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

Nathan Welch

Department of Statistics, University of Washington Seattle, WA, 98195, USA

Abstract

This report examines the findings published in *Statistical inference and computational efficiency for spatial infectious disease models with plantation data* **?**. This paper aims to conduct statistical inference for parameters associated with a simple individual level infectious disease model. Model parameters are estimated using the Metropolis sampling algorithm; however, the computation burden created by fitting even a simple model leads to prohibitively long computation time. Statistical and computational methods to overcome the computing challenge are reviewed as a result. (VM's comment: **Mention why this statistical problem is important in the abstract**)

1 Introduction

Disease propagation modeling is an expansive and active area of mathematical research (VM's comment: **add reference**). Statistical inference for parameters underlying these models is less mature. **?** set out to conduct statistical inference for a simple class of disease propagation models by estimating the risk that a susceptible individual contracts a disease from an infected member of the population. In these models, the risk of contracting the disease is modeled at the individual level rather than for the population as a whole. The goal is to formulate models that reflect changes in individual risk that correspond to the number of the infected individuals, their proximity to susceptible members of the population, and the duration a susceptible individuals exposure to those infected. While this model is conceptually convenient, the computational complexity and limitations of algorithms capable of fitting such a model create significant challenges. In (**?**), the authors appeal to standard likelihood methods and the Metropolis algorithm (VM's comment: **add Metropolis et al.**

reference) to estimate **parameters of** a simple individual level model (ILM) for disease propagation. The emphasis on basic components like the Metropolis sampler and simple disease model focuses readers on common challenges inherent with data for statistical inference with **infectious** disease data.

1.1 Literature Review

(VM's comment: **Again, I don't like this one paper per paragraph. You need better organization of your lit review**) **?** introduced ILMs in the context of disease spread among members within households. This early ILM assumes infected individuals are evenly dispersed evenly throughout the population. It also assumes complete data are available to fit these early models.

? surveys a number of models and the types of data sets these models can accommodate. He also summarizes the principle challenges that disease propagation presents to many statistical methods. Finally, this work highlights the distinction between studying disease spread from a mathematical perspective as apposed to a statistical inference point of view.

? considers inference for infection where transmission probabilities between individuals depend on distance; however, Gibson's work focuses on infection status when it is reported at only two time points. He also indicates that earlier work focused on limiting distributions among populations and threshold parameters to sustain or end an epidemic.

? demonstrates MCMC sampling for infectious disease models for two types of data. The first data set only takes into account the total number of infected households at the end of some set time. The second includes more information on within household transmission events. This second case was important as it demonstrated the utility and challenge of fitting a more physically plausible model to a data set with a complicated likelihood function.

? [It took a while to run this reference down. I'll add a review of it in the next draft.]

? investigates spatio-temporal point processes with data of the form $(x_i, t_i) : i = 1, \ldots, n$ over a set region A and time interval $[0, T]$. His approach forgoes the complexity of full likelihood inference and instead appeals to the partial likelihood function to carry out inference for the parameters of interest. This method reduces computation time as the partial-likelihood method effectively marginalizes nuisance parameters. Another key requirement needed to

apply Diggle's findings is that infection times and locations must both be known to fit a model using the procedure described. As a result, infection order in time is an important element of the data sets where one could apply the methods Diggle proposes.

? uses a Taylor series to approximate infection kernel function f to make a Bayesian approach computationally practical. This approach splits the Metropolis update step into two parts: one for *global* parameters updates are unchanged for most of the likelihood function and one for *local* parameter updates. This approximation reduces the computational burden, but introduces some uncertainty about parameter sensitivity to the Taylor center-points. Another restriction is that infection times are known throughout the study period.

? outlines a framework for MCMC with convoluted models. This work demonstrates the potential for fitting complex models using MCMC methods and highlights the reduced computation time brought by multi-core parallelization in particular.

1.2 Statistical Problem

?'s work adopts or adapts some portion of the from the previous section. (VM's comment: **This a very vague sentence**)

Disease propagation modeling is an expansive and active area of mathematical research. Statistical inference for parameters underlying these models is less mature, but computational advances in recent years make it possible to encode convoluted dependencies and draw inference on the parameters influencing disease propagation. Improving our understanding of these parameters will lead to more effective responses or interventions to disease outbreaks.

? set out to conduct statistical inference for a simple class of disease propagation models by estimating the risk that a susceptible individual contracts a disease from an infected member of the population. In these models, the risk of contracting the disease is modeled at the individual level rather than for the population as a whole. The goal is to formulate models that reflect changes in individual risk that correspond to the number of the infected individuals, their proximity to susceptible members of the population, and the duration a susceptible individuals exposure to those infected. While this model is conceptually convenient, the computational complexity and limitations of algorithms capable of fitting such a

model create significant challenges. In **?**, the authors appeal to standard likelihood methods and the Metropolis algorithm to estimate a simple individual level model (ILM) for disease propagation. The emphasis on basic components like the Metropolis sampler and simple disease model focuses readers on common challenges inherent with data for statistical inference with disease data.

1.3 Literature Review

? introduced ILMs in the context of disease spread among members within households. This early ILM assumes infected individuals are dispersed evenly throughout the population. It also assumes complete data are available to fit these early models. **?** surveys a number of models and the types of data sets these models can accommodate. His work highlights the distinction between studying disease spread from a mathematical perspective as apposed to a statistical inference point of view and summarizes the principle challenges that disease propagation presents to many statistical methods.

? builds on the idea that one does not have to choose between parameter inference and mathematical insight. In fact, the authors show that a basic Metropolis-Hastings algorithm clears the way for more realistic modeling assumptions and interdependencies. **?** provides an updated perspective and outlines several MCMC design patterns for convoluted models. **?** primarily contributes another example of these methods along with computational efficiencies beyond an MCMC framework.

While improved computation power made it possible to fit more complex statistical disease models, evaluating likelihood functions that include spatio-temporal components becomes challenging when there are more than a few observations. As the cost to evaluate the likelihood function grows with the number of observations, the utility of MCMC approaches declines. **?** proposes Approximately Bayesian Computation (ABC) that avoids calculating the likelihood by generating an approximation to the likelihood function with each pass through an MCMC implementation. **?** forgoes the complexity of full likelihood inference and instead appeals to the partial likelihood function to carry out inference for the parameters of interest. **?** uses a Taylor series to approximate infection kernel function to make a Bayesian approach computationally practical. These methods led to reduced computation

times compared to a full MCMC implementation, but each either require either completely observed data or are too closely associated with a particular model to be widely applicable.

1.4 Statistical Problem

?'s work adopts or adapts some portion of the from the previous section.

The authors use these contributions to model an insect infestation in a Guadeloupe sugar cane field over 30 weeks time. This data set is useful for showing how inference may proceed even when essential information needed for mathematical modeling and/or statistical inference are unknown or obfuscated.

Figure **??** summarizes the sugar cane data set. The black dots indicate locations of infected plants at the conclusion of the 30 week study period. Grey dots indicate the location of plants that were not infected. The plot on the right shows the cumulative number of infected plants for each inspection.

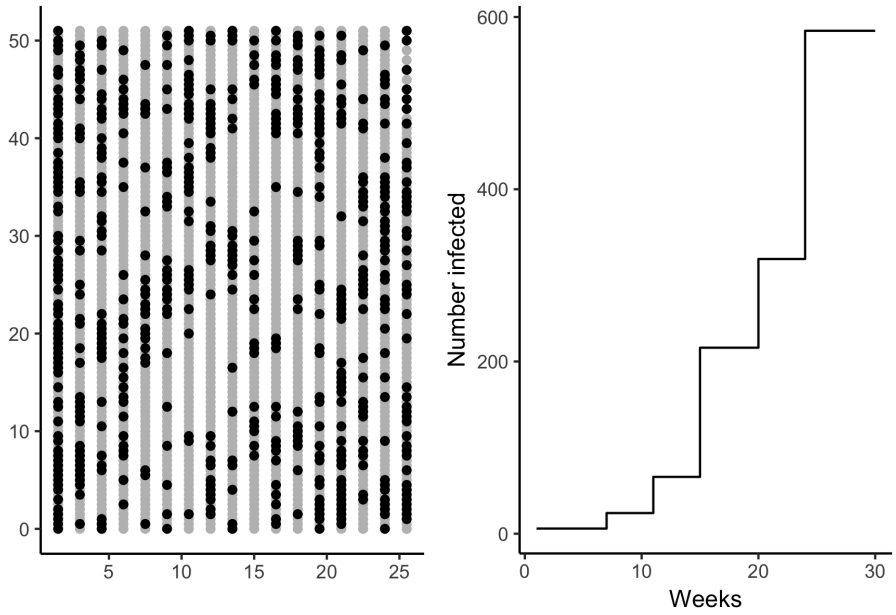

Figure 1: (left) Plant locations (meters) and infection status (black=infected) after 30 weeks; (right) number of infected plants at each inspection time

When modeling disease propagation among individuals who are either susceptible or already infected, the time of infection and duration of susceptible individuals' exposures to

infected members of the population are critical. These data are necessary to infer the rate at which the disease moves from the infected to the susceptible. However, scenarios that include granular infection times are exceptional. Such fine detail is typically reserved for extremely dangerous or damaging diseases such as avian flu or foot and mouth disease on farms in developed countries (**?**). Inference methods for less ominous outbreaks require either additional surveillance, different models, or some other post collection workaround. (VM's comment: **Actually, I don't know of any infectious disease surveillance program that can collect exact infection times.**)

The sugar cane data set includes the infection status of 1,742 plants at six times over a 30 week period. Each observation lists the plant location on a two-dimensional rectangular grid and whether it is infected at week 0, 6, 10, 14, 19, 23, and 30.

The authors selected a susceptible-infected (SI) framework for this data set. (VM's comment: **I don't like "framework for this data set" phrase**) Under this model, once a plant becomes infected, it remains infected for the duration of the study period. While this is a particularly simple model, it is plausible considering the way that such an infestation occurs for a 30 weeks time period. This simple model also keeps emphasis on the principle statistical challenge in this paper: the unknown infection times.

Intervals in which susceptible plants became infected are recorded, but the data do not include exact infection times. As a result, there is no way of knowing how long a plant remained in the susceptible and infected stages. **?**'s approach treats these unknown infection times as latent variables. The data do, however, bound infection times for the plants. If infection times are treated as latent variables, these bounds reduce the parameter space necessary to explore when sampling the true infection times.

With the model framework and likelihood function in hand, **?** take a Bayesian approach to approximate the posterior parameter space of the SI model. Weak priors are placed on the model parameters, and a basic Metropolis algorithm is used to sample the parameter space. This empirical parameter sample is then used to carry out basic inference and prediction. Simulated infestation paths resulting from the sampled parameters help assess the efficacy of the proposed model.

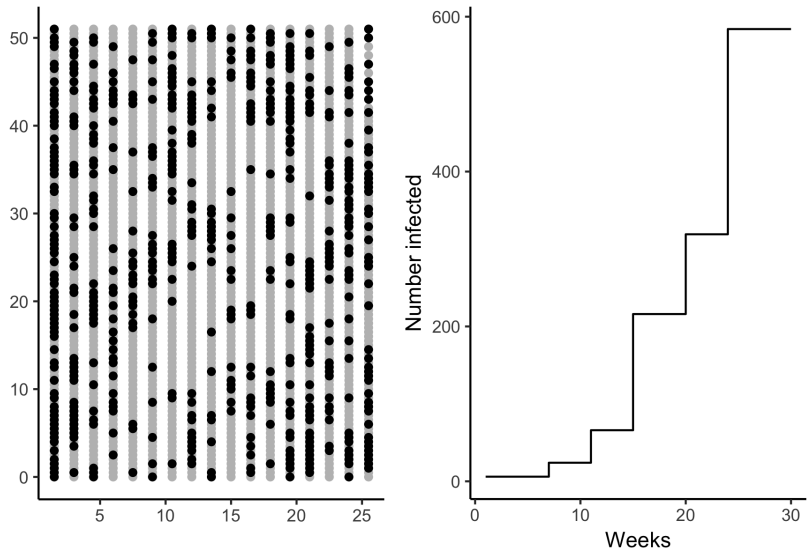

Figure 2: Replicated Figure 1

(VM's comment: **Add axis labels to the left plot: "grid dimension 1" and "grid dimension 2".**) Figure **??** summarizes the sugar cane data set. The black dots indicate locations of infected plants at the conclusion of the 30 week study period. Grey dots indicate the location of plants that were not infected. The plot on the right shows the cumulative number of infected plants for each inspection.

For this data set, it is reasonable to assume that once a plant becomes infected, it remains infected for the duration of the study period. A Susceptible-Infected (SI) model framework is appropriate for this situation and **?** provides a thorough introduction to this and other disease model frameworks. While an SI model is particularly simple, the model is plausible considering the way that an infestation proceeds for an aphid infestation of plants over 30 weeks time. Starting with a simple model also emphasizes the principle statistical components handled by **?**, e.g. unknown infection times.

Intervals in which susceptible plants became infected are recorded, but the data do not include granular infection time data. As a result, there is no way of knowing how long a plant occupied the susceptible and infected stages. This is a common experimental design, and the details of studies that include this *Type I censoring* with *right* and *interval censored* data are discussed at length in survival analysis texts such as **?**. While the data collection framework in **?** is standard, inference methods for *interval censored* modeling parameters

remains more nuanced.

Data augmentation is a common approach to overcoming this inference with unknown infection times. In this case, unknown infection times are modeled as latent variables. Including these latent variables makes the likelihood function tractable, but inflating the parameter space complicates the model fitting process. These factors point to a Bayesian approach to approximating the posterior parameter space of the SI model. **?** discusses an augmentation approach in the context of a *Metropolis-in-Gibbs* MCMC framework and the simulation setup described in the next section follows directly.

2 Methods

2.1 Model and Likelihood

Ǐnference for ILM model parameters begins with the likelihood function. Plausible models account for the locations of infected and susceptible plants along with the respective infection times and durations of exposures. However, time intervals in which a susceptible plant becomes infected offer the only temporal queues. As a result, lets treat infection time, τ_j, for plant j as a latent variable whose domain is restricted to the period in which the infection status of plant x_j changed.

The model must also account for the transmission rate differential induced by the physical distance separating infected and susceptible plants. The transmission rate between an infected plant far from a susceptible plant should be smaller than the rate between two plants next to one another. As a result, let the function $\theta f(x_i - x_j\sigma|\sigma)$ dente the rate of infection for plant i at time t and for a plant j infected prior to t.

Accounting for the rate of spontaneous infection μ, the cumulative intensity of infections at plant i and time t given the infection times for plants $j \in \{1, 2, \ldots, 1742\}$ is

(VM's comment: **Try to rewrite it now with CTMC generating process.**)

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

.

While this function increases with time, the SI construct assumed for these data suggest that once a plant becomes infected, it remains infected. With the rate of infection established, the likelihood of observing infection times $\{\tau_1, \ldots, \tau_N\}$ becomes

- Probability of not observing infections during each plant's time in the susceptible state
- Probability of observing infections at each τ_i

Approximating infection events with a Poisson process and noting that the rate parameter equals the mean, the probability that plant i remains uninfected up to time τ_i is

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

The density function associated with an infection at time τ_i, $\lambda(x_i, \tau_i)$, is the second component of the likelihood. The probability that a plant remains uninfected throughout the period under consideration follows the same Poisson process reasoning used as above. This leads to the following likelihood equation.

$$L_{\theta,\mu,\tau} \propto \left[\prod_{i;\tau_i \leq T} \exp\left\{-\int_0^{\tau_i} \lambda(x_i, t)dt\right\} \lambda(x_i, \tau_i)\right] \left\{\prod_{i;\tau_i > T} \exp\left\{-\int_0^{\tau_i} \lambda(x_i, t)dt\right\}\right\} \quad (1)$$

It remains to specify the infection kernel function, $f(x_i - x_j\sigma|\sigma)$. In this case, a radially symmetric bivariate Gaussian density is appropriate.

$$f(d|\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{||d||^2}{2\sigma^2}\right\} \quad (2)$$

? reasons that the Gaussian kernel is appropriate considering the movement of amphids is Brownian Motion. With that argument, σ^2 in (**??**) denotes the variance of the distance an amphid travels in one week.

After including a term capturing the rate of spontaneous infection μ, the hazard function at plant i and time t given the infection times for plants $j \in \{1, 2, \ldots, 1742\}$ is

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

Now consider the contribution that plant x_i makes to the likelihood if it is infected at time $0 < \tau_i < T = 30$ weeks. By definition of the hazard function, the instantaneous infection

rate at plant x_i is [**double check notation**]

$$\begin{aligned}
\lambda(x_i, t) &= \lim_{dt \to 0} \frac{Pr\{t \leq \tau_i < t + dt | \tau_i \geq t\}}{dt} \\
&= \lim_{dt \to 0} \frac{Pr\{t \leq \tau_i < t + dt\}}{Pr\{\tau_i \geq t\} dt} \\
&= \lim_{dt \to 0} \frac{F_i(t + dt) - F_i(t)}{(1 - F_i(t)) dt} \\
&= \frac{f_i(t)}{S_i(t)} \\
&= -\frac{d}{dt} \log S_i(t) \\
\implies S_i(t) &= \exp\left\{ -\int_0^t \lambda(x_i, s) ds \right\} = \exp\left\{ -\Lambda_i(t) \right\}
\end{aligned}$$

where S_i, f_i, and Λ_i denote the respective survival function, probability distribution function (pdf), and cumulative hazard for plant x_i. From this expression, it follows that

$$f_i(t) = \lambda(x_i, t) \exp\left\{ -\int_0^t \lambda(x_i, t) dt \right\}$$

establishing the pdf that plant x_i contributes to the likelihood function if $t < T = 30$ weeks. Plants that remain healthy throughout the study period contribute $S_i(t)$ to the likelihood. As a result, the likelihood function is

$$L(\theta, \mu, \tau) = \left[\prod_{i; \tau_i \leq T} \exp\left\{ -\int_0^{\tau_i} \lambda(x_i, t) dt \right\} \lambda(x_i, \tau_i) \right] \left[\prod_{i; \tau_i > T} \exp\left\{ -\int_0^{\tau_i} \lambda(x_i, t) dt \right\} \right] \quad (3)$$

The last component of the model to specify is the kernel function, $f(x_i - x_j | \sigma)$. In this case, a radially symmetric bivariate Gaussian density is appropriate.

$$f(x_i - x_j | \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i - x_j)^2}{2\sigma^2} \right\} \quad (4)$$

The Gaussian kernel is physically reasonable since the movement of aphids among plants can be modeled by Brownian Motion. With this argument, σ^2 in (**??**) denotes the variance of the distance an aphid travels in one week.

2.2 Prior Distributions

Weakly informative priors were used for the endemic infection rate (μ), the epidemic infection rate (θ), and the variance for distance disease travels in a week (σ^2). Prior parameters were specified using 95% prior prediction intervals for each parameter.

[I worked out the details for μ, and it's ready for writeup. θ and σ are still in process. Everything below is still the first draft/junk...]

- $\mu \sim$ Gamma(0.7, 0.004) yields a 95% prediction interval between 1 and 630. 583 infections were observed in the data.
- $\theta \sim$ Gamma(0.8, 10) leads to an infected plant requiring 1 to 16 weeks time to produce the first infection when surrounded by susceptible plants
- $\sigma \sim$ Gamma(0.5, 100) implies that nearly all aphids will have traveled less than 10 cm at the 2.5% quantile and around 500 meters at the 95.5% quantile

Placing μ and θ at the lower end of their ranges leads to few infections over 30 weeks time. Setting these too high leads to nearly universal infection. Extreme values of σ leads a to somewhat different result. Specifying σ too small would imply that the aphids driving the infection spread are incapable of moving between plants. Setting σ too large removes the spatial component of the model. **?** specified priors leading to physically plausible outcomes.

2.3 Inference

Exact infection times, τ_i, are unobserved. Only the intervals in which a plant became infected offer clues about the status of plants during the study period. Since each plant's infection status and location influence all other plants at each time step, there is no way to eliminate the dependency on the unknown infection times, τ. **?** follows a data augmentation and MCMC approach to resolve handle the issue. This process is known as a *non-centering parameterization*. **?** section 2.2 outlines a modeling framework nearly identical to the methods followed here.

In this case, priors are specified for the model parameters of interest, specifically θ, σ, and μ. Infection intervals, $\mathbf{Y} = \{Y_i : i = 1, \ldots, N\}$, are the data. The Metropolis algorithm summarized below leads to a posterior sample from $\pi(\mu, \theta, \sigma, \tau \mid \mathbf{Y})$.

1. Initialize the algorithm at τ_i^0, μ^0, σ^0, θ^0 for iteration $r = 0$
2. at iteration r, 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 $\mu^\star \sim N(\mu^{(r-1)}, \nu_\mu)$

4. set $\mu^{(r)} = \mu^\star$ with probability

$$\min\left\{1, \frac{L(\tau_1^{(r)}, \ldots, \tau_N^{(r)} \mid \mu^\star, \theta^{(r)}, \sigma^{(r)})\; p_\mu(\mu^\star)}{L(\tau_1^{(r)}, \ldots, \tau_N^{(r)} \mid \mu^{(r)}, \theta^{(r)}, \sigma^{(r)})\; p_\mu(\mu^r)}\right\}$$

   otherwise, set $\mu^{(r)} = \mu^{(r-1)}$

5. repeat steps 3 and 4 for θ and σ

6. for each $i = 1, \ldots, N$, propose a new $\tau_i^\star$ and accept with probability

$$\min\left\{1, \frac{L(\tau_1^{(r)}, \ldots, \tau_{i-1}^{(r)}, \tau_i^\star, \tau_{i+1}^{(r)}, \ldots, \tau_N^{(r)} \mid \mu^{(r)}, \theta^{(r)}, \sigma^{(r)})\; p_\mu(\mu^\star)}{L(\tau_1^{(r)}, \ldots, \tau_N^{(r)} \mid \mu^{(r)}, \theta^{(r)}, \sigma^{(r)})\; p_\mu(\mu^r)}\right\}$$

7. return to step 2 until a sufficiently large sample has been obtained

The implementation for this model uses normally distributed proposals with means at the previous θ, σ, μ, and τ_i values and $\nu_\theta = 0.005^2$, $\nu_\mu = 0.0005^2$, $\nu_\sigma = 0.05^2$, and $\nu_{\tau_i^{(r)}} = 1$. The authors select these standard deviations after assessing the mixing properties of test runs.

2.4 Implementation

The data include 583 unknown infection times. As a result, some portion of the likelihood function has to be computed 586 times for each pass through the sampler. Implementing the algorithm naively leads to prohibitively long computation times. The following sections review the steps taken to improve and benchmark computational performance enough to make the non-centering parameterization method viable for data with challenging interdependencies. The methods were implemented in the Julia.

2.4.1 Basic Algorithm

The basic implementation primarily executes the pseudo-code from the previous section with three exceptions. First, the distance matrix among the 1,742 plants is pre-computed and referenced throughout the implementation. The other exception concerns the $\tau^{(r)}$ update

routine. Many of the terms in the likelihood ratio are identical, eliminating the need to compute several terms in the likelihood function. Finally, eliminating impossible infection time sequences among the remaining pieces of the computation led to efficiencies in the compute time.

2.4.2 Parallel Algorithm

The parallel implementation distributes the computational burden two different ways. Both focus on computing updating τ. First, the subroutines of the likelihood ratio that cannot run asynchronously but still require significant computational resources are broken into pieces and computed on multiple cores. The other parallelization improvement results from recognizing that some updates are not possible given order that infections were known to have occurred. As a result, proposals that would reverse this known ordering are rejected immediately. More to come...

2.4.3 Improved Parallel Algorithm

Tbd...

2.4.4 Truncated Algorithm

Tbd...

2.4.5 Discrete Time Algorithms

Tbd...

3 Results

4 Discussion

References

Becker, N. (1989). Analysis of Infectious Diseas Data. Chapman and Hall-CRC, Boca Raton.

Brown, P., Chimard, F., Remorov, A., Rosenthal, J., and Wang, X. (2014). Statistical inference and computational efficiency for spatial infectious disease models with plantation data. Applied Statistics, 63(3):467–482.

Deardon, R., Brooks, S., Grenfell, B., Keeling, M., Tildesley, M., Savill, N., Shaw, D., and Woolhouse, M. (2006). Inference for individual-level models of infectious diseases in large populations. Staistical Sinica, 15:325–336.

Diggle, P. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. Statistical Methods in Medical Research, (15):325–336.
¡¡¡¡¡¡¡ HEAD

Gibson, G. J. (1997). Markov Chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. Applied Statistics, (46):215–233.
======= ¿¿¿¿¿¿¿ dfccd338026df8b27485897d67d1d004c05b1edf

Haber, M., Longini, J. I. M., and Cotsonis, G. (1988). Models for the statistical analysis of infectious disease data. Biometrics, (44):163–173.

Jewell, C., Kypraios, T., Neal, P., and Roberts, G. (2009). Bayesian analysis for emerging infectious diseases. Bayesian Analysis, (4):465–496.

Klein, J. P. and Moeschberger, M. L. (1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer.

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

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M., and Mollison, D. (2000). Analysis of infectious disease data from household outbreaks by Markov Chain Monte Carlo methods. Applied Statistics, (49):517–542.