Bayesian Estimation for Percolation Models of Disease Spread in Plant Populations

G. J. Gibson[1], W. Otten[2], J. A. N. Filipe[2,3], A. Cook[1, 4], G. Marion[4], C. A. Gilligan[2]

[1]Department of Actuarial Mathematics & Statistics, Heriot-Watt University, Riccarton, Edinburgh, EH14 4AS, UK

[2]Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK

[3]Current address: Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG

[4]Biomathematics & Statistics Scotland, James Clerk Maxwell Building, The King's Buildings, Edinburgh EH9 3JZ

Corresponding author:

Professor G J Gibson, Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS. Fax:+44 131 451 3249

e-mail: gavin@ma.hw.ac.uk

Keywords: spatio-temporal modeling, stochastic modeling, fungal pathogens, Bayesian inference, Markov chain Monte Carlo

AMS subject classification: 62P10, 62M99

Abstract

Statistical methods are formulated for fitting and testing percolation-based, spatio-temporal models that are generally applicable to biological or physical processes that evolve in spatially distributed populations. The approach is developed and illustrated in the context of the spread of *Rhizoctonia solani*, a fungal pathogen, in radish but is readily generalized to other scenarios. The particular model considered represents processes of primary and secondary infection between nearest-neighbour hosts in a lattice, and time-varying susceptibility of the hosts. Bayesian methods for fitting the model to observations of disease spread through space and time in replicate populations are developed. These use Markov chain Monte Carlo methods to overcome the problems associated with partial observation of the process. We also consider how model testing can be achieved by embedding classical methods within the Bayesian analysis. In particular we show how a residual process, with known sampling distribution, can be defined. Model fit is then examined by generating samples from the posterior distribution of the residual process, to which a classical test for consistency with the known distribution is applied, enabling the posterior distribution of the P-value of the test used to be estimated. For the *Rhizoctonia*-radish system the methods confirm the findings of earlier non-spatial analyses regarding the dynamics of disease transmission and yield new evidence of environmental heterogeneity in the replicate experiments.

1. Introduction

In the literature of spatio-temporal modeling there is a large body of work relating to the broad area of percolation. This is a natural approach for representing biological or physical processes that evolve through a spatially dispersed population when the dynamics of the process are determined by local interactions. It has been applied to *inter alia* the spread of forest fires (Cox & Durrett, 1981), the dynamics of interreacting species in ecosystems (Wissel, 2000) and to the spread of infectious diseases (Kuulasmaa, 1982; Bailey *et al.* 2000; Otten *et al.* 2004). It seems particularly suitable for the systems of study in this paper which deal with host-pathogen interactions in managed botanical populations that may naturally be represented on a lattice.

Research on percolation tends to focus on what may loosely be described as emergent behaviour. This deals with the large-scale and long-term, asymptotic properties of percolation processes, for example deriving critical values of parameters at phase transitions between extinction and persistence of a process. In the study of epidemic processes this work is important for characterizing conditions under which a disease may be expected to invade a population, or determining the level of control required to ensure ultimate extinction.

In order to inform studies that attempt to predict the dynamics of a system using percolation models it is necessary to obtain statistically valid estimates of key parameters. There is, however, relatively little research into statistical methodology for parameter estimation in percolation models. A common approach when modeling spatio-temporal systems is to replace a spatially-explicit model with a non-spatial model that is then fitted to non-spatial summaries of the data (*e.g.* Kleczkowski *et al.*, 1996, Gibson *et al.*, 1999). In the context of modeling epidemics these summaries may typically take the form of disease progress curves that record numbers of infectious or symptomatic individuals over time. More recently, the approach has been extended to fit semi-spatial models (Filipe *et al.*, 2004), which capture some of the spatial nature of epidemic processes by representing the evolution of low-order correlations in spatial pattern in addition to disease incidence. Such approaches, although convenient and useful, do not fully exploit experimental data that are explicitly spatial, such as those arising from the intensively mapped microcosm experiments considered in this paper. A further disadvantage of modeling only non-spatial summaries of spatial systems is that explicit spatial data cannot obviously be used to assess model fit. This makes it harder to distinguish between competing

models for a process, when the differences in models are not strongly reflected in the non-spatial summaries.

Here we describe and test methods for parameter estimation and model assessment for spatio-temporal models using extensive data from microcosm experiments involving the spread of a soil-borne disease through replicated seedling populations. Although the methods are applied to a specific laboratory-based system, they are applicable to a wide class of natural and semi-natural systems. Indeed, a closely related approach has been taken for fitting spatio-temporal models to disease spread in populations of citrus (Gibson 1997a, 1997b; Gottwald *et al.* 1999). The current study is distinct from this past work through its explicit treatment of times, rather than mere order, of infections and its focus on models with short range interactions, a specialisation which we exploit in the design of efficient algorithms for model fitting. The approach is fundamentally Bayesian and makes extensive use of modern computational methods – in particular Markov chain Monte Carlo (MCMC) methods – in order to take account of the incomplete nature of the experimental observations.

An important feature of the current study is the development of tests of model fit that take account of the spatial nature of the data. This we achieve by defining the notion of a stochastic residual for each individual in a population and then investigating the joint posterior properties of these. Since the sampling distribution of these residuals is known, we are able to test the model by carrying out a classical test of conformity to this distribution, treating the posterior distribution of a resulting P-value as a representation of evidence against the model fit. The idea of considering the posterior distribution of a classical test statistic, specifically a likelihood ratio, has recently been examined in other contexts by Aitkin et al. (2005).

Our aim in this paper, then, is to present this methodology and illustrate its use in the context of a particular host-pathogen system (*Rhizoctonia solani* Kühn that causes damping-off disease in a population of radish seedlings) described in section 2, for which a percolation approach is believed to be valid. Section 3 presents a spatio-temporal stochastic model for this disease which represents the spread of symptoms in the population as a percolation process and allows for dynamical changes in the rates of disease transmission of an earlier pseudo-spatial model (Filipe *et al.* 2004). Section 4 describes how the model parameters can be estimated in a Bayesian framework and gives a detailed description of the MCMC algorithms used to carry this out. In Section 5 we describe the process of generating latent stochastic residuals and how

to use them to investigate the posterior distribution of a P-value to quantify model fit. In section 6, we apply the techniques to spatio-temporal data of disease progress from the microcosm experiments and discuss the biological conclusions that can be drawn from the analyses. Finally, Section 7 discusses some further directions for the work.

2. Experimental systems

As a model system we consider microcosm experiments on epidemics of *R. solani* described more fully in Otten *et al*. (2003) and Filipe *et al*. (2004). The data record the observed development of damping-off disease on populations of radish seedlings (*cv*. Cherry Belle) caused by *R. solani*, a fungus that attacks hosts in the early stages of development, mature plants rapidly becoming resistant to parasitism (Deacon 1980). They were obtained for 13 replicate microcosms with low and high levels of initial inoculum (giving 26 experimental units). For each replicate, the data record daily the numbers and positions of diseased seedlings in a population of 414 plants arranged in an 18×23 rectangular array grown in clear plastic boxes. *Rhizoctonia solani* was added in the form of mycelial discs placed adjacent to randomly selected plants in each box. For low-inoculum and high-inoculum replicates, 15 discs and 45 discs were used, respectively. Host genotype and density, water availability, light and temperature were strictly and identically controlled in each replicate. Accordingly, differences in disease dynamics between treatments may be assumed to reflect differences in primary infection arising from different initial inoculum densities as well as demographic variability in disease transmission amongst replicates. The positions of damped-off seedlings were recorded daily for 17 days after seedling emergence. Figure 1 shows the evolution of a low-inoculum replicate and depicts the locations of symptomatic plants at a subset of the recorded times.

Filipe *et al.* (2004) developed a 'semi-spatial' stochastic model for this process. This was obtained by applying spatial moment-closure techniques from statistical physics (pair-wise approximation – see e.g. Filipe & Gibson (1998, 2001)) to an explicitly spatio-temporal model for the disease transmission process. Parameters in this non-spatial approximation were estimated for each treatment by maximum-likelihood fitting to the incremental numbers of infections (averaged over replicates) observed in the experiments. A potential limitation of the analysis of Filipe *et al.* (2004) is that by fitting to incremental numbers of diseased plants

averaged over replicates, within-treatment differences in replicates cannot be investigated. Furthermore, it does not make explicit use of the spatial nature of the data.

By contrast, the methods introduced in this paper fully exploit the availability of replicated spatial data to analyse the dynamics of a percolation process. We illustrate the application of the methods to answer epidemiologically important questions. These concern the analysis of treatment effects on transmission parameters, the identification of time-varying rates of host-to-host transmission, and the occurrence of variability in parameter estimates amongst replicate epidemics within treatments. Our prior expectation is that differences in disease dynamics between the low and high inoculum density treatments can be attributed to the different levels of inoculum *only* and not to systematic differences in transmission parameters between the treatments. In considering host-to-host transmission, we wish to determine whether or not the spatio-temporal analysis provides evidence supporting the time-varying rate of transmission, specifically a 'rise-and-fall', suggested by previous analyses (Otten *et al.* 2003, Filipe *et al.* 2004). Finally, and most importantly we demonstrate the power of the methods to detect possible variability between replicates and to assess the validity of the assumptions underlying the formulation of a given model.

3. A percolation based model for disease spread

In the model the population of seedlings is represented as being located at the vertices of a finite square lattice $\mathcal{L}$. The epidemic is initiated at time $t = 0$, when a subset of the sites $X_0 \subseteq \mathcal{L}$ is inoculated with the fungus. Each $\boldsymbol{x} \in X_0$ develops symptoms due to primary infection by this initial inoculum at a random time $T_x \sim \mathrm{Exp}(a)$ (if $\boldsymbol{x}$ is not already symptomatic due to infection via secondary infection from one of its four nearest neighbours). The random variables $\{T_x \mid \boldsymbol{x} \in X_0\}$ are assumed to be independent. Seedlings can also become symptomatic due to secondary infection. Specifically, if a site $\boldsymbol{x}$ develops symptoms at time t, then a neighbouring site $\boldsymbol{y}$ develops symptoms at $t + T_{\boldsymbol{xy}}$ (if it has not already become symptomatic due to infection from another source) where

$$T_{\boldsymbol{xy}} \sim \mathrm{Exp}(\phi(t)), \tag{3.1}$$

so that $E(T_{xy})$ is a function of the time at which x became symptomatic. We assume that the random variables $\{T_{xy} \mid y$ is a nearest neighbour of $x\}$ are *independent*, analogous to bond percolation. The function $\phi(t)$ reflects the formulation of Filipe *et al.* (2004). Specifically, we represent the biological hypothesis of a rise and fall in the secondary infection rate by

$\phi(t) = b_0 \exp\left(-b_1 \left(\log\left((t+4)/b_2\right)\right)^2\right)$. Other functions could equally well be proposed. The

model used here therefore has four free parameters. These are: a, the rate of primary infection from initial inoculum; b_0, which (loosely) represents the peak rate of secondary infection; b_1 which controls the range of variation in the secondary infection rate over time; and b_2 which determines the timing of the peak rate of secondary infection. In this paper we have changed our time origin from that used in Filipe *et al.* (2004) so that $t = 0$ here corresponds to the day of emergence of the seedlings (observed to be $t = 4$ by Otten *et al.* (2003)). Note, therefore, that b_2 corresponds to the timing of peak rate as measured on the scale used in Filipe *et al.* (2004). The replicates were generally observed until 17 days after emergence so that $t_{max} = 17$.

Realisations of this stochastic model for the onset of symptoms in the lattice can be readily simulated. Our main purpose in this paper is to use the above model to interpret experimental observations in terms of fundamental processes rather than to analyse the stochastic dynamics of the process. Therefore we do not present extensive simulations in this paper.

4. Bayesian fitting of percolation models

Suppose that we observe the population of plants in the lattice continuously from $t = 0$ to $t = t_{max}$ and record the precise times of onset for symptomatic seedlings during this period. In this case the observations would take the form of a lattice of times $\{t(x) \mid x \in \mathcal{L}\}$, precisely recorded for all sites displaying symptoms at some time during $[0, t_{max}]$ and censored for non-symptomatic sites. We suppose that there are N sites in the lattice of which k become symptomatic before time t_{max}. Denoting the observations by y, and the model parameter vector as $\theta = (a, b_0, b_1, b_2)$, we formulate a parameter likelihood $L(\theta \mid y)$ as follows.

Let $x_1, \ldots, x_k, x_{k+1}, \ldots, x_N$ denote the sites in the lattice arranged according to the order in which symptoms appear (with $x_{k+1}, \ldots, x_N$ representing any ordering of the unsymptomatic sites at t_{max}). It follows that y is observed if and only if ,

1.   for each site x_i, $i = 1, 2, .., k$, then the earliest time at which x_i becomes infected from a contacting source (primary infection or previously infected neighbour) is $t(x_i)$,

2.   for each site x_i, $i = k+1, \ldots, N$, the earliest time at which x_i becomes infected from a contacting source exceeds t_{max}.

It can be shown that the parameter likelihood can be expressed as a product of terms, one for each site, representing the contribution to the likelihood arising from constraints 1) or 2) above. That is

$$L(\boldsymbol{\theta} \mid y) = \prod_{i=1}^{k} f_1\big(t(x_i); t(x_1),...,t(x_{i-1}),\boldsymbol{\theta}\big) \prod_{i=k+1}^{N} f_2\big(t_{max}; t(x_1),...,t(x_k),\boldsymbol{\theta}\big) \tag{4.1}$$

For a given site $x_i \in \{x_1, \ldots, x_k\}$, with m previously infected neighbours x_{j_n}, $1 \leq n \leq m$, and primary inoculum

$$f_1\big(t(x_i); t(x_1),...,t(x_{i-1}),\boldsymbol{\theta}\big) = \left(a + \sum_{n=1}^{m} \phi\big(t(x_{j_n}),b\big)\right) e^{-at(x_i)} \prod_{n=1}^{m} \exp\big(-\phi\big(t(x_{j_n}),b\big)\big(t(x_i) - t(x_{j_n})\big)\big) \tag{4.2}$$

For $i > k$, if x_i has m previously infected neighbours and primary inoculum the corresponding factor is

$$f_2\big(t(x_i); t(x_1),..., t(x_{i-1}),\boldsymbol{\theta}\big) = e^{-at(x_i)} \prod_{n=1}^{m} \exp\big(-\phi\big(t(x_{j_n}),b\big)\big(t(x_i) - t(x_{j_n})\big)\big). \tag{4.3}$$

The corresponding expressions for sites without primary inoculum are similar but omit terms involving the parameter a.

An important property of the likelihood (4.1) – (4.3) is that the factor contributed by any site is a function of the history of that site and of its four nearest neighbours in the lattice only. It is perhaps worth noting that, although the likelihood depends on the history, the process itself is Markovian with the next event in time determined only by the current state of the system. This will be particularly useful for the MCMC computations carried out later, in that changes

proposed to the time of onset of symptoms for one site only affect the calculation of at most 5 terms in (4.1).

In practice however, the exact times of onset of symptoms are not observed, since the population is observed at discrete times t_1, …, t_r say. If a site is recorded as symptomatic for the first time at time t_j, then the observations merely constrain its time to lie within $[t_{j-1}, t_j]$. We denote these censored observations by y', and let Y' denote the set of all outcomes y (*i.e.* precise times in $(0, t_{max})$) that are consistent with the observation y'. Therefore the desired likelihood takes the form of an integral, namely

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}') = \int_{y \in Y'} L(\boldsymbol{\theta} \mid \boldsymbol{y}) dy \qquad (4.4)$$

This integral is not in general analytically tractable. Nevertheless we can apply MCMC methods within a Bayesian framework to solve the problem.

Parameter estimation using MCMC

This approach is now commonly used in the fitting of stochastic epidemic models to partial observations (Gibson & Renshaw 1998, O'Neill & Roberts 1999, O'Neill & Becker 2001). It entails treating the unobserved aspects of the process as nuisance parameters and investigating the posterior distribution of these jointly with the model parameters. Inference on model parameters can then be made from their marginal distributions.

Let Θ denote the model parameter space and assign a prior distribution, denoted $p(\boldsymbol{\theta})$, for $\boldsymbol{\theta}$. Then, given the censored observations y', the joint posterior density on $\Theta \times Y'$ is defined by

$$\pi(\boldsymbol{\theta}, \boldsymbol{y} \mid \boldsymbol{y}') \propto p(\boldsymbol{\theta}) L(\boldsymbol{\theta} \mid \boldsymbol{y}).$$

Since $L(\boldsymbol{\theta} \mid \boldsymbol{y})$ can be calculated using (4.1)-(4.4), then we know $\pi(\boldsymbol{\theta}, \boldsymbol{y} \mid \boldsymbol{y}')$ up to an unknown constant of proportionality and MCMC methods are particularly appropriate. We use a simple algorithm that uses mainly Metropolis-Hastings updates (Tierney 1994). The algorithm provides adequate performance in this context, although more sophisticated implementations could be carried out.

A Markov chain is defined on $\Theta \times Y'$. Each iteration of the chain involves proposing moves of two types to the current state vector $(\boldsymbol{\theta}, \boldsymbol{y})$: updates to the exact times of infection for

plants becoming symptomatic in [0, t_{max}], and updates to the parameters $\boldsymbol{\theta}$. These are effected as follows.

Updates to components of $\boldsymbol{y}$, are effected by considering each site that becomes infectious in [0, t_{max}] in turn. For the chosen site, a new infection time is proposed uniformly from the permissible range $[t_{j-1}, t_j]$ (where t_j denotes the observation time when the site was first observed to be infected), giving a new configuration $\boldsymbol{y}^p$. The new time is accepted with probability

$$p_{acc} = \min\left(1, \frac{L(\boldsymbol{\theta} \mid \boldsymbol{y}^p)}{L(\boldsymbol{\theta} \mid \boldsymbol{y})}\right).$$

Infectious sites are considered in a fixed order in contrast to random scan formulations.

Updates to parameters are effected by proposing and accepting or rejecting changes to the current values of a, b_0, b_1 and b_2 in turn. A new value of the parameter is proposed uniformly from a finite interval centred on the current parameter value, to give a new parameter vector $\boldsymbol{\theta}^p$. This new vector is accepted with probability

$$p_{acc} = \min\left(1, \frac{p(\boldsymbol{\theta}^p) L(\boldsymbol{\theta}^p \mid \boldsymbol{y})}{p(\boldsymbol{\theta}) L(\boldsymbol{\theta} \mid \boldsymbol{y})}\right).$$

Any proposed values of parameters that fall out of the range of the prior are naturally rejected since the prior density is zero at such points. The widths of the uniform windows for the proposed updates to each component of $\boldsymbol{\theta}$ are selected for each component, and a given replicate, on the basis of trial runs to give favorable mixing properties for the resultant chains. Several authors (*e.g.* Besag et al., 1995) indicate acceptance rates for Metropolis-Hastings updates in the range 0.2 – 0.5 as leading to efficient mixing.

We also consider an alternative implementation of a Markov chain sampler for this problem that considers a more complex state space to that above. The second approach allows an independent algorithm to be developed that can be compared to the first in order to check the validity of the posterior densities obtained. It may also have superior mixing properties when the period between successive observations is large. Specifically we extend the state space by considering not only the precise times of the onset of symptoms for each site becoming symptomatic in [0, t_{max}], but also an array $\boldsymbol{s}$ which denotes the source of infection for each symptomatic site. At most there are 5 possibilities for the source – primary infection or infection

from one of the (potentially infected) four neighbours. The new likelihood $L(\theta| y, s)$ is similar to that defined by (4.1) – (4.3). However the contribution from a symptomatic site (4.2) is different in that the summation, $\left(a + \sum_{n=1}^{m} \phi\left(t(x_{j_n}), b \right) \right)$ is replaced by a single term corresponding to the particular source of infection – a, if the infection is due to primary infection, or $\phi\left(t(x_{j_n}), b \right)$ if the infection is acquired from the neighbour at x_{j_n}.

The MCMC algorithm must also be altered for the new state space. When proposing a new infection time for a site, x, infected between $[t_{j-1}, t_j]$ the new time is now proposed uniformly from $[t_l, t_u]$ where t_l is the supremum of time of infection of the site infecting x and t_{j-1}, and t_u is the infimum of t_j and the times of infections of any sites infected by x. This ensures that the new time proposed will be consistent with the 'active' transmission pathways specified by s. Following updating of the time of infection of a site, the source is updated by selecting from the potential sources (primary inoculum if present, infection from nearest neighbours). This can be done via a Gibb's step. Primary infection and infection from a previously infected neighbour x_{j_n}, $n = 1, ..., m$, are selected with probability proportional to a and $\phi\left(t(x_{j_n}), b \right)$ respectively. With this second formulation, it is possible, for suitably chosen prior densities, to update the primary infection parameter, a, via a Gibb's step, although we do not do so in this paper.

The above methods can be applied to analyse a set of replicates jointly (as we do in section 6) simply by replacing the likelihood $L(\theta|y)$ by the joint likelihood $L^j(\theta/y) = \prod_i L(\theta/y_i)$ where y_i represents the infection times for the ith replicate, when updating components of θ.

Both formulations of the algorithm described above are implemented using the C programming language with the following minor modification to the above recipe. In all replicates a few infections that were not consistent with nearest-neighbour spread were encountered. This can be seen from Figure 1 where not all symptomatic plants are connected to a primary infection via a sequence of nearest-neighbour links. Consequently the models and algorithms above would be inapplicable due to the problem of vanishing likelihoods. There are several ways to accommodate this difficulty. The solution we choose is to introduce a further parameter δ, (fixed to be 10^{-7}) representing a primary infection rate presented to every plant in

the population. This additional source could be interpreted biologically as infection due to inadvertent contamination and the model can explain unconnected symptomatic plants as arising from this latter process. As δ is very small, the MCMC algorithm 'attributes' the minimum number of infections to this process (usually of the order of $10 - 30$ per replicate) in the pattern. We appreciate that this solution to handling non-connectivity is imperfect and discuss alternative approaches to resolving the problem in the final section of the paper.

5. Model checking using stochastic residuals

A principal aim of the paper is to describe and illustrate methods for assessing validity of modeling assumptions in percolation models and spatio-temporal models in general. We remark that there is a well established Bayesian methodology for model comparison (Draper, 1994). This approach would present considerable challenges in terms of implementation of the requisite computational algorithms (that would of necessity use reversible-jump MCMC (Green, 1995)), and the resulting model posterior probabilities are known to be sensitive to the choice of priors for model parameters. This presents a particular difficulty if model parameter spaces are non-nested. Instead we proceed to construct a set of stochastic residuals whose sampling distribution is known, and is independent of model parameters, and then to test the 'reconstructed' residuals for consistency with this known distribution using classical tests embedded in the Bayesian framework.

Using the notation of Section 4, suppose that we simulate the stochastic process in the following way. To each site $x \in \mathcal{L}$ we associate a random value r_x, drawn from a $U(0, 1)$ independently over sites. Now given the model parameters θ, we obtain a realization of the disease spread process from $\{r_x\}$ by the following algorithm, in which $X(t)$ denotes the set of infected sites at time t. Let Y_x be the random variable denoting the time at which site x becomes infected, then:

i) set $t = 0$, $X(t) = \varnothing$,

ii) for each uninfected site, x, calculate y_x where $r_x = P(Y_x > y_x \mid \theta, X(t)$, no other site infected in $[t, y_x])$,

iii) select the site $\boldsymbol{x}$ with minimum $y_{\boldsymbol{x}}$ to become the next infection and set $t = y_{\boldsymbol{x}}$ for this minimum value. Append the site $\boldsymbol{x}$ to the infected set $X(t)$, repeat (ii) and (iii) until all the sites are infected.

Now it can be easily verified that given $\boldsymbol{\theta}$, the above construction maps $\{r_{\boldsymbol{x}}\}$ to $\boldsymbol{y}$ by a 1-1 correspondence, $h_{\boldsymbol{\theta}}$, when complete infection of the population occurs. When the $y_{\boldsymbol{x}}$ are censored, due to observation terminating before total infection, then given the censored $\boldsymbol{y}$ we can define the set $h_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y})$, to be all residual processes $\{r_{\boldsymbol{x}}\}$ that would yield the censored times $\boldsymbol{y}$ under the mapping $h_{\boldsymbol{\theta}}$.

Now, returning to the Bayesian analysis of Section 4, it follows that since the conditional density of $\boldsymbol{r}$ given $\boldsymbol{\theta}$ and $\boldsymbol{y}$ is *uniform* on $h_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y})$, we can obtain the posterior density $\pi(\boldsymbol{r}|\boldsymbol{y})$ directly from the posterior $\pi(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{y})$. Under the assumption that the observations are generated by first drawing $\boldsymbol{\theta}$ from the prior and generating the infection times from the assumed model parameterised by $\boldsymbol{\theta}$, the residual process $\boldsymbol{r}$ is a realisation of a set of *i.i.d.* $U(0, 1)$ random variables. We consider the P-value from a standard test, $P(\boldsymbol{r})$, to quantify evidence against H_0: the $\{r_{ij}\}$ are *i.i.d.* $U(0, 1)$ and investigate the posterior density of $P(\boldsymbol{r})$ given $\boldsymbol{y}$. If the latter density suggests we would reject H_0 with high posterior probability, then we should question the assumed combination of prior and model as the mechanism for generating the observations. In carrying out this process, we omit from the classical analysis any residuals for any sites that have been infected via the spurious process (controlled by the fixed parameter δ).

The particular test used to assess the residuals is the Kolmogorov-Smirnov (K-S) test (Silvey 1970). Given a sample of $\boldsymbol{r}$ from the posterior, we order $\boldsymbol{r}$ from smallest to largest $r_{(1)}, \ldots, r_{(N)}$ where N' is the number of sites in the lattice minus those infected by the 'spurious' process. The test statistic used is

$$T(\boldsymbol{r}) = \max\{ |r(i) - i/N| \},$$

this being the modulus of the largest deviation between observed and theoretical distribution functions. A P-value for the test (*i.e.* the probability of observing a test statistic as or more extreme than $T(\boldsymbol{r})$) can be calculated approximately (see *e.g.* Glasbey *et al.,* 1986) as

$$P(\boldsymbol{r}) = 2\exp(-2NT(\boldsymbol{r})^2). \tag{5.1}$$

This approximation breaks down if $T(r)$ is small and indeed may give values exceeding unity, but it is accurate for small P-values (the frequency of which interests us most in detecting poor fit of the model).

6. Results

The methods of sections 4 and 5 are applied to analyse the data described in section 2 by fitting the percolation model of section 3. In the analyses that follow we assume that all parameters have independent uniform priors over finite regions of the positive real axis which are selected to be wide compared with the region over which the posteriors are informative. This is an attempt to represent prior ignorance of model parameters. Results for alternative choices of vague priors for parameters are broadly similar to those presented for the uniform case.

Figures 2 – 5 show histogram estimates of the posterior marginal densities of the model parameters with the high and low treatments analysed separately throughout. The particular Markov chain used is the second of the formulations in Section 4, in which the source of infection is represented in the components of the state vector. These graphs show the posterior densities when the 13 replicates for a treatment are analysed jointly (i.e. under the assumption of common parameters across replicates) and separately. Histogram estimates of the posterior densities for each parameter were obtained from 10^5 iterations of the chain (where each iteration involves proposing updates to all the parameters and all the components of *y*), following a burn-in period of 1000 iterations. Although we do not show details here, we remark that studying the autocorrelation of trace-plots and robustness of posterior inferences to starting values of parameters and infection times suggests that the chains mix well. Moreover, posterior inferences on parameters obtained using the first algorithm described in section 4, in which the infecting source is not represented in the state space, are identical to those shown here, providing further evidence that a valid picture of the posterior distribution has been obtained.

The posteriors for the joint analyses are very informative. There is little evidence of any difference in parameter values between treatments except for the case of b_0 whose value for the low-inoculum treatment appears to be smaller. However, we note that, within treatments, the posterior densities that arise from fitting the model to replicates separately vary substantially for all four parameters. For example in the high-inoculum case there is no value of the primary infection rate, a, estimated from a given replicate, that does not appear very extreme with respect

to the posterior density for at least one other replicate. This suggests that parameters differ among replicates within each treatment.

To investigate the validity of the model with common parameters amongst replicates within treatments we consider the predictive distribution of the number of infected individuals with model parameters fixed at their posterior marginal means for the joint analyses. Initial conditions (i.e. sites of primary inoculum and missing plants) are selected to match the experiment with an equal number of simulations carried out for each compatible initial configuration. Equal-tailed 95% intervals for the predicted number of infective plants $I(t)$ at the observation times are displayed in Fig. 6, along with the observed values of $I(t)$ for the 13 replicates for each treatment. Based on these, it could be suggested that the model with common parameters broadly captures the variability in the temporal evolution of $I(t)$ observed over replicates for each treatment.

In assessing model fit by comparing observed and predicted dynamics of $I(t)$ we have ignored all the spatial information in the data. By contrast the residual analysis of Section 5, which does take account of spatial information, tells a very different story, leading to different inferences about parameter variation amongst replicates. Tables 1(a) and (b) give descriptive summaries for the posterior distributions of the K-S P-values, as described in Section 5, when the replicates are analysed assuming common parameters within treatments. For each replicate, these summary statistics are based on a sample of 20,000 P-values obtained from the posterior by calculating the K-S P-values from the vector of residuals for each sample. From these it is clear that for nearly all replicates, the posterior density of the P-value is concentrated on low values, suggesting that a classical observer of the residual process would reject the hypothesis that it were i.i.d. U(0, 1) for most replicates with very high probability. For each replicate, this probability is shown for the commonly used 5% testing level. On the basis of these results we suggest that the model with common parameters within treatments may not be valid. We remark that the method of residual analysis is able to detect this lack of fit when measures of fit based on cruder summaries of the process (associated with goodness-of-fit to the temporal data alone) may fail.

Table 2 shows the corresponding posterior summaries when the replicates are analysed separately. Compared with the common-parameter analysis, the densities of P-values support

higher values and rejection at the 5% level becomes generally less likely. However, we note that there is still evidence of a lack of fit in that the posterior expectation of the number of replicates for which the classical observer rejects the hypothesis of *i.i.d.* U(0, 1) residuals at the 5% level is high – 6.9 and 5.3 out of 13 for High and Low treatments respectively. This is addressed further in the discussion.

From the observed disparity in posterior parameter densities between replicates seen in Figures 2-5 together with the improvement in measures of fit when the model is fitted to replicates separately we conclude that there is heterogeneity among replicates for the same treatment. From inspection of the marginal posterior densities for parameters for replicates separately, there is no evidence of systematic differences between the two treatments.

Our spatio-temporal analyses suggest qualitative conclusions that agree with earlier analyses of the data (Otten *et al.*, 2003; Filipe *et al.*, 2004). The secondary infection process obeys a 'rise-and-fall' in the rate. This is shown by the support for non-zero values of b_1 for most of the replicates. Where a pronounced variation is suggested, inspection of posteriors for b_2 suggests that the peak secondary rate seems most frequently to be located in the interval 5 to 15 days after planting (1 to 11 days after emergence of seedlings). In replicates where little variation in secondary infection rate appears to be taking place (evidenced by the posterior for b_1 supporting small values) little information can be gained on the timing of the peak.

7. Discussion

In this paper we have presented and illustrated methods for fitting spatio-temporal stochastic models with nearest-neighbour interactions to intensively mapped observations of epidemics. In particular, the paper provides a general methodology that can be adapted to allow percolation models to be used directly in the analysis and interpretation of observations of spatio-temporal processes. As such it has many potential applications beyond the botanical epidemics considered here. As with earlier work (*e.g.* Gibson, 1997; Gottwald *et al.* 1999), the approach uses Bayesian methods coupled with Markov chain techniques to explore parameter posterior densities. Moreover we have shown how, by defining a spatial residual process with fixed statistical properties, it is possible to assess the fit of a given model by incorporating classical

tests within the Bayesian framework. This approach to model assessment provides a simple alternative to Bayesian model comparison methods (*e.g.* Draper, 1994). The methods have been illustrated by fitting a continuous-time, spatial model for observed spread of damping-off symptoms in radish infected by *R. solani* to data from microcosm experiments, and assessing the fit of the model. The particular model fitted was formulated to reflect the modeling assumptions of earlier analyses (Filipe *et al.* 2004) but we stress at this point that all the methods presented can be readily adapted to a wide range of alternatives.

With regard to the particular system considered, our results agree with earlier analyses that suggest that there is little systematic difference between the two treatments (high and low levels of primary inoculum) in the experiment. This is consistent with biological intuition, since there is no *a priori* reason why changes in the amounts of initial inoculum would systematically affect parameters concerned with primary and secondary transmission in a spatially-explicit model. Our results, however, do show that there is variation in transmission of infection amongst replicates within treatments. This latter claim is substantiated by the wide disparity between marginal posterior densities of the parameters for the different replicates. Further evidence is provided by the residual analysis which suggests that a model with common parameters over replicates within a treatment fits poorly, while the fit improves greatly when parameters are allowed to vary between replicates. Earlier non-spatial analyses did not provide evidence of this variability and this illustrates the power of the present analysis which takes account of all spatial and temporal information in the data. The present analysis also confirms the findings of earlier work regarding the time-varying nature of the secondary infection rate for this host-pathogen system. There is some evidence for lack-of-fit to the model from the residual analysis of individual replicates. We do not see this as a cause for concern. The fitted model is specified by only four free parameters and the functions within it are selected on the basis of qualitative similarity with biological hypotheses and earlier formulations of related non-spatial models. On the other hand, the residual analysis takes account of the totality of a very rich set of data that includes disease history of every site in the lattice. It should not be surprising that some lack of fit is detected. Perhaps we should be surprised that the evidence against such a simple model provided by such a comprehensive data set is not stronger.

The occasional lack of connectivity in the disease patterns invalidates the strict nearest-neighbour assumption made in our basic model. In this paper we overcame this in an *ad hoc*

way, by extending the model to allow for non-nearest neighbour secondary or unexpected primary infections at a small, fixed rate. Alternatively, we could have retained the nearest-neighbour infection process by extending the model to include an additional state in which a plant is infective but not yet symptomatic with newly infected plants staying in this state for a random time until emergence of symptoms. This model is not contradicted by an unconnected pattern of symptomatic sites since the underlying pattern of infected sites may nevertheless be connected. Ongoing investigations with this model suggest that it has some merits. However, it may explain the observations poorly when 'spurious' infections are observed at large distances from other infections. In the latter case, a large number of intervening asymptomatic sites must be imputed in order to ensure connectivity of the infected sets. Moreover, the introduction of an unobservable asymptomatic state leads to sensitivity of posterior inferences to the choice of prior parameter density, not encountered for the simpler model. A second approach to coping with occasional lack of connectivity would be to allow longer range interactions in the disease transmission process as in Gibson (1997a,b), at the expense of the computational convenience of working with the nearest-neighbour model.

There are several avenues for future research leading from the present work. Clearly there is potential for extending the range of applications to other spatio-temporal processes most notably for ecological interactions involving competing species, as well as for other epidemiological models. Within the current system, the methods can be used in combination with further experimentation, to quantify effects of chemical and biological control or of different soil types on variability in transmission dynamics amongst replicate epidemics. It is also being used to analyse the spread of disease through mixtures of host species that differ in susceptibility (*cf.* Otten *et al.* 2005, for a description of the biological processes). Current work is also directed towards extending the percolation paradigm to incorporate more biological realism, in particular to allow for interference and synergy amongst infected individuals that challenge a common susceptible individual on a lattice.

Acknowledgements

8. References

Aitkin, M., Boys, R. J. & Chadwick, T. 2005. Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statistics and Computing*, **15**, 217-230.

Bailey, D. J., Otten, W. & Gilligan, C. A. 2000. Percolation, heterogeneity and the saprotrophic invasion of soil by the fungal plant pathogen *Rhizoctonia solani*. *New Phytol.* **146**, 535-544.

Besag, J., Green, P. J., Higdon, D. & Mengersen, K. 1995. Bayesian computation and stochastic computation (with discussion). Statistical Science, **10**, 3-66.

Deacon, J. W. 1980. *Introduction to modern mycology*. Blackwells Science, Oxford.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion), *J. Roy. Stat. Soc. B*, **57**, 45-97.

Filipe, J.A.N, Otten, W., Gibson, G. J. and Gilligan, C. A. 2004. Inferring the dynamics of a spatial epidemic from time series data. *Bulletin of Mathematical Biology*, **66**: 373-391.

Filipe, J. A. N. & Gibson, G. J. 1998. Studying and approximating spatio-temporal models for epidemic spread and control. *Phil. Trans. Roy. Soc. Lond . B*, **353**, 2153-2162.

Filipe, J. A. N. & Gibson, G. J. 2001. Comparing approximations to spatio-temporal models for epidemics with local spread. *B. Math. Biol.* **63**, 603-624.

Gibson, G. J. 1997a. Markov Chain Monte Carlo methods for fitting and testing spatio-temporal stochastic models in plant epidemiology. *Applied Statistics* **46**, 215-233.

Gibson, G. J. 1997b. Investigating mechanisms of spatio-temporal epidemic spread using stochastic models. *Phytopathology* **87**, 138-146.

Gibson, G.J. & Renshaw, E. 1998. Estimating parameters in stochastic compartmental models using Markov chain Monte Carlo methods. *IMA J. Math. Appl. Med. & Biol.* **15**, 19-40.

Gibson, G. J., Gilligan, C. A. & Kleczkowski, A. 1999. Predicting variability in biological control of a plant-pathogen system using stochastic models. *Proc. R. Soc. Lond. B* **266**, 1746-1753.

Glasbey, C. A. & Martin, R.J. 1986. Exploratory and confirmatory plots of single-channel records. *J. Neuroscience Methods,* **16**, 239-249.

Gottwald, T. R., Gibson, G. J., Garnsey, S. M. & Irey, M. 1999. Examination of the effect of aphid vector population composition on the spatial dynamics of citrus tristeza virus spread via stochastic modeling. *Phytopathology*, **89**, 603-608.

Kleczkowski, A., Bailey, D. J. & Gilligan, C. A. 1996. Dynamically generated variability in plant-pathogen systems with biological control. *Proc. R. Soc. Lond.* B **263**, 777-783.

O'Neill, P. D. and Roberts, G. O. 1999. Bayesian inference for partially observed stochastic epidemics, *J. Roy. Statist. Soc. A*, **162**, 121-129.

O'Neill, P.D. and Becker, N.G.. 2001. Inference for an epidemic when susceptibility varies. *Biostatistics*, **2**, 99-108.

Otten, W., Bailey, D. J. & Gilligan, C. A. 2004 Empirical evidence of spatial thresholds to control invasion of fungal parasites and saproptrophs. *New Phytol.* **163**, 125-132.

Otten, W., Filipe, J. A. N., Bailey, D. J. & Gilligan, C. A. 2003. Quantification and analysis of transmission rates for soil-borne epidemics. *Ecology* **84,** 3232-3239.

Otten, W., Filipe, J. A. N., & Gilligan, C. A. (2005) Damping-off epidemics, contact structure, and disease transmission in mixed-species populations. *Ecology* **86**, 1948-1957.

Silvey, S. D. 1970. *Statistical Inference*. Penguin Books Ltd, Harmondsworth.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701-1762.

Wissel, C. 2000. Grid-based models as tools for ecological research. In *The Geometry of Spatial Interactions: Simplifying Spatial Complexity*, eds. Diekmann, U., Law, R. & Metz, J. A. J., pp 94-115. Cambridge, UK: Cambridge University Press.

Table 1. Estimated mean, and quartiles for posterior distribution of *P*-values for a K-S test of the residuals when common parameters within treatments are assumed. These are estimated from 20,000 independent samples from the posterior distribution of the vector of residuals for each replicate.

a) High inoculum

mean	median	LQ	UQ	Pr(p<5%)	rep
0.013454	0.008933	0.003938	0.017715	0.9706	H1
0.000024	0.000006	0.000002	0.000020	1.0000	H2
0.000048	0.000015	0.000005	0.000046	1.0000	H3
0.002916	0.002039	0.001029	0.003757	1.0000	H4
0.136859	0.121301	0.070754	0.189353	0.1508	H5
0.075692	0.061481	0.033481	0.101138	0.4054	H6
0.014030	0.010434	0.005457	0.018893	0.9794	H7
0.000000	0.000000	0.000000	0.000000	1.0000	H8
0.000024	0.000009	0.000003	0.000026	1.0000	H9
0.552893	0.540598	0.388507	0.707525	0.0014	H10
0.000002	0.000000	0.000000	0.000002	0.9844	H11
0.000000	0.000000	0.000000	0.000000	1.0000	H12
0.086550	0.067288	0.035668	0.116220	0.3708	H13

b) Low inoculum

mean	median	LQ	UQ	Pr(p<5%)	rep
0.085966	0.063765	0.030051	0.118423	0.4036	L1
0.016515	0.011087	0.004239	0.022995	0.9412	L2
0.025164	0.018007	0.008519	0.034758	0.8724	L3
0.001911	0.001254	0.000602	0.002506	1.0000	L4
0.000126	0.000034	0.000009	0.000116	1.0000	L5
0.067819	0.052805	0.024010	0.097047	0.4804	L6
0.000045	0.000017	0.000005	0.000048	1.0000	L7
0.000004	0.000000	0.000000	0.000001	1.0000	L8
0.000030	0.000006	0.000001	0.000022	1.0000	L9
0.003341	0.001897	0.000737	0.004192	0.9998	L10
0.000001	0.000000	0.000000	0.000001	1.0000	L11
0.001243	0.000678	0.000257	0.001549	1.0000	L12
0.051377	0.027275	0.008475	0.069362	0.6560	L13

Table 2. Estimated mean, and quartiles for posterior distribution of *P*-values for a K-S test of the residuals when the model is fitted to each replicate separately. These are estimated from 20,000 independent samples from the posterior distribution of the vector of residuals for each replicate.

a) High inoculum

mean	median	LQ	UQ	pr{p<5%}	rep
0.052114	0.011119	0.001752	0.051376	0.7465	H1
0.206157	0.094937	0.023089	0.300549	0.3769	H2
0.281590	0.161915	0.047275	0.425640	0.2693	H3
0.031162	0.005034	0.000706	0.027090	0.8288	H4
0.213550	0.103440	0.022493	0.321352	0.3712	H5
0.382658	0.305408	0.099617	0.619855	0.1671	H6
0.412586	0.381779	0.178334	0.610587	0.0828	H7
0.056309	0.014313	0.002553	0.063408	0.7116	H8
0.040092	0.008516	0.001360	0.040168	0.7842	H9
0.254628	0.173918	0.045907	0.408351	0.2633	H10
0.170810	0.064701	0.014189	0.225905	0.4547	H11
0.020977	0.002134	0.000247	0.013154	0.8952	H12
0.007638	0.000434	0.000042	0.003514	0.9619	H13

b) Low inoculum

mean	median	LQ	UQ	pr{p<5%}	rep
0.118601	0.073594	0.024119	0.171492	0.3992	L1
0.312936	0.253385	0.098790	0.474637	0.1482	L2
0.172566	0.100250	0.022634	0.261131	0.3644	L3
0.593655	0.592876	0.348687	0.823331	0.0239	L4
0.058327	0.030834	0.010818	0.076999	0.6342	L5
0.178069	0.103703	0.024421	0.269454	0.3549	L6
0.055987	0.037735	0.015791	0.076855	0.5986	L7
0.290320	0.229458	0.075142	0.453747	0.1934	L8
0.185745	0.126875	0.045629	0.272254	0.2670	L9
0.206834	0.137491	0.033597	0.329708	0.3068	L10
0.026957	0.017073	0.006245	0.037063	0.8362	L11
0.037837	0.010746	0.002026	0.043177	0.7771	L12
0.153215	0.091455	0.022290	0.234767	0.3802	L13

List of figures

Figure 1. Evolution of pattern of infection for a low-inoculum replicate. Missing plants are denoted by O; sites of primary inoculum are denoted by +; infections first detected 0-4 days after emergence by ×; infections first detected 5-11 days after emergence by Δ; infections first detected 12-17 days after emergence are denoted by ◊. Several instances of a violation of nearest-neighbour connectivity can be discerned.

Figure 2. Estimated posterior densities for primary infection parameter, a, for high (upper graph) and low (lower graph) inoculum treatments. Dashed line indicates posterior density for assumption of common parameters within treatments and solid lines for separate estimates for each replicate.

Figure 3. Estimated posterior densities for secondary infection parameter, b_0, for high- (upper graph) and low- (lower graph) inoculum treatments. Dashed line indicates posterior density for assumption of common parameters within treatments and solid lines for separate estimates for each replicate.

Figure 4. Estimated posterior densities for secondary infection parameter, b_1, for high- (upper graph) and low- (lower graph) inoculum treatments. Dashed line indicates posterior density for assumption of common parameters within treatments and solid lines for separate estimates for each replicate.

Figure 5. Estimated posterior densities for secondary infection parameter, b_2, for high- (upper graph) and low- (lower graph) inoculum treatments. Dashed line indicates posterior density for assumption of common parameters within treatments and solid lines for separate estimates for each replicate.

Figure 6. Distribution of disease progress curves (t, I(t)) as predicted by model of section 3 with model parameters fixed at their posterior marginal means for the joint analyses for each treatment. Shaded area represents equal-tailed 95% (marginal) intervals for the predicted
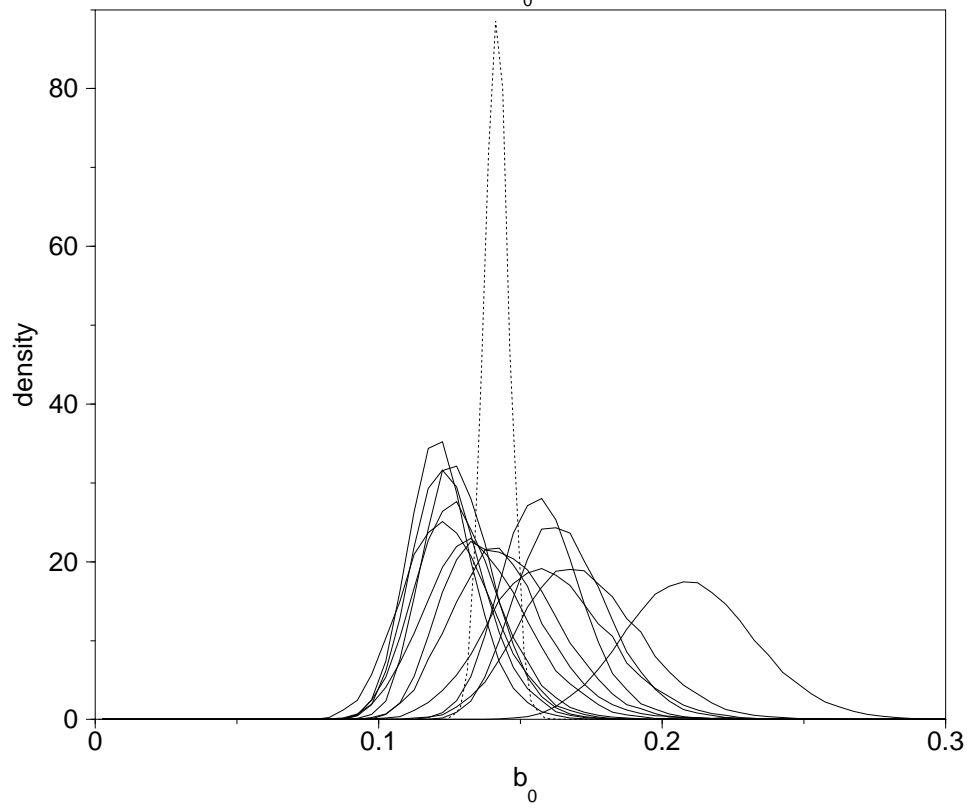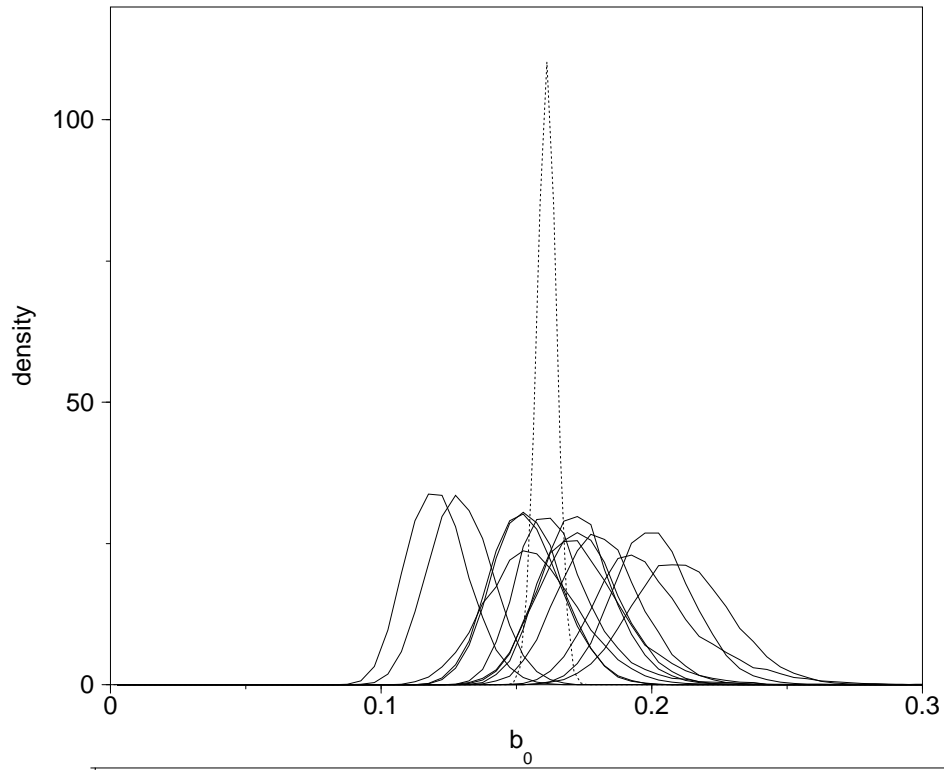
number of infective plants $I(t)$ at any given time. The observed values of $I(t)$ for the 13 replicates for each treatment are displayed and exhibit broad consistency with the predicted intervals.
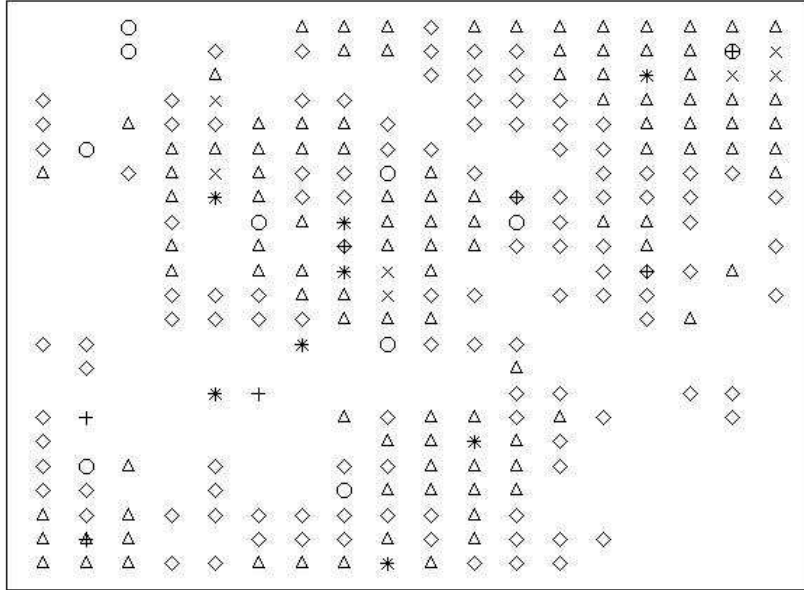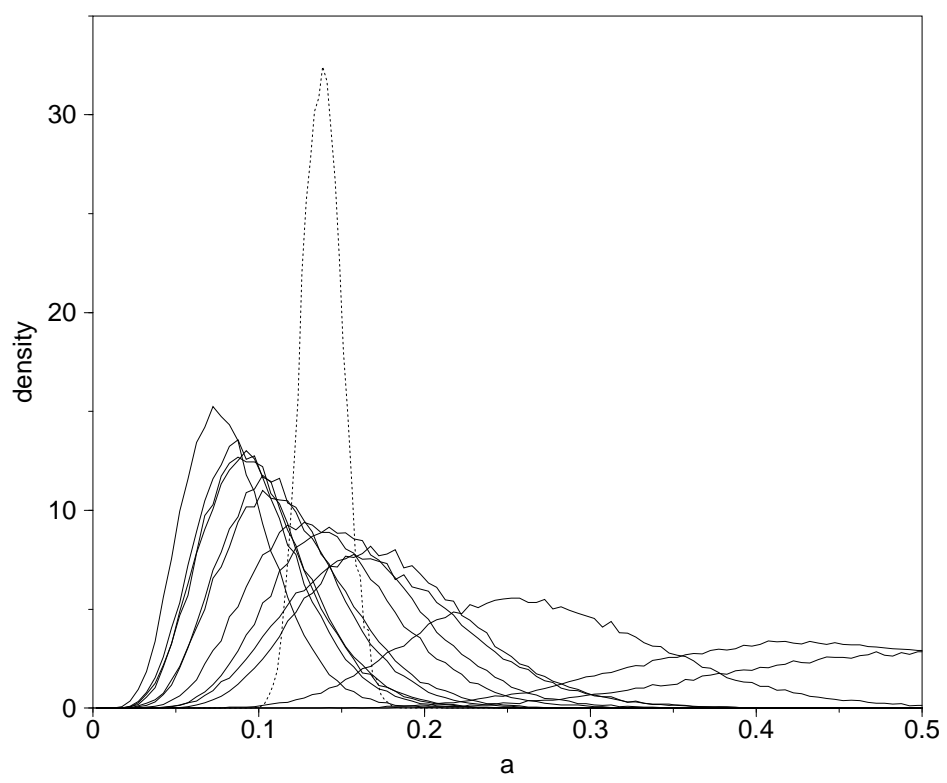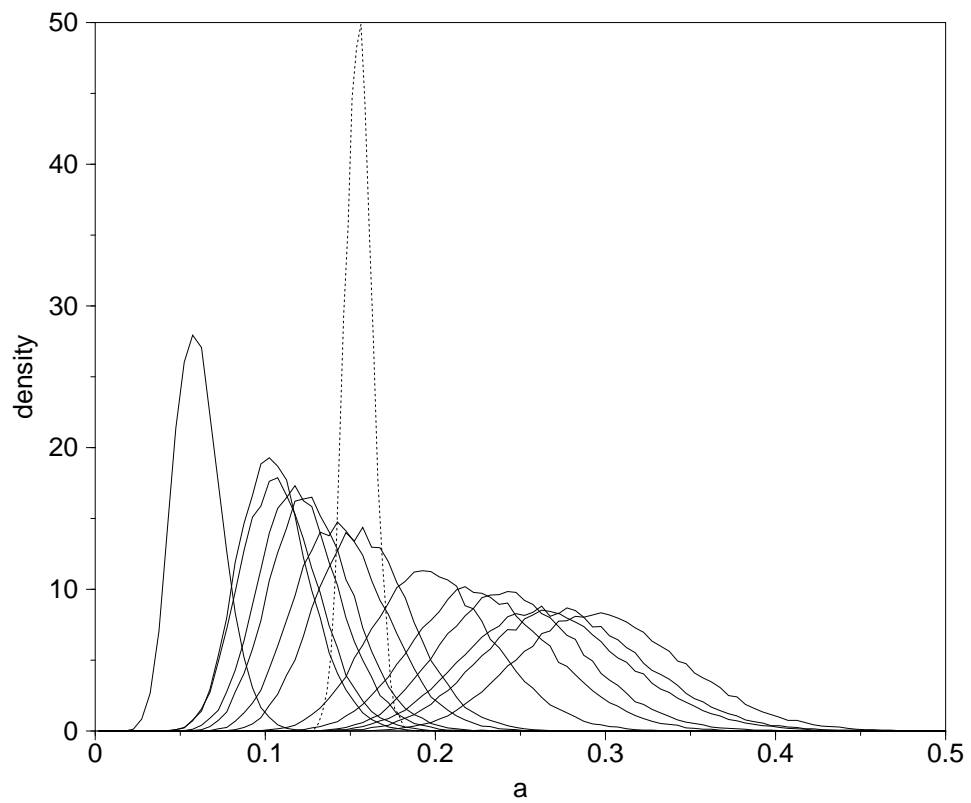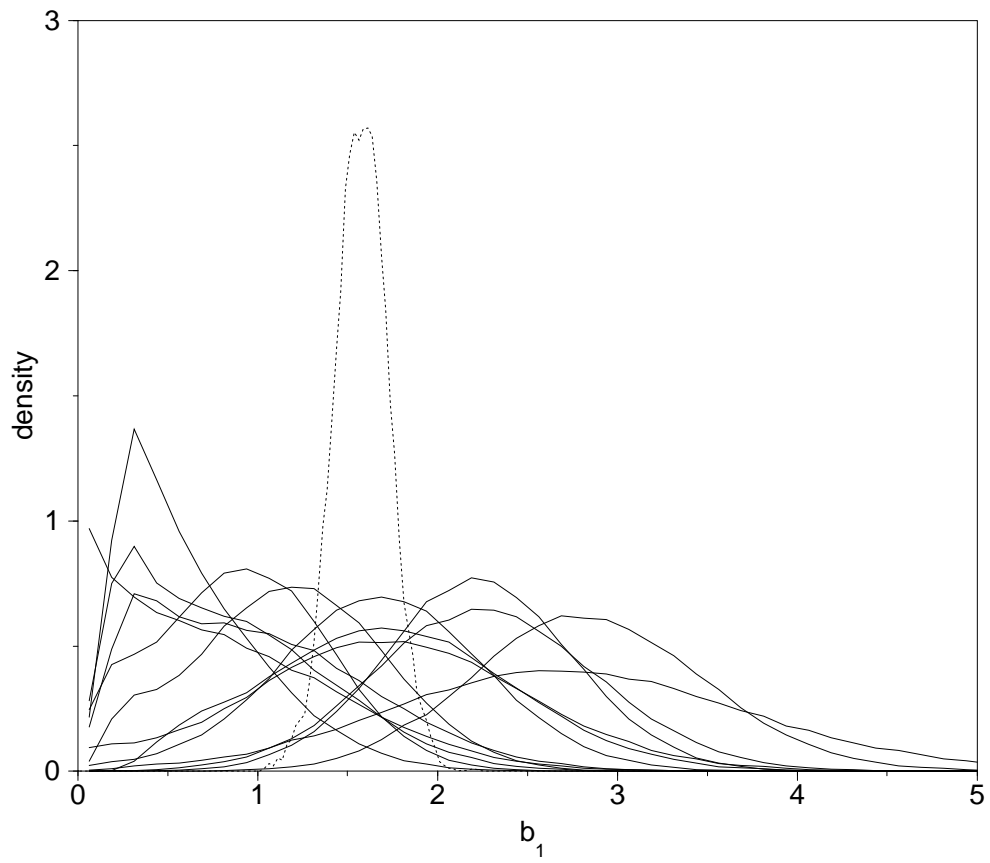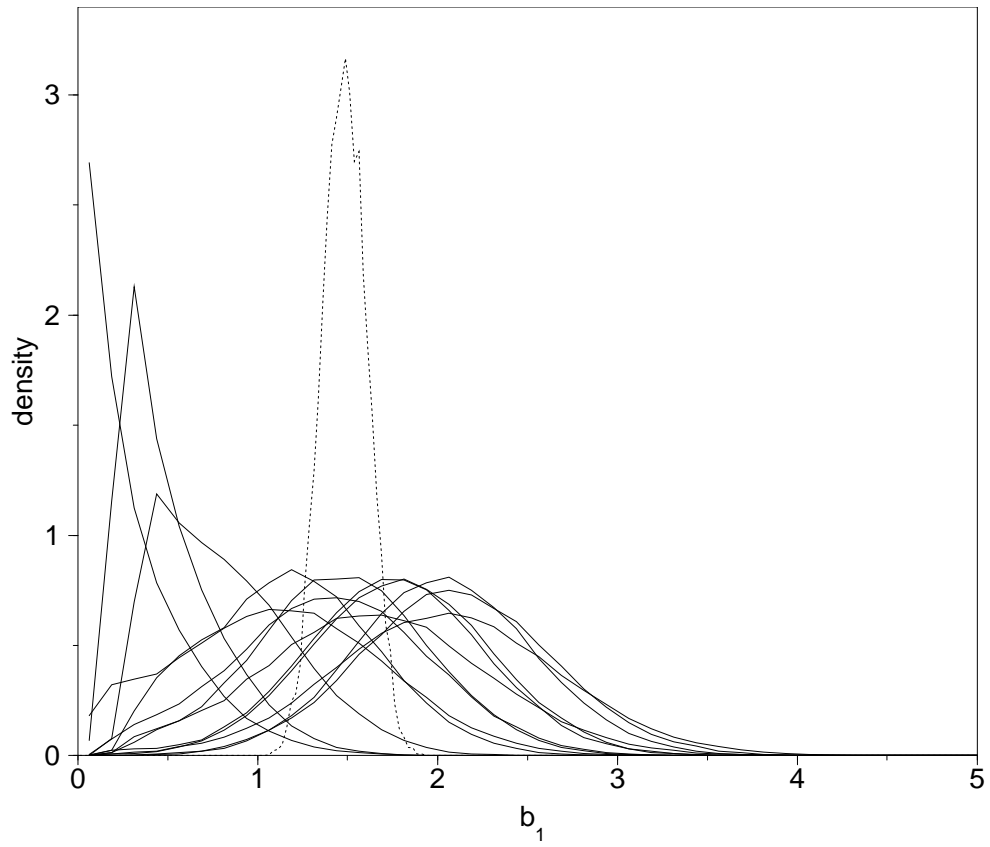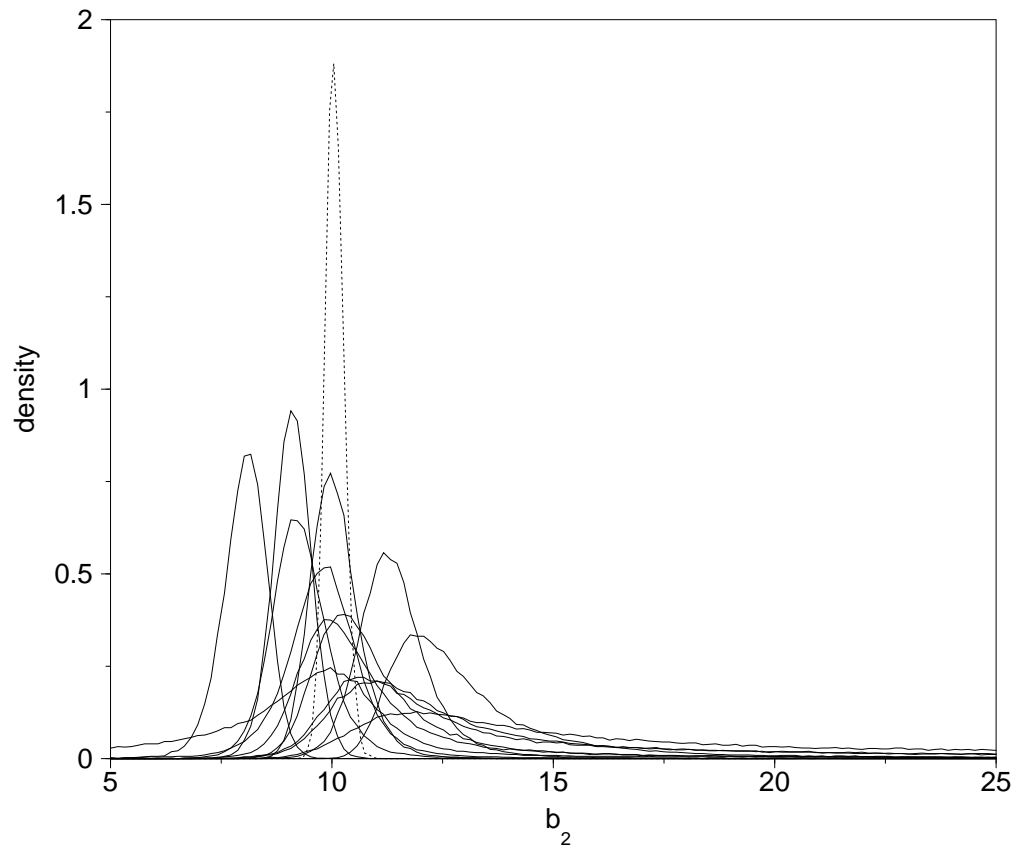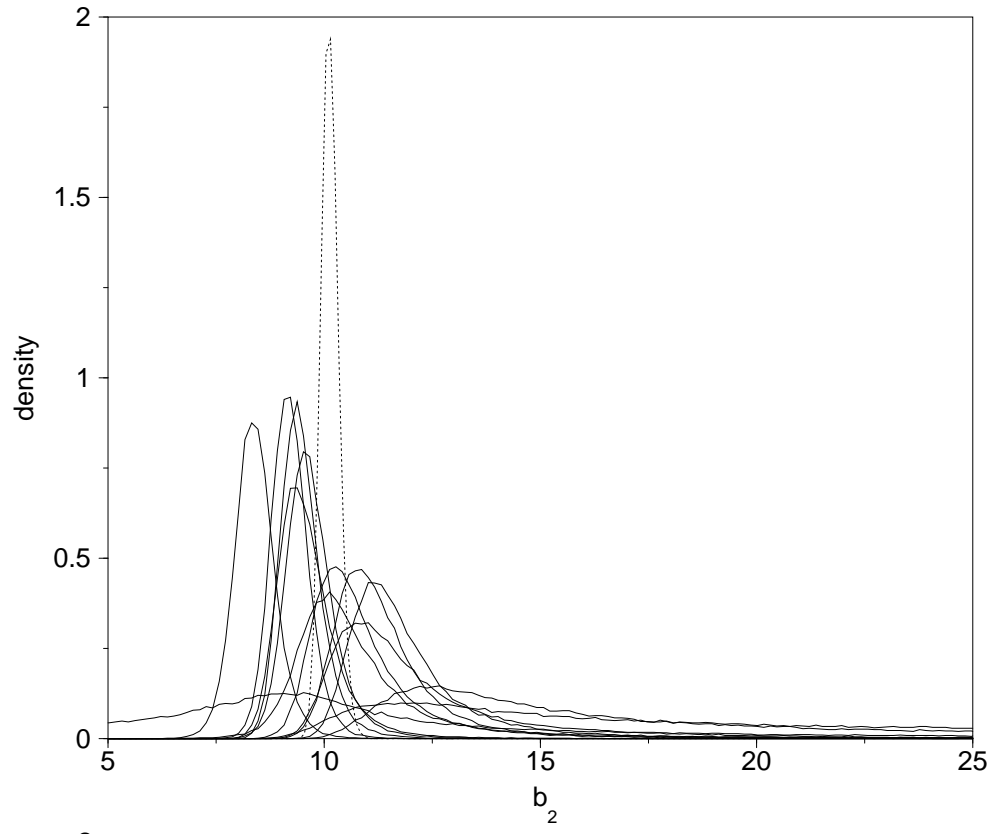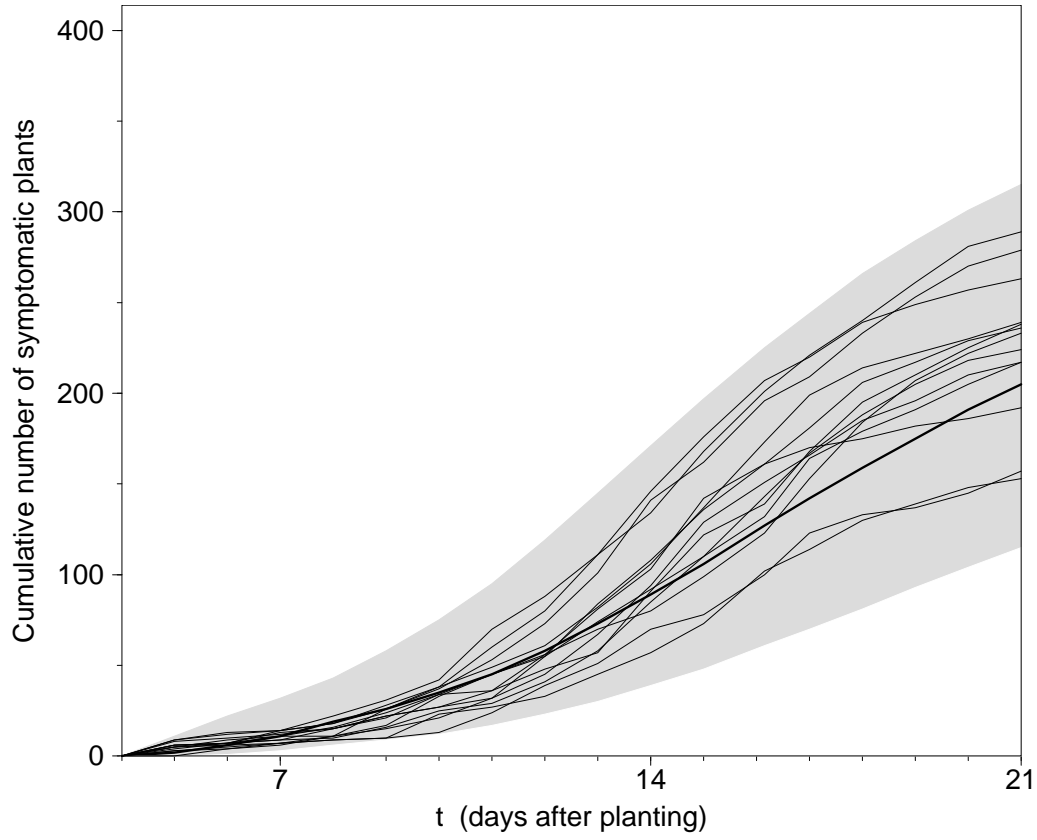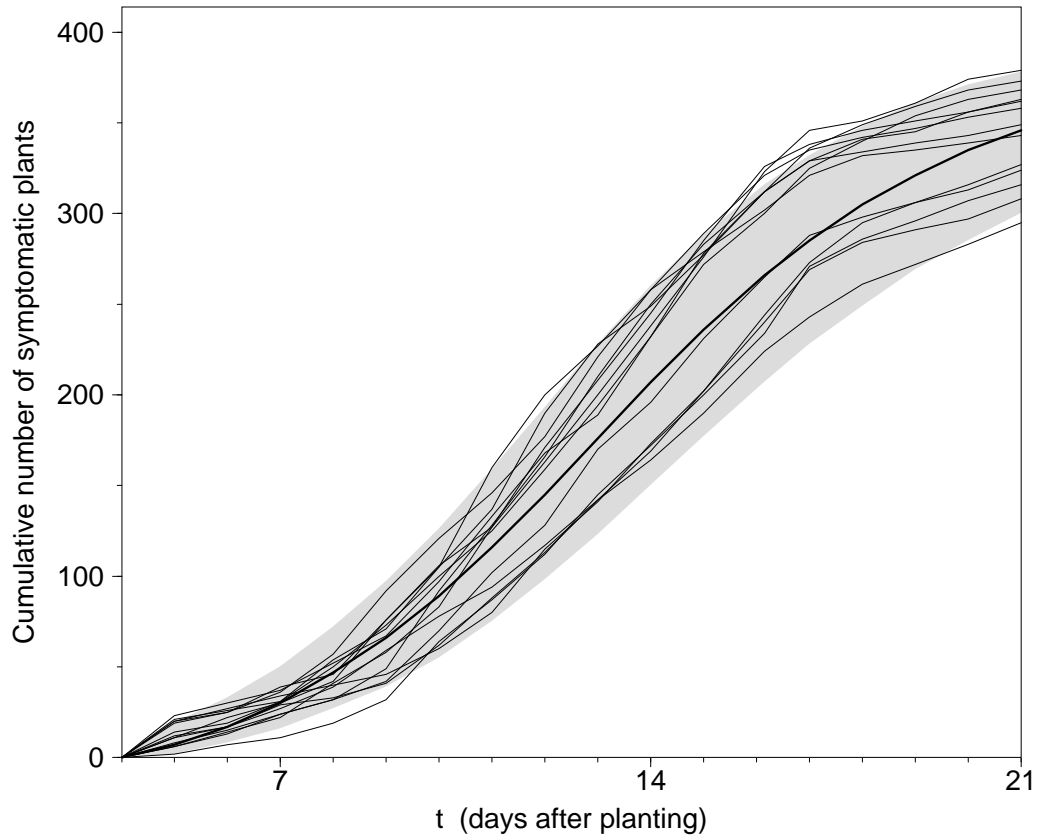
Footnotes

Affiliation of author

[1]Department of Actuarial Mathematics & Statistics, Heriot-Watt University, Riccarton, Edinburgh, EH14 4AS, UK

[2]Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK

[3]Current address: Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG

[4]Biomathematics & Statistics Scotland, James Clerk Maxwell Building, The King's Buildings, Edinburgh EH9 3JZ

Corresponding author: address for correspondence

Professor G J Gibson, Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, UK Fax:+44 131 451 3249; Tel: +44 131 451 3205

e-mail: gavin@ma.hw.ac.uk

Keywords

spatio-temporal modeling

stochastic modeling

fungal pathogens

Bayesian inference

Markov chain Monte Carlo