Statistical Inference and Computational Efficiency for Spatial Infectious-Disease Models with Plantation Data

by (in alphabetical order)

Patrick E. Brown[1], Florencia Chimard[2], Alexander Remorov[3],
Jeffrey S. Rosenthal[3], and Xin Wang[3]

Submitted August 2012, Revised May 2013

Abstract

This paper considers data from an aphid infestation on a sugar cane plantation, and illustrates the use of an individual-level infectious disease model for making inference on the biological process underlying these data. The data are interval censored, and the practical issues involved with the use of Markov Chain Monte Carlo algorithms with models of this sort are explored and developed. As inference for spatial infectious disease models is complex and computationally demanding, emphasis is put on a minimal, parsimonious model and speed of code execution.

With careful coding we are able to obtain highly efficient MCMC algorithms based on a simple random-walk Metropolis-within Gibbs routine. An assessment of model fit is provided by comparing the predicted numbers of weekly infections from the data to the trajectories of epidemics simulated from the posterior distributions of model parameters. This assessment shows the data have periods where the epidemic proceeds more slowly and more quickly than the (temporally homogeneous) model predicts.

Keywords

Markov Chain Monte Carlo; Spatial statistics; Individual-level models

[1] Cancer Care Ontario, 620 University Avenue, Toronto, Ontario, Canada M5G 2L7.

[2] Département de Mathématiques et Informatique, Université des Antilles et de la Guyane, 97159 Pointe-à-Pitre, Guadeloupe.

[3] Department of Statistics, University of Toronto, 100 St. George Street #6018, Toronto, Ontario, Canada M5S 3G3.

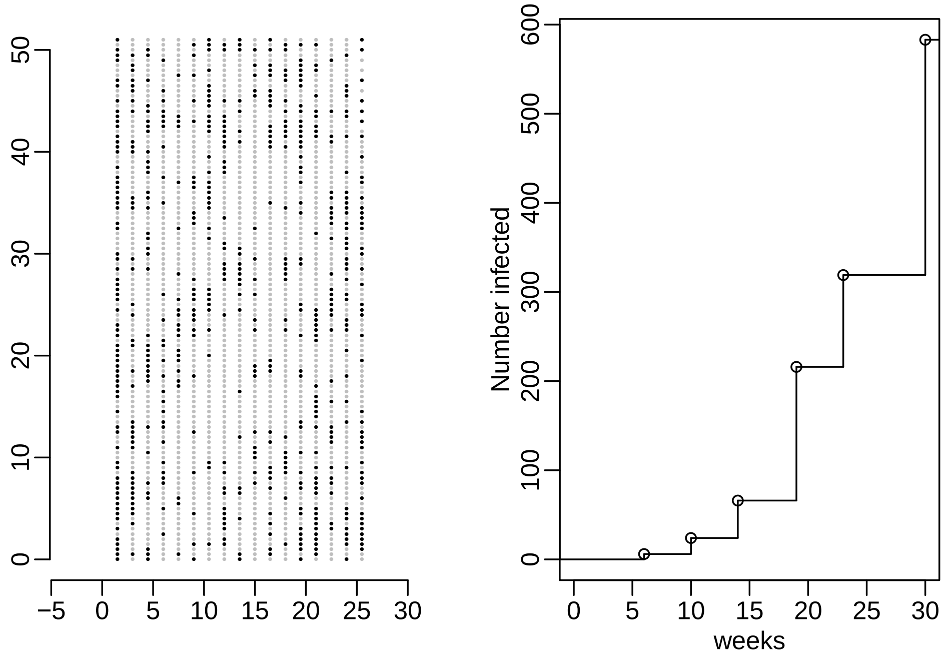
1 Introduction

Individual-level models (ILM) are a conceptually attractive way of quantifying and making inference on the characteristics of an infectious disease outbreak. The key feature of an ILM is that a susceptible individual has a probability of contracting the disease from each one of the infectious individuals. Statistical inference for ILMs is complicated by the fact that the infection events are not independent of one another, as individual i becoming infected increases the disease risk for those individuals whom i might transmit the infection to. This inherent dependence in infectious disease data is particularly problematic when inference is made on incompletely observed data, such as interval censored or aggregated observations. An explicit evaluation of the likelihood function would require integrating over all the unknown infection times, with each infection time affecting the distribution of the others. This is often impractical, and efficient algorithms for making inference on model parameters from interval censored event times is the crux of the problem in practical applications of ILMs.

This paper is motivated by a desire to understand the propagation and time evolution of an insect infestation among 1742 sugar cane plants on a particular experimental field on the Caribbean island of Guadeloupe, with the aim of yielding insights into possible control strategies. The goal has been to develop a computationally efficient and undemanding algorithm for performing statistical inference on an ILM with this dataset, with resulting emphases on parsimony of the model and comparing various implementations of the model-fitting algorithm. We find that by carefully improving and optimising the MCMC algorithm used, we are able to accurately estimate parameters and thus obtain a clear picture of the plant infection dynamics. Further insights are gained by assessing the fit of this parsimonious model to the observed data, and ways in which the biological process departs from the strict homogeneity assumptions of the model are identified.

1.1 The data and model

The insects present on the Guadeloupe plantation are aphids, small flying insects which can lay large numbers of eggs (or more properly nymphs) on stems and the underside of leaves. An egg on a sugar cane will take roughly 3 weeks to develop to an adult, following which it will lay eggs for another 3 weeks (Nuessly, 2005). The plants were inspected at week 0, 6, 10, 14, 19, 23, and 30, with the infection status of each plant recorded. Once a plant is infected, it remains infected (and alive) for the remainder of the study period. Figure 1a shows the locations of sugar canes on this Guadeloupe field, and Figure 1b shows the cumulative number of plants observed to be infected as the experiment progressed.



(a) Locations (in meters)

(b) Cumulative number of infections

Figure 1: Location of 1742 sugar canes which are infected (●) or uninfected (●) at the end of the study period, along with cumulative number of infections over time.

A two stage susceptible-infected model is the most basic of ILMs, and in its simplest form consists of a single parameter θ, being the rate of the Poisson process for an infected individual passing the infection to an individual who is susceptible. While this model is oversimplistic for most real-world applications, the fact that plants never recover from or die as a result of an aphid infection makes the model a reasonable starting point for the sugar cane data. The diffusion of the epidemic in space, with infected plants having a tendency to pass the infection to the plants closest to them, is key aspect of the research question considered and invites an enhancement to the model to accommodate spatial dependence. Doing so can be accomplished with a single additional parameter, σ, which combined with a parametric dispersion kernel $f(d;\sigma)$ gives the rate at which an infected plant located at x_i infects a susceptible plant at x_j as $\theta f(||x_i - x_j||;\sigma)$. These parameters can be interpreted as θ being the rate at which an infected plant produces adult aphids and σ relating to the distance an aphid is likely to travel during its lifetime.

A third parameter μ is added to the model as the rate of the Poisson process whereby a susceptible plant contracts the infection spontaneously and irrespective of the infection status of nearby plants. This parameter has been described in, for example, Meyer et al. (2011) as an *endemic* component whereas θ reflects the *epidemic* component of the infection transmission mechanism. The use of both θ and μ enhances the model's ability to inform control strategies as a large θ relative to μ would suggest propagation can be abated by adjusting plant spacing or treating infected plants. Conversely, a large μ would indicate infections are largely due to external factors and are less amenable to control measures.

This spatial infectious disease model for plants is fairly standard in the ecological literature, and is described in Chapter 7 of Keeling & Rohani (2008).

The assumptions behind this three parameter model are compatible with the sugar cane data in a number of respects: once infected plants remain so; topography of the field is flat and infection rates can plausibly be expected to depend only on distance between plants and not their locations; and infections are detectable nearly immediately following their occurrence via inspecting young leaves for nymphs. However, there are numerous ways in which the biological process would not be expected to conform to the model assumptions. First, a plant's infectivity is assumed to be constant over time following its infection. It might be expected that infectivity will increase over time as the colonisation of the plant progresses, due to either aphids arriving from other plants or as a result of nymphs on this plant maturing and reinfecting their host. Second, the the process is homogeneous in time, and it might be expected that weather and seasonal progression would affect the ability of nymphs to mature and aphids to disperse. Finally, there may be lags between a plant's exposure to infection (when the first egg is laid), an infection being observable on the plant, and the aphids resulting from that infection maturing and the plant being infectious. Diagnostic plots will be used to assess the validity of the model assumptions in light of the concerns above, and the feasibility of possible remedies is discussed.

1.2 Inference

Mathematical modelling of the spatial propagation of infectious diseases is a well established and active research area, engaging in simulation studies and in deriving the stationary distributions of increasingly complex individual-level infectious disease models (see Keeling & Rohani, 2008). Statistical inference for infectious disease models is a much smaller and less developed discipline, and early considerations of parameter estimation include Becker (1989) and Haber et al. (1988). Much of this early work considered non-spatial models where ILMs can be reduced to the distribution of case counts at fixed intervals. McKinley et al. (2009) compare computationally intensive Bayesian inference involving the full likelihood to approximate inference based on matching the case counts from simulated outbreaks to the observed case counts. They conclude that the latter is very effective when the data are completely observed and can still be informative under various types of missing data scenarios.

Inference for spatial models, where transmission probabilities depend on distances between individuals, was considered by Gibson (1997). Infection status at two time points were available to Gibson (1997), with Monte Carlo integration over the (unknown) order of infections used to approximate the likelihood function, and they note that an extension to multiple observation times is straightforward. Diggle (2006) uses a partial likelihood which considers only the ordering of events and not event times, the infection times are not interval censored but rather observed after a (constant) reporting delay. Deardon et al. (2010) use linear approximations to the infection kernel f to make full Bayesian or likelihood-based inference practical when infection times are observed, even when the datasets considered are large.

Here we consider a Markov Chain Monte Carlo (MCMC) algorithm for performing Bayesian inference on a spatial individual-level infectious disease model with interval-censored data. MCMC provides a natural and statistically efficient procedure for accounting for interval-censored data by treating the unknown infection times as latent variables for which posterior samples are drawn at each iteration. MCMC for infectious disease models was pioneered by O'Neill et al. (2000), and has since been used for models of increasing complexity as MCMC algorithms and processing power have improved. Jewell et al. (2009) use an MCMC for fitting a complex model involving a large number of parameters relating the probability of transmission of a disease between two farms to farm-level covariates. When the number of infected individuals is large, MCMC becomes computationally burdensome as the sampling of each of the infection times at every iteration can be time consuming. Developing an efficient implementation of an MCMC algorithm for the sugar cane data, able to perform inference in a reasonable amount of time on a workstation computer, is a central aim of this paper.

2 Methods

2.1 Model and Likelihood

We will begin by describing the model an likelihood for the scenario where the infection times τ_i are directly observed. This likelihood is then used in Section 3.2 to make inference on the model parameters using the interval censored data present in the sugar cane application.

Recall that at time t a susceptible plant located at s receives spontaneous infections with rate μ, and an infections from each of plant j infected prior to t with rate $\theta f(s - x_j; \sigma)$. The intensity $\lambda(s, t)$ of all infections arriving at s at time t is the sum of these individual intensities. Writing τ_i as the infection time for plant $i = 1 \dots 1742$, the rate of infection is

$$\lambda(x_i, t) = \mu + \sum_{j;\tau_j < t} \theta f(x_i - x_j; \sigma). \tag{1}$$

The infection rate is increasing in time, with increasing t resulting in a greater number of infected plants contributing to the intensity. Although the Poisson process assumption can result in multiple infections occurring in a plant, the first of these infections which moves the plant from the susceptible to the infected state and any subsequent infections are not observable.

The likelihood of observing a set of infection times $\tau = \{\tau_1 \dots \tau_N\}$ can be thought of as the product of 1) the probability of not observing infections during each plant's time in the susceptible state, and 2) the probability (or density) of observing infections at each of the τ_i. The Poisson process assumption dictates that the number of infections in a time interval is Poisson distributed with mean equal to the intensity function integrated over the period. Hence the first term in plant i's likelihood, the probability that no infections occur in the interval from zero to τ_i, is

$$\exp\left[-\int_0^{\tau_i} \lambda(x_i, u) du\right].$$

The second component of the likelihood, the density for the infection time τ_i, is simply $\lambda(x_i, \tau_i)$. Plants which are not infected by the final observation time T contribute a probability of $pr(\tau_i > T)$ to the likelihood, without the second term. The product over all plants results in the likelihood function

$$L(\mu, \theta, \sigma | \tau_1 \dots \tau_N) = \left(\prod_{i;\tau_i \leq T} \exp\left[- \int_0^{\tau_i} \lambda(x_i, t) dt \right] \lambda(x_i, \tau_i) \right) \left(\prod_{i;\tau_i > T} \exp\left[- \int_0^{T} \lambda(x_i, t) dt \right] \right). \quad (2)$$

Substituting in the intensity function from (1) gives

$$- \log L(\mu, \theta, \sigma | \tau_1 \dots \tau_N) = \sum_{i;\tau_i \leq T} \left(\tau_i \mu + \sum_{j;\tau_j < \tau_i} (\tau_i - \tau_j) \theta f(x_i - x_j; \sigma) \right) + \sum_{i;\tau_i > T} \left(T\mu + \sum_{j;\tau_j < T} (T - \tau_j) \theta f(x_i - x_j; \sigma) \right) - \sum_{i;\tau_i \leq T} \log \left[\mu + \sum_{j;\tau_j < \tau_i} \theta f(x_i - x_j; \sigma) \right]. \quad (3)$$

It remains to specify a parametric form for the infection kernel $f(d; \sigma)$, and we have chosen a radially symmetric bivariate Gaussian density with

$$f(d; \sigma) = (1/2\pi\sigma^2) \exp(-||d||^2/2\sigma^2).$$

The use of the Gaussian kernel motivated by the Gaussian being the stationary distribution of a Brownian motion. Writing $A_k(t)$ as the location of aphid k at time t, we assume that movements within a short time interval of length ϵ are normally distributed with $\text{var}[A(t+\epsilon) - A(t)] = \nu^2\epsilon$. An aphid born at time t_0 and location s_0 will have $pr[A_k(t_1) = a] = f[a - s_0, \nu^2(t_1 - t_0)]$, with σ^2 above being the stationary variance after one week of aphid movements. Introducing a further parameter to allow for heavier or lighter tails in the dispersion kernel, by generalising f to be a multivariate t-distribution, is also considered.

2.2 Prior Distributions

Weakly informative Gamma priors are used for the three parameters as follows: $\mu \sim \text{Gamma}(0.7, 0.004)$ and $\theta \sim \text{Gamma}(0.8, 10)$ (measured as infections per week), and $\sigma \sim \text{Gamma}(0.5, 100)$ (in meters). Interpreting these prior distributions is helped by the following prior 95% prediction intervals:

μ: the expected number of spontaneous infections is between 1 and 630 over the study period, recalling that the total number of infections observed is 583.

θ: an infected plant surrounded by susceptible plants has an average time to its first infection between 1 day and 16 months.

σ: 95% of aphids will have traveled less than 10cm at the 2.5% quantile of the prior distribution and 500m at the 95.5% quantile of the prior.

Having μ and θ at the lower end of their prior distributions would result in very few plants being infected over the 30 weeks, whereas values at the upper end of the priors would result in the entire plantation being infected within days. The range σ at the lower end of the prior would result in the infection being unable to spread between plants (which are 50cm apart). A value near the upper end would make the distribution of aphids flat over the 50 meter long plantation and the model would be effectively non-spatial. These priors thus allow for all parameter values which could create a plausible epidemic.

2.3 Inference

Recall that the infection times τ_i are unobserved, with the observed data $\mathbf{Y} = \{Y_i; i = 1 \ldots N\}$ consisting of vectors Y_i being plant i's status at each of the 6 occasions on which the plantation was surveyed. The τ_i are therefore interval censored, with each plant's infection occurring somewhere within the last occasion on which plant i was observed as susceptible and the first occasion where it was observed as infected. Closed form expressions for the likelihood of the observed Y_i are available in survival models which assume independence between observations, obtained by integrating out the τ_i. With infectious disease models each τ_i affects the distribution of every other τ_j, as evidenced by the double sums in (3), and the likelihood of the interval censored data is intractable.

Bayesian inference using MCMC is well suited to data of this sort, with the τ_i being treated as latent variables and accommodated through data augmentation (see e.g. Jewell et al., 2009). Prior distributions $p_\mu(\cdot)$, $p_\theta(\cdot)$, and $p_\sigma(\cdot)$ are specified for the three model parameters, and an MCMC algorithm produces samples from the posterior distribution $\pi(\mu, \theta, \sigma, \tau|\mathbf{Y})$. A random-walk Metropolis algorithm is used here to update each parameter and latent variable in sequence, using the following general algorithm.

1. Initialize the algorithm at iteration $r = 0$ with initial values $\tau_i^{(0)}$, $\mu^{(0)}$, $\sigma^{(0)}$, $\theta^{(0)}$;
2. At iteration r initially set $\tau_i^{(r)} = \tau_i^{(r-1)}$, $\mu^{(r)} = \mu^{(r-1)}$, $\sigma^{(r)} = \sigma^{(r-1)}$, $\theta^{(r)} = \theta^{(r-1)}$,
3. Simulate a proposal $\mu^* \sim N(\mu^{(r-1)}, \nu_\mu)$.
4. Set $\mu^{(r)} = \mu^*$ with probability

$$pr(\mu^{(r)} = \mu^*) = \min \left[1, \frac{L(\tau_1^{(r)} \ldots \tau_N^{(r)}; \mu^*, \theta^{(r)}, \sigma^{(r)}) p_\mu(\mu^*)}{L(\tau_1^{(r)} \ldots \tau_N^{(r)}; \mu^{(r)}, \theta^{(r)}, \sigma^{(r)}) p_\mu(\mu^{(r)})} \right]. \tag{4}$$

   otherwise set $\mu^{(r)}$ is unchanged from $\mu^{(r-1)}$.

5. Repeat steps 3 and 4 for θ and σ.
6. For each $i = 1 \ldots N$, propose a new τ_i^* and accept with probability

$$pr(\tau_1^{(r)} = \tau_1^*) = \min \left[1, \frac{L(\tau_1^{(r)} \ldots \tau_{i-1}^{(r)}, \tau_i^*, \tau_{i+1}^{(r)} \ldots \tau_N^{(r)}; \mu^{(r)}, \theta^{(r)}, \sigma^{(r)})}{L(\tau_1^{(r)} \ldots \ldots \tau_N^{(r)}; \mu^{(r)}, \theta^{(r)}, \sigma^{(r)})} \right]. \tag{5}$$

7. Return to step 2.

The θ, μ and σ have proposal distributions which are normally distributed, with mean equal to their previous value, and with standard deviations given by 0.005, 0.0005 and 0.05 respectively. New values τ_i^* are proposed from the proposal distribution $N(\tau_i^{(r)}, 1)$. The standard deviations of these proposal distributions were selected following visual assessment of chain mixing during a number of trial runs of the algorithm.

2.4 Implementation

There are 583 plants infected by the end of the study period and 583 unknown infection times to account for. The most straightforward implementation of the algorithm above would require calculating the likelihood function 586 times at each iteration (one likelihood for each of the 3 parameters and once for each of the 583 unknown infection times). Implementing the algorithm in this way is likely to result in unfeasibly long computational times, especially considering the long chains and heavy thinning often required to reduce the dependence in random-walk Metropolis MCMC of this type. A central aim of this paper is to explore the feasibility of several variations on this basic algorithm, investigating the possibility of exploiting computational efficiencies as an alternative to more sophisticated MCMC algorithms. These efficiencies include: pre-calculating and storing quantities used repeatedly; more efficient calculation of the likelihood ratio; truncating the kernel f; and performing as many calculations as possible in parallel. A detailed description of the algorithms used follows, and the C code for each of the algorithms appears on the journal web site.

2.4.1 Basic algorithm

The first algorithm is essentially as described above, direct evaluation of the likelihoods, though the two most obvious efficiencies are exploited. First, the distances between plants $||x_i - x_j||$ are pre-computed and stored, saving the time it would take to re-calculate these distances each time the likelihood is evaluated. Second, a number of terms in the likelihood ratios in (5) for updating the τ_i are identical in the numerator and denominator. Cancelling these terms results in a simplified expression for the likelihood ratio, as described in Appendix A, and resulted in much faster running times.

2.4.2 Parallel Algorithm

This algorithm involves using multiple computer cores to update the τ_i simultaneously to the greatest extent possible. The term involving i and j in the likelihood ratio in Appendix A for the updating of $\tau_i^{(r)}$ involves only $\tau_i^* - \tau_i^{(r)}$ and not $\tau_j^{(r)}$ unless the proposal τ_i^* would change the order of infection of i and j. When τ_i and τ_j are known to occur in different time intervals (having been first observed as infected at different times), any proposal which changes their ordering would be rejected and any changes to $\tau_k^{(r)}$ during the updating will not affect the updating of $\tau_i^{(r)}$. It is therefore possible to update infection times simultaneously when they occur in different intervals, and this parallel algorithm runs four parallel sets of updatings. The first three observation periods (which together have fewer infections than any of the subsequent three periods) are updated on one core, with each of the final three observations periods on separate cores (using four cores in total).

2.4.3 Improved Parallel Algorithm

This algorithm exploits two further efficiencies. First, the likelihood ratio for $\mu^{(r)}$ in (4) simplifies considerably, as described in Appendix A. Second, the values of $f(x_i - x_j; \sigma^{(r)})$ are pre-computed and stored as they are used multiple times at each iteration. Unlike the distances $||x_i - x_j||$, these values change at each iteration (or every iteration where σ changes), and the values are re-computed and stored periodically. Note that $f(x_i - x_j; \sigma^{(r)})$ only appears in the likelihood if one of i and j are infected during the study period. Thus only $600 \cdot (1742 - 600)$ values must be computed rather than 1742^2, and this computation is done in parallel on 4 cores.

2.4.4 Truncated Algorithm

Whereas the two parallel algorithms are mathematically equivalent to the Basic Algorithm, the Truncated Algorithm approximates the likelihood ratios in the hopes that the resulting loss of accuracy is negligible. The kernel $f(d; \sigma)$ is truncated with $f(d; \sigma) = 0$ when $||d|| > 4\sigma$, with the result that terms involving i and j in the likelihood are zero if $||x_i - x_j|| > 4\sigma$. The truncated approximation to the likelihood ratio can be computed quickly by computing and storing, for each plant i, the order from smallest distance to greatest distance of each of the other plants. Each double summation in the likelihood ratio proceeds with increasing distances between plants and ceases once a distance greater than 4σ is reached. The value at which f is truncated can of course be varied, a value of 4σ was chosen because the values set to zero are always less than 10^{-4}.

The Truncated Algorithm adds only the truncation approximation to the Basic Algorithm, and uses none of the efficiencies of the Parallel algorithms.

2.4.5 Discrete time algorithms

Two final algorithms, included more for comparison than an expectation that it will offer computational advantages over the other algorithms, approximate the likelihood by allowing infections at only a finite number of time points $\tilde{t}_0 \ldots \tilde{t}_M$. Using this approximation the likelihood can be written as a product over time with

$$\tilde{L}(\mu, \theta, \sigma | \tau_1 \ldots \tau_N) = \prod_{m=1}^{M} \left\{ \prod_{i;\tau_i > t_m} \exp[-(t_m - t_{m-1})\lambda_{im}] \prod_{i;T_k < \tau_i \le T_{k+1}} [1 - \exp(-(t_m - t_{m-1})\lambda_{im})] \right\}, \tag{6}$$

where

$$\lambda_{ik} = \mu + \theta \sum_{j;\tau_j < t_m} f(x_i - x_j; \theta). \tag{7}$$

When $\tau_i^{(r)} = t_m$, it is updated by proposing either $\tau_i^* = t_{m-1}$ or $\tau_i^* = t_{m+1}$ with equal probability. The likelihood ratios for τ_i^* are simpler than the continuous-time likelihoods and involve only the plants j with $\tau_j^{(r)} = \tau_i^{(r)}$ or $\tau_j^{(r)} = \tau_i^*$.

Two discrete time algorithms are implemented: a Basic Discrete Time Algorithm similar to the Basic Algorithm; and a Truncated Discrete Time Algorithm where the truncated kernel is used.

3 Results

3.1 Computing time

Table 1 shows the time taken for 100 MCMC iterations of each of the algorithms described above, using a quad-core 2GHz Opteron processor. Many of the results could be foreseen, with parallelizing reducing the time taken to update the τ substantially and improving the parallel algorithm by storing the $f(x_i - x_j; \sigma^{(r)})$ results in further time savings. The untruncated discrete time approximation is, unsurprisingly, considerably more computationally intensive than the continuous-time implementation. The magnitude of the reduction in computational time resulting from the truncation approximation is perhaps more surprising. The continuous time algorithm is sped up by a factor of nearly 50 and the discrete time algorithm improves from 15 minutes per 100 iterations to a more manageable 36 seconds.

Algorithm	θ	μ	τ	σ	total
Basic	28.15	14.27	160.92	28.99	232.33
Parallel	25.08	12.57	29.99	25.17	92.81
Improved	2.38	0.25	10.18	12.31	25.12
Truncated	1.07	1.07	1.76	1.13	5.03
Basic Disc	261.58	253.65	122.81	263.19	901.23
Trunc Disc	16.65	0.35	2.15	16.80	35.95

Table 1: CPU time, in seconds, for 100 iterations for the MCMC implementations listed in Section 2.4. Times taken to update each of the parameters and latent variable τ are shown separately with the total time appearing in the final column.

Random-walk Metropolis algorithms of the type used here often require thinning to produce independent samples, and many hundreds or thousands of iterations can be required to obtain accurate estimates. The results in Section 3.2 are from chains of 125,000 iterations with 5000 samples retained after burnin and thinning, and only the three algorithms quicker than 60 seconds per 100 iterations are able to produce a set of results in less than one day. Appendix B shows trace plots and correlations for the improved parallel algorithm and the other continuous-time algorithms (being mathematically equivalent or approximately so) had very similar mixing properties.

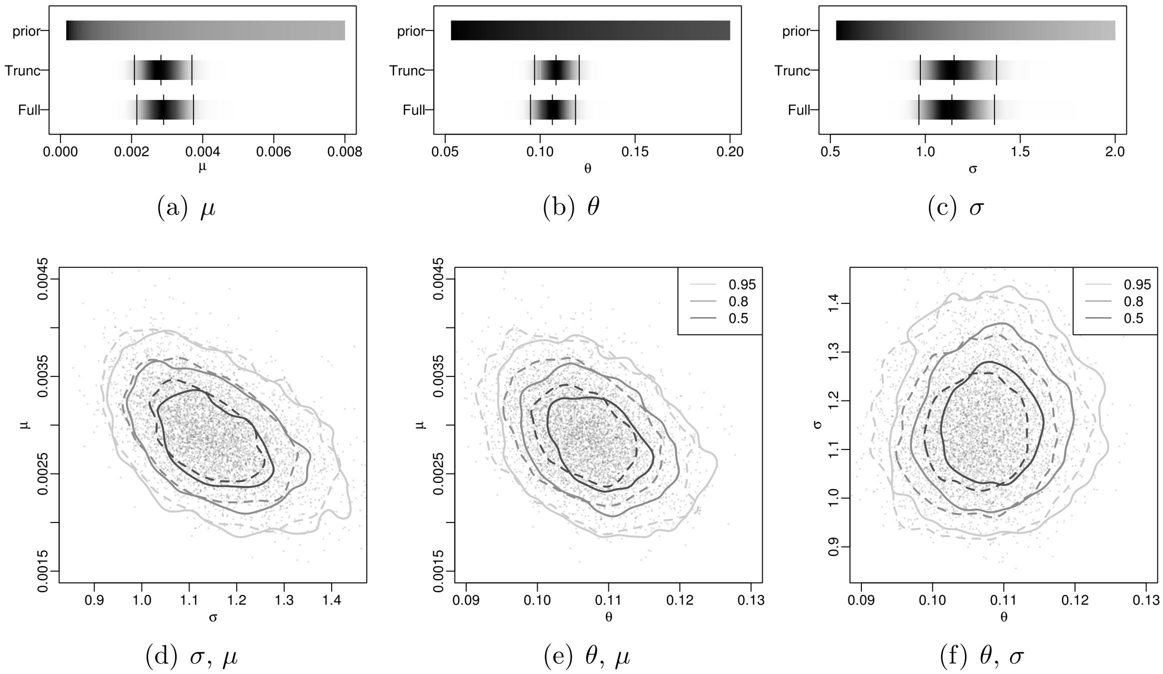
Our initial reaction to the performance of the Basic algorithm, at over 3.5 days per 5,000 samples retained, was to conclude that conventional MCMC was not powerful enough for this problem. We considered using far more complicated algorithms (e.g. particle MCMC, see Andrieu et al., 2010) in an effort to overcome this problem.

Our subsequent experience with parallelization, truncation, pre-sorting, and storing kernel values have led us to conclude that efficient coding results in inference on ILMs being possible with even the simplest of MCMC algorithms. The parallel implementation of the untruncated model would be expected to update the τ in one quarter the time of the basic untruncated model. The pre-sorting of plants by infection time, and having separate loops for infected and uninfected plants rather than the basic algorithm's single loop with a check for each plant's infection status, has resulted in yet further efficiency gains. All the algorithms pre-compute and store the distance matrix, and the improved parallel algorithm's storing of the matrix of $f(x_i - x_j; \sigma^{(r)})$ produces substantial time reductions for all parameters. The time taken to compute the $f(x_i - x_j; \sigma^{(r)})$ values is included in the σ column, and even with this step being parallelized the updating of σ is more time consuming than the updating of the τ. The improved parallel algorithm also computes the likelihood ratio for μ, which is considerably faster than evaluating the likelihood.

Truncating and pre-sorting the plants by distance introduces an approximation to the inference methodology, but improves the computational speed to the extent that no further improvements or coding efficiencies seem necessary. The gains from truncation will diminish as σ increases, however, and datasets exhibiting a large degree of spatial dependence might require parallelization and storing the matrix of kernel evaluations in addition to truncation.

3.2 Inference on model parameters

We next show and interpret the posterior distributions of model parameters, and assess the adequacy of the truncation approximation. Figure 2 shows the distribution of posterior samples from the improved parallel algorithm and from the truncated algorithm. Chains were run for 125,000 iterations, with the first 1000 iterations discarded as burnin, and subsequently thinned with one sample in 25 being kept thereafter, for a total of about 5000 iterations being retained. Figure 2(a-c) shows the marginal prior and posterior distributions of the three model parameters, with posterior distributions shown for both the truncated and untruncated continuous time model. Joint bivariate posterior samples for all parameter combinations are shown in Figure 2(d-f).



(a) μ

(b) θ

(c) σ

(d) σ, μ

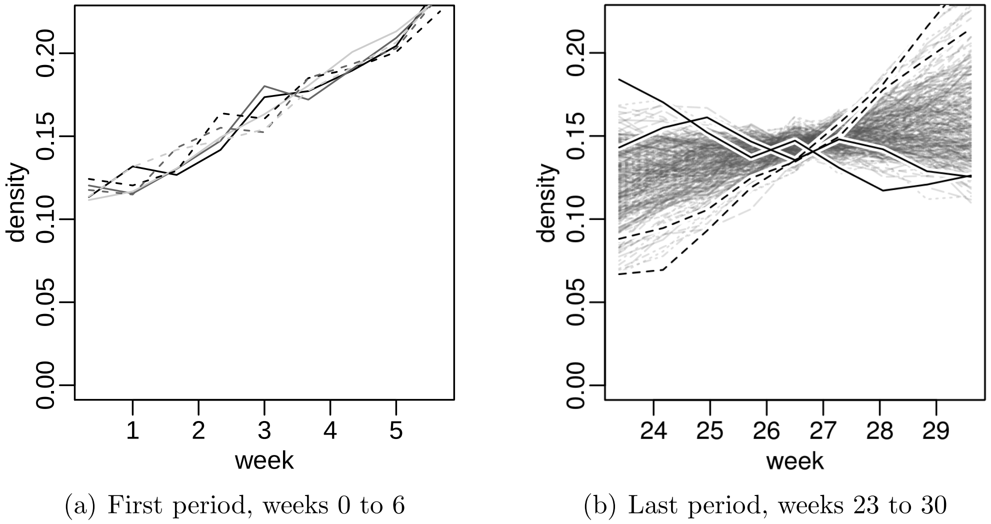
(e) θ, μ

(f) θ, σ

Figure 2: Prior and posterior distributions of: the endemic infection rate μ; the rate at which an infected plant produces aphids σ; and the spatial range of aphid movements σ. Shown are posterior means and 95% intervals (—), and bivariate confidence regions for the truncated (Trunc or —) and non-truncated (Full or - - -) algorithms. Also shown are bivariate posterior samples for the non-truncated algorithm (∘).

In Figure 2 there is some suggestion that the truncation approximation has increased θ and σ and reduced μ. This is particularly evident in Figure 2(f). This output thus provides some small argument against truncating, and the improved parallel untruncated continuous-time algorithm is shown for our remaining estimates. (In addition, in Appendix B we present trace plots and autocorrelation functions to illustrate that this algorithm is indeed mixing adequately following thinning.)

We next consider the infection times τ_i. Figure 3a shows the posterior distributions of the τ_i for 6 infections which occur in the first observation period, with Figure 3b doing the same for the 256 infections which occurred in the final period. The absence of infected plants at the start of the first period implies that the infections which do occur are ‘spontaneous’ infections due to μ, with identical posterior distributions for all 6 infections. In the final period, the susceptible plants in closer proximity to infected plants are infected near the start of the period (solid black lines), with the dashed black lines corresponding to plants which were likely infected following the infections of a number of its neighbours.



(a) First period, weeks 0 to 6

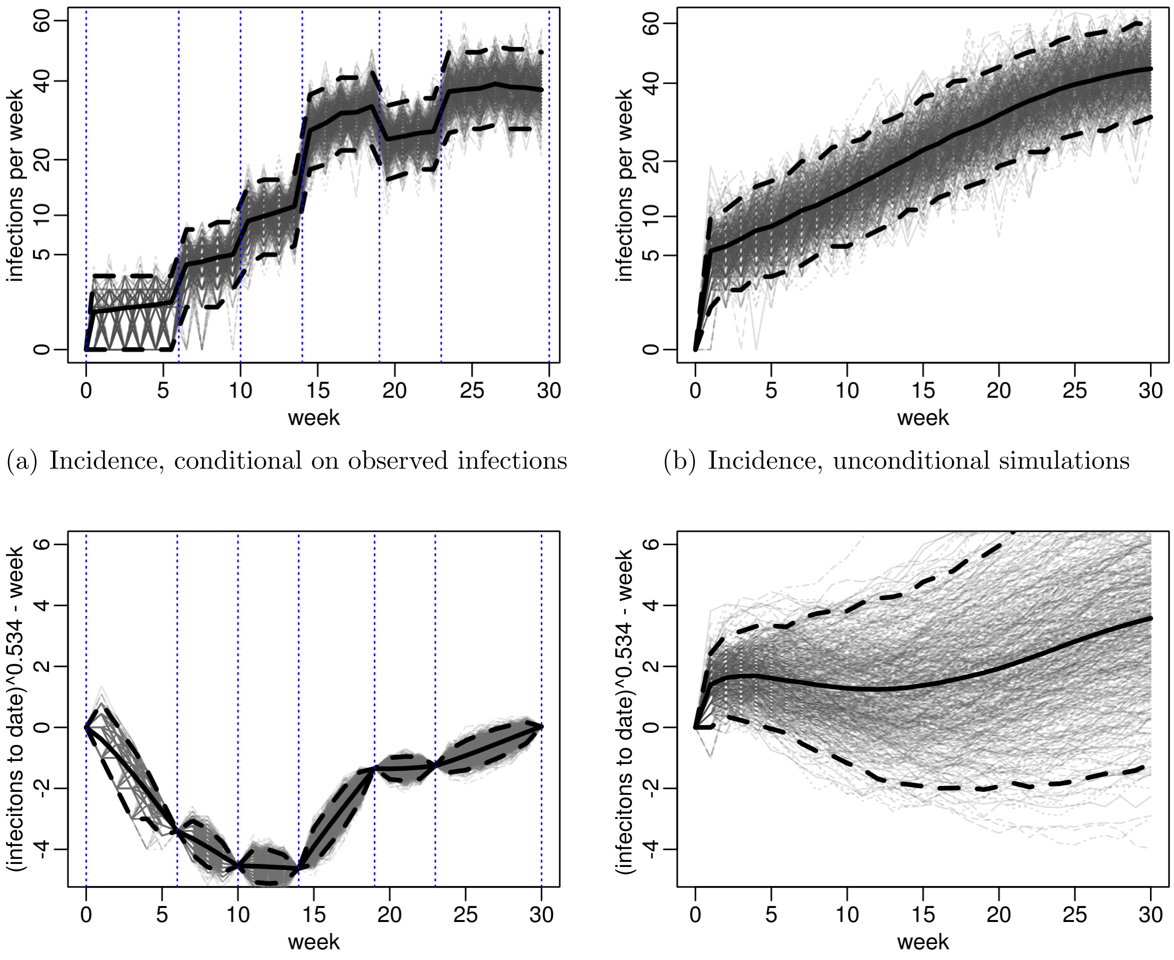
(b) Last period, weeks 23 to 30

Figure 3: Posterior densities for the time of infection for plants known to be infected during the first inspection period and the last inspection period. Each line represents the density of an individual plant's infection time.

Figure 4a shows posterior samples and 95% pointwise intervals for the number of new infections per week, conditional on the 6 observations. Figure 4b, by contrast, simulates new epidemics for each of the posterior samples of the model parameters, without reference to the observed infections. The black solid and dashed lines are posterior means and 95% intervals respectively, with the remainder of samples being in grey. The grey lines are semi-transparent with darker areas having a greater number of overlaid lines. This pair of graphs can be seen as a form of model diagnostics, with a good model fit being demonstrated when the data-driven samples in the former graph being similar to the model-driven samples in the latter. There are differences between the two graphs, however, primarily the higher initial infection activity in the unconditional simulations and the sharp discontinuity between weeks 14 and 19 in the data-driven posterior samples. Posterior means and intervals for the truncated model are shown in red, and for the most part coincide with the untruncated model.

Figures 4c and 4d are analogous plots showing the prevalence, or cumulative number of infections, transformed by exponentiating the counts to the power of $\log(30)/\log(583) \approx 0.534$ and subtracting the number of weeks since the start of the study. This transform was chosen because a horizontal line at $y = 0$ corresponds to prevalence increasing exponentially to 583 infections on week 30. The number of total infections is observed directly at the 6 observation times, with the width of the 95% intervals shrinking to zero on these occasions. The unconditional simulated epidemics in Figure 4d are, unsurprisingly, considerably more variable than the conditional samples.

The most apparent inconsistency between the unconditional simulations and the posterior distribution derived from the data is the slopes of the prevalences in Figures 4c and 4d. Exponential growth corresponds to horizontal lines, with the unconditional samples exhibit more rapid than exponential growth during the first three weeks and roughly exponential growth thereafter. The prevalence of aphids in the sugar cane dataset increases much more slowly than exponential growth before week 10 and substantially more rapidly than exponential between weeks 14 and 20.



(a) Incidence, conditional on observed infections

(b) Incidence, unconditional simulations

(c) Prevalence, conditional on observed infections

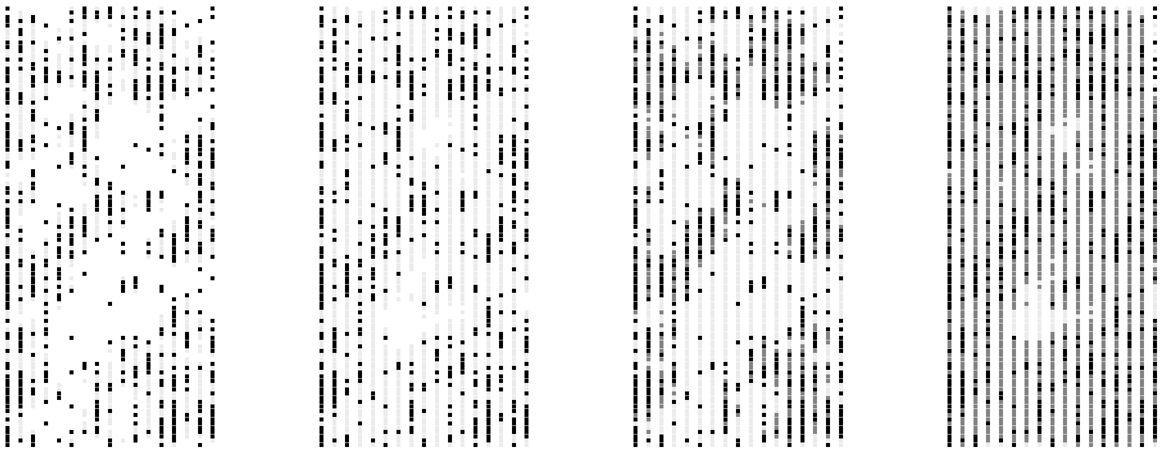
(d) Prevalence, unconditional simulations

Figure 4: Incidence (number of new infections per week) and prevalence (cumulative number of infections to date) as sampled from the posterior distribution conditional on the interval-censored infection times, and unconditional simulations using the posterior samples of model parameters. Shown are individual samples (—), posterior means (—) and 95% intervals (- - -) .

Finally, we turn to the question of longer-term prediction. Figures 9(a-d) forecast the epidemic past the 30 weeks for which data are observed, showing each plant's probability of being infected by weeks 35, 40, 50 and 60. Notice the plants close to infected plants are forecast to become infected first, and by week 60 nearly all plants are likely to be infected.

4 Discussion

This exploration of an aphid infestation on a sugar cane plantation had the dual aims of: using a simple ILM to yield insights into the underlying biological process; and of better



(a) week 35 (b) week 40 (c) week 50 (d) week 60

Figure 5: Probabilities forecast at week 30 of each plant being infected by week 35, 40, 50 and 60. Probabilities are 0-19% ( ), 20-79% (●), 80-99% (●), 100% (●)

understanding the practical issues related to the use of MCMC with ILMs. Addressing the first aim, it has been shown that aphids can infect plants up to 2σ (on the order of two to three meters) away from the host plant. The plants are currently spaced 0.5m apart in the y-direction; increasing the spacing by 10% or 20% would be unlikely to have a demonstrable effect on the aphid spread and reducing the number of plants by a factor of two or three would likely be required before the epidemic were to be slowed substantially. The weekly number of spontaneous infections, or more likely infections caused by long-range phenomenon, amongst 2000 susceptible plants is Poisson distributed with mean approximately 6 and unlikely to exceed 15. A plantation of 2000 sugar canes could be kept aphid-free by having the capacity to detect and treat 15 spontaneous infections per week sufficiently quickly that the nymphs on these infected plants were unable to mature and spread the epidemic.

The model applied to the sugar cane data is overly simplistic, and comparing the observed data to simulated time trajectories from the fitted model suggests this simple model is unable to reflect the temporal dynamics of the underlying biological process. The epidemic starts much more slowly than the fitted model would predict, as evidenced by the downward slope in Figure 4c, followed by faster growth than the model allows for with a subsequent levelling off from week 20. One possible cause of this phenomenon is that the assumption of time homogeneity is incorrect, and seasonal or meteorological factors were particularly conducive to the spread of the infestation in weeks 14 to 20. A second potential explanation is the failure of the model to account for a possible time lag between infection of a plant and the plant becoming infectious. An SEI model (Susceptible-Exposed-Infectious) would introduce an additional state of 'exposed but not yet infectious', and a model more general still would allow for a gradual increase in a plant's infectivity over time as nymphs from the original infection mature and re-infect the host plant. The slowing of the infection rate between weeks 20 and 25 would suggest that time inhomogeneity rather than a time lag in infectivity is more likely. SEI models do exhibit slowing of the infection rate when the number of susceptible plants available for infection decreases, though 85% of plants are still susceptible in week 20 so a change in infectivity during this period is the more likely explanation.

An assessment of the spatial aspects of the model assumptions (radially symmetric Gaussian infection kernel, spatial homogeneity) has not been presented, and there are no established and widely recognised methods for accomplishing this. While the plots comparing conditional and unconditional simulations of case counts over time can assess the assumptions related to the temporal dynamics, it is not clear what a spatial analog of these plots would be. One options would be to introduce additional parameters into the infection kernel f to allow for more general profiles. We have implemented a multivariate-t density for f to check the robustness of the results to the very weak tails in a Gaussian kernel. This analysis produced conditional and unconditional simulations of infection counts which were indistinguishable from the Gaussian kernel, though the fitted t-density kernel had heavier tails and a sharper peak (see Figure 8). Directional effects could be assessed by using a kernel with elliptical contours as opposed to the circular contours used here. This would involve two additional parameters (ratio of rotation and angle of major to minor axis lengths), likely worsening the chain mixing and requiring a higher number of iterations. A yet more complex algorithm would allow the data to choose between possible kernels (perhaps including a kernel with finite support) with a reversible jump MCMC (Green, 1995).

The outcome from the second goal of the paper, an exploration of the computational considerations related to the use of MCMC for ILMs, is that random-walk Metropolis MCMC is entirely feasible for use with populations of thousands of individuals (if programmed very carefully). The initial effort at improving the simple Metropolis-within-Gibbs algorithm involved storing the distance matrix and simplifying the acceptance probabilities for the infection times τ_i, improvements which dramatically lowered the time taken per iteration but still proved unacceptably slow. A decision had to be made between either using a more sophisticated MCMC algorithm or creating a more efficient implementation of the existing algorithm. The dependence structure inherent in ILMs would cause problems for many of the more sophisticated methods for Bayesian inference: the non-Gaussian distribution of the latent variables precludes the use of Integrated Nested Laplace Approximations (see Rue et al., 2009); the lack of conditional independence of the observations would complicate the calculations of manifolds for the use with Reimann Manifold Hamiltonian MCMC (see Girolami & Calderhead, 2011); the lack of a closed form for the conditional distributions negates some of the advantages of particle Gibbs (see Andrieu et al., 2010).

Two avenues were identified for creating an efficient and practical implementation of the algorithm: truncation of the kernel; and parallelization. Truncation of f has a disadvantage with respect to parallelization in that it introduces an approximation into the algorithm. Also, the benefits from truncation will decrease as the spatial dependence parameter σ increases and the number of pairs of plants within 4σ of one another grows. Advantages of the truncation algorithm include its lower computational times and relative ease in coding in comparison with the parallel algorithm. Extensions of the model with an 'exposed but not observed' state could not be parallelized as implemented here, as it would not always be known which time interval an infection event occurred. Similarly, having infectivity depend on time since exposure would preclude parallelizing as the infectivity of a plant would depend not only on it's infection status at the beginning of each observation period but also on the exact time of exposure.

The conclusion to be drawn from this paper's comparison of different MCMC implementations is that efficient coding and a truncation approximation enables fairly standard MCMC methods to be used to fit spatial ILMs to moderately large datasets. There are a number of further MCMC techniques, such as adaptive scaling (see Roberts & Rosenthal, 2009), which might perhaps, if implemented carefully, improve the mixing and convergence of chains in problems such as these. Implementing the simple algorithms used in this paper, however, would be feasible for a non-specialist such as a numerate and computer-literate infectious disease epidemiologist.

References

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(3), 269–342.

Becker, N. G. (1989). *Analysis of infectious disease data*, volume 33. Chapman & Hall/CRC.

Deardon, R., Brooks, S., Grenfell, B., Keeling, M., Tildesley, M., Savill, N., Shaw, D., & Woolhouse, M. (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, *20*, 239–261.

Diggle, P. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Statistical methods in medical research*, *15*(4), 325–336.

Gibson, G. (1997). Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(2), 215–233.

Girolami, M. & Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(2), 123–214.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732.

Haber, M., Longini, Ira M., J., & Cotsonis, G. A. (1988). Models for the statistical analysis of infectious disease data. *Biometrics*, *44*(1), pp. 163–173.

Jewell, C., Kypraios, T., Neal, P., & Roberts, G. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian analysis*, *4*(3), 465–496.

Keeling, M. & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton Univ Press.

McKinley, T., Cook, A. R., & Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, *5*(1).

Meyer, S., Elias, J., & Höhle, M. (2011). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*.

Nuessly, G. S. (2005). Featured creatures: yellow sugarcane aphid. http://entnemdept.ufl.edu/creatures/field/bugs/yellow_sugarcane_aphid.htm.

O'Neill, P., Balding, D., Becker, N., Eerola, M., & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *49*(4), 517–542.

Roberts, G. O. & Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, *18*(2), 349–367.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392.

A Likelihood ratios

In this appendix, we compute log acceptance probabilities and likelihood ratios used in our MCMC algorithms.

The likelihood ratio for updating τ_i to τ_i^* is

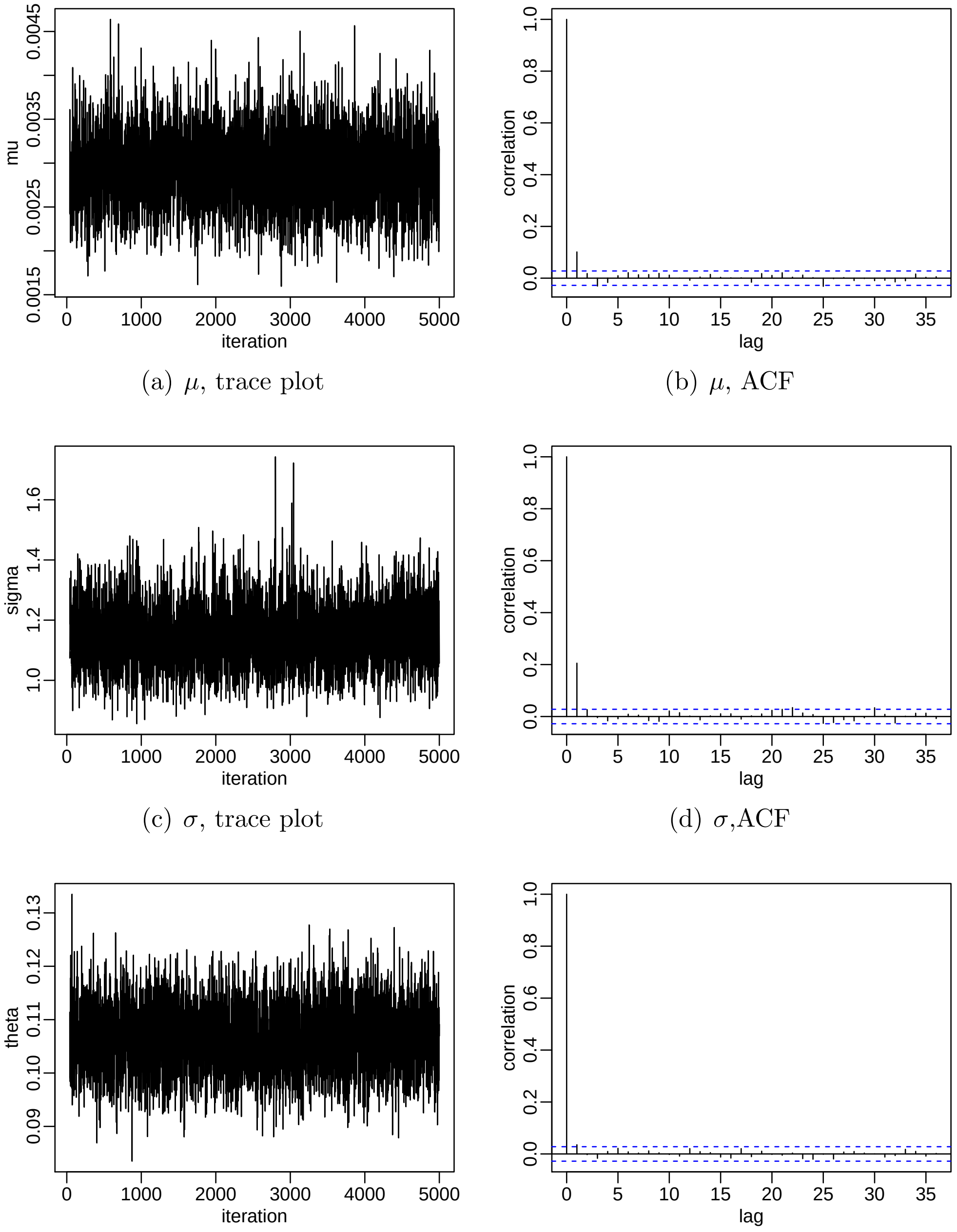
$$\begin{aligned}
&\log L(\mu,\theta,\sigma,\tau_1\ldots\tau_{i-1},\tau_i^*,\tau_{i+1}\ldots\tau_N) - \log L(\mu,\theta,\sigma,\tau_1\ldots\tau_N) = \mu(\tau_i-\tau_i^*)+\\
&\quad\sum_{j;\tau_j<\min(\tau_i,\tau_i^*)}(\tau_i-\tau_i^*)\theta f(x_i-x_j;\sigma)+\sum_{j;\tau_j>\max(\tau_i,\tau_i^*)}(\tau_i^*-\tau_i)\theta f(x_i-x_j;\sigma)+\\
&\quad\sum_{j;\tau_i<\tau_j<\tau_i^*}(2\tau_j-\tau_i-\tau_i^*)\theta f(x_i-x_j;\sigma)+\sum_{j;\tau_i^*<\tau_j<\tau_i}(\tau_i+\tau_i^*-2\tau_j)\theta f(x_i-x_j;\sigma)+\\
&\qquad\sum_{i;\tau_i\le T}\log\left[\mu+\sum_{j;\tau_j<\tau_i}\theta f(x_i-x_j;\sigma)\right]-\log\left[\mu+\sum_{j;\tau_j<\tau_i^*}\theta f(x_i-x_j;\sigma)\right].
\end{aligned}$$

From (3) the log of the acceptance probability for the μ updates is

$$\begin{aligned}
&\log L(\mu^*,\theta,\sigma,\tau)-\log L(\mu,\theta,\sigma,\tau)=\sum_{i;\tau_i\le T}\tau_i(\mu-\mu^*)+||\{i;\tau_i>T\}||T(\mu-\mu^*)-\\
&\qquad\sum_{i;\tau_i\le T}\log\left[\mu+\sum_{j;\tau_j<\tau_i}\theta f(x_i-x_j;\sigma)\right]-\log\left[\mu^*+\sum_{j;\tau_j<\tau_i}\theta f(x_i-x_j;\sigma)\right].
\end{aligned}$$

B MCMC Convergence

In this appendix, we present trace plots and autocorrelation functions (ACFs) for our continuous, untruncated parallel MCMC algorithm. They are produced by chains of 125,000 iterations, discarding the first 1000 iterations as burn-in and retaining only every 25th sample thereafter. They illustrate that MCMC convergence is indeed taking place, as indicated by both the rapid mixing of the trace plots and the rapid decay of the ACF plots.



(a) μ, trace plot

(b) μ, ACF

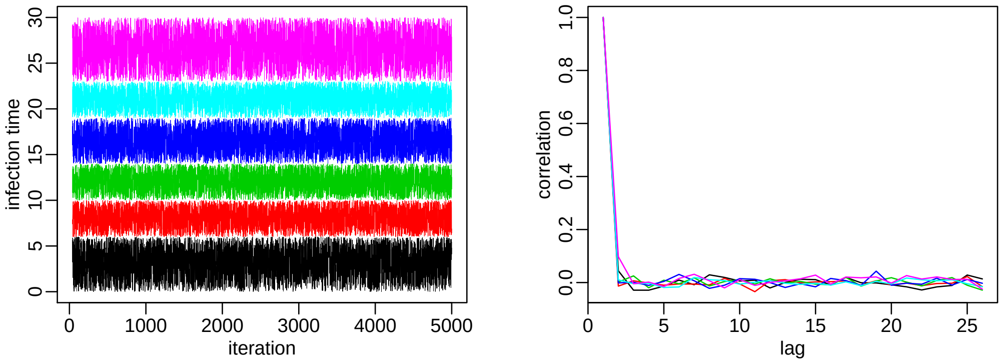
(c) σ, trace plot

(d) σ,ACF

(e) θ, trace plot

(f) θ, ACF

Figure 6: Trace plots and Autocorrelation functions for μ, σ and θ.



(a) Traces (b) ACF

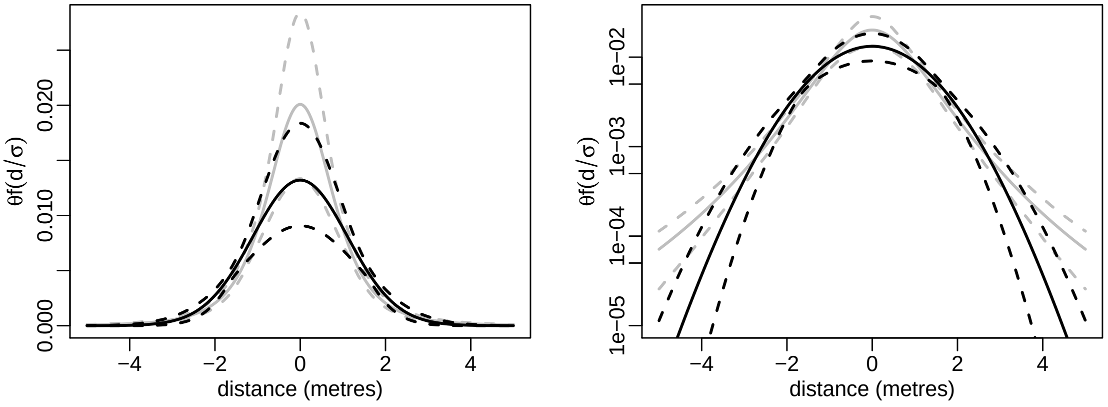
Figure 7: Trace plots and ACF for infection times τ_i for six selected infected plants.

C t density infection kernel

The following are a summary of the results from using a multivariate-t density for the infection kernel. Posterior distributions are shown in Table 2 with Figure 8 showing the posterior mean and quantiles of the infection kernel $\theta f(d;\sigma)$.

	σ	100 μ	θ	df
mean	1.96	0.27	0.11	2.99
2.5%	1.24	0.19	0.10	2.18
50%	1.86	0.27	0.11	2.75
97.5%	3.29	0.35	0.13	4.94

Table 2: Posterior means and quantiles for the model parameters using a multivariate-t density for the infection kernel. The final parameter 'df' is the degrees of freedom for the t-density.

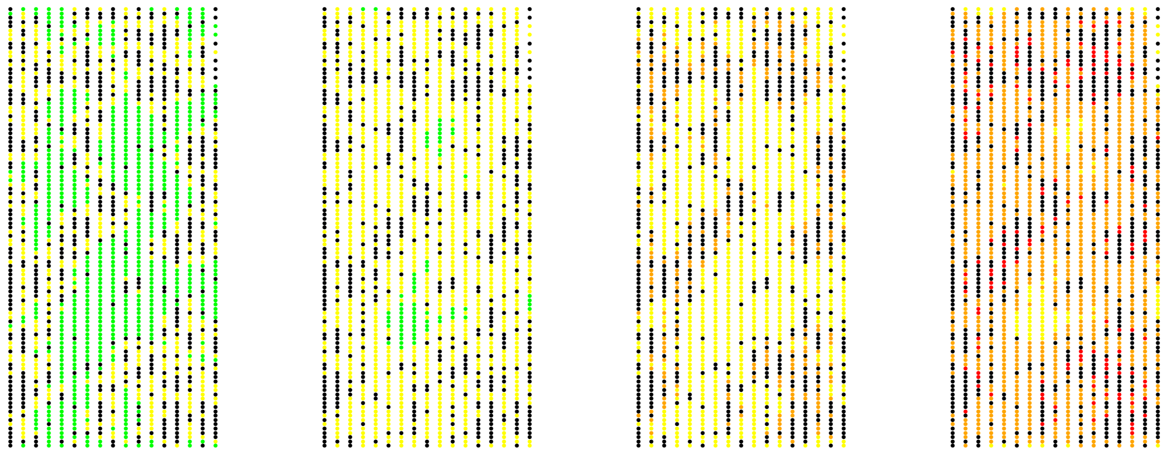


(a) Natural scale

(b) Log scale

Figure 8: Posterior means and 95% posterior credible intervals for the scaled infection kernel $\theta f(d, \sigma)$ using a Gaussian infection kernel (—) and multivariate t-distribution kernel (—).

D Colour figures



(a) 35 weeks

(b) 40 weeks

(c) 50 weeks

(d) 60 weeks

Figure 9: Forecast probabilities of each plant being infected by 35, 40, 50 and 60 weeks. Probabilities are 0-19% (🟢), 20-79% (🟡), 80-95% (🟠),95-<100% (🔴), 100% (⚫)