PLOS ONE

RESEARCH ARTICLE

Parameterizing Spatial Models of Infectious Disease Transmission that Incorporate Infection Time Uncertainty Using Sampling-Based Likelihood Approximations

Rajat Malik[1]*, **Rob Deardon**[1,2,3]*, **Grace P. S. Kwong**[3]*

1 Department of Mathematics & Statistics, University of Guelph, Guelph, Ontario, Canada, **2** Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, Canada, **3** Department of Mathematics & Statistics, University of Calgary, Calgary, Alberta, Canada

* rmalik@uoguelph.ca (RM); robert.deardon@ucalgary.ca (RD); grace.kwong@ucalgary.ca (GPSK)

Citation: Malik R, Deardon R, Kwong GPS (2016) Parameterizing Spatial Models of Infectious Disease Transmission that Incorporate Infection Time Uncertainty Using Sampling-Based Likelihood Approximations. PLoS ONE 11(1): e0146253. doi:10.1371/journal.pone.0146253

Editor: Gui-Quan Sun, Shanxi University, CHINA

Received: July 9, 2015

Accepted: December 15, 2015

Published: January 5, 2016

Data Availability Statement: The data we utilized, called the 2001 UK foot and mouth disease farm-level data set, was obtained from the United Kingdom Government Department of Environment, Food and Rural Affairs (DEFRA), and were given permission by them to use the data for our paper. We do not, however, have permission to redistribute the data. If any readers wish to obtain the data, they may contact Dr. Deardon or DEFRA directly at defra.helpline@defra.gsi.gov.uk.

Funding: This research was funded by the Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) and by the Natural Sciences and

Abstract

A class of discrete-time models of infectious disease spread, referred to as individual-level models (ILMs), are typically fitted in a Bayesian Markov chain Monte Carlo (MCMC) framework. These models quantify probabilistic outcomes regarding the risk of infection of susceptible individuals due to various susceptibility and transmissibility factors, including their spatial distance from infectious individuals. The infectious pressure from infected individuals exerted on susceptible individuals is intrinsic to these ILMs. Unfortunately, quantifying this infectious pressure for data sets containing many individuals can be computationally burdensome, leading to a time-consuming likelihood calculation and, thus, computationally prohibitive MCMC-based analysis. This problem worsens when using data augmentation to allow for uncertainty in infection times. In this paper, we develop sampling methods that can be used to calculate a fast, approximate likelihood when fitting such disease models. A simple random sampling approach is initially considered followed by various spatially-stratified schemes. We test and compare the performance of our methods with both simulated data and data from the 2001 foot-and-mouth disease (FMD) epidemic in the U.K. Our results indicate that substantial computation savings can be obtained—albeit, of course, with some information loss—suggesting that such techniques may be of use in the analysis of very large epidemic data sets.

Introduction

Modeling the spread of infectious diseases is a research area of great importance to public health and agriculture. Particularly, studies involving data-driven spatial models have recently been used in a number of applications. Work by [1], for example, illustrate the importance of incorporating spatial and temporal data in the mathematical modeling of infectious diseases. Studies have also investigated factors that influence disease persistence/extinction such as

Engineering Research Council of Canada (NSERC). Computer equipment was provided by the Canada Foundation for Innovation via the Leading Edge Fund grant, "Centre for Public Health and Zoonoses (CPHAZ)."

Competing Interests: The authors have declared that no competing interests exist.

infection rate (e.g., [2]). In addition, disease control policies and vaccination policies can be better developed as a result of the understanding of the spread of infectious disease gained from using such models [3]. The increase in computational power over the last several years and the availability of spatio-temporal data have been key factors driving growth in this area of statistics [4].

Significant computational prowess is required for models utilizing large-scale spatial data, such as recent studies performed on the 2001 foot-and-mouth disease (FMD) epidemic in the U.K. (e.g., [5–8]). These can all be considered examples of modeling infectious disease dynamics at the individual-level, although the individuals of concern may vary between studies (e.g., plants, humans, farms, etc.). [8] describe a framework of discrete-time individual-level models (ILMs) that are capable of modeling the spread of infectious diseases in such disease systems. These models can incorporate various heterogeneities within a population; for example, [9] consider the spread of human influenza allowing for vaccination status, age, and the disease status of fellow household occupants in the model. Although ILMs are flexible and intuitive, inference for these and other similar models, especially when dealing with large data sets, can be computationally prohibitive. In fact, even for moderately sized populations, obtaining meaningful results can require running these models for a considerable amount of time.

For such models, parameter estimation is typically carried out in a Bayesian framework via Markov chain Monte Carlo (MCMC) methods (e.g., [10]), wherein the likelihood function (a primary source of the computational problem) is calculated numerous times. An obvious way to reduce the extent of this problem would be to make a simplifying assumption, such as homogeneous mixing (e.g., [11]). However, by allowing for heterogeneity within the population, we hope to draw more sound inferences.

Numerous studies have focused on speeding up the likelihood calculation to reduce the time required to carry out parameter estimation in such models. For example, [8] introduce an approach using a Taylor series expansion of the non-linear spatial infection kernel, allowing for the decomposition of a substantial part of the likelihood function into a small parameter-dependent part and a larger data-dependent part. [12] expand on this by exploring the use of a piecewise linear kernel to carry out the linearization. In both cases, time-saving ensues from the fact that the data-intensive component of the likelihood does not require re-calculation at each step of the MCMC algorithm. However, the resulting model is an approximation of the true model we might actually want to fit. These approaches are also limited to situations where infection event histories of individuals are assumed known.

[13] explore several variations of a random-walk Metropolis algorithm to achieve computational efficiency in the context of infectious disease modeling. Their approach involves pre-calculating and storing quantities that are used repeatedly (something vital to the approaches of [8] and [12]), performing calculations in parallel, and refining the calculation of the likelihood ratio in the Metropolis-Hastings MCMC algorithm.

Other approaches to decreasing computation time in the context of fitting infectious disease models to data—in these cases, homogeneous-mixing models—are based around so-called approximate Bayesian computations, as explored in [14] and [15], and pseudo-marginal approaches, as discussed by [16]. In such approaches, explicit likelihood calculation is completely avoided. An alternative approach is given by [17], not within the context of inference for infectious disease transmission models, but for mixture models. They explore methods basing inference on carefully selected subsamples of data, constructed to provide the most relevant information to the parameters of interest.

In this paper, we consider an approach similar in nature to that of [17] that replaces the likelihood calculation with a faster likelihood approximation based upon data sampling. We also

include infection time uncertainty in our modeling via a data augmented Bayesian analysis. The method works by selecting samples from the infected set of individuals at every discrete time point when calculating the infection rate for susceptible individuals, thereby avoiding the need to use the entire infectious set. We show how the resulting approximated likelihood function-based analysis can require significantly less time to carry out and compare the approximate posterior inference to the full Bayesian analysis. Whereas the aforementioned approximate methods of [8] and [12] for parameterizing ILMs cannot be used in a data augmented framework in which infection times are considered unknown, we show how our approach, with careful algorithm development, can allow such uncertainty in the analysis. This is of obvious importance in infectious disease systems because infection event times are very rarely observed with any certainty in practice [18]. We begin by considering a simple random sampling (SRS) approach, followed by a series of spatially-stratified sampling schemes. As a proof of concept, we test our methodology through the use of both simulated data and a relatively small subset of the 2001 UK FMD epidemic data. Note that we consider infectious disease models in a susceptible-infectious-removed (SIR) framework, although extension of the methods to other frameworks (e.g., $\mathcal{SEIR}$) would be relatively straight forward.

Our paper is structured as follows. The *Methodology* section summarizes the general ILM framework of [8], the specific models used in this paper, and the MCMC algorithm used to carry out a full Bayesian MCMC analysis. Our algorithms are presented in the *Sampling-Based Likelihood Approximations* section. The *Epidemic Data* section describes the data used to test our methods and the *Results* section presents our findings. The *Discussion* section concludes this paper and presents possible avenues of future work.

Methodology

General Model

We utilize the modeling framework of [8], which defines a class of flexible discrete-time disease transmission models that include covariate information at the individual level. With a finite population of a total of n individuals (each individual represented as $i = 1, \ldots, n$), we observe epidemic data at discrete time points, $t = 1, \ldots, t_{\max}$, where $t_{\max}$ is the last observation time. Under a susceptible-infectious-removed ($\mathcal{SIR}$) framework, each individual i is in only one of these three states at any given time t. If $i \in \mathcal{S}_t$, then i is susceptible to the disease and has not yet contracted it at time t; if $i \in \mathcal{I}_t$, then i has contracted the disease and can now infect others at time t; and if $i \in \mathcal{R}$, then i has been removed from the population at time t; e.g., due to recovery combined with acquiring immunity. Once an individual is in this final state, they cannot become infected again or transmit the disease to others. Over the course of the epidemic, individuals move through the three states in the order $\mathcal{S} \rightarrow \mathcal{I} \rightarrow \mathcal{R}$.

As described by [8], the general ILM calculates the probability a susceptible individual i will become infectious to the disease at time t, and this is given by

$$P_{it}(\boldsymbol{\theta}) = 1 - \exp\left[\left\{-\Omega_S(i)\sum_{j\in\mathcal{I}_t}\Omega_T(j)\kappa(i,j)\right\} - \epsilon(i,t)\right], \quad (1)$$

where $\Omega_S(i)$ is a susceptibility function that includes risk factors for individual i contracting the disease; $\Omega_T(j)$ is a transmissibility function describing risk factors for individual j transmitting the disease to others; $\kappa(i, j)$ is an infection kernel describing shared risk factors between susceptible and infectious individuals; $\epsilon(i, t)$ describes external infectious pressure not explained by the rest of the model and is commonly referred to as the 'sparks term'; and $\boldsymbol{\theta}$ is the set of ILM parameters we want to estimate.

Spatial ILM

In this section, we present a simplified version of the general ILM such that $\Omega_S(i) = \alpha$, $\Omega_T(j) = 1$, and $\kappa(i,j) = d_{ij}^{-\beta}$. Here, d_{ij} represents the Euclidean distance between susceptible i and infectious j and β represents the power law rate of decay. We also set the sparks term, $\epsilon(i, t) = 0$. We refer to this model as the Spatial ILM. Under this model, the probability of infection for susceptible i at time t is given by

$$P_{it}(\boldsymbol{\theta}) = 1 - \exp\left[-\alpha \sum_{j \in \mathcal{I}_t} d_{ij}^{-\beta}\right], \qquad \alpha > 0, \beta > 0. \tag{2}$$

FMD-ILM

We also modify the general ILM in order to model data from the 2001 U.K. FMD epidemic. Using a simplified version of the model found in [8] and by modeling at the farm-level, we can determine the probability that susceptible farm i is infected at time t using

$$P_{it}(\theta) = 1 - \exp\left[\left(-\left(\alpha_s N_i^s + \alpha_c N_i^c\right) \sum_{j \in \mathcal{I}_t} \left(\phi_s N_j^s + \phi_c N_j^c\right) d_{ij}^{-\beta}\right) - \epsilon\right], \qquad \alpha_c > 0, \phi_s > 0, \phi_c > 0, \beta > 0, \epsilon > 0, \tag{3}$$

where α_s and α_c are susceptibility parameters and ϕ_s and ϕ_c are transmissibility parameters, for sheep and cattle, respectively. The terms N_x^s and N_x^c represent the number of sheep and cattle on farm x, respectively. To avoid identifiability issues, we set $\alpha_s = 1 \times 10^{-7}$, which is an arbitrary constant reference level and is not estimated. Once again, the power-law kernel, $\kappa(i,j) = d_{ij}^{-\beta}$, is used with d_{ij} being the Euclidean distance between farms. The sparks terms is set as a constant such that $\epsilon(i, t) = \epsilon$, which represents a constant infectious pressure from outside the study area. We refer to Model 3 as our FMD-ILM.

Bayesian Computation

Our parameter estimation is here carried out under a Bayesian framework. Assuming known infection and removal times, the likelihood function for ILMs is the product of all infection and non-infection events over the entire observed epidemic period ($t = 1, \ldots, t_{\max}$), and is given by

$$\pi(\mathbf{x}|\boldsymbol{\theta}) = \prod_{t=1}^{t_{\max}} \left[\prod_{i \in \mathcal{S}_{t+1}} (1 - P_{it}(\boldsymbol{\theta})) \prod_{i \in \mathcal{I}_{t+1} \setminus \mathcal{I}_t} P_{it}(\boldsymbol{\theta})\right], \tag{4}$$

where $\mathbf{x}$ is the observed epidemic data set (including the infection times), $\mathcal{S}_{t+1}$ is the set of all susceptible individuals at time $t + 1$, and $\mathcal{I}_{t+1} \setminus \mathcal{I}_t$ is the set of newly infectious individuals at time $t + 1$. Using our likelihood function and by placing a prior, $\pi(\boldsymbol{\theta})$, on our parameter set, $\boldsymbol{\theta}$, we can obtain the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$, up to a constant of proportionality. To explore the posterior distribution, we can use the random-walk Metropolis Hastings (RWMH) algorithm [19–21].

We assume here a disease system such as foot-and-mouth as seen in the U.K. in 2001, in which the disease is reported after infection and then individuals are later removed from the population through some intervention (see Fig 1). Thus, we assume that removal times are known and fixed (although this assumption could quite easily be relaxed—see Discussion). However, we do not assume to know when individuals become infectious and so utilize

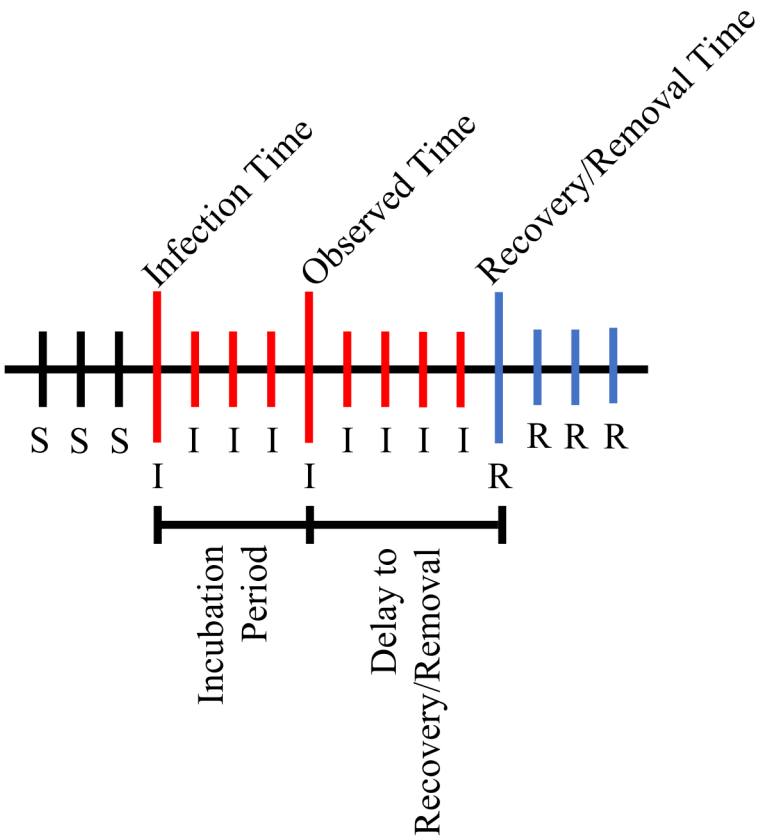

Fig 1. Average infectious period under the simulation study. Illustration of the average infectious period under the simulation study. The average incubation period is 3 days, and the average delay to disease recovery and removal from the population is 4 days. The 'S' symbol indicates the individual is susceptible to the disease at that time point and the 'R' symbol indicates the individual has recovered from the disease and has been removed from the population at that time point.

doi:10.1371/journal.pone.0146253.g001

Bayesian data augmentation, treating the infection times as unknown nuisance parameters (see also sections *MCMC Algorithm* and *Data Augmentation*). We determine the infection time indirectly by inferring the incubation period (the time between infection and disease reporting/diagnosis/observation) of each individual. Given the aforementioned assumption that removal times are known, the incubation period thus defines the infection time for each individual. The incubation period, plus a further delay until removal (e.g., through quarantine or animal culling), thus define the infectious period.

We denote the unknown incubation periods $\boldsymbol{\mathcal{Z}} = (\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_v)$, where v is the number of infected individuals in $[1, t_{\max}]$, and assume $\mathcal{Z}_c \overset{\text{i.i.d}}{\sim} \text{DExp}(\lambda_z)$ (discretized exponential distribution), where the rate parameter, λ_z, is also to be estimated. We augment the model parameter set to include $\boldsymbol{\mathcal{Z}}$. Under the spatial ILM, our parameter vector is thus $\boldsymbol{\theta}^+ = \{\alpha, \beta, \lambda_z, \boldsymbol{\mathcal{Z}}\}$; and under the FMD-ILM, the parameter set is $\boldsymbol{\theta}^+ = \{\alpha_c, \phi_s, \phi_c, \beta, \epsilon, \lambda_z, \boldsymbol{\mathcal{Z}}\}$.

In general, we define an augmented parameter set, $\boldsymbol{\theta}^+ = (\boldsymbol{\theta}, \boldsymbol{\mathcal{Z}})$, and assuming independence between $\boldsymbol{\theta}$ and $\boldsymbol{\mathcal{Z}}$ derive the posterior distribution up to proportionality as

$$\begin{aligned} \pi(\boldsymbol{\theta}^+|\mathbf{x}^-) &\propto \pi(\mathbf{x}^-|\boldsymbol{\theta}^+)\pi(\boldsymbol{\theta}^+) \\ &= \pi(\mathbf{x}^-|\boldsymbol{\theta}, \boldsymbol{\mathcal{Z}})\pi(\boldsymbol{\theta})\pi(\boldsymbol{\mathcal{Z}}) \\ &= \pi(\mathbf{x}^-, \boldsymbol{\mathcal{Z}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \end{aligned} \tag{5}$$

where $\mathbf{x}^-$ is the epidemic data not including infection times and $\pi(\boldsymbol{\theta}^+|\mathbf{x}^-)$ is sampled from using Metropolis-Hastings MCMC. Here, we use an independence sampler to update λ_z and $\boldsymbol{\mathcal{Z}}$ and random-walk updates for other model parameters. Note that we are only indirectly estimating the infectious period distribution and not using any prior information on the infectious period. This framework could easily be changed, of course, to fit the requirements of other disease systems.

MCMC Algorithm

Here, we outline our MCMC algorithm to update the data augmented parameter set, $\boldsymbol{\theta}^+$, and obtain realizations from the posterior distribution, $\pi(\boldsymbol{\theta}^+|\mathbf{x}^-)$, in order to carry out a full gold-standard Bayesian analysis. For our MCMC procedure, we break down our augmented parameter set and let Θ contain the same set of parameters as $\boldsymbol{\theta}$ but without rate parameter λ_z; i.e., $\Theta = \{\theta_1, \ldots, \theta_b\}$, where $b = |\Theta|$ (the number of parameters in Θ). Hence, for our MCMC algorithm below, we specify our augmented parameter set as $\Phi = (\Theta, \lambda_z, \boldsymbol{\mathcal{Z}})$. There are a total of $v = |\boldsymbol{\mathcal{Z}}|$ parameters in $\boldsymbol{\mathcal{Z}}$, and a total of $d = b + v + 1$ parameters in Φ. We define θ_w as the w^{th} parameter in Θ and $\mathcal{Z}_c$ as the c^{th} parameter in $\boldsymbol{\mathcal{Z}}$. Let r be a counter for the number of MCMC iterations and let λ_z^r represent the r^{th} iteration of λ_z. The MCMC algorithm is as follows:

1. Let $r = r + 1$.
2. Update $\boldsymbol{\mathcal{Z}}$:
   a. Let $c = 1$.
   b. Given the current position, $\mathcal{Z}_c^r$, generate a new value, $\mathcal{Z}_c^{r+1} \sim \text{DExp}(\lambda_z^r)$, using the inverse transform method.
   c. Calculate the acceptance probability,
   $$\mathcal{A} = \min\left(1, \frac{\pi(\Theta^r, \lambda_z^r, \mathcal{Z}_1^{r+1}, \mathcal{Z}_2^{r+1}, \cdots, \mathcal{Z}_c^{r+1}, \mathcal{Z}_{c+1}^r, \mathcal{Z}_{c+2}^r, \cdots, \mathcal{Z}_v^r|\mathbf{x})}{\pi(\Theta^r, \lambda_z^r, \mathcal{Z}_1^{r+1}, \mathcal{Z}_2^{r+1}, \cdots, \mathcal{Z}_c^r, \mathcal{Z}_{c+1}^r, \mathcal{Z}_{c+2}^r, \cdots, \mathcal{Z}_v^r|\mathbf{x})}\right).$$
   d. With probability $\mathcal{A}$, accept $\mathcal{Z}_c^{r+1}$. Otherwise, set $\mathcal{Z}_c^{r+1} = \mathcal{Z}_c^r$.
   e. Let $c = c + 1$.
   f. If $c \leq v$, then repeat from step 2b.
   g. If $c > v$, then continue to step 3.
3. Update Θ:
   a. Let $w = 1$.
   b. Given the current position, θ_w^r, generate a new value such that $\theta_w^{r+1} = \theta_w^r + s$, where $s \sim U[-g_w, g_w]$, $g_w \in \mathbb{R}^+$, and $\theta_w^{r+1} > 0$.
   c. Calculate the acceptance probability,
   $$\mathcal{A} = \min\left(1, \frac{\pi(\theta_1^{r+1}, \theta_2^{r+1}, \cdots, \theta_w^{r+1}, \theta_{w+1}^r, \theta_{w+2}^r, \cdots, \theta_b^r, \lambda_z^r, \boldsymbol{\mathcal{Z}}^{r+1}|\mathbf{x})}{\pi(\theta_1^{r+1}, \theta_2^{r+1}, \cdots, \theta_w^r, \theta_{w+1}^r, \theta_{w+2}^r, \cdots, \theta_b^r, \lambda_z^r, \boldsymbol{\mathcal{Z}}^{r+1}|\mathbf{x})}\right).$$
   d. With probability $\mathcal{A}$, accept θ_w^{r+1}. Otherwise, set $\theta_w^{r+1} = \theta_w^r$.
   e. Let $w = w+1$.

f. If $w \leq b$, then repeat from step 3b.

g. If $w > b$, then continue to step 4.

4. Update λ_z via the independence sampler:

   a. Given the current position, λ_z^r, generate a new value such that $\lambda_z^{r+1} \sim \Gamma(f_1, f_2)$, where f_1 is a shape parameter and f_2 is a rate parameter.

   b. Calculate the acceptance probability,

$$\mathcal{A} = \min\left(1, \frac{\pi(\Theta^{r+1}, \lambda_z^{r+1}, \boldsymbol{\mathcal{Z}}^{r+1}|\mathbf{x})}{\pi(\Theta^{r+1}, \lambda_z^{r}, \boldsymbol{\mathcal{Z}}^{r+1}|\mathbf{x})}\right).$$

   c. With probability $\mathcal{A}$, accept λ_z^{r+1}. Otherwise, set $\lambda_z^{r+1} = \lambda_z^r$.

5. Repeat from step 1 until a sufficiently large sample of realizations has been obtained.

Sampling-Based Likelihood Approximations

As stated previously, the full likelihood calculation for ILMs can be computationally taxing. Our focus here is on the key problem of calculating the infectious pressure,

$$\mathcal{X}_{it} = \sum_{j \in \mathcal{I}_t} \Omega_T(j)\kappa(i,j),$$

for each individual $i \in \mathcal{S}_{t+1}$ and $i \in \mathcal{I}_{t+1} \backslash \mathcal{I}_t$ at each time point for which data are observed. The problem worsens when we attempt to incorporate infection time (or incubation period) parameters via data augmentation. In doing so, we increase the number of parameters and, thus, the number of parameter updates in each MCMC iteration.

We propose to alleviate this problem by estimating $\mathcal{X}_{it}$ by sampling from the infectious set $\mathcal{I}_t$ at each time point that data are observed. We begin this section by outlining the need to organize all infectious individuals into a matrix that can be updated in an efficient manner as the incubation period (and, thus, infection time) parameters are updated as part of the data-augmented MCMC. We then detail our sampling algorithms in such a data-augmented context. The two algorithms considered here allow for an SRS approach and a spatially-stratified sampling scheme, respectively, for sampling from the $\mathcal{I}_t$ sets.

Simple Random Sampling Algorithm

For the SRS method, calculating each $P_{it}(\boldsymbol{\theta})$ (or $1 - P_{it}(\boldsymbol{\theta})$) in the likelihood is achieved by replacing the full set of infectious individuals $\mathcal{I}_t$ with a set $\hat{\mathcal{I}}_t$ obtained through SRS with replacement from the set $\mathcal{I}_t$, and scaling by the empirical sampling proportion, $\hat{\rho}_t$. This method is shown to severely reduce the computational time required to calculate P_{it} in the likelihood function. When $\mathcal{I}_t$ is 'small' (i.e., $|\mathcal{I}_t| \leq q|$), we do not sample and use the entire infectious set. Here, we set $q = 10$ because the time savings would be negligible for $q \leq 10$ simply due to the overhead associated with sampling. The infectious pressure is approximated as

$$\mathcal{X}_{it} \simeq \hat{\mathcal{X}}_{it} = \begin{cases} \sum_{j \in \mathcal{I}_t} \Omega_T(j)\kappa(i,j) & |\mathcal{I}_t| \leq q \\ \hat{\rho}_t^{-1} \sum_{j \in \hat{\mathcal{I}}_t} \Omega_T(j)\kappa(i,j) & |\mathcal{I}_t| > q \end{cases}$$

and thus the approximation of our original probability of infection is

$$P_{it}(\boldsymbol{\theta}) \simeq \hat{P}_{it}(\boldsymbol{\theta}) = 1 - \exp\left[\left\{-\Omega_S(i)\hat{\mathcal{X}}_{it}\right\} - \epsilon(i,t)\right], \tag{6}$$

which we refer to as the SRS-ILM. Initially, we will assume that all infection times (as well as removal times) are known.

We now define notation relevant to our infection matrix, $\mathcal{M}$, of dimension $n \times t_{\max}$. The elements of $\mathcal{M}$ take the form of integer identification numbers for each farm and thus, each column, $\mathcal{M}[\cdot, t_1, \ldots, t_{\max}]$ consists of an arbitrary ordering of farm IDs indicating their infection times, followed by a series of zeros in the remaining elements of the column. We also store the length of each column of the matrix up until the presence of empty cells, defined as $\ell_t = |\mathcal{M}[\cdot, t]|$. We use the notation $\mathcal{M}[B, C]$ to represent the farm ID located in row B and time column C in matrix $\mathcal{M}$. We also use the notation $\mathcal{DU}[a, b]$ to refer to a discrete uniform on $[a, b]$, $a, b \in \mathbb{Z}$, i.e., a distribution consisting of equally sized point masses on all integers within the interval $[a, b]$. To calculate the likelihood function, we then use the following algorithm:

1. Let $\hat{\mathcal{L}}_t = 0 \ \forall \ t = 1, \ldots, t_{\max}$ and set $t = 1$.
2. If $\ell_t \leq q$, calculate the full likelihood component for time t,

$$\hat{\mathcal{L}}_t = \prod_{i \in \mathcal{S}_{t+1}} (1 - P_{it}(\boldsymbol{\theta})) \prod_{i \in \mathcal{I}_{t+1} \setminus \mathcal{I}_t} P_{it}(\boldsymbol{\theta}),$$

   and go to step 6.
   Else, if $\ell_t > q$, let $\xi = \rho_t \ell_t$ and continue to step 2.
3. Let $c = 0$ and $\hat{v}_t$ be a set of "empty" vectors of length to be determined by the algorithm.
4. Let $c = c + 1$.
5. If $c \leq \xi$, then simulate $U \sim \mathcal{DU}[1, \ell_t]$, let $\hat{v}_t[c] = \mathcal{M}[t, U]$, and return to step 3.
   If $c > \xi$, then let $\hat{\mathcal{I}}_t$ be the set containing all $c - 1$ elements of $\hat{v}_t$ and continue to step 5.
6. Calculate the approximated likelihood component for time t,

$$\begin{aligned} \hat{\mathcal{L}}_t \quad = \quad & \prod_{i \in \mathcal{S}_{t+1}} \exp\left[\left\{-\Omega_S(i)\hat{\mathcal{X}}_{it}\right\} - \epsilon(i,t)\right] \times \\ & \prod_{i \in \hat{\mathcal{I}}_{t+1} \setminus \hat{\mathcal{I}}_t} \left[1 - \exp\left\{\left(-\Omega_S(i)\hat{\mathcal{X}}_{it}\right) - \epsilon(i,t)\right\}\right], \end{aligned}$$

   where $\hat{\mathcal{X}}_{it} = \hat{\rho}_t^{-1} \sum_{j \in \hat{\mathcal{I}}_t} \Omega_T(j)\kappa(i,j)$, as before, and $\hat{\rho}_t = \frac{c-1}{\ell_t}$.
7. Let $t = t+1$.
   If $t < t_{\max}$, then go to step 1.
   Else, if $t = t_{\max}$, then calculate the approximated likelihood function,

$$\hat{\pi}(\mathbf{x}|\theta) = \prod_{t=1}^{t_{\max}} \hat{\mathcal{L}}_t.$$

Spatially-Stratified Sampling Algorithm

In this section, we consider grouping individuals into strata based on their x–y coordinates. From here, we sample only a proportion of the infectious set from each stratum at each time point t when calculating $\hat{P}_{it}(\boldsymbol{\theta})$. Let k represent the index for each stratum up to a total of m strata and let

$$\hat{\mathcal{Z}}_{itk} \simeq \begin{cases} \sum_{j\in\mathcal{I}_{tk}} \Omega_T(j)\kappa(i,j) & |\mathcal{I}_{tk}| \leq q \\ \hat{\rho}_{tk}^{-1} \sum_{j\in\hat{\mathcal{I}}_{tk}} \Omega_T(j)\kappa(i,j) & |\mathcal{I}_{tk}| > q \end{cases}$$

be the estimate of the infectious pressure exerted on susceptible individual i from stratum k at time t. Here, $\hat{\rho}_{tk}$ is the empirical sampling proportion of the sampled infectious set for the stratum, $\mathcal{I}_{tk}$ is the complete set of infectious individuals in strata k at time t, and $\hat{\mathcal{I}}_{tk}$ is the randomly sampled set of infectious individuals obtained via SRS (with replacement) from strata k with empirical sampling proportion $\hat{\rho}_{tk}$. The sum of infectious pressures from each stratum exerted on individual i at time t is referred to as the total infectious pressure and calculated as

$$\hat{\mathcal{Z}}_{it} = \sum_{k=1}^{m} \hat{\mathcal{Z}}_{itk}.$$

As before, for small infectious sets, i.e., $|\mathcal{I}_{tk}| \leq q$, we use the entire infectious set and do not sample. Under a spatial-stratification scheme, we use $q = 5$. Thus, the approximation of the probability of infection is

$$P_{it}(\boldsymbol{\theta}) \simeq \hat{P}_{it}(\boldsymbol{\theta}) = 1 - \exp\left[\left\{-\Omega_S(i)\hat{\mathcal{Z}}_{it}\right\} - \epsilon(i,t)\right], \tag{7}$$

which is substituted into our likelihood function. We refer to this model as the SSS-ILM.

We consider a three-dimensional infection matrix, $\mathcal{Q}$, with dimensions $t_{\max} \times m \times n$ that contain elements corresponding to integer identification numbers for each farm. We use the notation $\mathcal{Q}[B, C, D]$ to refer to the farm ID located at time B, stratum C, and cell D within matrix $\mathcal{Q}$. We also define a two-dimensional matrix, $\mathcal{W}$, with dimensions $t_{\max} \times m$ that contain the number of infectious individuals in each stratum at every time point (up until the presence of empty cells). Thus, for each combination of $t = 1, \ldots, t_{\max}$ and $k = 1, \ldots, m$, $\mathcal{W}[t,k] = |\mathcal{Q}[t,k,\cdot]|$, where $\mathcal{W}[t,k]$ represents the number of infectious individuals at time t, in stratum k. We use the following algorithm to calculate the approximated likelihood function under the spatial stratification scheme:

1. Let $\hat{\mathcal{L}}_{tk} = 0 \;\forall\; t = 1, \ldots, t_{\max}$ and $k = 1, \ldots, m$.
   Set $t = 1$, $k = 1$ and $\ell_{tk} = \mathcal{W}[t,k]$.
2. If $\ell_{tk} \leq q$, calculate the likelihood component for strata k at time t,
$$\hat{\mathcal{L}}_{tk} = \prod_{i\in\mathcal{S}_{(t+1),k}} \left(1 - P_{it,k}(\boldsymbol{\theta})\right) \prod_{i\in\mathcal{I}_{(t+1),k}\setminus\mathcal{I}_{tk}} P_{it,k}(\boldsymbol{\theta}),$$
   and go to step 7.
   Else, if $\ell_{tk} > q$, let $\xi = \rho_{tk}\ell_{tk}$ and continue to step 2.
3. Let $c = 0$ and $\hat{v}_{tk}$ be a set of "empty" vectors of length to be determined by the algorithm.
4. Let $c = c + 1$.

5. If $c \leq \xi$, then simulate $U \sim \mathcal{DU}[1, \ell_{tk}]$, let $\hat{v}_{tk}[c] = v_{tk}[U]$, and return to step 3.
   If $c > \xi$, then let $\hat{\mathcal{I}}_{tk}$ be the set containing all $c - 1$ elements of $\hat{v}_{tk}$ and continue to step 5.

6. Calculate the approximated likelihood component for strata k at time t,

$$\hat{\mathcal{L}}_{tk} = \prod_{i \in \mathcal{S}_{(t+1),k}} \exp\left[\left\{-\Omega_S(i)\hat{\mathcal{Z}}_{itk}\right\} - \epsilon(i, tk)\right] \times \prod_{i \in \hat{\mathcal{I}}_{(t+1),k} \setminus \hat{\mathcal{I}}_{tk}} \left[1 - \exp\left\{\left(-\Omega_S(i)\hat{\mathcal{Z}}_{itk}\right) - \epsilon(i, tk)\right\}\right],$$

   where $\hat{\mathcal{Z}}_{itk} \simeq \hat{\rho}_{tk}^{-1} \sum_{j \in \hat{\mathcal{I}}_{tk}} \Omega_T(j)\kappa(i, j)$, as defined earlier, and $\hat{\rho}_{tk} = \frac{c-1}{\ell_{tk}}$.

7. Let $k = k + 1$.
   If $k \leq m$, then go to step 1.
   Else, if $k > m$, continue and calculate the approximated likelihood function for time t,

$$\hat{\mathcal{L}}_t = \prod_{k=1}^{m} \hat{\mathcal{L}}_{tk}.$$

8. Let $t = t+1$.
   If $t \leq t_{\max}$, then reset $k = 1$ and go to step 1.
   Else, if $t > t_{\max}$, then calculate the approximated likelihood function:

$$\hat{\pi}(\mathbf{x}|\theta) = \prod_{t=1}^{t_{\max}} \hat{\mathcal{L}}_t.$$

Data Augmentation

Because infection times/incubation periods are unknown and can change during the incubation period MCMC update, their infectious period can become longer or shorter; recall that, in our framework, the removal times remain constant. Thus, as each individual's infection time changes, we continually need to update our infection matrix to reflect the current infection times. For computational reasons, however, it is vital that the infection matrix, $\mathcal{Q}$, be updated in as efficient a manner as possible (and certainly not reconstructed from scratch) as these data-augmented parameters change.

As an example, say individual i_3's current infectious period is from $t_2 \rightarrow t_4$, and we are carrying out simple random sampling (i.e., no stratification). During the update process, the infection time increases by one and now i_3 is only infectious during the period of $t_3 \rightarrow t_4$. Below, we illustrates the matrix update process, and use 0s to represent empty cells. Each column represents a time point and the number of individuals infected at each time is displayed underneath the matrix. The first matrix shows the current infection times (before the update) for all individuals, including i_3. The second matrix shows that at time column t_2, i_3 is removed from its current position and replaced with a temporarily empty cell. The final matrix shows that i_5, which is the *last* individual in the t_2 column, is moved to i_3's old position (now i_5's new position) and i_5's old position is replaced with an empty cell. At each state, the number of elements in each column is also updated.

State before update:

$$
\begin{array}{c|cccccccc}
 & \mathbf{t=1} & \mathbf{t=2} & \mathbf{t=3} & \mathbf{t=4} & \mathbf{t=5} & \dots & \mathbf{t=t_{max}-1} & \mathbf{t=t_{max}} \\
 & i_1 & i_1 & i_2 & i_2 & i_2 & \dots & i_{n-1} & i_n \\
 & 0 & i_2 & i_3 & i_3 & i_6 & \dots & i_n & 0 \\
 & 0 & \boxed{i_3} & i_4 & i_4 & i_7 & \dots & 0 & 0 \\
 & 0 & i_4 & i_5 & i_6 & i_8 & \dots & 0 & 0 \\
 & 0 & i_5 & i_6 & i_7 & i_9 & \dots & 0 & 0 \\
 & 0 & 0 & i_7 & i_8 & 0 & \dots & 0 & 0 \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\
 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
|\mathcal{I}_t| & \mathbf{1} & \boxed{\mathbf{5}} & \mathbf{6} & \mathbf{6} & \mathbf{5} & \dots & \mathbf{2} & \mathbf{1}
\end{array}
$$

Intermediate step: (removing i_3 from the t_2 column)

$$
\begin{array}{c|cccccccc}
 & \mathbf{t=1} & \mathbf{t=2} & \mathbf{t=3} & \mathbf{t=4} & \mathbf{t=5} & \dots & \mathbf{t=t_{max}-1} & \mathbf{t=t_{max}} \\
 & i_1 & i_1 & i_2 & i_2 & i_2 & \dots & i_{n-1} & i_n \\
 & 0 & i_2 & i_3 & i_3 & i_6 & \dots & i_n & 0 \\
 & 0 & \boxed{0} & i_4 & i_4 & i_7 & \dots & 0 & 0 \\
 & 0 & i_4 & i_5 & i_6 & i_8 & \dots & 0 & 0 \\
 & 0 & i_5 & i_6 & i_7 & i_9 & \dots & 0 & 0 \\
 & 0 & 0 & i_7 & i_8 & 0 & \dots & 0 & 0 \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\
 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
|\mathcal{I}_t| & \mathbf{1} & \boxed{\mathbf{*}} & \mathbf{6} & \mathbf{6} & \mathbf{5} & \dots & \mathbf{2} & \mathbf{1}
\end{array}
$$

Final state of the matrix: (following movement of last element of t_2 column)

$$\begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ |\mathcal{I}_t| \end{array} \begin{array}{c} \mathbf{t=1} \\ \\ i_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mathbf{1} \end{array} \begin{array}{c} \mathbf{t=2} \\ \\ i_1 \\ i_2 \\ \boxed{i_5} \\ i_4 \\ \boxed{0} \\ 0 \\ \vdots \\ 0 \\ \boxed{\mathbf{4}} \end{array} \begin{array}{c} \mathbf{t=3} \\ \\ i_2 \\ i_3 \\ i_4 \\ i_5 \\ i_6 \\ i_7 \\ \vdots \\ 0 \\ \mathbf{6} \end{array} \begin{array}{c} \mathbf{t=4} \\ \\ i_2 \\ i_3 \\ i_4 \\ i_6 \\ i_7 \\ i_8 \\ \vdots \\ 0 \\ \mathbf{6} \end{array} \begin{array}{c} \mathbf{t=5} \\ \\ i_2 \\ i_6 \\ i_7 \\ i_8 \\ i_9 \\ 0 \\ \vdots \\ 0 \\ \mathbf{5} \end{array} \begin{array}{c} \dots \\ \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \end{array} \begin{array}{c} \mathbf{t=t_{max}-1} \\ \\ i_{n-1} \\ i_n \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mathbf{2} \end{array} \begin{array}{c} \mathbf{t=t_{max}} \\ \\ i_n \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mathbf{1} \end{array}$$

Then, if the change in infection time results in an individual being infected at a time point at which they were not previously, then they are simply added at the first zero in the particular column. So, for example in this case, if at the next MCMC iteration i_3's infection time changed such that they returned to having an infectious period from $t_2 \rightarrow t_4$, then a new matrix would be formed:

New state of the matrix:

$$\begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ |\mathcal{I}_t| \end{array} \begin{array}{c} \mathbf{t=1} \\ \\ i_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mathbf{1} \end{array} \begin{array}{c} \mathbf{t=2} \\ \\ i_1 \\ i_2 \\ i_5 \\ i_4 \\ \boxed{i_3} \\ 0 \\ \vdots \\ 0 \\ \boxed{\mathbf{5}} \end{array} \begin{array}{c} \mathbf{t=3} \\ \\ i_2 \\ i_3 \\ i_4 \\ i_5 \\ i_6 \\ i_7 \\ \vdots \\ 0 \\ \mathbf{6} \end{array} \begin{array}{c} \mathbf{t=4} \\ \\ i_2 \\ i_3 \\ i_4 \\ i_6 \\ i_7 \\ i_8 \\ \vdots \\ 0 \\ \mathbf{6} \end{array} \begin{array}{c} \mathbf{t=5} \\ \\ i_2 \\ i_6 \\ i_7 \\ i_8 \\ i_9 \\ 0 \\ \vdots \\ 0 \\ \mathbf{5} \end{array} \begin{array}{c} \dots \\ \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \end{array} \begin{array}{c} \mathbf{t=t_{max}-1} \\ \\ i_{n-1} \\ i_n \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mathbf{2} \end{array} \begin{array}{c} \mathbf{t=t_{max}} \\ \\ i_n \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mathbf{1} \end{array}$$

By following these methods, we avoid the problem of ending up with zeros in the middle of columns representing each time point (which would require searching, keeping track of where non-zeros are, or recompiling the matrix), and we only require sampling from the first $|\mathcal{I}_t|$ of each column of the matrix. In our MCMC update, infection times, and thus periods, increment or decrement only by one time point at each MCMC iteration, and are considered in single parameter updates. However, this scheme can easily be extended to allow for updates of larger magnitude, or indeed block updates.

Under the spatial stratification sampling schemes, the same method of adding to the first non-zero of a column, and switching a zero in the middle of a non-zero section of a column with the last non-zero of said column, is used. However, we are of course now working with columns in the 3-dimensional rather than 2-dimensional $\mathcal{Q}$.

Epidemic Data

To demonstrate the effectiveness of our sampling methods, we apply them to real and simulated data. Here, we describe these data in some detail, beginning with the simulated epidemic data and followed by the 2001 U.K. FMD epidemic.

Simulated Data

Using the Spatial ILM, we generated ten epidemics. The chosen population is of size $n = 625$ spread out evenly on a 25×25 grid, 1 unit apart in the x and y planes. Our susceptibility parameter is set to $\alpha = 1.4$ and our power law spatial parameter is set to $\beta = 2.3$. We generated the incubation period from an exponential distribution with rate parameter $\lambda_z = \frac{1}{3}$, giving an average incubation period of 3 days. The period from disease diagnosis to disease recovery and removal from the population was also generated from an exponential distribution with rate parameter $\lambda_w = \frac{1}{4}$, resulting in an average delay period of 4 days. Thus, the total infectious period, on average, is 7 days. Fig 1 illustrates the average infectious and non-infectious periods. In our modeling, we assume we know removal times but not the incubation period and thus estimate it via data augmentation. There is also an implicit assumption that the observation time occurs before removal.

FMD Data

We also implement our sampling-based parameterization methods on data from the 2001 U.K. FMD epidemic. We used a subset of the epidemic data, which was from the county of Cumbria located in North West England and consisted of 1,636 individual farms. According to [22], sheep and cattle farms accounted for almost all cases of the FMD outbreak in 2001 in the U.K. We consider farms to be the "individual"-level at which we are modeling and use cattle and sheep populations within farms as covariates in our model. We treat the disease diagnosis times recorded by veterinarians and epidemiologists who were on the ground during the outbreak as observed infection times. The times when animals were culled were also recorded and we treat these as the removal times in our modeling framework. We estimate the incubation periods (and the infection times indirectly) through Bayesian data augmentation. For some farms, disease diagnosis times were not recorded and so we assume these farms transition from state $\mathcal{S} \rightarrow \mathcal{R}$ on their cull date. In total, 730 infections were recorded in our data set. Refer to, for example, [8], for a more detailed description of the U.K. 2001 data set.

Note that most models for FMD assume an $\mathcal{SEIR}$ framework, with a latent, non-infectious state before infectiousness. We simplify our model for the purposes of illustration of our method, and—as mentioned in the discussion—extension to an $\mathcal{SEIR}$ framework would be relatively straightforward.

Priors

For both data sets, all ILM parameters, except for λ_z, are assigned independent, vague marginal priors under the assumption of weak prior knowledge. The marginal priors chosen here are positive, half-normal distributions with mode 0 and a 'large' variance of 10^5.

The marginal prior choice for λ_z, the incubation period rate parameter, is a Gamma distribution such that $\lambda_z \sim \Gamma(3, 9)$. This prior suggests an average incubation period of 3 days, which is the same as the incubation period from the simulated data. For the FMD epidemic, this prior may, of course, be misspecified because we do not know the actual incubation period.

Results

In this section, we present the results of our analyses. All computations were performed on an Apple Mac Pro with two 6-core Intel Xeon 2.93 GHz processors with 12 GB of RAM.

Simulation Study

Figs 2 and 3 illustrate the posterior means and 95% credible intervals for each ILM parameter, for each of the 10 epidemics simulated. Averages of these results over the ten epidemic data sets are shown in S1 Table.

As expected, bias for all parameters decreases as the sampling proportion, ρ, increases under the SRS technique. Posterior variance also decreases as ρ increases, leading to tighter credible intervals. It also appears to be more difficult to estimate the spatial parameter β using an SRS scheme with precision approaching that seen under the full MCMC analysis than is the case with α and λ_z.

Introducing spatial stratification in our sampling scheme appears to lead to increased posterior accuracy (Fig 3). Under these results (shown for $\rho = 0.10$ and $\rho = 0.50$), as the number of strata, m, increases, posterior variance and bias decrease and, thus, credible intervals are tighter. The posterior means also tend to be closer to the true parameter values. We also observe that spatial stratification is less sensitive to different ρ values tested and more sensitive to different values for m. For example, in Fig 3, the posterior results for α under $m = 4$ show almost no change for $\rho = 0.10$ versus $\rho = 0.50$. However, increasing the number of strata to $m = 9$ does increase posterior accuracy compared to $m = 4$. Under both values of ρ, we are able to obtain more posterior accuracy under the spatial stratification scheme than the simple random sampling scheme, demonstrating the advantage that including spatial stratification presents.

However, although spatial stratification appears more desirable in terms of posterior approximation, we must also consider the computation time required for the MCMC to run. Table 1 displays the average computation time (in hours) of each of our sampling methods. We observe that it takes approximately 92.64 hours to run 20,000 MCMC iterations of the true model without any data sampling. However, if we introduce SRS and set $\rho = 0.25$, we notice a drastic reduction in computation time; the MCMC takes only 36.48 hours to run. The computation time increases when ρ increases, as expected. The same is true if we consider spatial stratification in our sampling scheme. We see that smaller values of m yield faster run times, but the posterior approximation improves with larger m. Obviously, in practice, a trade off between posterior accuracy and computation time would be required.

FMD Model Fitting Results

Figs 4 and 5 show the results of implementing our sampling methods when fitting the data augmented FMD-ILM (tabulated results are given in S2 Table). Under the SRS approach, we see that posterior means for each ILM parameter tend to approach the posterior mean estimate

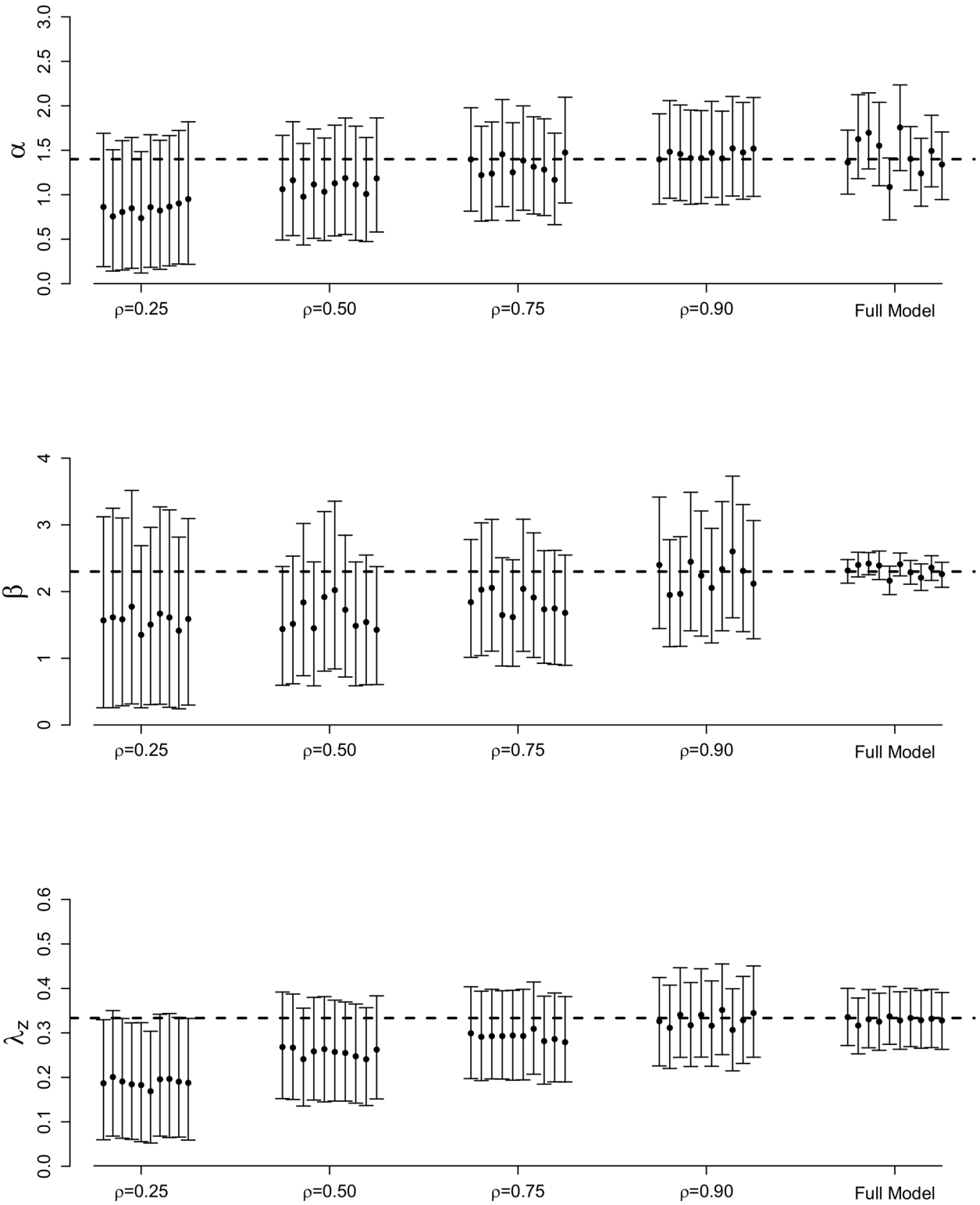

Fig 2. Posterior results for full MCMC and SRS methods. Posterior means and 95% credible intervals for α, β, and λ_z for the full MCMC and SRS methods for 10 different epidemics simulated from the data augmented spatial ILM with varying sampling proportions. The dashed, horizontal lines represent the true parameter values: $\alpha = 1.4$, $\beta = 2.3$, and $\lambda_z = \frac{1}{3}$, with a population of size $n = 625$.

doi:10.1371/journal.pone.0146253.g002

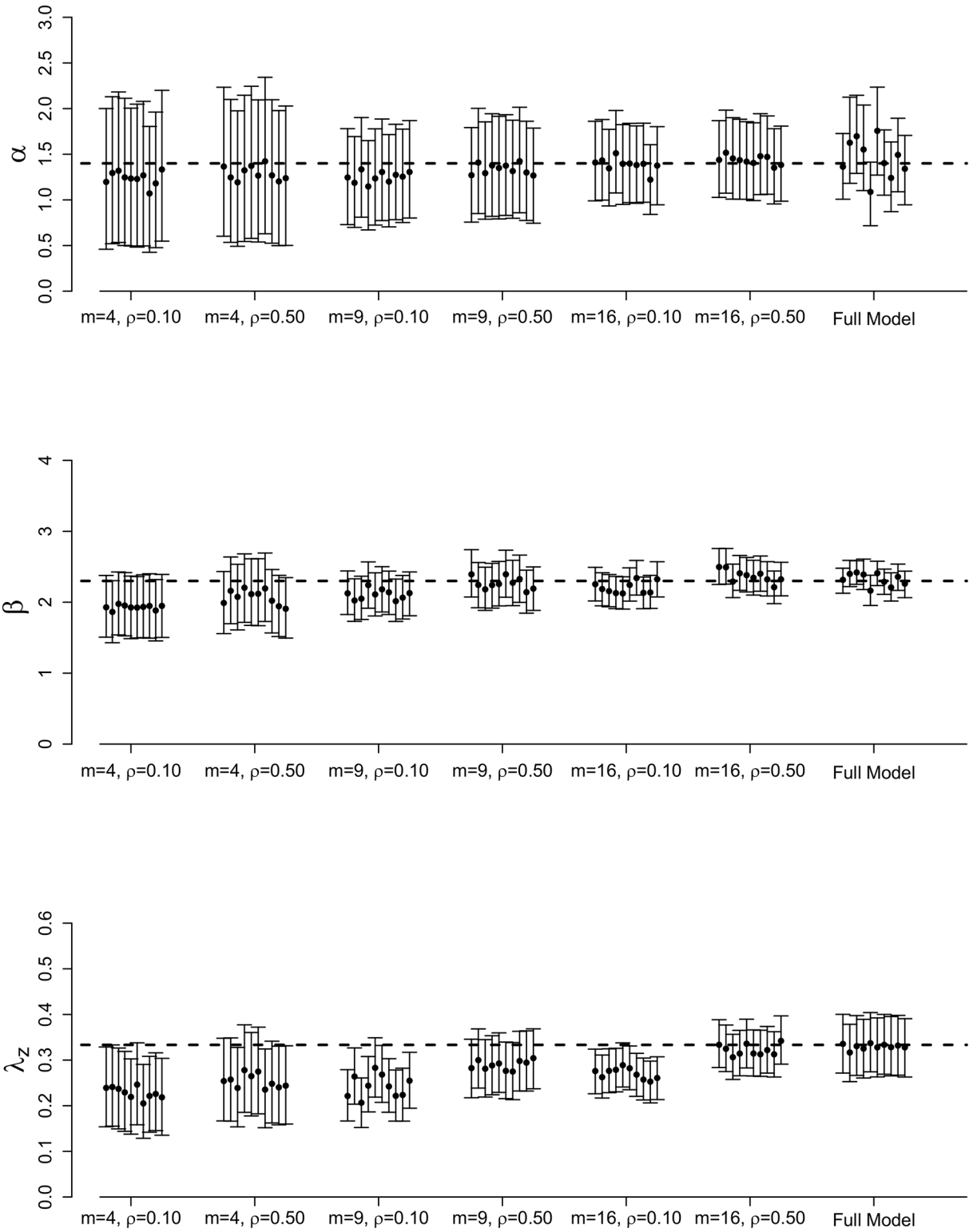

Fig 3. Full posterior results for full MCMC and spatial stratification methods. Posterior means and 95% credible intervals for α, β, and λ_z for the full MCMC and spatial stratification methods for 10 different epidemics simulated from the data augmented spatial ILM with varying values for m and ρ. The dashed, horizontal lines represent the true parameter values: $\alpha = 1.4$, $\beta = 2.3$, and $\lambda_z = \frac{1}{3}$, with a population of size $n = 625$.

doi:10.1371/journal.pone.0146253.g003

Table 1. Average computation time for the simulation studies.

ρ	m	Computation Time (hours)
—	—	92.64
0.25	—	36.48
0.50	—	47.76
0.75	—	59.28
0.90	—	75.12
0.10	4	46.88
0.50	4	56.16
0.10	9	66.24
0.50	9	71.52
0.10	16	82.32
0.50	16	88.56

Average computation run times (in hours) of fitting the data augmented spatial ILM, SRS-ILM, and the SSS-ILM to 10 different simulated epidemics. These epidemics were simulated using ILM parameters α = 1.4, β = 2.3, $\lambda_z = \frac{1}{3}$, and n = 625.

doi:10.1371/journal.pone.0146253.t001

under the full model as ρ is increased, similar to the findings of the simulation study. In general, we also observe that posterior variance decreases as ρ increases, resulting in tighter credible intervals closer to those under the full full model (which are generally the most narrow). Once again, these results mimic those seen in the simulation study.

Considering the spatially stratified schemes, we draw similar conclusions to those seen in the simulation study. Posterior accuracy tends to increase, and posterior variances decrease, as the number of strata, m, increases. In fact, for m = 16 we obtain posterior estimates that are very close to those seen under the full model for parameters ϕ_s, ϕ_c, and λ_z. Credible interval widths for ϕ_s and β change negligibly with regards to choice of m. Further, although posterior variance decreases as m increases, the credible intervals obtained under SRS with m = 4 (the lowest number of strata) contain those seen under the full MCMC analysis, suggesting good approximate inference.

Comparing the two sampling schemes, we see that spatial stratification tends to yield more accurate results. For example, if we compare SRS at ρ = 0.50 with stratification even at m = 4 (also sampled with ρ = 0.50), we find that all parameters under the spatial stratification scheme provide very similar, or more accurate, posterior means and tighter credible intervals. As m increases, as we have also seen, these SRS-based results improve even further.

Once again, however, a key aspect in addition to modeling accuracy is reduction in computation time. Table 2 provides the run times (in hours) for each FMD data analysis. With the full model taking approximately 249.12 hours to run for 20,000 MCMC iterations, we notice significant time savings at ρ = 0.25 and ρ = 0.50 using SRS and at m = 4 using spatial stratification. The time savings are much lower and of more questionable benefit for larger ρ and m. Further, and once again, time savings achieved using these sampling methods would be expected to be greater for larger data sets (see discussion).

Discussion

In this paper, we introduced sampling algorithms to help speed up the likelihood calculation for ILMs in a Bayesian MCMC framework. Unlike other proposed methods (e.g., [8, 12]), ours incorporates data augmented MCMC to allow for uncertainty about infection times into our analysis. We test the usefulness of our methods by comparing ILM parameter estimation under

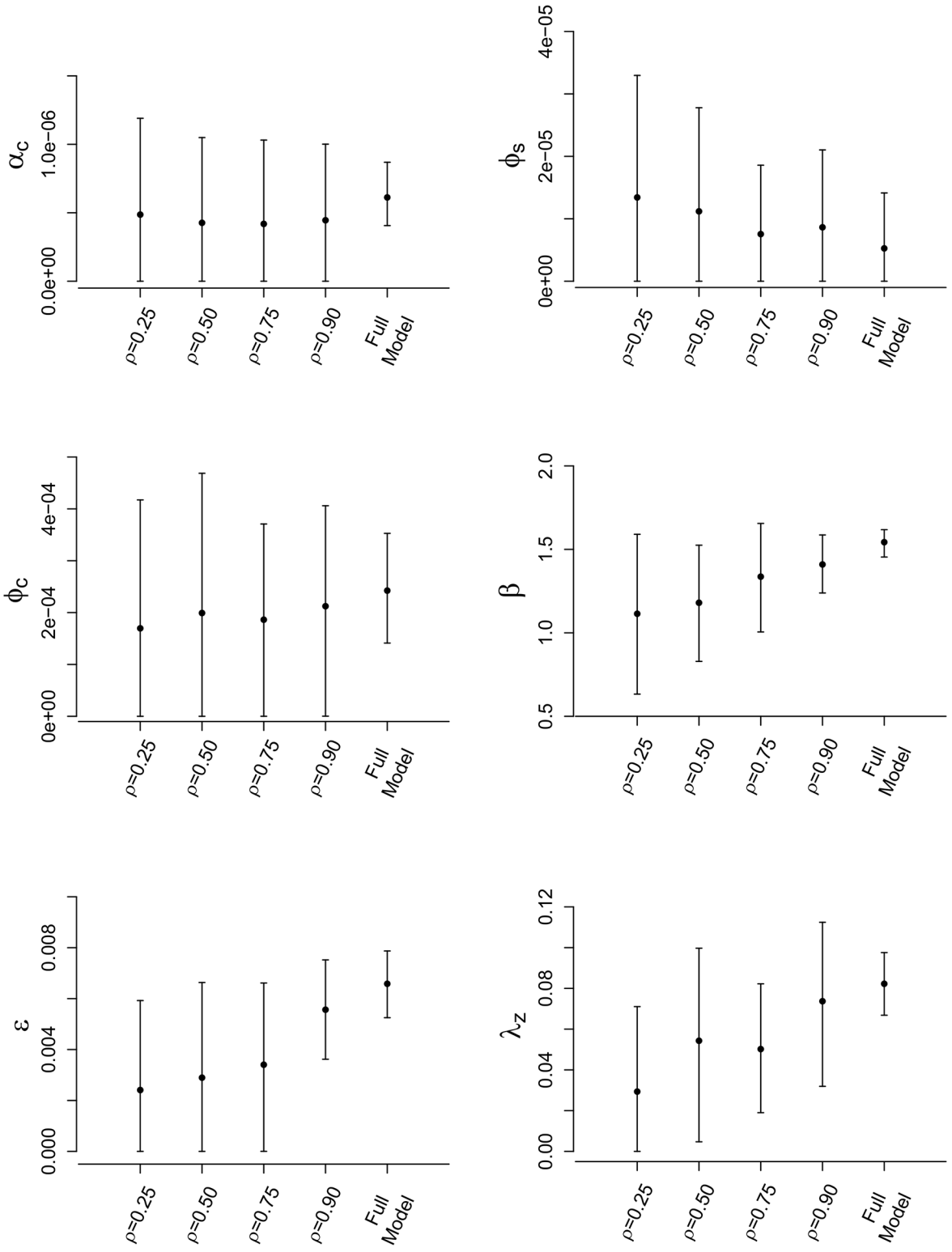

Fig 4. Posterior results for FMD-ILM using the SRS method. Posterior means and 95% credible intervals for all parameters of the data augmented FMD-ILM under the SRS method. The results are compared to the full model to assess accuracy.

doi:10.1371/journal.pone.0146253.g004

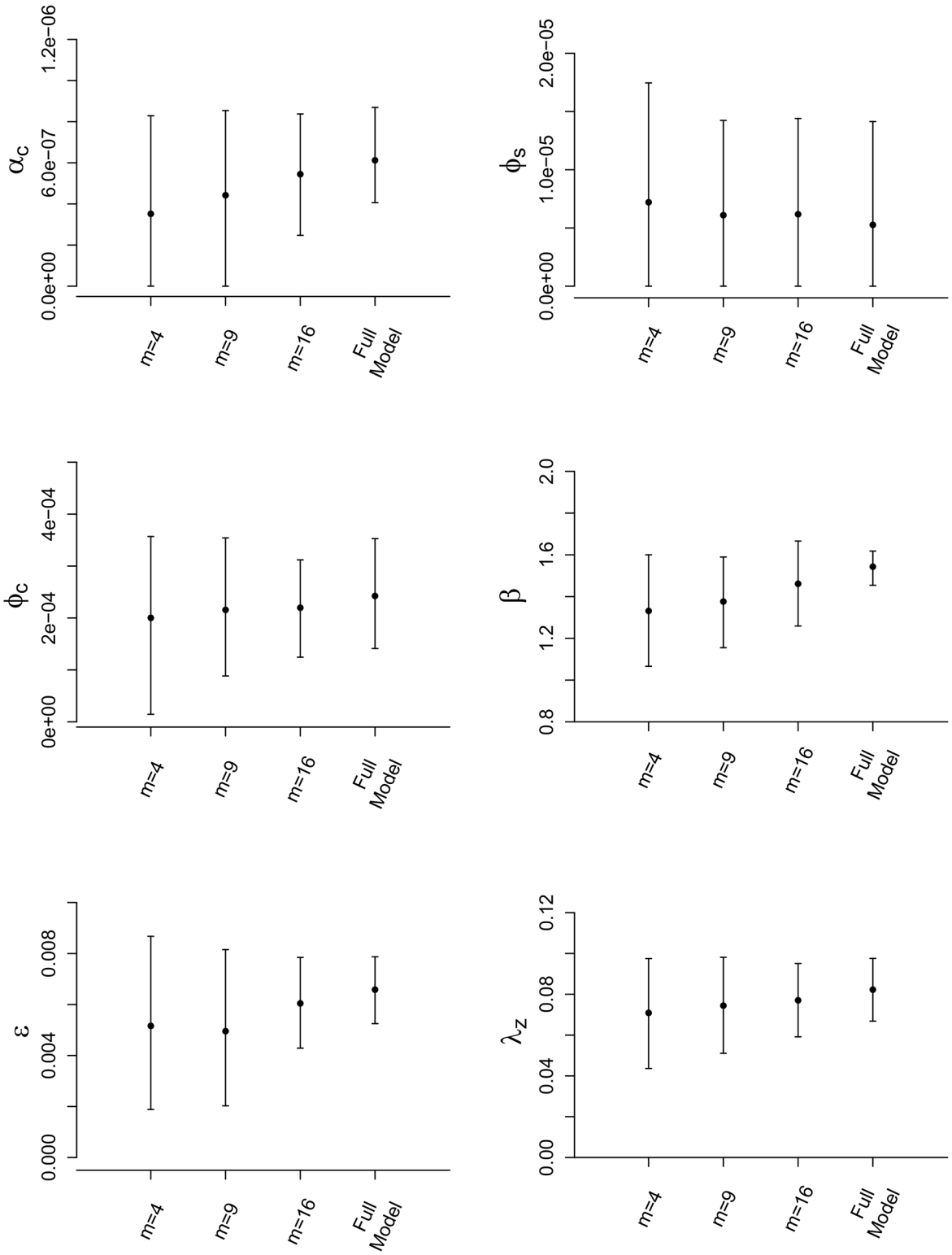

Fig 5. Posterior results for FMD-ILM using the spatial stratification method. Posterior means and 95% credible intervals for all parameters of the data augmented FMD-ILM under the spatial stratification method. We sampled $\rho = 0.50$ from each stratum. The results are compared to the full model.

doi:10.1371/journal.pone.0146253.g005

Table 2. Computation times for the FMD-ILM.

ρ	m	Computation Time (hours)
—	—	249.12
0.25	—	60.24
0.50	—	84.72
0.75	—	164.88
0.90	—	227.76
0.50	4	108.24
0.50	9	185.04
0.50	16	232.80

Computation run times (in hours) of fitting the data augmented FMD-ILM to the FMD data using various sampling techniques. Models produced 20,000 realizations from the posterior.

doi:10.1371/journal.pone.0146253.t002

the full Bayesian analysis via a simulation study and using data from the 2001 FMD epidemic in the U.K. Our results show that overall, we were able to obtain fairly accurate (though less precise) results using sampling-based likelihood approximations compared to the results obtained under the full likelihood analysis. In terms of computation run times, we found significant savings could be made by using data sampling. Because the problem of repeated likelihood calculations under the full model is increased drastically with the inclusion of data augmentation, this is a result of key importance. However, we found using larger values of ρ or m can drastically reduce the time saving benefit over the full MCMC analysis.

In our studies, only two sampling techniques were considered. Possible future work could involve investigating other sampling procedures that might provide stronger inferential conclusions. For example, our spatial stratification technique consisted of dividing the population into equally sized cells/strata and then sampling from each cell with equal sampling proportions. This would seem intuitively sensible when the population is spread across a grid, as was the case in our simulation study. This may be reasonable for some crop diseases or perhaps if points on the grid represent regions or cells (e.g., consider the modeling of fire spread by [23]), but such a population layout would be quite unrealistic in most situations. (It was, of course, adequate for the main aim of this paper, which was to illustrate the facilitation of faster likelihood calculations via data sampling).

In most populations, some natural clustering of individuals tends to take place (e.g., there tend to be high density clusters of farms in regions where infrastructural and/or environmental conditions are suitable for the type of farming in question). In such situations, spatial strata could, for example, be based upon some spatial clustering method applied to the population data. Alternatively, for a population in which some sort of contact network, or series of such networks, were being used as a prime risk factor in the model, clustering based on the network (s), using say partitioning around medoids (PAM) [24], could be considered as a way of defining strata from which to sample.

Here, we assumed that the sampling proportion was invariant to time and/or stratum. However, it might be useful to allow the sampling proportion to vary according to one or both. We might also possibly want to place more weight on sampling at times when the epidemic intensity is highest. Alternatively, in the case of spatially stratified sampling, we might want to avoid sampling from some strata with very low epidemic intensity. Such methods could possibly provide faster computation concurrent with more accurate model parameterization.

There are also, of course, many other options for carrying out approximate inference when computational efficiency is a driving factor. For example, [25] use a Gaussian process emulator

method based on mapping key summary statistics from model simulations to the parameter space. In a similar vein, the aforementioned so-called approximate Bayesian computational methods used by, for example, [14] and [15], can be employed. These are also based on comparing salient summary statistics from observed and simulated data. A systematic comparison of all of these different approaches would be of obvious interest.

Our study used a $\mathcal{SIR}$ modeling framework. We could extend the analysis presented here to a $\mathcal{SEIR}$ framework to investigate disease exposure times. In our modeling, we accounted for incubation by treating it as a period when infected individuals have not been diagnosed yet but can pass on the disease to others. Introducing an exposed state would indicate an individual has contracted the disease but cannot pass it on to others until they reach the infectious state, regardless of confirmation of disease diagnosis. Additionally, we assumed knowledge of when individuals were removed from the population; however, this would not be the case for most diseases (e.g., human influenza). In a future study, we can also explore scenarios where removal times are unknown and instead estimated through data augmentation. The modeling framework used in this paper was also set in discrete time. The time saving sampling used here can also be applied in a (more natural, arguably) continuous time modeling framework.

We have demonstrated as a proof of concept that, for these relatively small datasets, our sampling-based likelihood approximations can result in a significant decrease in computation time. The time savings using these sampling algorithms would be even more beneficial in large-scale problems involving massive data sets compared to a full Bayesian analysis. A natural avenue of possible future work would be to apply these techniques to much larger data sets. Of course, these techniques would only really be worth using for large data sets in which a full Bayesian analysis was computationally prohibitive, in which case the priority would likely to be to get some sort of 'rough and ready' inference done as quickly as possible, rather than worry too much about the quality of posterior approximation. However, some degree of thought would have to be given to the choice of ρ and the stratification methods used in order to achieve parametrization of a reasonable quality in a feasible time frame. Further work on the use of some sort of adaptive scheme, based initially on a quick pilot study over sampling proportions and stratification schemes, might also therefore be of interest.

Supporting Information

S1 Table. Summary statistics for the simulation studies. Summary statistics from the simulation studies comparing model parameter estimation across our different sampling schemes. The results are averaged over 10 different epidemics simulated from the data augmented spatial ILM with parameter values $\alpha = 1.4, \beta = 2.3, \lambda_z = \frac{1}{3}$, and $n = 625$. Here, CIs are the mean credible interval limits.
(PDF)

S2 Table. Summary statistics for modeling the FMD-ILM. Summary of results from fitting the data augmented FMD-ILM to the FMD data. We compare the results across our different sampling methods. Note that for spatial stratification, we sample $\rho = 0.50$ from each stratum.
(PDF)

Author Contributions

Conceived and designed the experiments: RM RD. Performed the experiments: RM. Analyzed the data: RM RD GPSK. Wrote the paper: RM RD.

References

1. Sun GQ. Pattern formation of an epidemic model with diffusion. Nonlinear Dynamics. 2012; 69 (3):1097–1104. doi: 10.1007/s11071-012-0330-5
2. Sun GQ, Liu QX, Jin Z, Chakraborty A, Li BL. Influence of infection rate and migration on extinction of disease in spatial epidemics. Journal of Theoretical Biology. 2010; 264(1):95–103. doi: 10.1016/j.jtbi.2010.01.006 PMID: 20085769
3. O'Neill PD. Introduction and snapshot review: relating infectious disease transmission models to data. Statistics in Medicine. 2010; 29(20):2069–2077. doi: 10.1002/sim.3968 PMID: 20809536
4. Shekhar S, Evans MR, Kang JM, Mohan P. Identifying patterns in spatial information: a survey of methods. WIREs Data Mining and Knowledge Discovery. 2011; 1:193–214. doi: 10.1002/widm.25
5. Chis Ster I, Ferguson NM. Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. PLoS ONE. 2007; 2(6):e502. doi: 10.1371/journal.pone.0000502 PMID: 17551582
6. Chis Ster I, Singh BK, Ferguson NM. Epidemiological inference for partially observed epidemics: the example of the 2001 foot and mouth epidemic in Great Britain. Epidemics. 2009; 1(1):21–24. doi: 10.1016/j.epidem.2008.09.001 PMID: 21352749
7. Jewell CP, Kypraios T, Neal P, Roberts GO. Bayesian analysis for emerging infectious diseases. Bayesian Analysis. 2009; 4(2):191–222.
8. Deardon R, Brooks SP, Grenfell BT, Keeling MJ, Tildesley MJ, Savill NJ, et al. Inference for individual level models of infectious diseases in large populations. Statistica Sinica. 2010; 20(1):239–261. PMID: 26405426
9. Malik R, Deardon R, Kwong GPS, Cowling BJ. Individual-level modeling of the spread of influenza within households. Journal of Applied Statistics. 2014; 41(7):1578–1592. doi: 10.1080/02664763.2014.881787
10. Gamerman D, Lopes HF. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. 2nd ed. Chapman & Hall/CRC Texts in Statistical Science. CRC Press; 2006.
11. Daley DJ, Gani J. Epidemic Models: An Introduction. Cambridge University Press; 2001.
12. Kwong GPS, Deardon R. Linearized forms of individual-level models for large-scale spatial infectious disease systems. Bulletin of Mathematical Biology. 2012; 74(8):1912–1937. doi: 10.1007/s11538-012-9739-8 PMID: 22718395
13. Brown PE, Chimard F, Remorov A, Rosenthal JS, Wang X. Statistical inference and computational efficiency for spatial infectious disease models with plantation data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2013.
14. McKinley T, Cook AR, Deardon R. Inference in epidemic models without likelihoods. The International Journal of Biostatistics. 2009; 5(1): Article 24. doi: 10.2202/1557-4679.1171
15. Toni T, Welch D, Strelkowa N, Ipsen A, Strumpf MP. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface. 2009; 6:187–202. doi: 10.1098/rsif.2008.0172
16. McKinley TJ, Ross JV, Deardon R, Cook AR. Simulation-based Bayesian inference for epidemic models. Computational Statistics & Data Analysis. 2014; 71:434–447. Available from: http://www.sciencedirect.com/science/article/pii/S016794731200446X. doi: 10.1016/j.csda.2012.12.012
17. Manolopoulou I, Chan C, West M. Selection sampling from large data sets for targeted inference in mixture modeling. Bayesian Analysis. 2010; 5(3):429–450. doi: 10.1214/10-BA517
18. Cauchemez S, Ferguson NM. Methods to infer transmission risk factors in complex outbreak data. Journal of the Royal Society Interface. 2012; 68(9):456–469. doi: 10.1098/rsif.2011.0379
19. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. Journal of Chemical Physics. 1953; 21(6):1087–1092. doi: 10.1063/1.1699114
20. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970; 57(1):97–109. doi: 10.1093/biomet/57.1.97
21. Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. The American Statistician. 1995; 49(4):327–335. doi: 10.1080/00031305.1995.10476177
22. Anderson I. Foot and mouth disease 2001: Lessons to be learned inquiry. London: The Stationary Office; 2002.
23. Vrbik I, Deardon R, Feng Z, Gardner A, Braun J. Using individual-level models for infectious disease spread to model spatio-temporal combustion dynamics. Bayesian Analysis. 2012; 7(3):615–638. doi: 10.1214/12-BA721

24. Park HS, Jun CH. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications. 2009; 36(2):3336–3341. doi: 10.1016/j.eswa.2008.01.039

25. Jandarov R, Haran M, Bjørnstad O, Grenfell B. Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2014; 63(3):423–444. doi: 10.1111/rssc.12042