Bayesian Analysis (2009) **4**, Number 4, pp. 465–496

Bayesian Analysis for Emerging Infectious Diseases

Chris P Jewell[*], Theodore Kypraios[†], Peter Neal[‡] and Gareth O. Roberts[§]

Abstract. Infectious diseases both within human and animal populations often pose serious health and socioeconomic risks. From a statistical perspective, their prediction is complicated by the fact that no two epidemics are identical due to changing contact habits, mutations of infectious agents, and changing human and animal behaviour in response to the presence of an epidemic. Thus model parameters governing infectious mechanisms will typically be unknown. On the other hand, epidemic control strategies need to be decided rapidly as data accumulate. In this paper we present a fully Bayesian methodology for performing inference and online prediction for epidemics in structured populations. Key features of our approach are the development of an MCMC- (and adaptive MCMC-) based methodology for parameter estimation, epidemic prediction, and online assessment of risk from currently unobserved infections. We illustrate our methods using two complementary studies: an analysis of the 2001 UK Foot and Mouth epidemic, and modelling the potential risk from a possible future Avian Influenza epidemic to the UK Poultry industry.

1 Introduction

Animal disease epidemics in agriculture are potentially devastating in terms of economics, sociology, and public health. Although there are many diseases that are endemic in UK herds and flocks, there are some from which policy dictates we must remain free (Defra 2007a). These are the so-called *notifiable* diseases, and are chosen on the basis of their current absence, and risk to the agricultural economy and public health. The disadvantage to maintaining a disease-free status is that the population, in the absence of vaccination, has no prior exposure to the disease and hence a chance to build an acquired immunity. Such populations are therefore highly susceptible to the accidental introduction of disease which, if the transmission mechanisms allow, will run unchecked on a national scale.

Responding to such accidental outbreaks is a problem faced by governmental authorities, and the design of national-scale disease control measures prior to an outbreak is a high priority. Nevertheless, since epidemic outbreaks are rare and tend to behave differently to one-another, a high degree of flexibility must be maintained within control-policies to adapt to unexpected features of an epidemic in progress.

[*] Department of Mathematics and Statistics, Lancaster University, UK, mailto:chris.jewell@warwick.ac.uk

[†] School of Mathematical Sciences, University of Nottingham, UK, mailto:theodore.kypraios@nottingham.ac.uk

[‡] School of Mathematics, University of Manchester, UK, mailto:P.Neal-2@manchester.ac.uk

[§] Department of Statistics, University of Warwick, UK, mailto:Gareth.O.Roberts@warwick.ac.uk

DOI:10.1214/09-BA417

This paper presents a unified Bayesian approach that can be used to inform policy adaptation during disease outbreaks. First we develop a generic modelling framework at the population level. Here, individual elements of the 'population' can be chosen according to the specific application at hand (eg individual animal, barn, farm, etc.). Our methodology is illustrated by two extended case studies of recent popular interest: the 2001 UK Foot-and-Mouth epidemic (FMD), and the prospect of a future Highly Pathogenic Avian Influenza epidemic (HPAI) within the UK poultry industry. In both cases, the population consists of UK farms of the relevant type. Before describing the motivating examples in detail, we give a brief background to parametric inference for infectious disease spread along with motivation for the key aspects of the methodology to be used in the analysis.

Excellent reviews of classical approaches to analyzing epidemic data, such as martingale methods and generalized linear models are given in Becker (1989) and Becker and Britton (1999). However, statistical inference for infectious disease models has taken off with the introduction of MCMC Gibson (1997); Gibson and Renshaw (1998); O'Neill and Roberts (1999). This is primarily because infectious disease data usually suffers from the complaint that the data which is available is incomplete, making the evaluation of the model likelihood difficult. Therefore, data augmentation (imputation) forms a key aspect of any MCMC algorithm to analyze such data. Data augmentation is usually a very straightforward procedure. However, there is often a strong dependence between the model parameters and the imputed data which can lead to slow convergence of the MCMC algorithm. Non-centered reparameterizations (see, Papaspiliopoulos et al. (2003)) have been developed to improve the convergence rates of MCMC algorithms where there is a strong dependence between the model parameters and the imputed data. This approach has been successfully applied to epidemic models in Neal and Roberts (2005) and Kypraios (2007) and we give an outline and refinement of the methodology in Section 2.5.

As with many statistical models, neither the full posterior distribution of the model parameters nor the conditional posterior distributions of the model parameters follow standard probability distributions. Therefore a Gibbs sampler algorithm is inappropriate and a random walk Metropolis (rwM) algorithm is a natural candidate. Suppose that there are p model parameters and that θ (a p-dimensional vector) is the current value of the model parameters based upon the MCMC algorithm. Then in the (Gaussian) rwM algorithm we propose new values θ' for the model parameters from a multivariate Gaussian distribution with mean θ and covariance matrix Σ. The choice of Σ is crucial to the efficiency of the MCMC algorithm. Often if p is small, Σ is fine tuned through pilot runs of the algorithm. However this takes time and if $p \geq 10$ this can become infeasible in conjunction with real-time analysis of data. Real-time data analysis is crucially important in seeking to control emerging diseases. A solution is offered by adaptive MCMC (see, Haario et al. (2001), Roberts and Rosenthal (2007)) which updates Σ based upon the performance of the algorithm. This produces an efficient MCMC algorithm which automatically tunes itself, thus making the algorithm applicable for analysing data in real-time. This is further assisted by the parallelization of the computer code. Parallel computing is a computational rather than statistical issue, but does have a large bearing on the size of data sets which can be analyzed in a reasonable time frame. Therefore this is an important issue for MCMC practitioners to be aware of, even if it is not always an issue.

The paper is structured as follows. In section 3 the motivating examples Foot-and-Mouth disease and Avian influenza are introduced in detail. However, first in section 2 we introduce the methodology. Whilst the methodology is described with the motivating examples in mind, it is sufficiently general to be applied to a whole host of agricultural epidemics. The results of the analysis FMD and HPAI are given in section 3 with concluding discussion in section 4.

2 Generic Methodology

Before we describe our specific implementations for FMD and HPAI, we will discuss a generic methodology applicable to all disease epidemics. We therefore begin by describing the epidemic model, followed by the construction of a likelihood expression for the model parameters. Having derived the likelihood, we consider the practical problems that arise from the fact that our observation of the epidemic process is necessarily incomplete. Throughout we highlight the novel aspects of the model and their analysis.

2.1 Introduction

The main motivation for the research was the risk analysis of emerging infectious diseases through UK farms. Therefore the key aspect of the analysis is obtaining an understanding of the transmission mechanism between individual farms. As a result we model the disease on the macroscopic farm level rather than the microscopic individual level. The epidemic model is defined as follows:

Consider a total population of size N farms. We assume that at any given time point each farm i can be in one of four states:

- **S**usceptible premises (S) do not have the disease and are able to be infected by it.
- **I**nfected premises (I) have the disease and are able to infect susceptible premises. Their infectivity increases as a function of time.
- **N**otified premises (N) have been detected as having the disease and are subject to government-imposed movement restrictions. However, they are still capable of infecting susceptibles by other means, such that their infectivity is curbed at a lower level.
- **R**emoved premises (R), in the case of both FMD and HPAI, have had their animals culled and therefore play no further part in the epidemic.

The only transitions in state we allow are: from susceptible to infected, from infected to notified, and from notified to removed. There are two features which we shall impose on the model which are worthy of extra comment at this stage. Firstly, we allow the infection rate of a farm to vary with time. This corresponds to the within-farm epidemic leading to the infection of more and more individual animals over time. We are implicitly assuming that, within the interval between a farm's infection and notification, that the farm is infectious the infectivity is increasing and not significantly affected by the death or recovery of animals. For FMD and HPAI this assumption is reasonable. Often when modeling the disease at an animal level the infection period is divided into a latent period (ie infected but not yet infectious) followed by an infectious period with a constant rate of infectivity. However, the time varying infectivity is more natural/biologically plausible at the macroscopic level than assuming a step function for the infectivity of farms (latent/infectious period). Secondly, it is assumed that a farm is infected only once from outside sources. This implies that we treat the herd or flock of animals present on a farm as a single entity. This is a reasonable assumption when the intra-herd/flock infection rate dominates the inter-farm infection rate which is the case for FMD and HPAI.

We assume that the epidemic is observed up to a certain time, say T_{obs}. Denote by n_I, n_N and n_R, the total number of premises who got infected, notified and removed by time T_{obs}, respectively. The exact relationship between n_I, n_N and n_R will depend upon the culling regime imposed. Finally, note that we assume that only notified farms are removed (culled). The methodology can easily be adapted to other cases such as the culling of contiguous premises.

2.2 Infectious Process

We assume that an infectious farm i makes infectious contacts with a given farm j at the points of a time inhomogeneous Poisson point process with transmission rate at time t:

$$\beta_{ij}(t) = T_{ij} h(t - I_i), \tag{1}$$

where I_i denotes the point of time at which farm i becomes infected, T_{ij} is a measure of the association/infectivity between farms i and j and $h(\cdot)$ represents how a farm's infectivity changes over time. In particular, we choose

$$\begin{aligned} 0 < h(s) \leq 1 \qquad & s > 0 \\ h(s) = 0 \qquad & s \leq 0. \end{aligned}$$

The function $h(\cdot)$ is problem-specific and the choice of this function is discussed later. In particular in Section 3, $h(s) = 1$ $(s > 0)$, a step function, and $h(s) = e^{\nu s}/(\mu + e^{\nu s})$ $(s > 0)$ are used for FMD and HPAI, respectively.

Let $\mathcal{I}_t$, $\mathcal{N}_t$ and $\mathcal{S}_t$ denote the sets of infected, notified and susceptible farms, at time

t, respectively. Then, at time t

$$T_{ij} = \begin{cases} \beta_{ij}, & i \in \mathcal{I}_t,\ j \in \mathcal{S}_t \\ \beta^*_{ij}, & i \in \mathcal{N}_t,\ j \in \mathcal{S}_t, \end{cases} \tag{2}$$

where

$$\begin{aligned} \beta_{ij} &= q(i;\zeta)s(j;\xi)\left\{K(i,j;\psi) + \mathbf{p}^T\mathbf{r}_{ij} + \boldsymbol{\beta}^T\mathbf{c}_{ij}\right\} \\ \beta^*_{ij} &= \gamma q(i;\zeta)s(j;\xi)K(i,j;\psi), \end{aligned} \tag{3}$$

and $q(i;\zeta)$ and $s(j;\xi)$ denote the baseline infectivity of farm i and the baseline susceptibility of farm j, respectively. This depends upon covariates/parameters ζ and ξ. Throughout this paper ζ and ξ will depend upon the total number of animals of each type on each farm.

The vectors $\mathbf{r}_{ij}$ and $\mathbf{c}_{ij}$ are vectors of covariates which describe the (business) relationships between farms i and j providing potential sources of infectious contacts. Let $\mathbf{r}_{ij}$ represent covariates with known contact frequencies such as, for example, the total number of weekly feed deliveries with the vector $\mathbf{p}$ denoting the corresponding probabilities of infection associated with each type of contact. Conversely, $\mathbf{c}_{ij}$ represents covariates with unknown contact frequencies, and takes the values 0 or 1 depending on whether a possible route of contact exists. Then $\boldsymbol{\beta}$ denotes the corresponding overall infection rate for each type of contact link.

Also, environmental factors could be important in the spread of the disease – rodents, wild birds and walkers all being potential transmission sources. Therefore $K(i,j;\psi)$ represents the environmental transmission rate between farms i and j and depends upon parameters ψ. Throughout we take $K(i,j;\psi)$ to be a function of the Euclidean distance between farms i and j. Whilst it is assumed that the control measures eliminate business risk, the above model assumes that environmental risks can only be reduced, not eliminated, by control measures. Assuming that γ takes values between 0 and 1, $1-\gamma$ then represents the reduction in this risk.

Finally, we assume that each farm j has an underlying risk of infection with transmission rate ϵ. This captures unexplained infections caused by, for example, migratory birds.

2.3 Infectious Periods

The notification and culling (removal) dates of farms are recorded. However, it is not known when infection of the farm occurred. Therefore $D_i = N_i - I_i$, the time from infection to notification of farm i, is a random variable. In the so-called GSE (homogeneous mixing Markov model, see for example Bailey (1975)), D_i would be taken to be exponentially distributed, but in our context any non-negative probability distribution can be used. Let $f_D(x)$ $(x \geq 0)$ denote the probability density function (pdf) of D and

$$F_D(x) \quad = \quad \int_x^\infty f_D(y)\,dy \qquad (x \geq 0). \tag{4}$$

The choice of D will be problem specific.

2.4 Likelihood

Having described the epidemic process, we turn to the question of statistical inference. Let $\mathbf{I}$, $\mathbf{N}$ and $\mathbf{R}$ denote the infection, notification and removal times, respectively. Note that up to time T_{obs}, $\mathbf{N}$ and $\mathbf{R}$ are known, whilst both the total number n_I and the times of $\mathbf{I}$ of infections are unknown. Let $\boldsymbol{\theta} = (\psi, \boldsymbol{\beta}, \gamma, \epsilon, \zeta, \xi)$ and $\boldsymbol{\alpha}$ denote the infection and infectious period (*i.e.* D – time from infection to notification) parameters, respectively. The likelihood of the data (the epidemic outbreak) given the model parameters can then be expressed as the product of the infection terms and the infectious period terms. Before giving the likelihood, however, we introduce some notation. We label the premises that become infected up to time T_{obs} by $i = 1, 2, \ldots, n_I$ and the remainder by $i = n_I + 1, n_I + 2, \ldots, N$. We adopt the notation that if premise j is never infected then $I_j = N_j = R_j = \infty$. A premises j just prior to becoming infected receives *infectious pressure* from the premises in the sets $\mathbf{Y}_{j-}$ and $\mathbf{Y}^*_{j-}$.

$$\begin{aligned}\mathbf{Y}_{j-} &:= \{i : I_i < I_j \leq N_i\}\\ \mathbf{Y}^*_{j-} &:= \{i : N_i < I_j \leq R_i\}\end{aligned}$$

Thus the likelihood is given by

$$\begin{aligned}L(\mathbf{I},\mathbf{N},\mathbf{R}|\boldsymbol{\theta},\boldsymbol{\alpha}) \quad &\propto \quad \prod_{j\neq\kappa}^{n_I}\left(\epsilon + \sum_{i\in\mathbf{Y}_{j-}}\beta_{ij}(I_j) + \sum_{i\in\mathbf{Y}^*_{j-}}\beta^*_{ij}(I_j)\right)\\ &\times \quad \exp\left\{-\int_{I_\kappa}^{T_{obs}}\left(\sum_{i\in\mathcal{S}_t}\epsilon + \sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{S}_t}\beta_{ij}(t-I_i) + \sum_{i\in\mathcal{N}_t}\sum_{j\in\mathcal{S}_t}\beta^*_{ij}(t-I_i)\right)dt\right\}\\ &\times \quad \prod_{i=1}^{n_I} f_D(N_i - I_i), \qquad (5)\end{aligned}$$

where κ denotes the label of the initial infective farm with I_κ its corresponding infection time. The parameter ϵ represents the (additive) unexplained background infectious pressure. The term in the exponent is termed the *total infectious pressure* and we shall denote this by S.

In our examples, we shall use independent Gamma priors for each of the parameters, *i.e.* parameter μ has prior $f(\mu) \sim Ga(\lambda_\mu, \nu_\mu)$. Thus

$$f(\boldsymbol{\theta},\boldsymbol{\alpha}|\mathbf{I},\mathbf{N},\mathbf{R}) \propto L(\mathbf{I},\mathbf{N},\mathbf{R}|\boldsymbol{\theta},\boldsymbol{\alpha}) \times \prod_{\mu\in\boldsymbol{\theta},\boldsymbol{\alpha}} f(\mu). \qquad (6)$$

For the infection parameters $\boldsymbol{\theta}$, none of the parameters have standard conditional distributions. Our approach uses a simple random walk Metropolis (within Gibbs) algorithm. The infectious period parameters will be studied in detail shortly.

The first step before we give the MCMC algorithm is to simplify (5). In particular, we can rewrite, S, the integral on line two of (5) as a double sum which is both illuminating and easy to calculate. For $t \geq 0$, let $H(t) = \int_0^t h(s)\,ds$. Then since $\beta_{ij}(t) = \beta_{ij}h(t)$,

$$\begin{aligned} S &= \int_{I_\kappa}^{T_{obs}} \left(\sum_{i\in\mathcal{S}_t} \epsilon + \sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{S}_t} \beta_{ij}(t-I_i) + \sum_{i\in\mathcal{N}_t}\sum_{j\in\mathcal{S}_t} \beta^\star_{ij}(t-I_i) \right) dt \\ &= \epsilon \sum_{i=1}^{N} \{T_{obs} \wedge I_i - I_\kappa\} + \sum_{i=1}^{n_I}\sum_{j=1}^{N} H(\{T_{obs}\wedge N_i \wedge I_j\} - \{I_i \wedge I_j\})\beta_{ij} \\ &\quad + \sum_{i=1}^{n_N}\sum_{j=1}^{N} \{H(\{T_{obs}\wedge R_i \wedge I_j\} - \{I_i\wedge I_j\}) - H(\{T_{obs}\wedge N_i\wedge I_j\} - \{I_i \wedge I_j\})\}\beta^\star_{ij} \end{aligned} \tag{7}$$

2.5 Non-centering

For missing data problems, MCMC convergence is often significantly improved by the adoption of a *non-centered* parameterization (see Papaspiliopoulos et al. (2003) for a review). Such parameterizations orthogonalize prior structure, and work particularly well when components of the missing data are poorly identified by observations.

Since $\mathbf{I}$ is unknown and can be treated as a parameter in the model, a standard (*centered*) Metropolis-within-Gibbs MCMC procedure would carry out the following steps.

1. For each i, update $\theta_i|\boldsymbol{\alpha}, \boldsymbol{\theta}_{i-}, \mathbf{I}, \mathbf{N}, \mathbf{R}$.
2. For each i, update $\alpha_i|\boldsymbol{\theta}, \boldsymbol{\alpha}_{i-}, \mathbf{I}, \mathbf{N}, \mathbf{R}$.
3. For each i, update $I_i|\boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{I}_{i-}, \mathbf{N}, \mathbf{R}$.

A non-centered construction for our epidemic model with missing infection times can be constructed as follows. We focus on non-centering of the infectious period $D_i = N_i - I_i$. Subscripts will be omitted for clarity, but it is understood that this construction can be carried out for some or all infection events.

We introduce U, *apriori* independent of $\boldsymbol{\alpha}$, such that $D \overset{D}{=} \phi(U;\boldsymbol{\alpha})$. Thus, for $1 \leq i \leq n_I$, we can reparameterise the model in terms of $\mathbf{U}$ rather than $\mathbf{I}$ with

$$I_i = N_i - D_i = N_i - \phi(U_i, \boldsymbol{\alpha}) \tag{8}$$

For this parameterization, updating $\boldsymbol{\alpha}$ conditional on $\mathbf{U}$ simultaneously results in updating $\mathbf{I}$. $\mathbf{I}$ is also updated by changing $\mathbf{U} = (U_1, U_2, \ldots, U_{n_I})$ for fixed $\boldsymbol{\alpha}$.

It turns out to be important to use a compromise between centering and non-centering, resulting in so-called *partial non-centered methods* (see Papaspiliopoulos et al. (2003); Neal and Roberts (2005); Kypraios (2007)). There are a number of ways of constructing these methods. The approach we adopt here chooses, at each iteration at random, a collection of infectious periods to be centered, and others non-centered. Let μ denote the probability that an infectious period is non-centered. For $1 \leq i \leq n_I$, let

$$Z_i = \begin{cases} 1 & \text{with probability } \mu, \\ 0 & \text{with probability } 1-\mu. \end{cases} \tag{9}$$

Set $\mathcal{A} = \{i : Z_i = 1\}$ (non-centered individuals) and $\mathcal{B} = \{i : Z_i = 0\}$ (centered individuals) with for $i \in \mathcal{A}$, $I_i = \phi(U_i, \boldsymbol{\alpha})$. We have the following partially non-centered MCMC algorithm:

```
1. Update $\mathbf{Z}$ and hence, $\mathcal{A}$ and $\mathcal{B}$ using (9).
2. For each $i$, update $\theta_i|\boldsymbol{\theta}_{i-}, \boldsymbol{\alpha}, \mathbf{U}^{\mathcal{A}}, \mathbf{I}^{\mathcal{B}}, \mathbf{N}, \mathbf{R}$.
3. For each $i$, update $\alpha_i(\mathbf{I}^{\mathcal{A}})|\boldsymbol{\alpha}_{i-}, \boldsymbol{\theta}, \mathbf{U}^{\mathcal{A}}, \mathbf{I}^{\mathcal{B}}, \mathbf{N}, \mathbf{R}$.
4. For each $i$, update $I_i = \phi(U_i, \boldsymbol{\alpha})|\boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{I}_{i-}, \mathbf{N}, \mathbf{R}$.
```

2.6 Unknown number of infections

The above approach is applicable if n_I is known and all the notification times are known. For example, analysing a past epidemic such as the 2001 FMD outbreak. However, if the epidemic is in progress, all that is known is that $n_I = n_N + m$ where n_N denotes the total number of notified premises and $0 \leq m \leq N - n_N$. Therefore we include n_I as a parameter in the model and have to incorporate MCMC moves for the addition/deletion of infection times. We call premises which are infected but not notified by time T_{obs}, *occult* premises. The MCMC algorithm is therefore modified to the following:

```
1. For each $i$, update $\theta_i|\boldsymbol{\alpha}, \boldsymbol{\theta}_{i-}, \mathbf{I}, \mathbf{N}, \mathbf{R}$.
2. For each $i$, update $\alpha_i|\boldsymbol{\theta}, \boldsymbol{\alpha}_{i-}, \mathbf{I}, \mathbf{N}, \mathbf{R}$.
3. For each $i$, update $I_i|\boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{I}_{i-}, \mathbf{N}, \mathbf{R}$.
4. Propose an occult infection: $\boldsymbol{I} \rightarrow \{\boldsymbol{I} + \boldsymbol{o}\}$
5. Propose to delete an occult infection: $\boldsymbol{I} \rightarrow \{\boldsymbol{I} - \boldsymbol{o}\}$
```

Note the above algorithm is based upon an extension of the centered algorithm but could just as easily be added to the partial non-centered algorithm described in section 2.5. The last two steps involve changing the dimension of $\boldsymbol{I}$, with a corresponding change of dimension of $\boldsymbol{S}$, such that whenever we add an infection, $[\boldsymbol{I}]$ increases by one, and $[\boldsymbol{S}]$ decreases by one. The reverse is true for deletion, while for moving an infection time, both $[\boldsymbol{I}]$ and $[\boldsymbol{S}]$ remain constant. These, therefore, constitute reversible jump moves the proposal of which is implementation-specific. This will be discussed further in section 3.2, where, since the infectious period distribution is assumed to be known, a partially non-centered algorithm is not applicable.

3 Motivating data examples

There are many plausible statistical approaches to analysing epidemic data, many involving aggregation of the data to simplify the analysis (see for example, Andersson and Britton 2000; Ferguson et al. 2001a,b; Keeling et al. 2001; Diggle 2006; Scheel et al. 2007). The merits of any methodology are necessarily specific to the application at hand. Here we argue that a population level approach is often appropriate since it offers the opportunity to model detailed population structure which can often explain much of the stochasticity in the epidemic.

In our two examples, the 'population' consists of a collection of animal-keeping premises. Of course in both cases, a finer population model is in theory possible by considering individual animals. Clearly such an approach is inappropriate in our cases due to the unavailability of data at this fine scale. Moreover, within-farm epidemics take place on much quicker timescales than between-farm epidemics so little would be gained by working on a within-farm level.

Clearly a likelihood-based population-level approach will be infeasible when the population is extremely large. For populations of 80000 or more, serious computational and algorithmic complications need to be overcome. One possibility in this case is to use diffusion approximations to epidemic trajectories for which inference can be carried out in a fully Bayesian framework.

Whilst we shall not use informative priors in either of our examples, both offer opportunities for the incorporation of substantial expert opinion from professionals in the appropriate agriculture sector. So it is beneficial to be able to work in a fully Bayesian framework.

3.1 2001 UK Foot and Mouth Disease Outbreak

Background

Foot and Mouth Disease (FMD) is caused by a highly transmissible *Aphthovirus* affecting cloven-hoofed animals. Although FMD is rarely lethal to adult livestock, it causes vesicular lesions on the mouth and feet and often leads to a significant drop in milk production in dairy cattle and very slow weight gain in other livestock (Alexandersen et al. 2003).

Great Britain experienced a severe FMD epidemic in 2001 which lasted 7 months and led to the slaughter of around 6 million animals. Following the outbreak, the country was, for some time, restricted in its participation in the international trade of live animals as well as other products that could transmit the FMD virus. According to the UK National Audit Office (2002), the direct cost to the public sector was estimated at over £3 billion, and to the private sector at over £5 billion.

In order to analyse this epidemic, we use publicly available data on the epidemic from the UK Government Department of Environment, Food, and Rural Affairs (DE-

FRA)[1]. The data contain a variety of premises-level information for each of the 2026 infected farms in the UK (see Figure 1). Information on the non-infected premises was also provided by DEFRA from the June 2000 Agricultural and Horticultural Census. Approximately $134,000$ livestock farms were included in this dataset.

Cumbria in the north-west of England and Devon in the south-west were the two most severely affected counties in the country: we restrict our attention to Cumbria. This enables us to demonstrate non-centered parameterization with a manageable dataset, leaving our large-dataset methodology for the HPAI example. After cleaning, we obtain premises-level information on the geographical location and the number of animals of two different main species (cattle or/and sheep) of (initially susceptible) 5378 farms within Cumbria. At the end of the epidemic 1021 farms had been infected, and for these the slaughter date is also available. We should note that during the epidemic some farms (eg. dangerous contacts (DC), contiguous premises (CP)) were culled without knowledge of their infection status, being part of DEFRA's control policies in order to prevent the spread of the epidemic. Information on these premises (eg CPs and DCs) was not available to us.

Transmission Model

We define the component functions to build an epidemic model to capture the dynamics of the 2001 UK Foot and Mouth disease outbreak. The available data provide us with a point location for each premises and the total number of animals of each species on the premises. Taking into account that the 2001 Foot and Mouth disease was primarily confined to cows and sheep and following Keeling et al. (2001) we include in the transmission mechanism these two species only. Denote by n_i^c and n_i^s the number of cattle and sheep respectively for a given premises i. The rate at which an infectious premises i makes infectious contacts with a given premises j is modeled as follows:

$$\begin{aligned}\beta_{ij} &= \delta q(i;\zeta,\chi)s(j;\xi,\chi)K(i,j;\psi) \\ &= (\zeta\cdot(n_i^c)^\chi + (n_i^s)^\chi)\cdot\left(\xi\cdot(n_j^c)^\chi + (n_j^s)^\chi\right)\cdot K(i,j;\psi),\end{aligned} \tag{10}$$

where δ denotes the baseline infection rate between sheep and assumed to be multiplicative in this model. Therefore, the parameter ϵ which appears in Equation (5) is set equal to zero. The parameters ζ and ξ represent the relative infectiousness and susceptibility, respectively, of cattle to sheep. Note that given the available data information, the vectors of covariates which describe the business relationships between farms i and j, $\boldsymbol{r}_{ij}$ and $\boldsymbol{c}_{ij}$ are set to null, whilst following Keeling et al. (2001) and Diggle (2006), we choose $h(\cdot)$ to be a step-function, *ie.* $h(s) = 1\ (s > 0)$. A step-function $h(\cdot)$ is chosen primarily for its simplicity but is reasonably realistic given the relative rapid appearance of FMD symptoms.

[1] 'DataForModellersOct03.xls, available from (Defra 2007a)

Specific modeling details

An important issue regarding the transmission kernel is whether or not the Euclidean metric is the most appropriate distance measure, especially when an outbreak takes place in a geographical area with rich landscape such as hills, mountains and lakes. Regarding the FMD outbreak in the UK, Savill et al. (2006) showed that the Euclidean distance metric between infectious and susceptible premises is a better predictor of transmission risk than the shortest and quickest routes via road, except where major geographical features intervene. Therefore, they concluded that a simple spatial transmission kernel based on Euclidean distance suffices in most regions, probably reflecting the multiplicity of potential transmission routes during the epidemic.

Therefore, due to the lack of geographical information on the landscape of Cumbria, the difficulty of obtaining metrics such as minimum walking distances and taking into account the results presented in Savill et al. (2006) we adopt the Euclidean metric. For reasons of robustness, it is prudent to adopt a heavy-tailed transmission kernel. Therefore if $\rho(i, j)$ denote the Euclidean distance between farms i and j, the environmental spread is modeled by taking a Cauchy-type kernel

$$K(\rho_{i,j}, \psi) = \frac{\psi}{\rho_{ij}^2 + \psi^2}.$$

We adopt a simpler version of the full SINR model assuming that each farm i at any given time point can be either susceptible, infected or removed. Thus the model can be seen as a heterogeneously mixing stochastic SIR model. In order to illustrate the performance of the non-centering methodology in such a context, we assume that $D \sim Ga(a, b)$ with $a = 4$, fixed, and b an unknown parameter to be estimated. Note that this leads to a bell-shaped distribution where the mean (or the mode) and the variance depends only on b, see Kypraios (2007) for more details. Whilst, the choice of $a = 4$ is somewhat arbitrary, it is supported by the model fit analysis performed, see section 3.1 below.

MCMC Algorithm

We wish to make inference for $(\boldsymbol{\theta}, b)$ where $\boldsymbol{\theta} = (\delta, \zeta, \xi, \chi, \psi)$, so we are assuming that the infection times of each of the infected premises are considered to be unknown. Following Subsection 2.5 we adopt a partially non-centered approach. We reparameterise D_i as $D_i \overset{D}{=} b \cdot U_i$ where $U \sim Ga(a, 1)$ with U and b *a priori* independent. Furthermore, since we are making inference on b, we use a Gamma prior. We therefore implement the following MCMC algorithm:

1. `Choose` i `uniformly at random and update` $I_i | \boldsymbol{I}_{i-}, b, \boldsymbol{\theta}, \boldsymbol{R}$.
2. `Update` $b | \boldsymbol{I}, \boldsymbol{\theta}, \boldsymbol{R}$.
3. `Update` $\boldsymbol{Z}$ `and hence,` $\mathcal{A}$ `and` $\mathcal{B}$ `using (9).`

```
4. Update $b|\boldsymbol{\theta}, \mathbf{U}^{\mathcal{A}}, \mathbf{I}^{\mathcal{B}}, \mathbf{R}$ using the partially non-centered algorithm.
   (Note that for $i \in \mathcal{A}$, $I_i$ is also updated).

5. Update $\boldsymbol{\theta}|b, \mathbf{I}, \mathbf{R}$.
```

Note that the above algorithm is the partially non-centered algorithm of section 2.5 with the inclusion of a draw of b from its conditional distribution,

$$\pi\left(b|\boldsymbol{I}, \boldsymbol{R}, a\right) \sim Ga\left(an_I + \lambda_b, \sum_{i=1}^{n_I}(R_i - I_i) + \nu_b\right), \tag{11}$$

step 2. Step 2 is included because it improves the mixing of the MCMC algorithm for minimal computational cost. We chose to non-center 25% of the infection times (i.e. $\mu = 0.25$) at each iteration as this was found to produce an efficient algorithm (step 3) according to a pilot study which was carried out to seek for the "optimal" choice of μ. Step 4 is performed using a Metropolis-Hastings algorithm with b proposed from (11). Finally, step 1 is repeated a number of times in each iteration to improve the mixing of the algorithm. The model parameters $\boldsymbol{\theta}$ are updated in block, step 5, using a multiplicative random walk Metropolis algorithm.

An Illustrative Example

The following example is taken from Kypraios (2007) and the interested reader is referred there for more details. A dataset has been simulated consisting of $N = 500$ initially susceptibles and one initially infective individual uniformly located in a square $[0, 1] \times [0, 1]$. A distance-dependent infection rate and a Gamma infectious period with known shape parameter have been considered. Assuming that we only observed the removal times of each individual, a centered and a partially non-centered algorithm (similar to the one described in the previous section 3.1) were implemented such that to obtain samples from the posterior distribution of the scale parameter of the Gamma distribution. Figure 2 shows that the non-centered algorithm performs significantly better than the centered algorithm.

Results

In this section we present a Bayesian analysis of the 2001 FMD outbreak in Cumbria by using the model described in section 3.1, using a partially non-centered approach as explained in 3.1. Non-informative priors were chosen for all the parameters.

In Figure 3, we present the posterior distribution of b, the scale parameter for the infectious period distribution. Directly as a result of this we give the posterior mean infectious period. It agrees with the assumptions made in Keeling et al. (2001), although in Keeling et al. (2001) they considered an SEIR-type model where the exposed and the infectious periods are assumed to be known and fixed without any variability or differences between premises.

Figure 4 depicts the posterior distribution of the spatial transmission kernel. Posterior uncertainty is illustrated through the superposition of kernels drawn from the posterior. The kernel is shown on a log scale (see Figure 5) and the modal value agrees well with other literature (see for example, Keeling et al. 2001). An interesting feature of the kernel is the fact that there appears to be little infectious pressure exerted over a distance of more than about 4km, a result which is also in association with other published work (see for example, Keeling et al. 2001; Deardon et al. 2007; Diggle 2006).

Having obtained the marginal distributions of the relative infectiousness and susceptibility of cattle to sheep (see Figures 6) we can infer that each individual cow was more likely to transmit the disease and also likely to be more susceptible to the disease than each individual sheep. Moreover, there is strong evidence for a non-linear effect of the number of different species in each farm (see Figure 7). These results are qualitatively similar to those reported in Keeling et al. (2001), Diggle (2006) and Deardon et al. (2007) although the former only looks at the case where $\chi = 1$.

For illustrative purposes we turn our attention to the infectivity $q(i; \zeta, \chi)$ and the susceptibility $s(i; \xi, \chi)$ for farms of typical size. An average medium-size farm which has only cattle, an average medium-size farm which has only sheep as well as an average large- and small-size farm with cattle and sheep have been considered (see Table 1 and Table 4). An interesting result from these tables is that a typical medium-size sheep farm is more infectious than a typical medium-size farm with cattle (only). In addition, a medium-size cattle and a medium-size sheep farm are both similarly susceptible to infection. Furthermore, both tables reveal that there is a significant risk to susceptible farms from small holdings with a handful of animals. Therefore, we can conclude that all animal holdings, however small, play an important role in the spread of the disease.

Model fit

It is important to assess the appropriateness of the model proposed to the epidemic data. For practical purposes, measures of model fit based on predictive accuracy are important in particular applications. Implementation of these measures are easily applied and are widely used for epidemics (see for example Lekone and Finkenstädt (2006); Cauchemez and Ferguson (2008)).

Instead, we focus on using non-centered residuals to assess model fit (Papaspiliopoulos 2003). The suitability of the infectious period distribution is straightforward and quick to check by examining the distribution of the non-centred variables U_i. If the model fits well, then the population of U_i's ranging over premises and MCMC iterations should be approximately Gamma$(4, 1)$ distributed, as specified in Section 3.1. Figure 8 shows a good agreement between these two distributions. In addition, to check our assumption of a fixed $a = 4$, we ran the algorithm with a as an unknown parameter. The posterior mode of a was 3.76, with the fixed value well within the support.

3.2 High Pathogenicity Avian Influenza

Background

In 1996, High Pathogenicity Avian Influenza H5N1 (HPAI) emerged as a disease of geese in southern China and began to spread throughout south-east Asia (Xu et al. 1999; Claas et al. 1998). In 2005, outbreaks were recorded in eastern Europe with multiple cases in wild birds and a large outbreak in domestic poultry in Romania. Sporadic cases have since appeared further west in France, Denmark, Germany, Sweden, and more recently the UK (European Centre for Disease Surveillance 2006).

The poultry industry is a large economic force in the UK, with 1.5 million tonnes of meat produced in 2004 representing 40% of the primary meat market. The possibility of an HPAI epidemic therefore poses a significant economic risk (Defra 2007b). In addition, it has been shown that a significant human health risk exists in that people with high exposure to the virus have been infected with high mortality (Claas et al. 1998; WHO 2007). Therefore the implications of an outbreak of HPAI are very serious, and hence, its classification as a notifiable disease which means that any outbreak in the UK must be eliminated.

For the Poultry Industry, covariate data was obtained from DEFRA in the form of the Great Britain Poultry Register (GBPR) and Poultry Network Data (PND). After cleaning, these data give premises-level information on the geographical location, production type, and commercial contacts of 8636 poultry premises within Great Britain. In our analysis only production-level premises are included since the breeding sector operates extremely high biosecurity and is considered to present negligible risk in the event of an outbreak. This assumption is further strengthened by the industry supposition that all live-bird movement (apart from transport to slaughterhouses) would immediately cease in the event of an unfolding epidemic[2]. Furthermore, since the production-level relies on a low-frequency "all-in-all-out" system with birds arriving from the breeding sector and going directly to the slaughterhouse upon departure, market-mediated transmission as seen for FMD in the cattle, sheep, and pig industries is not considered.

In comparison to the livestock industry, the UK poultry industry is highly structured. Besides spatial proximity, potential infectious contact within the industry is governed by production type and managemental networks. Firstly, the GBPR currently characterizes 11 production types (Table 3) all of which are relatively isolated from each other, and may have differential susceptibilities to disease. Secondly, large integrated production companies, representing about 60% of the broiler premises, have only limited contact with the independent producers that make up the remainder of the industry. Premises belonging to the same company may, due to the sharing of equipment and personnel, transmit disease between themselves, but be relatively isolated from other companies and the independent sector. We have also identified that transmission of disease may occur via third-party feed-mills and slaughterhouses (Figure 9). From the PND, it is possible to estimate the frequencies of feed mill and slaughterhouse contacts, and of the number of farms each feed mill or slaughterhouse services. Finally, there is the

[2] Confirmed at a meeting with poultry industry representatives, 01/03/2007

possibility of disease transmission by other networks such as egg collection and catching teams (employed to "thin" populations of housed birds prior to final depopulation). However, scant data is available for such networks and consultation with the poultry industry indicates that these networks operate on a local scale and may, therefore, be modeled spatially. Since there has been no major outbreak of HPAI in the UK thus far, a simulated outbreak in the Great Britain poultry industry (Figure 10) is used to show how our statistical methodology can be applied to perform inference on such an epidemic in real-time, and stochastic simulation used over the results to provide a quantitative risk analysis.

Transmission model

We now define our component functions to provide an inference mechanism for HPAI. The covariate data give information about production type ($s_{pt}(j)$), distance between premises ($\rho(i,j)$), the company affiliation (if any) ($\boldsymbol{C}^{CP}$), and the rate of lorry contacts for feed mill ($\boldsymbol{R}^{FM}$) and slaughterhouse ($\boldsymbol{R}^{SH}$) networks. For the poultry industry, it is not clear how flock size might affect either premises level infectivity or susceptibility. Without any further information, it would be impossible to identify infectivity from susceptibility; we therefore use only production-type susceptibility on the susceptible premises, $s_{pt}(j)$. This represents categorical data and is therefore modeled as a multiplicative factor with 11 levels according to the production types in the dataset (Table 3). We choose susceptibility to be a property of the major production-type on each premises j relative to broilers, such that for broilers the corresponding parameter is set to 1. Thus we attempt to identify industry sectors that might be at higher risk during an outbreak. For this data, in contrast to FMD, we use the full SINR model to give:

$$\beta_{ij} = s_{pt}(j)\left(p_1 r_{ij}^{FM} + p_2 r_{ij}^{SH} + \beta_1 c_{ij}^{CP} + \beta_2 e^{-\delta\rho(i,j)}\right) \tag{12}$$

and

$$\beta_{ij}^{\star} = s_{pt}(j)\left(\gamma\beta_2 e^{-\delta\rho(i,j)}\right). \tag{13}$$

The distance kernel, $K(i,j;\psi)$ is taken to be exponential in this case. We therefore perform inference on $\boldsymbol{\theta} = \{\boldsymbol{s}_{pt}, p_1, p_2, \beta_1, \beta_2, \gamma, \delta\}$, where $\boldsymbol{s}_{pt}$ is the vector of production-type susceptibilities.

Infectious Period

For the infectious period, we use a Gumbel distribution with pdf

$$f_D(x) = abe^{bx - a\left(e^{bx}-1\right)} \tag{14}$$

and

$$F_D(x) = e^{-a\left(e^{bx}-1\right)}. \tag{15}$$

This allows flexibility of modeling, and interpretations of a and b as the rate of detection, and rate of development of clinical signs throughout the flock, respectively. These parameters are estimated from literature and expert opinion and consequently for the analysis a and b are assumed to be known, fixed constants.

The farm-level infectivity function, $h(\cdot)$, is specified as:

$$h(s) = \frac{e^{\nu s}}{\mu + e^{\nu s}}, \tag{16}$$

satisfying the conditions in Section 2.2 and which is easily integrated to

$$H(t) = \int_0^t h(s)ds = \frac{1}{\nu}\log\left(\frac{\mu + e^{\nu t}}{\mu + 1}\right). \tag{17}$$

MCMC Algorithm

For HPAI, we require an MCMC algorithm that is not only robust to the problem, but that is also fast enough to support the real-time aspect of the inference. We use the MCMC algorithm specified in section 2.6 with certain implementation-specific modifications. Step 1 uses a multisite update of $\boldsymbol{\theta}$. We use multiplicative random walk for this 95% of the time, and additive random walk 5% of the time to allow components of $\boldsymbol{\theta}$ to "escape" from small values. Step 2 is omitted since we are not making inference on the parameters of the infectious period distribution. However, the remaining steps are worthy of further explanation.

Firstly, steps $3, 4$ and 5 are not executed sequentially in a deterministic fashion. Rather, we choose each step with equal probability a number of times per iteration of the MCMC. In the following descriptions, we denote a move by $\boldsymbol{I} - t + s$, an addition by $\boldsymbol{I} + s$, and a deletion by $\boldsymbol{I} - t$.

Step 3 **Move an infection time**: We choose an infection time to move from a discrete uniform distribution, $\mathrm{U}[1, [\boldsymbol{I}]]$, and propose a replacement infection time drawn from the distribution D defined either in (14) if the notification time is known, or in (19) if not. We accept the proposed value with probability

$$1 \wedge \frac{L(\boldsymbol{I} - \{t\} + \{s\} | \boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})}{L(\boldsymbol{I} | \boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})} \times Q,$$

where

$$Q = \begin{cases} \frac{f_D(N_t - I_t)}{f_D(N_t - I_s)} & \text{infection known} \\ \\ \frac{F_D(T_{obs} - I_t)}{F_D(T_{obs} - I_s)} & \text{infection occult.} \end{cases}$$

Step 4 **Add an infection**: We choose an occult infection from the susceptibles using a discrete uniform distribution U$[1, [\boldsymbol{S}]]$. An infection time is then drawn from the distribution in Equation 19. We accept such an addition with probability:

$$1 \wedge \frac{L(\boldsymbol{I} + \{s\}|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})}{L(\boldsymbol{I}|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})} \times \frac{[\boldsymbol{S}]}{(m+1) \cdot \tilde{g}(T_{obs} - I_s)},$$

where m is the number of previously added infections prior to the addition and $T_{obs} - I_s$ is sampled from a non-negative pdf $\tilde{g}(\cdot)$. The choice of $\tilde{g}(\cdot)$ is discussed below.

Step 5 **Delete an infection**. We choose an infection time to delete from a discrete uniform distribution over the premises that have been previously added. We accept such a move with probability:

$$1 \wedge \frac{L(\boldsymbol{I} - \{t\}|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})}{L(\boldsymbol{I}|\boldsymbol{N}, \boldsymbol{R}, \boldsymbol{\theta})} \times \frac{m \cdot \tilde{g}(T_{obs} - I_t)}{[\boldsymbol{S}] + 1}$$

where m is the number of previously added infections prior to the deletion.

In order to efficiently propose occult infection times for the MCMC, we would like to sample from:

$$\tilde{g}(x) \propto F_D(x) \qquad (x \geq 0) \tag{18}$$

Due to the specification of (15) it is non-trivial to simulate directly from a distribution with pdf proportional to $F_D(\cdot)$. For algorithmic speed we prefer to avoid rejection sampling in favor of a truncated Normal approximation, obtained by a 2nd order Taylor Series expansion of (15), giving:

$$T - I_i \sim \mathrm{N}\left(-\frac{1}{b}, \frac{1}{ab^2}\right), \qquad T - I_i \geq 0 \tag{19}$$

Simulation studies showed that such an independence sampler achieves an acceptance probability greater than 0.5.

Adaptive MCMC

One of the barriers to implementing such an algorithm for real-time inference is the time needed to tune random-walk proposal densities. Multisite updating of the transmission parameters ($\boldsymbol{\beta}$) was chosen to minimise the time-consuming need of fully recalculating the likelihood, but manually adjusting the order-16 proposal variance-covariance matrix would, conversely, be a very long process. To relieve this issue, we use the adaptive proposal scheme of Haario et al. (2001). This is implemented following work by Roberts and Rosenthal (2007), and Andrieu and Moulines (2006) on theory to guarantee ergodicity of the resulting Markov chain, and allows the proposal density to adapt for optimal scaling as the chain converges. The proposal density, therefore, is:

$$Q_n(x, \cdot) = (1 - \xi)N(x, (2.38)^2\boldsymbol{\Sigma}_d/d) + \xi N(x, (0.1)^2\boldsymbol{I}_d/d), \tag{20}$$

where Σ_d is the d-dimensional empirical variance-covariance matrix of the current posterior density with $d = 16$ and ξ is a small positive constant which following Roberts and Rosenthal (2007) we take $\xi = 0.05$.

Parallel Computing

Finally, to speed up the calculation of the likelihood, domain-decomposition parallelization of the sums in (7) was achieved using a shared-memory architecture with an implementation of the OpenMP standard (Dagum and Menon 1998). This was chosen over a distributed architecture since the dependent nature of the epidemic data requires high levels of inter-process communication; the high-bandwidth busses on a multiprocessor mainboard being many times faster than network interconnects. We were able to achieve a 10-fold speedup in algorithm runtime on a 8 dual-core Sun X4600 server running Linux, giving a final runtime of 4.5 hours for 100000 iterations with this dataset.

Results

In order to demonstrate our methods, we use a simulated Avian Influenza epidemic in the UK poultry industry. We start by using a stochastic simulation on the dataset of 8636 premises, creating an epidemic that lasts for 77 days and infects a total of 375 premises. We then choose to observe this epidemic at three time points: 14, 25, and 50 days after the first notification (Table 4, Figure 10). At each observation time, we use the data available to perform risk-prediction, and show how this is refined as the epidemic progresses.

We begin with posterior parameter distributions, and Figure 11 plots the prior and posteriors for β_2 at three observation times. This illustrates how posterior uncertainty and prior dependence recedes during the course of the epidemic.

We also consider the Bayes predictive probability (risk) that a given farm becomes infected during the current epidemic. This is estimated by off-line forward model simulation under parameter values drawn randomly from the posterior distribution of the parameters at the current time. This is an example of a dynamic quantity which varies as the epidemic evolves (as well as due to the changing parameter posterior distributions). These are shown pictorially in Figure 12, and demonstrate how the risk estimate progresses with time and amount of information available.

For control purposes, it may be of interest to know which premises would present a high risk to the remaining susceptibles if they themselves were to be infected. We define R_i to be the expected number of further premises a premises, i, would infect were it to be the index infection in a hypothetical infection where all other farms started as susceptibles, and conditional on all parameters. R_i plays the role of a farm-specific basic reproduction number. For each premises we calculate the posterior probability: $P(R_i > 1)$.

However $P(R_i > 1)$ does not discriminate between farms which infect "large R_i farms" and those which do not. Thus within our very heterogenous population, $P(R_i > 1)$ does not necessarily indicate the degree of risk posed by a premises to the population. Instead, we prefer to take a bootstrap expectation of the epidemic size that would result if each premises were to be the index case for the epidemic. To illustrate this, we have taken two premises from our dataset with differing $P(R_i > 1)$, and have run stochastic simulations over the joint posterior for the transmission and removal rates (Table 5).

In terms of active disease control surveillance, a more efficient policy might be to target limited resources to suspect premises. Of primary concern is the number of occult infections present at any one time as a measure of how much control effort will be needed when these infections are detected. The reversible jump algorithm allows us to assign a probability of being infected to each apparently susceptible premises at each observation time, thus identifying high-occult probability locations. The probabilities can then be plotted in a similar fashion to Figure 12 (results not shown) or monitored in an appropriate way. Of statistical interest are the posteriors of the number of occults at each time point (Figure 13). Early in the epidemic, the algorithm overestimates the number of occults, consistent with the high degree of uncertainty in the parameter posteriors, and the influence of conservatively chosen priors. However, this effect disappears later in the epidemic.

4 Discussion

In this paper particular attention has been paid to the generality of the modelling framework, the computational and algorithmic efficiency of the MCMC algorithms particularly for incomplete epidemics where risk assessment crucially depends upon the possible presence of occult infections. We also allow for two different types of network: those which describe a non-specific connectivity (such as company contacts), and those which explicitly describe risk caused by particular activities (eg feed lorry deliveries, slaughter house visits, etc.). However, there are many important issues we have not considered in any detail.

We have not focused here on the evaluation of control strategies, although our methodology provides a natural framework for doing this for specific applications. From our on-line simulation study, two important issues emerge. Control strategies are most influential very early in the course of an epidemic. It is therefore important to be able to subsume all available expert information into prior distributions for parameters at the outset, so as to be well-informed about parameters early on to guide the evolution of control policy. Secondly, data from the early stages of an epidemic are often augmented by *contact tracing* information which can often identify disease transmission pathways, and hence add substantially to the available information on model parameters governing the transmission of infections. We are currently extending our methodology to incorporate this information into our likelihood-based framework.

More fundamentally is the issue of network data quality, particularly with respect to contact rate information. In our HPAI example, we use network data that is derived from a questionnaire sample of approximately 30% of the premises registered in the GBPR. From this, we have simulated contact rates on the feed mill and slaughter house networks for the remaining 70% of premises. This will inevitably lead to inaccuracies in our predictions when working at the individual level. An important extension of our methodology will, therefore, be to allow for uncertainty in the network specification and therefore reflect this in the prediction.

We have shown in our FMD example that use of non-centred variables for assessment of model fit is a convenient method for outbreak data. Building on this, model choice for epidemic models in a fully Bayesian framework could be easily performed using standard methodology (eg reversible jump MCMC or marginal likelihood), and part of our ongoing work is to address this.

In summary, we have introduced a unified modelling inference and prediction methodology for emerging infectious disease epidemics within a Bayesian framework. We have applied this to two important agricultural contexts, demonstrating complementary situations in which our approach can be applied. It is therefore anticipated that our methodology could be used for informing control strategy in future epidemics in a wide range of populations.

References

Alexandersen, S., Zhang, Z., Donaldson, A. I., and Garland, A. J. M. (2003). “The pathogenesis and diagnosis of foot-and-mouth disease.” *J. Comp. Pathol.*, 129(1): 1–36. 473

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*, volume 151 of *Lecture Notes in Statistics*. New York: Springer-Verlag. 473

Andrieu, C. and Moulines, É. (2006). “On the ergodicity properties of some adaptive MCMC algorithms.” *Annals of Applied Probability*, 16: 1462–1505. 482

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*. Hafner Press [Macmillan Publishing Co., Inc.] New York, second edition. 469

Becker, N. G. (1989). *Analysis of infectious disease data*. Monographs on Statistics and Applied Probability. London: Chapman & Hall. 466

Becker, N. G. and Britton, T. (1999). “Statistical studies of infectious disease incidence.” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(2): 287–307. 466

Cauchemez, S. and Ferguson, N. (2008). "Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London." *J. R. Soc. Interface*, 5: 885–897. 477

Claas, E., Osterhaus, A., van Beek, R., De Jong, J., Rimmelzwaan, G., Senne, D., Krauss, S., Shortridge, K., and Webster, R. (1998). "Human influenza A (H5N1) virus related to a highly pathogenic avian influenza virus." *The Lancet*, 351: 472. 478

Dagum, L. and Menon, R. (1998). "OpenMP: an Industry Standard API for Shared-memory Programming." *IEEE Comput. Sci. Eng.*, 5: 46–55. 482

Deardon, R., Brooks, S. P., Grenfell, B., Keeling, M. J., Tildesley, M. J. S. S. J., Shaw, D., and Woolhouse, M. E. J. (2007). "Inference for individual-level models of infectious diseases in large populations." *Submitted*. 477

Defra (2007a). "Defra Website." [Online; Accessed 24-04-2007].
URL `http://www.defra.gov.uk` 465, 474

— (2007b). "Eggs and poultry facts and statistics." [Online; Accessed 12-01-2007].
URL `http://www.defra.gov.uk/foodrin/poultry/statistics/index.htm` 478

Diggle, P. J. (2006). "Spatio-temporal point processes, partial likelihood, foot and mouth disease." *Stat. Methods Med. Res.*, 15(4): 325–336. 473, 474, 477

European Centre for Disease Surveillance (2006). "Weekly surveillance report." *Euro Surveill.*, 11(51): pii=3098.
URL `http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=3098`
478

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001a). "The Foot-and-Mouth Epidemic in Great Britain: Pattern of Spread and Impact of Interventions." *Science*, 292(5519): 1155–1161. 473

— (2001b). "Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain." *Nature*, 413: 542–547. 473

Gibson, G. (1997). "Markov chain Monte Carlo methods for ftting spatiotemporal stochastic models in plant epidemiology." *Applied Statistics*, 46(2): 215–233. 466

Gibson, G. and Renshaw, E. (1998). "Estimating parameters in stochastic compartmental models using Markov Chain methods." *IMA J. Math. Appl. Med. Biol.*, 15: 19–40. 466

Haario, H., Saksman, E., and Tamminen, J. (2001). "An adaptive metropolis algorithm." *Bernoulli*, 7(2): 223–242. 466, 481

Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J., and Grenfell, B. T. (2001). "Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape." *Science*, 294(5543): 813–818. 473, 474, 476, 477

Kypraios, T. (2007). "Efficient Bayesian Inference for Partially Observed Stochastic Epidemics and A New class of Semi−Parametric Time Series Models." Ph.D. thesis, Department of Mathematics and Statistics, Lancaster University, Lancaster. Available from `http://www.maths.nott.ac.uk/personal/tk/files/Kyp07.pdf`. 466, 471, 475, 476

Lekone, P. and Finkenstädt, B. (2006). "Statistical Inference in a Stochastic Epidemic SEIR Model with Control Intervention: Ebola as a Case Study." *Biometrics*, 62: 1170–1177. 477

Neal, P. and Roberts, G. (2005). "A case study in non-centering for data augmentation: stochastic epidemics." *Stat. Comput.*, 15(4): 315–327. 466, 471

O'Neill, P. D. and Roberts, G. O. (1999). "Bayesian inference for partially observed stochastic epidemics." *J. Roy. Statist. Soc. Ser. A*, 162: 121–129. 466

Papaspiliopoulos, O. (2003). "Non-centered parametrisations for hierarchical models and data augmentation." Ph.D. thesis, Department of Mathematics and Statistics, Lancaster University, Lancaster. 477

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). "Non-centered parameterizations for hierarchical models and data augmentation." In *Bayesian statistics, 7 (Tenerife, 2002)*, 307–326. New York: Oxford Univ. Press. Editors J. M. Bernardo and M. J. Bayarri and J. O. Berger and A. P. Dawid and D. Heckerman and A. F. M. Smith and M. West. 466, 471

Roberts, G. and Rosenthal, J. (2007). "Coupling and Ergodicity of adaptive Markov chain Monte Carlo algorithms." *Journal of Applied Probability*, 44: 458–475. 466, 481, 482

Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E., Brooks, S. P., and Grenfell, B. T. (2006). "Topographic determinants of foot and mouth disease transmission in the UK 2001 epidemic." *BMC Vet Res*, 2. Available at doi:10.1186/1746-6148-2-3. 475

Scheel, I., Aldrin, M., Frigessi, A., and Jansen, P. A. (2007). "A stochastic model for infectious salmon anemia (ISA) in Atlantic salmon farming." *J. R. Soc. Interface*, 4: 699–706. 473

UK National Audit Office (2002). "The 2001 outbreak of foot and mouth disease." Report by the Comptroller and auditor general, HC 939, Session 2001-2002, London: The Stationery Office. 473

WHO (2007). "Avian Influenza." [Online; Accessed 12-01-2007].
URL `http://www.who.int/csr/disease/avian_influenza/en/index.html` 478

Xu, X., Subbarao, K., Cox, N., and Guot, Y. (1999). "Genetic Characterization of the Pathogenic Influenza A/Goose/Guangdond/1/96 (H5N1) Virus: Similarity of Its Hemagglutinin Gene to Those of H5N1 Viruses fromthe 1997 Outbreaks in Hong Kong." *Virology*, 261: 15. 478

Acknowledgments

Thanks to Dr RM Christley and Dr SE Robinson, Department of Clinical Studies, Faculty of Veterinary Science, University of Liverpool; Streamline Computing Ltd for the loan of the Sun X4600 server, and Dr Mike Pacey for his help and advice with it; Dr Miles Thomas, DEFRA Central Scientific Laboratory, for providing the Foot-and-Mouth data; DEFRA for providing covariate information on the GB Poultry Industry.

Appendix

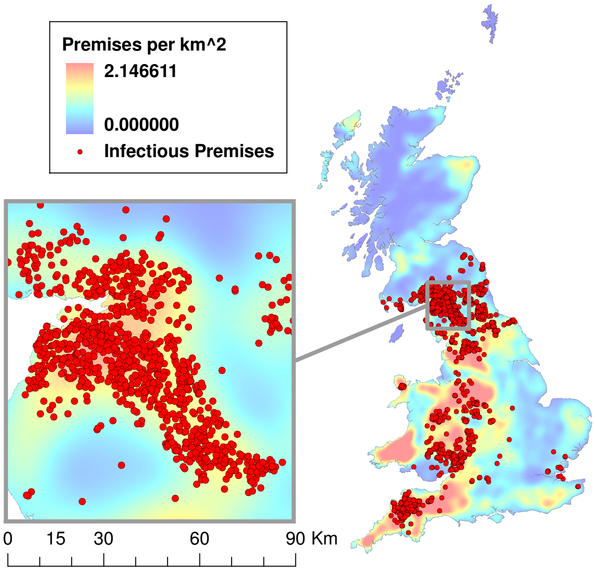


Figure 1: The spatial distribution of farms in Great Britain during the 2001 UK Foot and Mouth Disease Epidemic. Infected premises are superimposed (red dots). Inset, Cumbria region.

Cows	Sheep	2.5% Quantile	Median	97.5 % Quantile
100	0	3.17	6.63	14.05
0	1000	5.91	9.53	15.35
50	500	8.08	13.04	21.36
2	6	2.70	3.64	5.31

Table 1: Posterior quantiles of farm's infectivity

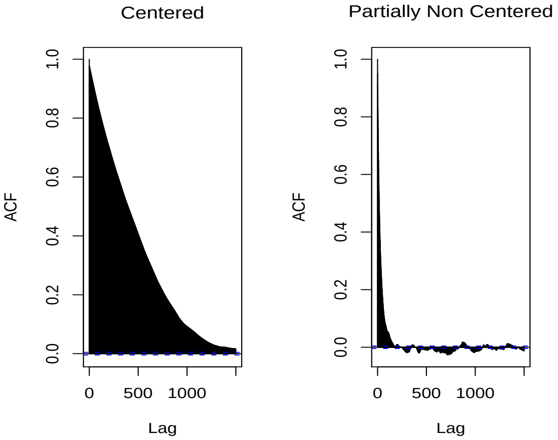


Figure 2: Comparison of ACFs of the scale parameter between the standard (centered) and the non-centered algorithm for simulated example.

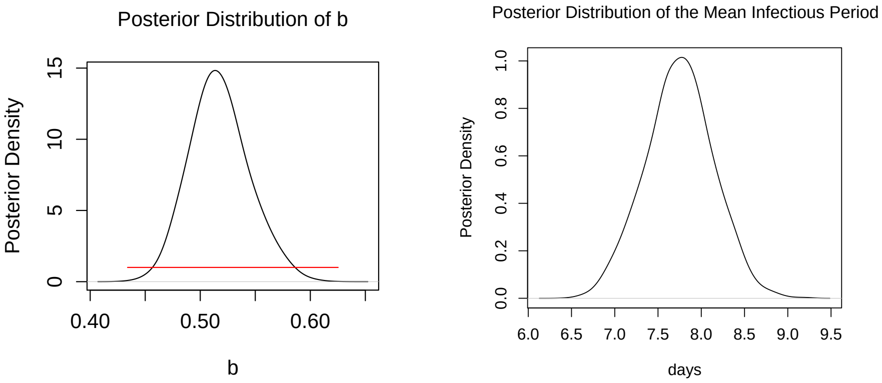


Figure 3: Posterior distribution of the parameter b (left) and the average infectious period of a single farm $4.0/b$ (right). The red/horizontal line indicates the prior chosen for b.

Cows	Sheep	2.5% Quantile	Median	97.5 % Quantile
100	0	6.10	10.43	17.58
0	1000	5.91	9.53	15.35
50	500	10.21	15.93	24.75
2	6	3.76	4.70	5.94

Table 2: Posterior quantiles of farm's susceptibility

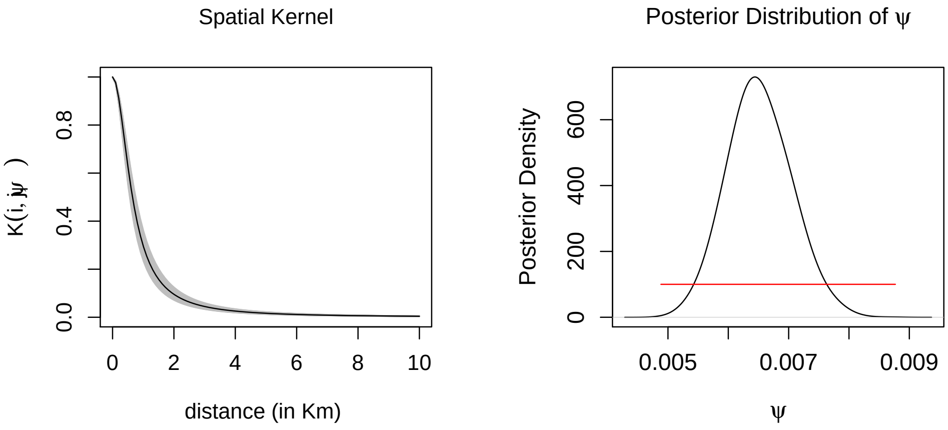


Figure 4: The marginal posterior distribution of ψ (right) and a 95% HPDR of $K(i, j; \psi)$ (left). The black/solid line of the latter refers to the modal shape of the kernel based on the posterior mode of parameter ψ.

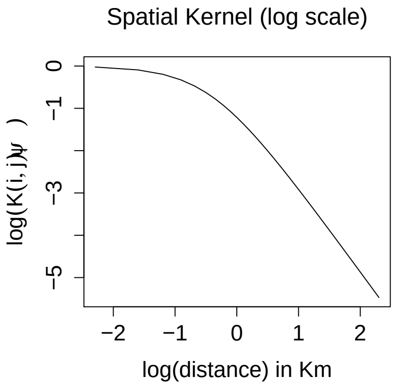


Figure 5: The spatial kernel $K(i, j; \psi)$ on log-scale drawn by using the modal value of the posterior distribution of ψ.

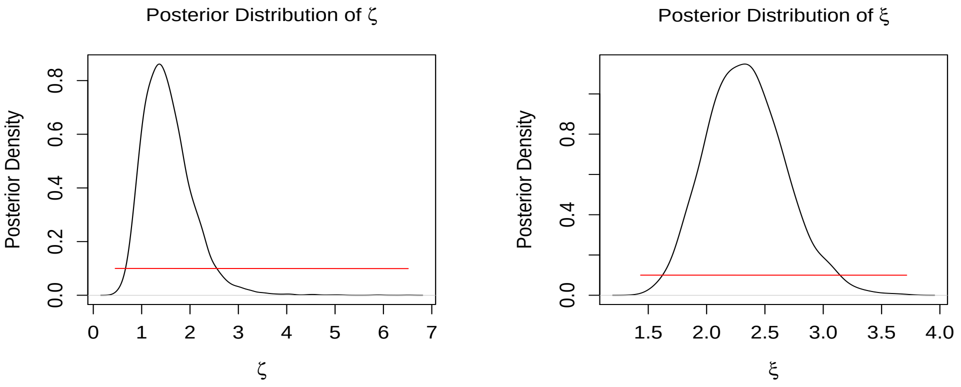


Figure 6: Posterior distribution of the parameters ζ (left) and ξ (right). The horizontal lines refer to the prior distributions.

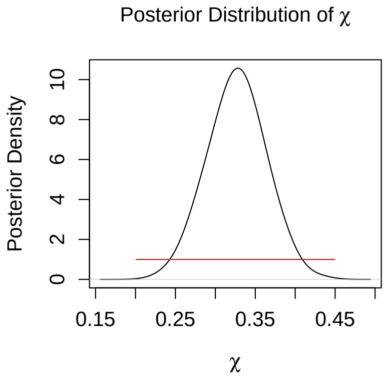


Figure 7: Posterior distribution of the parameters χ.

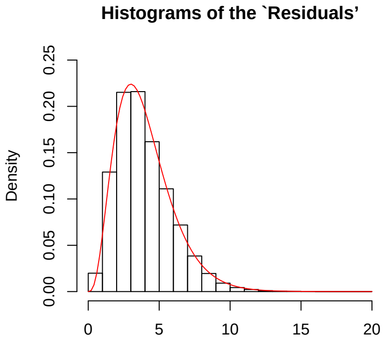


Figure 8: Posterior distribution of non-centered variables U_i for the FMD example. The red/solid line, showing a Gamma$(4, 1)$, indicates good model fit.

Production Type	Number of premises	% total dataset
Broilers	1416	16.4
Chicken Layers	3733	43.2
Turkey	668	7.7
Duck Meat	239	2.8
Duck Layers	85	1.0
Goose Meat	57	0.7
Goose Layers	17	0.2
Pheasant	2011	23.3
Partridge	376	4.4
Quail Layers	34	0.4
Total	8636	100

Table 3: Production-type distribution within the dataset.

Time / days	# notified infections	occult infections
14	10	15
25	61	13
50	290	40

Table 4: The state of the epidemic at each observation time

$\mathcal{P}(R_i > 1)$	$\mathcal{E}_{i,\boldsymbol{\beta},\gamma}$ [epidemic size]
0.01	3
0.95	184

Table 5: Expected epidemic size starting at premises with different R_is.

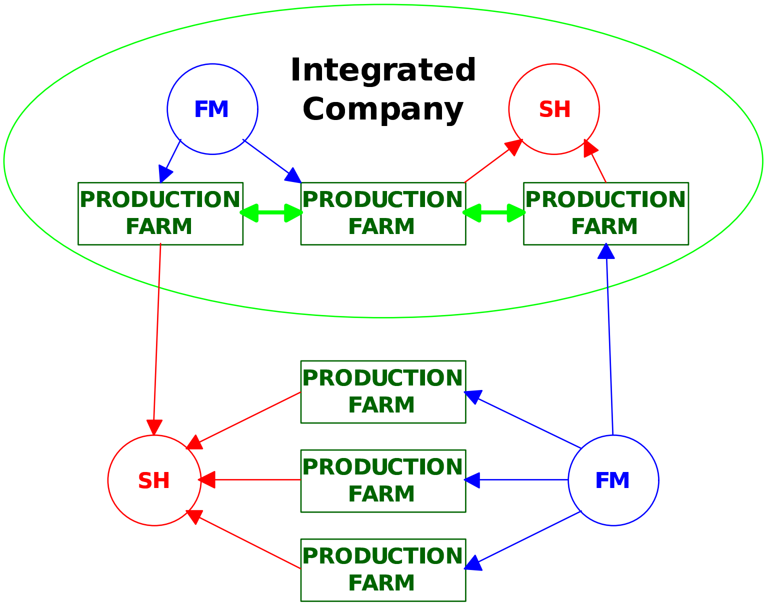


Figure 9: Simplified diagram of the three identified contact networks in the poultry industry: feed-mills (blue), slaughterhouses (red), company (green).

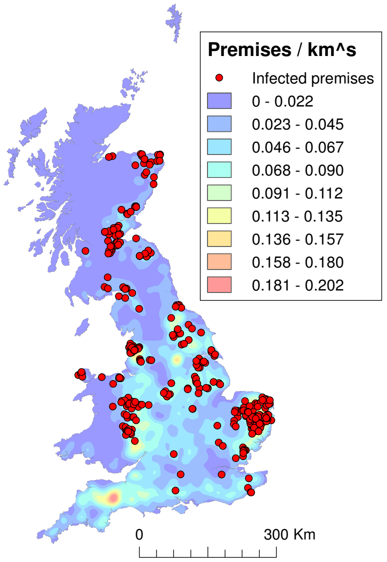


Figure 10: Simulated outbreak of HPAI in GB poultry showing infected farms, and premises density

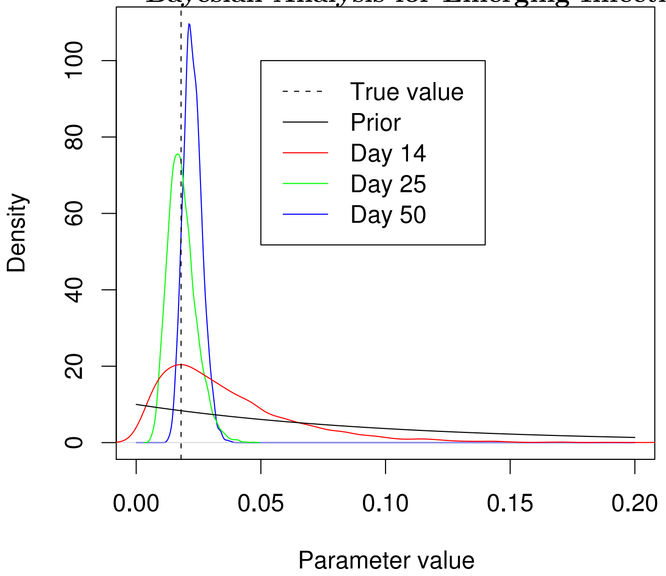


Figure 11: The prior and posteriors for β_2 at the three time points during the epidemic

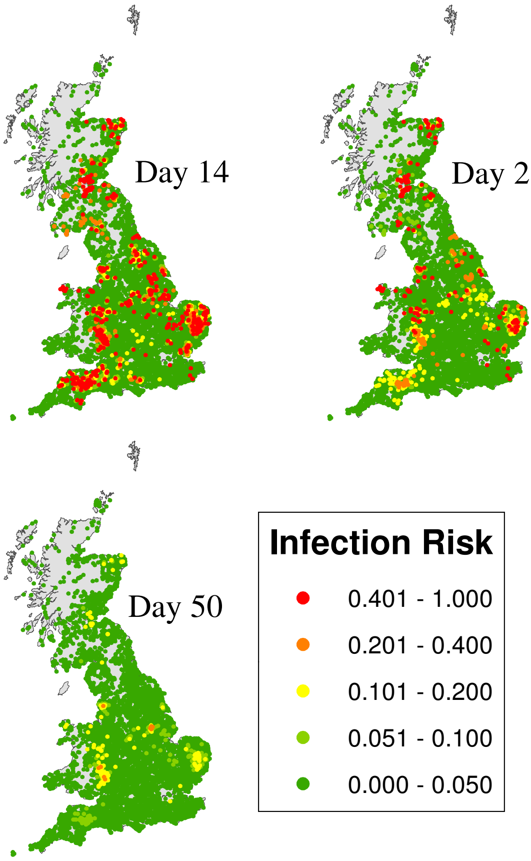
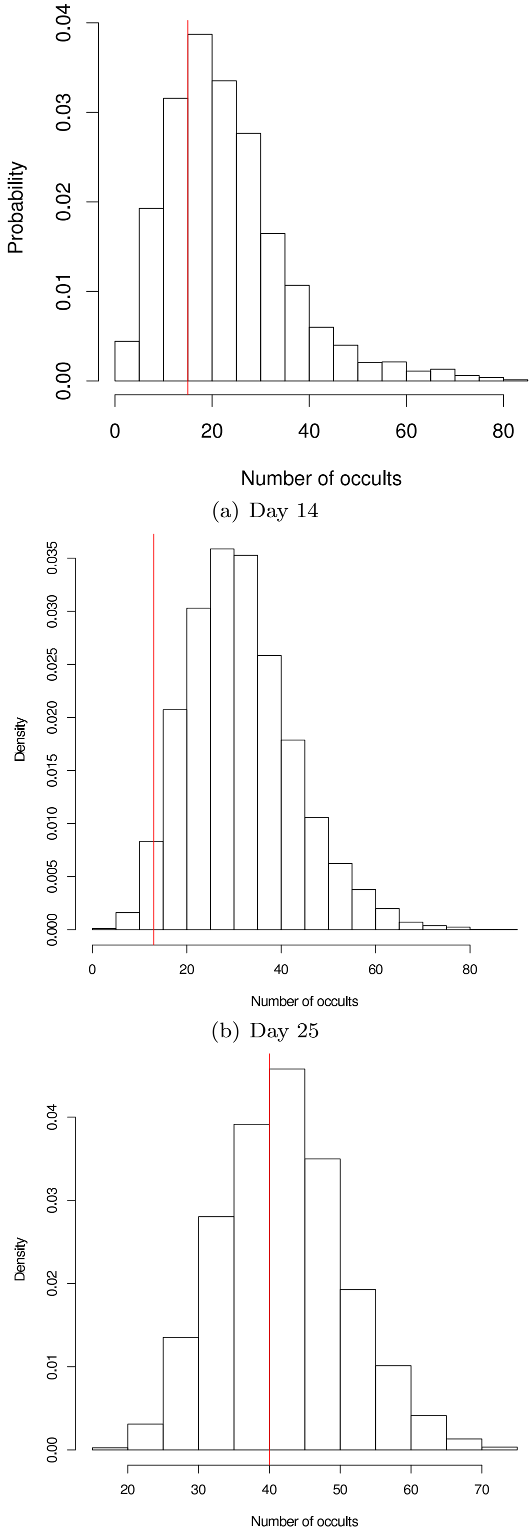


Figure 12: Spatial distribution of predicted risk to the population



(a) Day 14

(b) Day 25

(c) Day 50

Figure 13: Posterior distributions of the number of occult infections. The red/vertical line denotes the true number, determined from the simulated data (T=10:, T=25:, T=50).