in our proposed model.

5.4 Model summary

In the previous sections, we argued, after close examination of inter-event distances and inter-event times between pairs of plants, that the process by which plants trigger the occurrence of new plants may be modeled as exponentially distributed over time and space. Let α and β be the parameters of the temporal and spatial clustering components, respectively, of the model. For the purpose of data summary and description, it is useful to introduce a further parameter p governing the proportion of plants that are direct descendants of previous plants and those plants that are part of the background immigration. The result is a

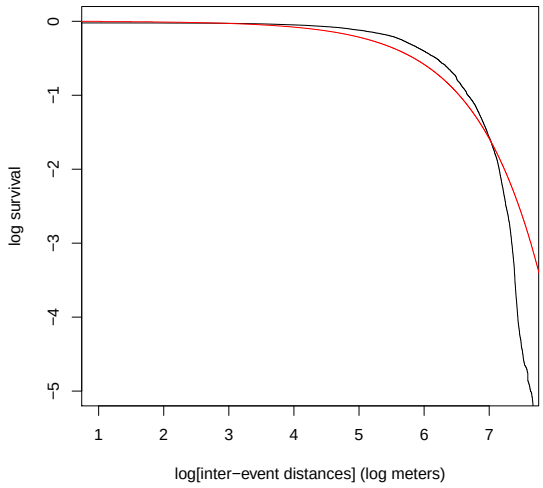

Figure 5.7: A survival plot for the inter-event distances (black). The survival curve of the fitted exponential distribution function (red) has rate $\lambda = 1.44$.

conditional intensity function

$$\lambda_\theta(t,x,y) = (1-p)\mu(x,y) + p \sum_{\{i:t_i<t\}} g(t-t_i, x-x_i, y-y_i) \tag{5.5}$$

where the triggering function g is given by

$$g(t,x,y) = \frac{\alpha\beta}{\pi} \cdot e^{-\alpha t - \beta(x^2+y^2)}, \tag{5.6}$$

where $\frac{\alpha\beta}{\pi}$ is a normalizing constant so that g integrates to unity. One may thus write out the full model as

$$\lambda(t,x,y|H_t) = (1-p)\mu(x,y) + \frac{p\alpha\beta}{\pi} \sum_{\{i:t_i<t\}} e^{-\alpha(t-t_i)-\beta\{(x-x_i)^2+(y-y_i)^2\}}, \tag{5.7}$$

where p represents the expected proportion of events attributed to triggering by other events, rather than occurrences attributable to background immigration.

CHAPTER 6

Analysis of Parametric Model

6.1 Maximum likelihood estimation

For estimation, we proceed in much the same manner as Musmeci and Vere-Jones (1992). We first estimate the non-homogeneous background rate $\mu(x, y)$ via a two-dimensional Guassian kernel smoother over all plant locations, using the default normal reference bandwidth based on Silverman (1986) and described in Venables and Ripley (2002).

The parameters $\boldsymbol{\theta} = \{\alpha, \beta, p\}$ are estimated simultaneously by maximizing the log-likelihood

$$logL = \sum_{i=1}^{n} log\lambda(t_i, x_i, y_i) - \iint_A \int_0^\infty \lambda(t, x, y) dt dx dy \tag{6.1}$$

where standard errors of the parameters may be obtained in standard fashion, via the square root of the diagonal elements of the inverse Hessian of the log-likelihood function (Ogata 1978, see e.g.). The integral in the right side of the

equation simplifies to

$$
\begin{aligned}
&\iint_A \int_0^\infty \lambda(t,x,y)dtdxdy \\
&= \iint_A \int_0^\infty \left[(1-p)\mu(x,y) + \frac{p\alpha\beta}{\pi}\sum_{i=1}^{n} e^{-\alpha(t-t_i)-\beta||x-x_i,y-y_i||^2}\right]dtdxdy \\
&= \iint_A \int_0^\infty (1-p)\mu(x,y)dtdxdy + \iint_A \int_0^\infty \frac{p\alpha\beta}{\pi}\sum_{i=1}^{n} e^{-\alpha(t-t_i)-\beta||x-x_i,y-y_i||^2}dtdxdy \\
&= (1-p)\mathrm{T}\iint_A \mu(x,y)dxdy + p\sum_{i=1}^{n}\iint_A \int_0^\infty \frac{\alpha\beta}{\pi} e^{-\alpha(t-t_i)-\beta||x-x_i,y-y_i||^2}dtdxdy \\
&= (1-p)\mathrm{T}\iint_A \mu(x,y)dxdy + p\mathrm{N}, \qquad (6.2)
\end{aligned}
$$

where T represents the entire time window, A is the space window, and N is the total number of observed plants. The log-likelihood equation becomes

$$logL = \sum_{i=1}^{n} log\lambda(t_i,x_i,y_i) - (1-p)\mathrm{T}\iint_A \mu(x,y)dxdy - p\mathrm{N}. \qquad (6.3)$$

We obtain the estimates $\hat{\boldsymbol{\theta}} = \{\hat{\alpha} = 0.0761, \hat{\beta} = 0.0292, \hat{p} = 0.5767\}$, with corresponding asymptotic standard errors 0.0045, 0.0022 and 0.0193, respectively.

The estimate of p suggests that about 58% of observed plants are triggered by other observed plants while the rest are immigrants triggered by unobserved plants or by other unknown sources. The estimated proportion of background events over time is shown in Figure 6.1; not surprisingly, this proportion is initially very high, as the initial events are necessarily attributed to immigration, rather than triggering, according to the model (5.7).

Note that the estimate of the parameter β governing the exponential spatial decay is very small, suggesting that the clustering nature of these plants is much more uniform and less dramatically clustered and long-tailed compared to the approximately power-law spatial decay of earthquakes and their aftershocks.

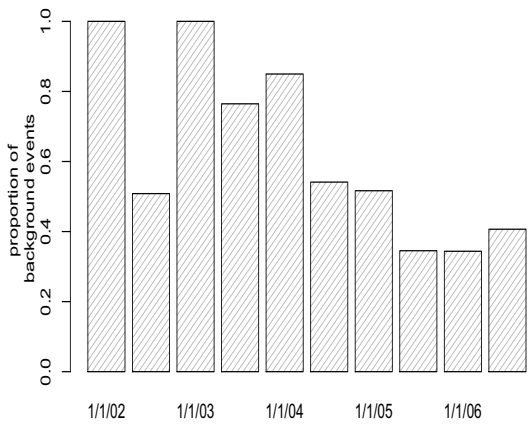

Figure 6.1: Estimated proportion of background events over time.

Although there are no strict rules in choosing an appropriate bandwidth parameter h, the user must find a balance between a value that is too high or too low. This is important as it affects the kernel estimates. If h is chosen too low, the estimate may be too noisy and capture too much information of meaningless clusters that do not generalize well to a random realization of the data. If h is chosen too high, the estimate may be too smooth and information on important features of the data may be lost. The estimated spatial background rate $\hat{\mu}(x, y)$, shown in Figure 6.2, shows the general trend of plants attributed to immigration by model (5.7).

6.2 Errors in estimated birth times

Errors in the birth time estimates appear to have relatively minimal impact on the parameter estimates. For instance, after adding iid normally distributed random errors with mean 0 and standard deviation 4 weeks to the estimated birth times, the estimates of the parameters in model (5.7) remained quite robust. One hun-

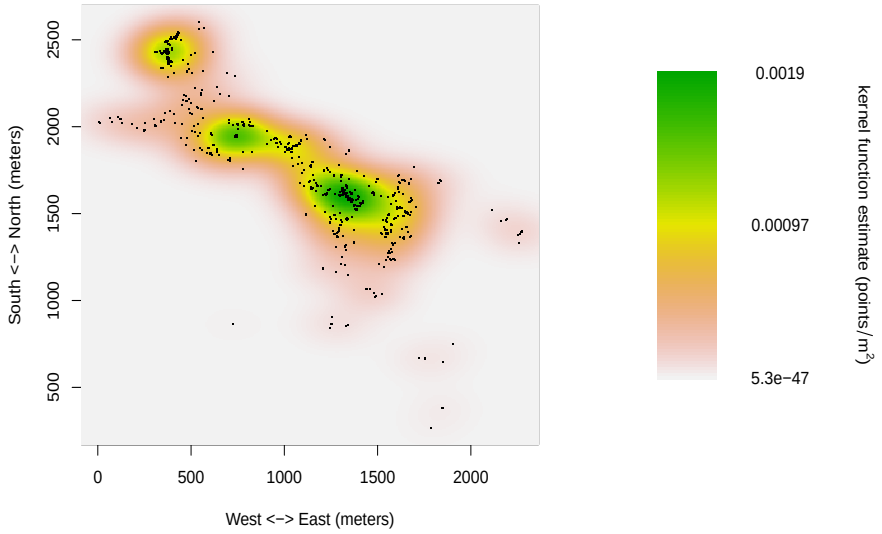

Figure 6.2: Estimated background rate $\hat{\mu}(x, y)$ obtained by kernel smoothing, along with all observed plants (black points).

dred new sets of birth times were simulated and the parameters were re-estimated after each set. Table 6.1 shows summary statistics of the distribution of these parameter estimates, given as $\tilde{\boldsymbol{\theta}} = \{\tilde{\alpha}, \tilde{\beta}, \tilde{p}\}$. The estimates $\tilde{\alpha}$ are expectedly less than $\hat{\alpha}$ because of the increased variability in estimated birth times.

6.3 Residual Analysis

6.3.1 Thinned residuals

To assess the goodness-of-fit of the parametric point process model, one may perform a residual analysis such as the random thinning procedure described in Schoenberg (2003). If the model fits well to the data, the output are thinned residuals that should closely follow a homogeneous Poisson process inside the observation window.

Unfortunately, simple thinning typically has low power when the estimated

	$\tilde{\alpha}$ $(\frac{plants}{week})$	$\tilde{\beta}$ $(\frac{plants}{m^2})$	$\tilde{p}$
Minimum	0.06578	0.02854	0.5751
1st Quartile	0.06830	0.02909	0.5778
Median	0.06900	0.02932	0.5787
Mean	0.06906	0.02931	0.5787
3rd Quartile	0.06986	0.02952	0.5800
Maximum	0.07192	0.03006	0.5821
sd	0.00118	0.00033	0.0016

Table 6.1: Summary statistics of estimated parameters after giving random errors to the original estimated birth times.

conditional intensity is volatile, as is the case here, because few points remain after thinning (Clements et al. 2011). Inversely, a method consisting of superposition alone may have the adverse affect of too many remaining points. A more powerful approach than either thinning or superposition is via an approach that blends both thinning and superposition, where one thins observed points in areas of high intensity and superposes simulated points in areas of low intensity, resulting in a homogeneous point process if the model for λ used in the thinning and superposition is correct. With this hybrid method, called super-thinning by Clements et al. (2011), the user may specify the overall rate of the resulting residual point process.

6.3.2 Super-thinned residuals

In super-thinning, if one wishes the resulting residual process to be Poisson with rate k under the null hypothesis that the estimated conditional intensity of the original process is correct, then one first keeps each observed point (t, x, y) in the catalog independently with probability $\min\{k/\hat{\lambda}(t, x, y), 1\}$ and subsequently

superposes a simulated Poisson process with rate $\max\{0, k-\hat{\lambda}(t,x,y)\}$. The result is a homogeneous Poisson process with rate k if and only if the model $\hat{\lambda}$ for the conditional intensity is correct, and hence the resulting super-thinned residuals can be assessed for homogeneity as a way of evaluating the model. In particular, any clustering or inhibition in the residual points indicates a lack of fit. The parameter k is chosen by the user and governs the extent to which the original points in the data are thinned. If the threshold k is too high, then too many of the super-thinned points are simulated, and thus the resulting residual plot has low power. Similarly, if the threshold is too low, then too few residual points remain after super-thinning, so that again the resulting residuals have low power at detecting deficiencies in the model. Clements et al. (2011) suggest the median of λ as an appropriate choice of k.

The spatial homogeneity of the residual points may be assessed for instance using Ripley's K-function (Ripley 1981), which is the average number of points within r of any given point divided by the overall rate λ, and is typically estimated via

$$\hat{K}(r) = AN^{-2} \sum_{i<j,\ ||x_i - x_j|| < r} s(\mathbf{x}_i, \mathbf{x}_j),$$

where A is the area of the observation region, N is the total number of observed points, and $s(\mathbf{x}_i, \mathbf{x}_j)^{-1}$ is the proportion of area of the ball centered at $\mathbf{x}_i$ and passing through $\mathbf{x}_j$ that falls within the observation region (see Ripley (1981) or Cressie (1993)). For a homogeneous Poisson process in $\mathbb{R}^2$, $K(r) = \pi r^2$, and hence one expects similar behavior for the residuals under the null hypothesis that the model used in rescaling the points is correct.

Figure 6.3 shows the super-thinned residuals from one iteration of the super-thinning procedure. In this particular iteration, 22 newly generated points were superposed while 18 of the original observed points were retained during thinning. Figure 6.3 also shows a summary of the K-functions for 1000 sets of residuals.

The shaded region captures the middle 95% for K-functions corresponding to 1000 realizations of super-thinned residuals, and the dotted lines indicates the middle 95% for K-functions corresponding to 1000 realizations of a true homogeneous Poisson process of the same rate. The shaded region matches closely with the dotted lines, indicating that the model fits well to the data.

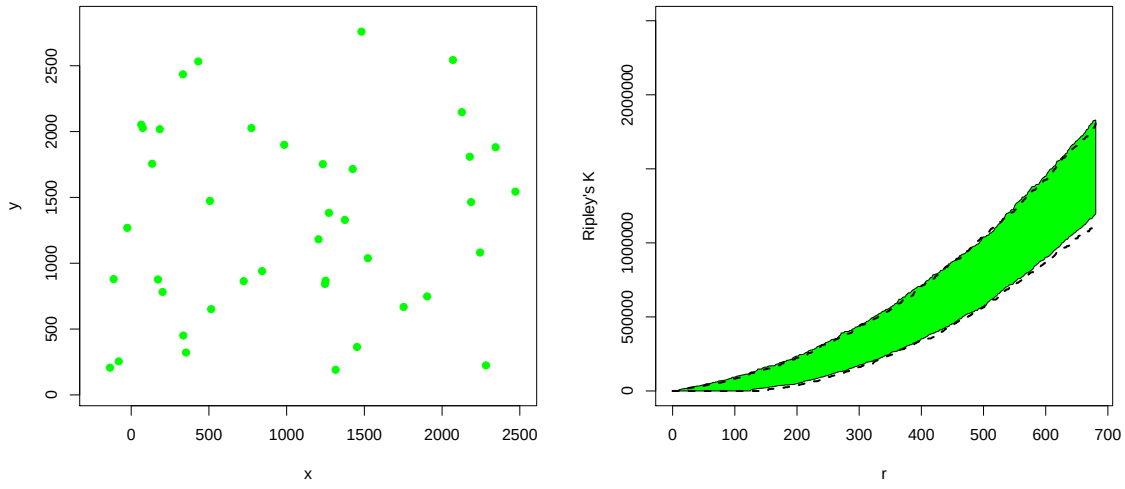

Figure 6.3: Left: One realization of super-thinned residuals with rate $\lambda = 40$. Right: Ripley's K-functions; dotted lines represent 95% interval bounds for homogeneous Poisson processes and shaded region represents 95% interval bounds for super-thinned residuals.

6.4 Simulations

Using the parameters estimated by maximum likelihood, the ETAS model (5.7) for red bananas may be simulated repeatedly. One sample simulation shown in Figure 6.4, along with the observed data, is a realization over the identical spatial range as the data and the same 194-week temporal window. The simulation consists of 783 plants, while the data contain 788 observed plants. Simulations

such as that in Figure 6.4 appear to be consistent with the observed data.

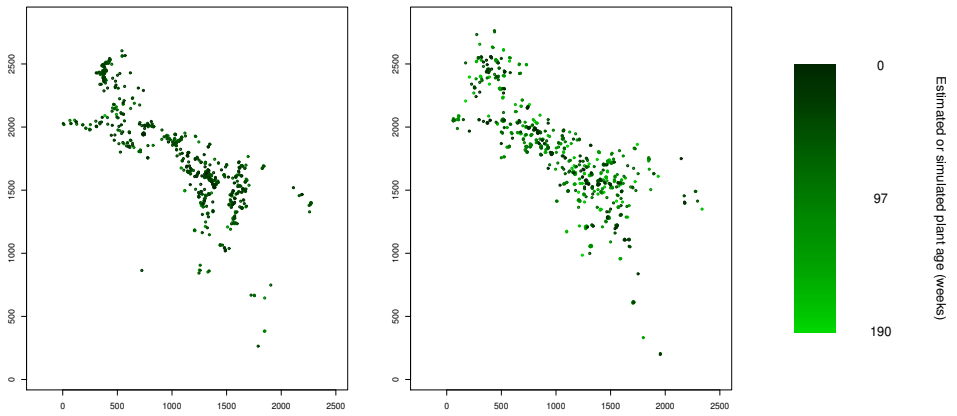

Figure 6.4: Left: Original data points. Right: Simulated points based on model (5.7). Darker circles correspond to more recent points.

6.5 Future range expansion

Simulations of model (5.7) may also be used for forecasting future spread of the red banana plants. From 100 simulations of model (5.7), the mean amount of time of the first occurrence of a red banana plant in each block (or cell) of a 20×20 grid is illustrated in Figure 6.5, with each simulation spanning 100 years. Figure 6.6 gives these values in weeks for each cell. Cells without a value were never invaded during any of the simulations, and thus have mean initial invasion times greater than 100 years (about 5200 weeks). Figure 6.7 shows the mean number of blocks (or pixels) invaded by the plants over time, according to these 100 simulations. One sees from Figures 6.5 and 6.7 that, after the initial explosive burst, the spread becomes more gradual over succeeding years, according to the fitted model. Although simulations are only observed inside the 20×20 grid, in reality, invasion may spread into blocks outside of the observed space window after many years. These figures simply illustrate spread within the observed space, and

the fixed space window may explain the flatter part of the curve to the right of 50 years in Figure 6.7.

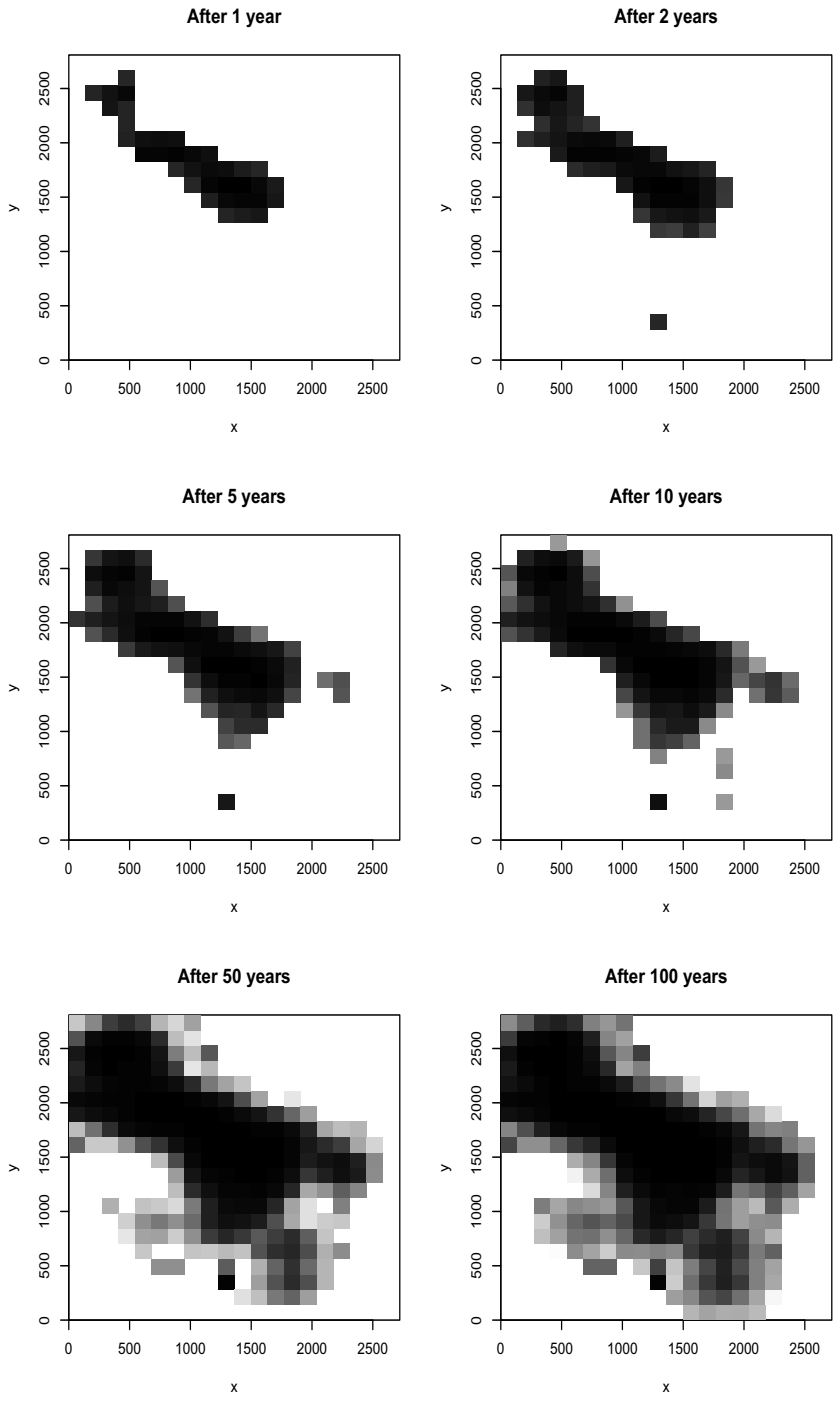

Figure 6.5: Invasion within space window over time. Darker blocks indicates earlier mean time of first invasion.

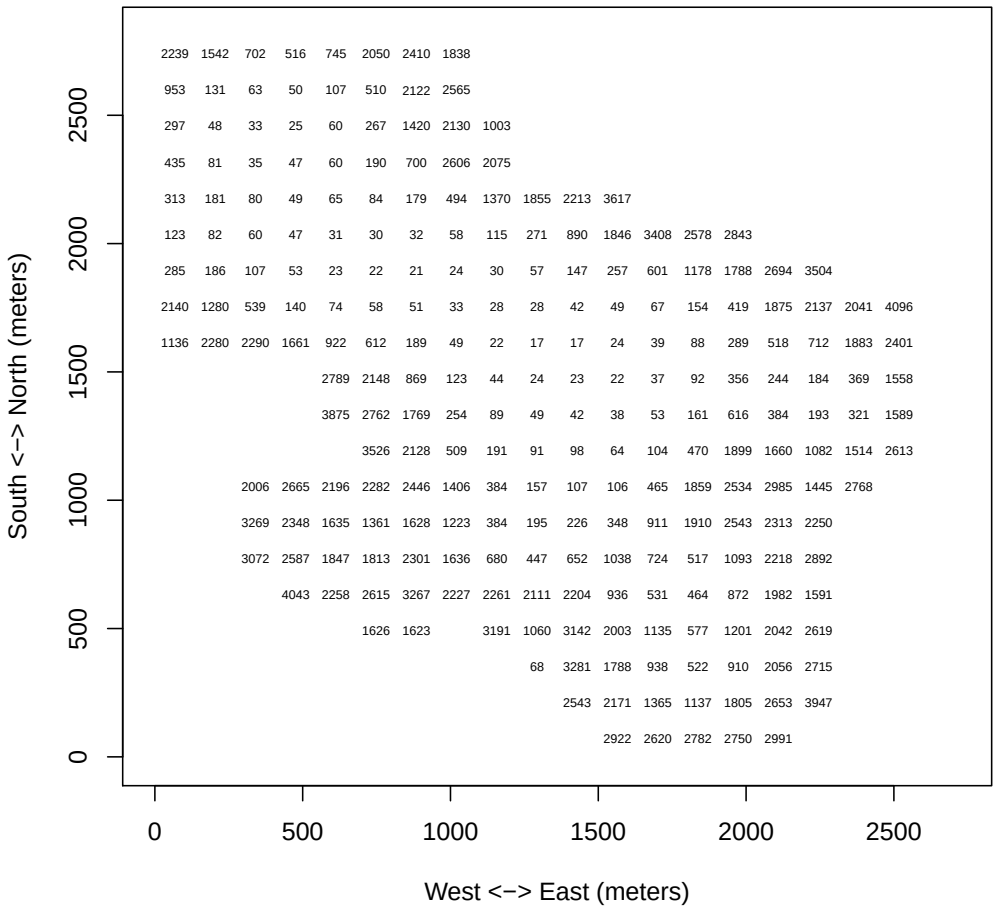

Figure 6.6: Mean number of weeks of first invasion occurrence in a cell. The space window is divided into a 20×20 grid of cells. Cells without a number indicates an average first invasion time of greater than 100 years.

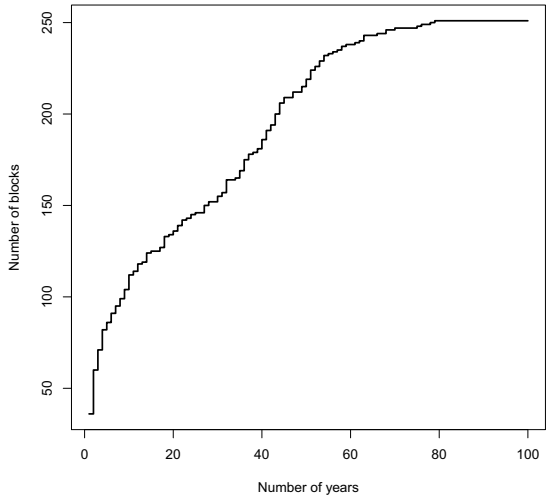

Figure 6.7: Cumulative number of blocks invaded over time. The horizontal axis indicates the number of years after the first simulated plant birth. The space window is divided into a 20×20 grid shown in Figure 6.5, for a total of 400 possible invasions.

CHAPTER 7

Non-parametric Methods

7.1 Stochastic declustering

In seismology, the important problem of distinguishing mainshocks from aftershocks has led to the development of numerous methods of declustering earthquake catalogs (e.g. Reasenberg 1985). The stochastic declustering method of Zhuang et al. (2002; 2004) uses the fitted ETAS model as a starting point, and subsequently computes the probabilities ρ_{ij}, according to the ETAS model, that event j (in the case of Zhuang et al., each event is an earthquake) was triggered by event i, for all i, j with $i < j$. Using this probabilistic branching structure, one may simulate iid Bernoulli draws with the corresponding probabilities ρ_{ij} and assign a branching structure. Some events will not be assigned a trigger and will be labelled background events in such an assignment, and the collection of these events represents a *declustered* catalog. Using the method of Zhuang et al. (2002, 2004), one could in principle construct numerous realizations of declustered catalogs.

7.2 Non-parametric stochastic declustering

Marsan and Lengliné (2008) introduced a stochastic declustering algorithm that requires no a priori model parameterization. This iterative procedure converges quickly to assign triggering probabilities ρ_{ij}, indicating the probability that event i is the actual parent of event j. To review the method of Marsan and Lengliné

(2008), the algorithm may be outlined as follows:

Given a linear conditional intensity function in space-time-magnitude,

$$\lambda(\mathbf{x}, t, m) = \mu + \sum_{t_i<t} \lambda_i(|\mathbf{x} - \mathbf{x}_i|, t - t_i, m_i), \tag{7.1}$$

and assuming that the mean-field response solely depends on magnitude, initialize the algorithm with guesses for the triggering rates $\lambda_i(|\mathbf{x} - \mathbf{x}_i|, t - t_i, m_i)$ and the uniform background rate μ. Then repeat the following two steps until convergence is reached:

I. Estimate the triggering weights (triggering probabilities)

$$\hat{\rho}_{ij} = \begin{cases} a_j \lambda(|\mathbf{x}_j - \mathbf{x}_i|, t_j - t_i, m_i), & \text{if } t_i < t_j \\ 0, & \text{otherwise} \end{cases} \tag{7.2}$$

and the background weights (background probabilities)

$$\hat{\rho}_{0j} = a_j \mu \tag{7.3}$$

where the normalization coefficients a_j are such that

$$\hat{\rho}_{0j} + \sum_{i=1}^{j-1} \hat{\rho}_{ij} = 1 \tag{7.4}$$

II. Update the "a posteriori" triggering rates by computing

$$\lambda(|\Delta\mathbf{x}|, \Delta t, m) = \frac{1}{N_m \cdot \delta t \cdot S(|\Delta\mathbf{x}|, \delta r)} \sum_{i,j \in A} \hat{\rho}_{ij} \tag{7.5}$$

where A is the set of pairs such that $|\mathbf{x}_j - \mathbf{x}_i| = |\Delta\mathbf{x}| \pm \delta r$, $m_i = m \pm \delta m$, and $t_j - t_i = t \pm \delta t$, with discretization parameters $\delta r, \delta t, \delta m$. N_m is the number of events such that $m_i = m \pm \delta m$, and $S(|\Delta\mathbf{x}|, \delta r)$ is the spatial disk of radii $|\Delta\mathbf{x}| \pm \delta r$.

Update the "a posteriori" background rate by computing

$$\mu = \frac{1}{T \cdot S} \sum_{j=1}^{N} \hat{\rho}_{0j} \tag{7.6}$$

where $T \cdot S$ is the space-time window containing all N events.

7.3 Modified non-parametric algorithm for invasive plant species

A non-parametric method is especially beneficial to the problem of declustering a sequence of invasive species, since different species may have very different branching patterns. Here, we apply the non-parametric stochastic declustering method of Marsan and Lengliné (2008) to the red banana data and compare the estimated triggering function with that of the parametric ETAS model. The general algorithm of Marsan and Lengliné (2008) must be modified in two ways, however, in order to accommodate the red banana data. First, since the analysis of the spread of red bananas contains no magnitude information, all dependencies on magnitude are removed. Second, since the background rate is spatially inhomogeneous, the background term μ is given as the product of a constant term μ_t and the local density $\mu_s(\mathbf{x})$, conditioned on position relative to the background events.

Given a linear conditional intensity function in space-time,

$$\lambda(\mathbf{x}, t) = \mu(\mathbf{x}) + \sum_{t_i<t} \lambda_i(|\mathbf{x} - \mathbf{x}_i|, t - t_i), \tag{7.7}$$

and initial guesses for the triggering rates $\lambda_i(|\mathbf{x} - \mathbf{x}_i|, t - t_i)$ and the background rate $\mu(\mathbf{x})$, the following two steps are repeated until convergence is reached:

I. Estimate the triggering probabilities

$$\hat{\rho}_{ij} = \begin{cases} a_j \lambda(|\mathbf{x}_j - \mathbf{x}_i|, t_j - t_i), & \text{if } t_i < t_j \\ 0, & \text{otherwise} \end{cases} \tag{7.8}$$

and the background probabilities

$$\hat{\rho}_{0j} = a_j \mu(\mathbf{x}_j) \tag{7.9}$$

where the normalization coefficients a_j are such that

$$\hat{\rho}_{0j} + \sum_{i=1}^{j-1} \hat{\rho}_{ij} = 1 \tag{7.10}$$

II. Update the "a posteriori" triggering rates by computing

$$\lambda(|\Delta\mathbf{x}|, \Delta t) = \frac{1}{\delta t \cdot S(|\Delta\mathbf{x}|, \delta r)} \sum_{i,j \in A} \hat{\rho}_{ij} \tag{7.11}$$

where A is the set of pairs such that $|\mathbf{x}_j - \mathbf{x}_i| = |\Delta\mathbf{x}| \pm \delta r$ and $t_j - t_i = t \pm \delta t$, with discretization parameters δr and δt. $S(|\Delta\mathbf{x}|, \delta r)$ is the spatial disk of radii $|\Delta\mathbf{x}| \pm \delta r$.

Update the "a posteriori" background rate by computing $\mu(\mathbf{x}) = \mu_t \cdot \mu_s(\mathbf{x})$ using the following substeps:

Compute

$$\mu_t = \frac{1}{T} \sum_{j=1}^{N} \hat{\rho}_{0j} \tag{7.12}$$

and

$$\mu_s(\mathbf{x}) = \sum_{j=1}^{N} h_j(|\Delta\mathbf{x}|) \hat{\rho}_{0j} \tag{7.13}$$

where

$$h_j(|\Delta\mathbf{x}|) = \frac{\sum_{i,j \in A} \hat{\rho}_{0i} \hat{\rho}_{0j}}{S(|\Delta\mathbf{x}|, \delta r) \sum_{i,j} \hat{\rho}_{0i} \hat{\rho}_{0j}} \tag{7.14}$$

where T is the duration of the time series containing all N events, and A is the set of pairs such that $|\mathbf{x}_j - \mathbf{x}_i| = |\Delta\mathbf{x}| \pm \delta r$.

7.4 Results

Compared with parametric estimates, the non-parametric estimate of (7.7) appears to be similar in both temporal and spatial branching structures. The similarity in temporal clustering is apparent in Figure 7.1, which compares the survival plots of the distribution of parent-child inter-event times obtained from both the non-parametric and parametric models. The branching structure obtained from the non-parametric model follows the parametric model quite closely up to about a 26-week time distance between pairs of plants, during which approximately 84% of first-generation triggered events occur. The triggering becomes slightly more

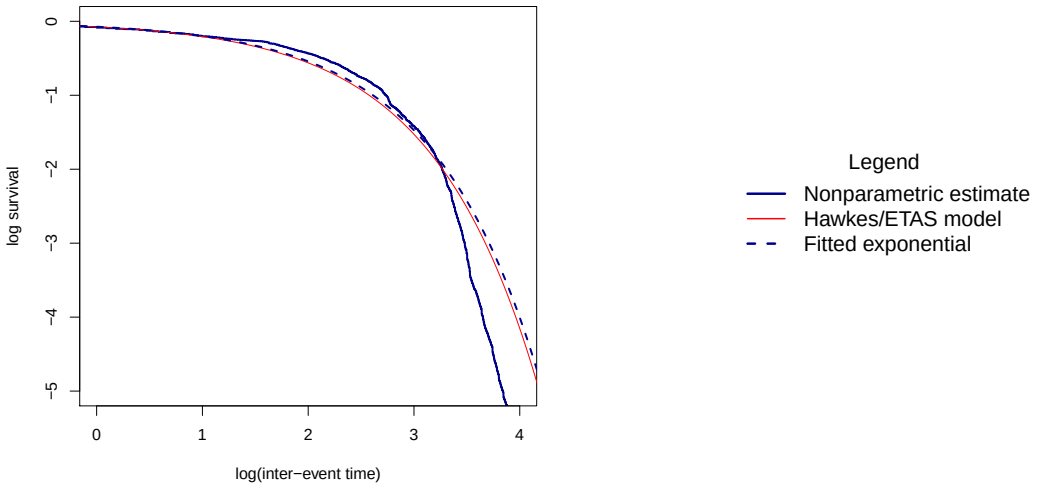

Figure 7.1: Log-survival function of inter-event times between pairs of plants. The solid blue curve represents the cumulative sum of the probabilities ρ_{ij} for the specified inter-event time after applying the non-parametric method discussed in Section 7.3. The red Hawkes/ETAS curve is the exponential decay function with rate 0.0761, obtained from the parametric estimate of the branching model. The dashed blue curve corresponds to an exponential distribution, fitted to the non-parametric estimates, with rate 0.0734.

intense for offspring occurrences beyond 26 weeks of the parent, resulting in an accelerated decay in the temporal distribution of pair distances. More than 99% of first-generation triggered events occur within 43 weeks of the parent according to the non-parametric model, compared to 60 weeks from the parametric model.

The spatial clustering estimated by the non-parametric estimates of the triggering function indicate a higher degree of clustering at smaller inter-event distances, compared with the histogram of inter-event distances of observed pairs of plants used for the parametric model formulation. According to the parametric estimates, approximately 60% of first-generation triggered events occur within 30

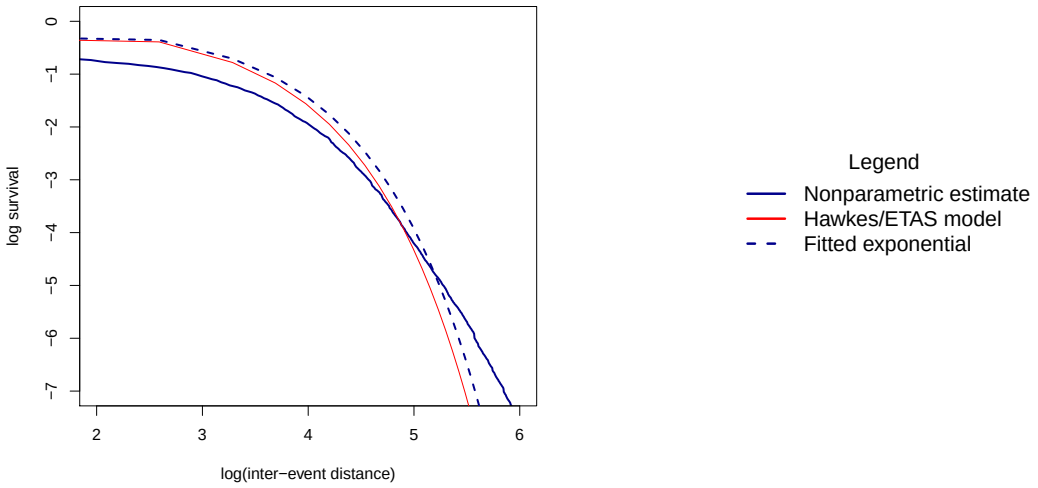

Figure 7.2: Log-survival functions of inter-event distances between pairs of plants. The solid blue curve represents the cumulative sum of the probabilities ρ_{ij} for the specified inter-event distance, after applying the non-parametric model discussed in Section 7.3. The red Hawkes/ETAS curve is the exponential decay function with rate 0.0292, obtained from the parametric estimate of the branching model. The dashed blue curve corresponds to an exponential distribution, fitted to the non-parametric estimates, with rate 0.0265.

meters from the parent and over 99% occur within 158m. On the other hand, our non-parametric model estimates 60% of triggered events occur within just 15 meters away from the parent, and 99% within 170m. Figure 7.2 shows this slightly slower decay in the spatial distribution of pair distances after an initial burst of triggering at very small distances.

The non-parametric estimates indicate that the clustering of the red banana plants is relatively more confined to smaller spatial scales. Indeed, the non-parametric estimates identify 31% (standard error = 1.97%) of the data as background events and 69% (standard error = 2.97%) as triggered events, compared to

42% and 58% for background and triggered events, respectively, according to the parametric estimates of the modified ETAS model. Since the total rate, integrated over the entire space-time window, is approximately n, the standard error is computed by assuming that the total number of triggered events is approximately a Poisson random variable with mean $0.69n$ and standard deviation $\sqrt{0.69n}$. The standard error for the triggering rate is thus $\sqrt{0.69n}/n$ or $\sqrt{0.69/n}$. Similarly, the standard error for the background rate is computed as $\sqrt{0.31/n}$.

The probabilities ρ_{ij}, giving the probability that event i triggered event j, can be used to infer, probabilistically, the location of events relative to their parent. From Figure 7.3, triggered events are more likely to occur very close to the parent's location, and have an overall tendency to appear towards the northwest and southeast directions. The anisotropy exhibited here is likely attributed to the Rio Puerto Viejo running along this direction, which is an obvious potential mechanism for seed dispersal.

The probabilities ρ_{0j}, giving the probability that event j is a background event, can be used to find the most likely locations of background, or immigrant, events. Figure 7.4 particularly highlights several distinct locations of high background activity, especially in the northwest and central portions of the observation region. These are high likelihood targets for where the red banana plants may be immigrating into the region, most likely transported by birds and other animals; this is consistent with previous observations in Figure 6.2.

Figures 7.5 and 7.6 show histograms of the inter-event times and inter-event distances, respectively, when the non-parametric estimator described in Section 7.3 is used to decluster the data probabilistically. That is, according to these non-parametric estimates, each observed plant is assigned some probability of being an immigrant as well as a probability of being triggered by plant i, for each plant i originating before it, and one may translate these probabilities into a sequence of immigrant and triggered plants as in Zhuang et al. (2002). When one limits

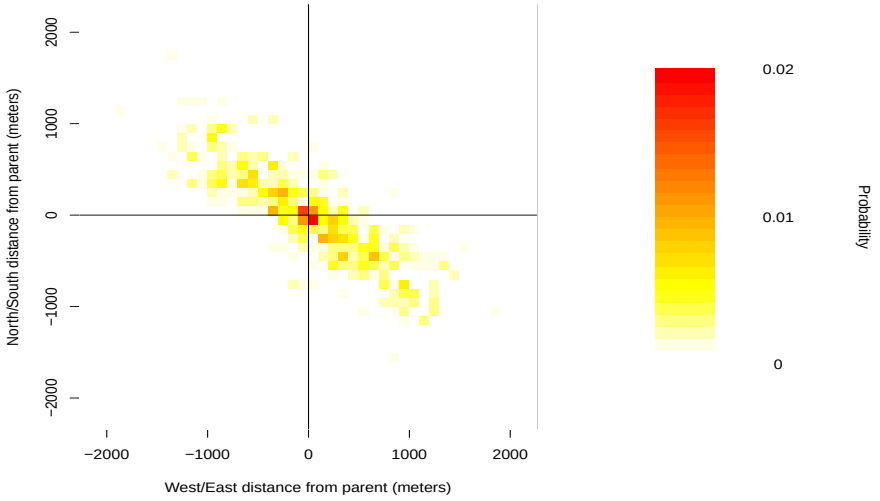

Figure 7.3: Locations of triggered plants relative to their parents, weighted by the likelihood of the parent-offspring relationship estimated by the non-parametric estimator. The center coordinate (0,0) marks the location of a parent, and a darker red cell indicates a higher density of first-generation offspring in that cell. Each cell is 100×100 square meters.

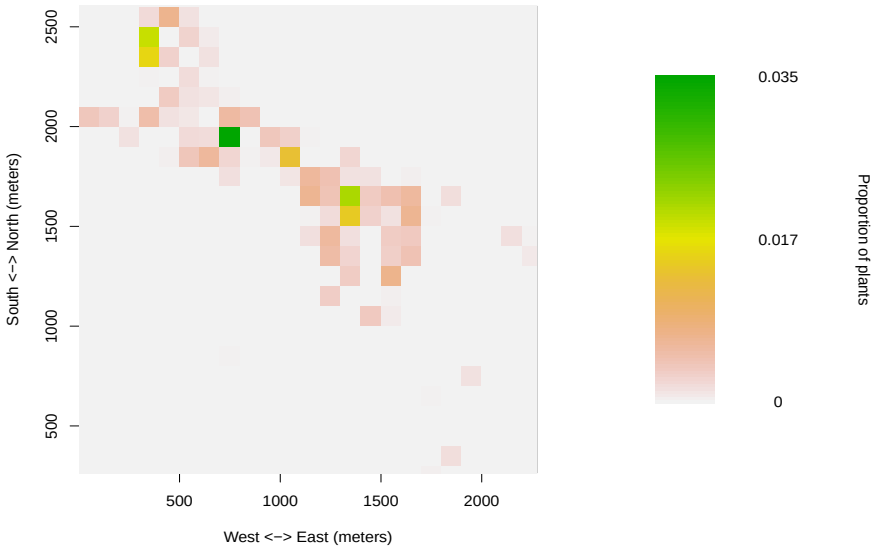

Figure 7.4: Locations of immigrant plants, weighted by the likelihood of each plant being an immigrant as estimated by the non-parametric estimator. Darker green cells indicate a higher probability of observing an immigrant plant. Each cell is 100×100 square meters.

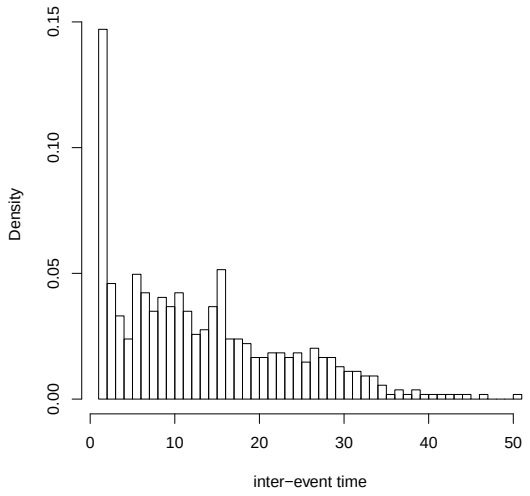

Figure 7.5: Histogram of inter-event times from one realization of a declustered sequence of red banana plants, using the triggering probabilities ρ_{ij} obtained from the non-parametric method discussed in Section 7.3.

attention purely to the sequence of immigrant plants in one such realization, the inter-event times and distances are clearly far less clustered than the raw data, as evidenced in Figures 7.5 and 7.6.

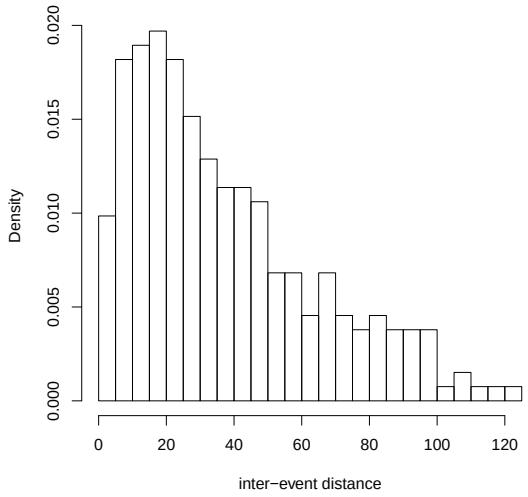

Figure 7.6: Histogram of inter-event distances from one realization of a declustered sequence of red banana plants, using the triggering probabilities ρ_{ij} obtained from the non-parametric method discussed in Section 7.3.

CHAPTER 8

Discussion

While there have been many attempts to model invasive plant spread, methods treating the data as a spatial-temporal point pattern have been elusive. The advantage of point pattern analysis is that it allows one to analyze and simulate the exact locations of individual plants, and to estimate conditional intensities and branching structure in greater detail.

The ETAS model has been widely used to study the spatial and temporal spread of earthquakes and aftershocks. In this dissertation, we were able to use ETAS in an ecological setting and apply the model to an invasive species of red banana plants. The ETAS model is a branching point process model that requires the exact locations of the points, as well as the origin times of the points. Unlike earthquakes, however, the origin or birth times of the red banana plants considered here are generally unknown. While they may be estimated using the weekly height measurements of the plants, the birth times of the plants are likely estimated with substantial error, and the improved estimation of these birth times is an important topic for future consideration. We showed, however, that the parameter estimates remained quite robust even after adding a random error with mean 0 and standard deviation 4 weeks to the estimated birth times.

Maximum likelihood estimates are commonly used due to their nice asymptotic properties (Ogata 1978), but sometimes can be very difficult to calculate if there is no closed form solution to the likelihood equation. Thus, one must typically use numerical optimization methods. However, there can be convergence

problems in the estimation of the parameters if the log-likelihood is extremely flat. Veen and Schoenberg (2008) showed that the EM-type algorithm may be more stable and robust than Newton-Raphson methods in the face of extreme flatness of the log-likelihood function, which may occur when estimating models with large numbers of parameters and when data are sparse. In the present study, the dataset appears to be sufficiently large to estimate the three parameters in the given model stably, but in the future, if more complex models are desired, the EM-method of estimation may be of interest.

After specifying the model (5.7) and presenting the model parameters from the likelihood estimation, we were able to simulate point patterns that appear to agree with the original data. Note that the background rate and clustering density estimates reported here may be slightly underestimated because of missing data. In particular, since missing plants are potentially more likely to occur disproportionately in locations further away from those where plants were observed, one may conjecture that in particular the parameters β and p may be slightly overestimated in the present study. The result of the super-thinning procedure and analysis of the K-functions, however, implies that the model may be slightly under-predicting the clustering in the data. Further development of the model and comparison to other models of invasive species, in terms of estimation and predictive power, will be examined in future work.

It is important to note that the ETAS model attributes all clustering of the red banana plants in space and time to a birth mechanism where one plant essentially causes another to occur nearby. However, it is plausible that much of the clustering is due to spatial inhomogeneity, since the plants are more likely to occur in some locations than others due to the compatibility with the plant's survival needs, or due to environmental covariates excluded from the model such as distance from rivers, sunlight, soil quality, elevation, and temperature. More realizations of the process would need to be observed in order for one to discriminate between

the clustering due to the red banana plants reproducing and that due to spatial inhomogeneity and excluded variables, and the measurement and inclusion of some of these excluded variables into an ETAS model is also another important subject for future research. Unlike earthquakes, red banana plants are affected greatly by human interactions, since human decisions are not only responsible for the sighting and recording of the plants, but also the decisions about the termination of plants and about ecological and forest management decisions that may affect the reproduction and longevity of the plants, and this further complicates the interpretation of the fitted parameters in the ETAS model. Note also that the ETAS model is by assumption stationary in time, meaning that while clusters of plants (or earthquakes) can occur, in the long run the process is neither explosive (with a number of occurrences per spatial-temporal unit approaching infinity) nor does the process terminate entirely. This seems reasonable in the case of earthquakes, but is perhaps objectionable in the case of a particular species of plant, especially one that spreads as robustly as the red banana plant. A more plausible model for red banana plants might possibly have a variable or even finite immigration process and may have a triggering density that is explosive in the short-term and that decays over time, ultimately becoming sub-critical, since otherwise these plants would be located everywhere. On the other hand, estimating such a model using observations over only a short time span may be difficult. The extension to non-stationary point process models that alternate between sub- and super-criticality is an important subject for future research. While the model presented here seems to adequately describe the behavior of the red banana plant over the time period considered, non-stationary models might be preferable given observations over longer time spans. Methods for the estimation of non-stationary or covariate-influenced versions of ETAS models are important methods for future research, as such methods appear to be lacking in the literature at present. Nevertheless, for the space-time window considered here, our

results can perhaps be interpreted as reasonable summaries of the spatial-temporal spreading characteristics of the red banana plant in Costa Rica.

The space-time model presented here for invasive plants does not contain marks or magnitude information, unlike the ETAS model commonly used to describe earthquakes. Recorded covariates for each plant, including number of flowers, amount of sunlight, distance from a main water source, soil quality, elevation and temperature were excluded from model formulation mainly due to the very large amount of missing data or poor quality measurements. The inclusion of such covariates into branching models is important for future research, however. In the present analysis, we consider purely the branching structure of invasive plants using only the precise locations and estimated origin times of the plants, and our focus is mainly on comparing the behavior and performance of current parametric and non-parametric estimates.

One purpose of such a non-parametric approach is that it may possibly give insight to or justify an underlying parametric model previously used to study plant spread activity and seed dispersal. Here, however, we find important differences between the non-parametric estimates and the parametric conditional intensity estimates. In particular, the clustering appears to be stronger and sharper, i.e., more intense at smaller spatial-temporal scales, compared with the parametric estimates used previously. Such parametric estimates were based largely on examination of histograms of inter-event times and distances between *all* observed plants. The current results show that such histograms can result in a significant underestimation of clustering. We also find that red banana plants not only occur predominantly along the diagonal running from the northwest to the southeast portion of the observation region, but that in addition, the triggering of offspring appears to run predominantly in this direction as well. The results here make sense in light of the Rio Puerto Viejo running along this direction, and also may be used to obtain estimates of the most likely targets where immigration of red

banana plants may be occurring.

Note that the model (4.2) assumes that the triggering function g contributes linearly to the total conditional intensity λ. Multiplicative and other non-linear forms may be important considerations in future work.

Bibliography

Begon, M., Townsend, C. R., and Harper, J. L. (2006), *Ecology: From Individuals to Ecosystems*, Blackwell Publishing, 4th ed.

Chu, A., Schoenberg, F. P., Bird, P., Jackson, D. D., and Kagan, Y. Y. (2011), "Comparison of ETAS Parameter Estimates across Different Global Tectonic Zones," *Bulletin of the Seismological Society of America*, 101, 2323–2339.

Clements, R., Schoenberg, F. P., and Veen, A. (2011), "Evaluation of space-time point process models using super-thinning," In review.

Colautti, R., Bailey, S., van Overdijk, C., Amundsen, K., and MacIsaac, H. (2006), "Characterised and Projected Costs of Nonindigenous Species in Canada," *Biological Invasions*, 8, 45–59, 10.1007/s10530-005-0236-y.

Console, R., Murru, M., and Falcone, G. (2010), "Probability gains of an epidemic-type aftershock sequence model in retrospective forecasting of M $\geq$ 5 earthquakes in Italy," *Journal of Seismology*, 14, 9–26, 10.1007/s10950-009-9161-3.

Cressie, N. (1993), *Statistics for Spatial Data, Revised ed.*, New York: Wiley.

Cressie, N. and Wikle, C. K. (2011), *Statistics for spatio-temporal data*, Oxford: Wiley.

Daley, D. and Vere-Jones, D. (2003), *An Introduction to the Theory of Point Processes*, Springer, 2nd ed.

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.

Diggle, P. and Zheng, P. (2005), "Nonparametric estimation of spatial segregation

in a multivariate point process: bovine tuberculosis in Cornwall, UK," *Applied Statistics*, 54, 645–658.

Diggle, P. J. (2006), "Spatio-temporal point processes, partial likelihood, foot and mout disease," *Statistical Methods in Medical Research*, 15, 325–336.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. (1996), "Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology," *Transactions of the Institute of British Geographers*, 21, pp. 256–274.

Gosper, C. R., Stansbury, C. D., and Vivian-Smith, G. (2005), "Seed dispersal of fleshy-fruited invasive plants by birds: contributing factors and management options," *Diversity and Distributions*, 11, 549–558.

Gutenberg, B. and Richter, C. F. (1944), "Frequency of Earthquakes in California," *Bulletin of the Seismological Society of America*, 34, 185–188.

Halpern, C. B., Antos, J. A., Rice, J. M., Haugo, R. D., and Lang, N. L. (2010), "Tree invasion of a montane meadow complex: temporal trends, spatial patterns, and biotic interactions," *Journal of Vegetation Science*, 21, 717–732.

Harte, D. (2012), "Bias in fitting the ETAS model: a case study based on New Zealand seismicity," *Geophysical Journal International*, to appear.

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

Helmstetter, A. and Sornette, D. (2003), "Predictability in the Epidemic-Type Aftershock Sequence model of interacting triggered seismicity," *J. Geophys. Res.*, 108.

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

Howe, H. F. and Westley, L. C. (2009), *Ecology of Pollination and Seed Dispersal*, Blackwell Publishing Ltd., chap. 9, 2nd ed., pp. 262–283.

Keeling, M. J. (2005), "Models of foot-and-mouth disease," *Proceedings of The Royal Society B*, 272, 1195–1202.

Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornwell, S. J., Kappey, J., Wilesmith, J., and Grenfell, B. T. (2001), "Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape," *Science*, 294.

Law, R., Illian, J., Burslem, D., Gratzer, G., Gunatilleke, C., and Gunatilleke, I. (2009), "Ecological information from spatial patterns of plants: insights from point process theory," *The Journal of ecology*, 97.

Lonsdale, W. M. (1993), "Rates of Spread of an Invading Species–Mimosa Pigra in Northern Australia," *Journal of Ecology*, 81, pp. 513–521.

Marco, D. E., Montemurro, M. A., and Cannas, S. A. (2011), "Comparing short and long-distance dispersal: modelling and field case studies," *Ecography*, 34, 671–682.

Marsan, D. and Lengliné, O. (2008), "Extending Earthquakes' Reach Through Cascading," *Science*, 319, 1076–1079.

Moller, J. and Waagepetersen, R. P. (2004), *Statistical Inference and Simulation for Spatial Point Processes*, Chapman and Hall/CRC.

Musmeci, F. and Vere-Jones, D. (1992), "A space-time clustering model for historical earthquakes," *Annals of the Institute of Statistical Mathematics*, 44, 1–11.

Ogata, Y. (1978), "The asymptotic behavior of maximum likelihood estimators for stationary point processes," *Annals of the Institute of Statistical Mathematics*, 30, 243–261.

— (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.

Ogata, Y., Jones, L. M., and Toda, S. (2003), "When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern California," *Journal of Geophysical Research (Solid Earth)*, 108, 2318.

Pejchar, L. and Mooney, H. A. (2009), "Invasive species, ecosystem services and human well-being," *Trends in Ecology and Evolution*, 24, 497–504.

Perrins, J., Fitter, A., and Williamson, M. (1993), "Population Biology and Rates of Invasion of Three Introduced Impatiens Species in the British Isles," *Journal of Biogeography*, 20, pp. 33–44.

Perry, G. L. W., Miller, B. P., and Enright, N. J. (2006), "A comparison of methods for the statistical analysis of spatial point patterns in plant ecology," *Plant Ecology*, 187, 59–82, 10.1007/s11258-006-9133-4.

Peters, D. P. C. (2004), "Selection of models of invasive species dynamics," *Weed Technology*, 18, 1236–1239.

Pimentel, D., Pimentel, M., and Wilson, A. (2007), *Plant, Animal, and Microbe Invasive Species in the United States and World*, vol. 193 of *Ecological Studies*, Springer Berlin Heidelberg.

Pimentel, D., Zuniga, R., and Morrison, D. (2005), "Update on the environmental and economic costs associated with alien-invasive species in the United States," *Ecological Economics*, 52, 273 – 288, integrating Ecology and Economics in Control Bioinvasions.

Pysek, P. and Hulme, P. E. (2005), "Spatio-temporal dynamics of plant invasions: Linking pattern to process," *Ecoscience*, 12, 302–315.

Pysek, P. and Prach, K. (1993), "Plant Invasions and the Role of Riparian Habitats: A Comparison of Four Species Alien to Central Europe," *Journal of Biogeography*, 20, 413–420.

Reasenberg, P. (1985), "Second-Order Moment of Central California Seismicity, 1969-1982," *Journal of Geophysical Research*, 90, 5479–5495.

Rice, J., Halpern, C., Antos, J., and Jones, J. (2012), "Spatio-temporal patterns of tree establishment are indicative of biotic interactions during early invasion of a montane meadow," *Plant Ecology*, 213, 555–568, 10.1007/s11258-012-0021-9.

Riley, S. (2007), "Large-Scale Spatial-Transmission Models of Infectious Disease," *Science*, 316.

Ripley, B. (1981), *Spatial Statistics*, New York: Wiley.

Schoenberg, F. P. (2003), "Multi-dimensional residual analysis of point process models for earthquake occurrences," *Journal of the American Statistical Association*, 98, 789–795.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1st ed.

Sinnott, R. W. (1984), "Virtues of the Haversine," *skytel*, 68, 158.

Skellam, J. G. (1951), "Random Dispersal in Theoretical Populations," *Biometrika*, 38, pp. 196–218.

Sornette, D. and Werner, M. J. (2005), "Apparent clustering and apparent background earthquakes biased by undetected seismicity," *Journal of Geophysical Research (Solid Earth)*, 110, 9303.

Stoyan, D. and Penttinen, A. (2000), "Recent applications of point process methods in forestry statistics," *Statistical Science*, 15, 61–78.

Thompson, J. D. (1991), "The Biology of an Invasive Plant," *BioScience*, 41, 393–401.

Tiampo, K. F. and Shcherbakov, R. (2012), "Seismicity-based earthquake forecasting techniques: Ten years of progress," *Tectonophysics*, 522–523, 89 – 121.

Utsu, T. (1961), "Statistical study on the occurrence of aftershocks," *The Geophysical Magazine*, 521–605.

Veen, A. and Schoenberg, F. P. (2008), "Estimation of Space-Time Branching Process Models in Seismology Using an EM-Type Algorithm," *Journal of the American Statistical Association*, 103.

Venables, W. and Ripley, B. (2002), *Modern Applied Statistics with S*, Springer, 4th ed.

Vere-Jones, D. (2009), "Some models and procedures for space-time point processes," *Environmental and Ecological Statistics*, 16, 173–195, 10.1007/s10651-007-0086-0.

Vere-Jones, D. and Zhuang, J. (2008), "Distribution of the largest event in the critical epidemic-type aftershock-sequence model," *Physical Review E*, 78, 047102.

Wang, H.-H., Grant, W. E., Swannack, T. M., Gan, J., Rogers, W. E., Koralewski, T. E., Miller, J. H., and Taylor, J. W. (2011a), "Predicted range expansion of Chinese tallow tree (Triadica sebifera) in forestlands of the southern United States," *Diversity and Distributions*, 17, 552–565.

Wang, Q., Jackson, D. D., and Kagan, Y. Y. (2011b), "California Earthquake Forecasts Based on Smoothed Seismicity: Model Choices," *Bulletin of the Seismological Society of America*, 101, 1422–1430.

Werner, M. J., Helmstetter, A., Jackson, D. D., and Kagan, Y. Y. (2011), "High-Resolution Long-Term and Short-Term Earthquake Forecasts for California," *Bulletin of the Seismological Society of America*, 101, 1630–1648.

Wiegand, T. and Moloney, K. A. (2004), "Rings, circles, and null-models for point pattern analysis in ecology," *Oikos*, 104, 209–229.

Zhuang, J. (2011), "Next-day earthquake forecasts for the Japan region generated by the ETAS model," *Earth, Planets, and Space*, 63, 207–216.

Zhuang, J. and Ogata, Y. (2004), "Analyzing earthquake clustering features by using stochastic reconstruction," *Journal of Geophysical Research*, 109.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002), "Stochastic Declustering of Space-Time Earthquake Occurrences," *Journal of the American Statistical Association*, 97, 369–380.