Received: 8 September 2020 | Accepted: 18 May 2021

DOI: 10.1111/rssc.12515

ROYAL STATISTICAL SOCIETY
Series C Applied Statistics

ORIGINAL ARTICLE

A Bayesian nonparametric analysis of the 2003 outbreak of highly pathogenic avian influenza in the Netherlands

Rowland G. Seymour[1] | **Theodore Kypraios**[1] | **Philip D. O'Neill**[1] | **Thomas J. Hagenaars**[2]

[1]School of Mathematical Sciences, University of Nottingham, Nottingham, UK

[2]Wageningen Bioveterinary Research (WBVR), Lelystad, The Netherlands

Correspondence
Theodore Kypraios, School of Mathematical Sciences, University of Nottingham, Nottingham, UK.
Email: theodore.kypraios@nottingham.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: EP/N50970X/1

Abstract

Infectious diseases on farms pose both public and animal health risks, so understanding how they spread between farms is crucial for developing disease control strategies to prevent future outbreaks. We develop novel Bayesian nonparametric methodology to fit spatial stochastic transmission models in which the infection rate between any two farms is a function that depends on the distance between them, but without assuming a specified parametric form. Making nonparametric inference in this context is challenging since the likelihood function of the observed data is intractable because the underlying transmission process is unobserved. We adopt a fully Bayesian approach by assigning a transformed Gaussian process prior distribution to the infection rate function, and then develop an efficient data augmentation Markov Chain Monte Carlo algorithm to perform Bayesian inference. We use the posterior predictive distribution to simulate the effect of different disease control methods and their economic impact. We analyse a large outbreak of avian influenza in the Netherlands and infer the between-farm infection rate, as well as the unknown infection status of farms which were pre-emptively culled. We use our results to analyse ring-culling strategies,

and conclude that although effective, ring-culling has limited impact in high-density areas.

KEYWORDS

avian influenza, Bayesian nonparametrics, disease control, epidemic models, Gaussian processes

1 | INTRODUCTION

Diseases of livestock and farmed poultry, such as avian influenza or foot and mouth disease, pose serious public and animal health risks, as well as having a considerable impact on both the domestic and international farming economy. Authorities are keen to control the spread of such diseases as quickly as possible to reduce the health risks, but must also consider other stakeholders, such as farmers, and the economic consequences of intervention.

In 2003 a serious outbreak of a highly pathogenic avian influenza A/H7N7 virus took place among poultry farms in the Netherlands. Over the course of 3 months, more than 30 million birds were culled, 90 people developed influenza-like symptoms, with six confirmed cases, and one fatality occurred (Koopmans et al., 2004). The Dutch authorities implemented a culling strategy to control the disease, whereby animals were culled on farms where the pathogen was detected, and pre-emptively culled on farms within a certain distance from the site of detection. For convenience we shall refer to farms as *naturally* culled or *pre-emptively* culled in the obvious manner. The culling strategy took place alongside strict biosecurity measures and a ban on the transportation of poultry goods (Directorate-General for Health and Consumers, 2003). In the data set we use there is a total of 5397 Dutch poultry farms, including 241 infected farms and 1232 pre-emptively culled farms. The approximate locations of the farms are shown in Figure 1.

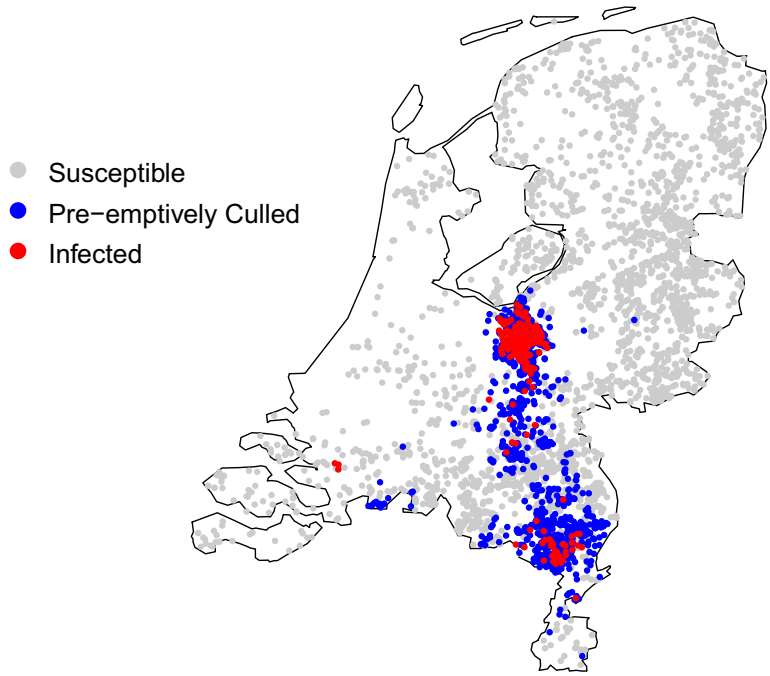

FIGURE 1 A map of the poultry farms in the Netherlands with their status at the end of the outbreak

There is a clear spatial element to the spread of the disease; for example, there are two distinctive clusters of infected farms, within which there appears to be local transmission. However, analysing the disease spread is challenging due to the fact that the times at which infections occurred are not observed. For farms which were confirmed to be infected, the date of poultry culling was recorded, but the date on which poultry on the farm were first infected was unobserved. The infection status of pre-emptively culled farms is considered uncertain, since the absence of clinical suspicion at the time of culling would not necessarily rule out the presence of the pathogen.

Various data were collected during the outbreak. The particular data set that we shall focus on consists of the spatial coordinates of all poultry farms in the Netherlands, plus the culling dates and identities of all farms that were either naturally or pre-emptively culled. There are several previous approaches to modelling data from this outbreak. In Stegeman et al. (2004), the authors construct a model based on a generalized linear model proposed in Becker (1989), where the number of new infections per day is assumed to follow a Binomial distribution. However, the infection rate is assumed to be constant between all farms, which is a questionable assumption given the clear spatial element to the spread of the disease. In Boender et al. (2007), the authors use a type of generalized linear model which allows for spatial variation in spread of the disease, and propose several plausible forms for the infection rate as a function of distance. It is assumed that pre-emptively culled farms are never infected, and that unobserved events such as infections occur at known times, obtained by simple assumptions motivated by expert opinion. Models are fitted using maximum likelihood methods and the Akaike information criterion is used to choose between them. In Backer et al. (2015), the authors take a different approach modelling both within- and between-farm transmission. They model within-farm transmission using an SEIR (Susceptible-Exposed-Infective-Removed) model whose parameters are taken from the literature (see, e.g., van der Goot et al., 2005). Transmission between farms is assumed to depend on the number of infectious animals and the distance between farms via a monotonically decreasing function in a similar fashion to the approach taken by Boender et al. (2007). The outbreak has also been studied from a public health and veterinary perspective, analysing the symptoms both humans and poultry display, see, for example, Fouchier et al. (2004); Koopmans et al. (2004); Elbers et al. (2004).

Although some previous modelling approaches attempt to capture the spatial variation in the infection rate, they rely on making strict parametric assumptions about the infection rate as a function of the distance between farms; such functions are commonly called distance kernels. The choice of a particular distance kernel may not accurately represent the underlying process and can lead to incorrect predictions which, in consequence, can have a significant impact on formulating policy decisions with regards to optimal disease control measures such as culling. Our approach removes the need to make such assumptions by modelling the infection rate nonparametrically. We do this by treating the infection rate as an unknown function with a transformed Gaussian process (GP) prior distribution. This allows us to make more general assumptions about the type of function, for example how smooth it is, whether it is continuous, or if it is monotonic, rather than its exact shape. Furthermore, previous modelling approaches assume that the times at which farms were infected are known. In this paper we relax this assumption, by adopting a data-augmentation approach within a Bayesian framework in which we treat infection times as additional parameters. We make inference for the infection rate function using a Markov Chain Monte Carlo (MCMC) algorithm, which also allows us to infer the unobserved infection times, and to estimate the probability of any pre-emptively culled farm having been infected. We anticipate that the proposed framework is suitable for analysing completed major outbreaks among populations in which there is a clear spatial component in the infection rate.

The paper is structured as follows. In Section 2 we describe the available data, define our stochastic transmission model in detail and derive an augmented likelihood function assuming that the epidemic process is fully observed. In Section 3 we present our Bayesian nonparametric approach by specifying

a transformed GP prior distribution for the infection rate function and the prior distributions for the other model parameters. We also describe an efficient MCMC algorithm to sample from the posterior distribution of the parameters given the observed data. In Section 4 we demonstrate the proposed models and methods via an application to simulated data and the avian influenza data set. We also illustrate how our methods can be used to assess control strategies. We finish in Section 5 with brief conclusions and a discussion of our methods.

2 | METHODOLOGY

2.1 | Data

The data set contains the geographical locations of 5397 poultry farms in the Netherlands at the time of the outbreak. The data set lacks reliable information on small non-commercial flocks, as most of these are exempt from registration. For that reason, and because an earlier analysis showed that such 'backyard flocks' played only a marginal role in this epidemic (Bavinck et al., 2009), we discarded all flocks with fewer than 500 animals from the data set. For each farm, the data specifies its status at the end of the outbreak, describing whether or not it had contracted the virus, had been culled due to confirmed infection, or had been culled pre-emptively. For farms which were culled we have the date on which this occurred. After the removal of farms with fewer than 500 animals, the data set contains 4466 farms. Of these, 233 farms were confirmed to be infected and consequently culled, while 1232 farms were pre-emptively culled. Table 1 illustrates the available information for each farm in the data set.

2.2 | Stochastic epidemic model

We construct our model based on the standard SIR (Susceptible-Infective-Removed) epidemic model in continuous time; see, for example, Bailey (1975); Andersson and Britton (2000). Consider a population consisting of N farms. We assume that initially all farms are disease free apart from one which contains animals infected via some external source. At any time t, a farm is either *susceptible* to the disease, *infected* with the disease and infectious, or *removed* as the animals on the farm have been culled. The model dynamics can be separated into two processes: the infection process and the removal process. The infection process is governed by a rate function $\beta(d)$, where d denotes the Euclidean distance between two farms.

TABLE 1 An example of the available information on each farm. Farms 1 and 2 were confirmed to be infected and culled in consequence. Farm 3 was culled pre-emptively and farm 4 was not culled. Farm geographical locations are provided in terms of x and y coordinates

Farm ID	x	y	Culling date	Pre-emptively culled
1	5.25	52.13	5 May	×
2	5.59	54.49	10 April	×
3	4.99	55.00	2 May	✓
4	5.50	51.40	—	—
⋮	⋮	⋮	⋮	⋮

We assume an infectious farm infects a given susceptible farm that is d km away according to a Poisson process with rate $\beta(d)$. The processes governing different pairs of farms are assumed to be independent. For the removal process, once a farm is infected it is assumed to be infectious for a time which follows a Gamma distribution, $\Gamma(\lambda, \gamma)$, which has mean λ/γ and variance λ/γ^2. The infectious periods of different farms are assumed to be independent. Note that the infectious period of a farm is the time between infection and culling as a result of infection being detected, rather than the time period during which animals would be infectious in the absence of any intervention.

To account for the fact that some farms are pre-emptively culled by the authorities as a disease control measure, we introduce pre-emptive culling times. We make no attempt to explicitly model the culling strategy, since in practice such strategies may change over time or not always be carried out as originally intended. Instead, we assume that pre-emptive cullings are deterministic events. If, under the disease control strategy, a farm is pre-emptively culled at time t, then the farm becomes removed at time t irrespective of whether it is currently susceptible or infectious. From this time, it can no longer infect other farms or be infected. We shall refer to culling events that are not pre-emptive as natural cullings. The epidemic continues until there are no more infected farms.

2.3 | Likelihood

Recall that the observed data consist of culling times, which can be pre-emptive or not, and farm locations. To fit our model to such data in a Bayesian framework requires the likelihood of the observed data given the model parameters. However, such a likelihood is intractable in practice since its computation involves integrating over all unknown infection events; see, for example, O'Neill and Roberts (1999); Jewell et al. (2009). We therefore proceed by deriving a likelihood based on full observation of the epidemic process, and use a data-augmentation MCMC algorithm as described in Section 3.

Let N denote the total number of poultry farms in the Netherlands and n the number of ever-infected farms. We denote the infection and culling times for farm j by i_j and r_j respectively, where culling may be pre-emptive or natural. We label the infected farms $1, \ldots, n$ by their culling date (i.e. $r_1 \leq r_2 \leq \ldots \leq r_n$) and the remaining farms $n+1, \ldots, N$ arbitrarily. We denote by ω the label of the initially infected farm.

We denote by $\mathbf{i} = \{i_1, \ldots, i_{\omega-1}, i_{\omega+1}, \ldots, i_N\}$ the set of all infection times excluding the initial infection time i_ω. If farm j was not infected, its infection time is set to be $i_j = \infty$. We account for pre-emptive culling by defining $r_j = \min(r_j^p, r_j^c)$, where r_j^p and r_j^c denote, respectively, the pre-emptive and natural culling time of farm j. We consider the times r_j^p to be deterministic, and set $r_j^p = \infty$ if farm j was not pre-emptively culled. For farms which were not culled at all, we set $r_j^p = r_j^c = \infty$, hence $r_j = \infty$. The sets $\mathbf{r}^c = \{r_1^c, \ldots, r_N^c\}$ and $\mathbf{r}^p = \{r_1^p, \ldots, r_N^p\}$ denote the set of natural and pre-emptive culling times respectively.

We require the following sets based on the infection status of the farms during the outbreak. Set $\mathcal{A}$ consists of the farms that remained susceptible to the disease throughout the course of the epidemic and were not culled, set $\mathcal{B}$ is the set of farms that were infected with the virus and naturally culled in consequence, set $\mathcal{C}$ is the set of farms that were infected but were culled pre-emptively, and finally set $\mathcal{D}$ consists of the farms that were not infected but still pre-emptively culled. These sets are shown in Table 2. Note that if a farm has been pre-emptively culled, we are unable to distinguish whether it belongs to set $\mathcal{C}$ or $\mathcal{D}$ unless its infection status is known.

The likelihood function consists of three parts: a contribution from farms avoiding infection, a contribution from farms being infected and a contribution from farms remaining infectious until culled.

TABLE 2 The infectious status of each farm at the end of the outbreak

Set	Infected	Culled	Pre-emptively culled
$\mathcal{A}$	×	×	×
$\mathcal{B}$	✓	✓	×
$\mathcal{C}$	✓	✓	✓
$\mathcal{D}$	×	✓	✓

For a farm k in either set $\mathcal{A}$, $\mathcal{B}$ or $\mathcal{C}$, the probability it avoids infection from infectious farm j, until either j is removed or k is infected, is

$$\psi_{j,k} = \exp\{-\beta(d_{j,k})((r_j \wedge i_k) - (i_j \wedge i_k))\},$$

where $\beta(d)$ is the infection rate for a pair of farms that are d km apart, and $a \wedge b = \min\{a, b\}$. The difference in minimum times is the amount of time during which farm j is able to infect k. If farm k is in set $\mathcal{D}$ we must take into account its pre-emptive culling time, $r_k = r_k^p$, and the corresponding probability is given by

$$\psi_{j,k} = \exp\{-\beta(d_{j,k})((r_j \wedge r_k) - (i_j \wedge r_k))\}.$$

When farm j is infected, the set of farms that are able to infect j is

$$\mathcal{Y}_j = \{k\colon i_k < i_j < r_k\},$$

so the event that j is infected contributes to the likelihood function through the overall hazard rate of infection given by

$$\phi_j = \sum_{k \in \mathcal{Y}_j} \beta(d_{k,j}).$$

For the removal process, the likelihood contribution is given by

$$\prod_{j \in \mathcal{B}} p(r_j - i_j \mid \lambda, \gamma) \prod_{j \in \mathcal{C}} S(r_j - i_j \mid \lambda, \gamma),$$

where $p(x|\lambda, \gamma)$ is the probability density function of a $\Gamma(\lambda, \gamma)$ distribution evaluated at x and $S(x|\lambda, \gamma)$ is the survivor function

$$S(x \mid \lambda, \gamma) = \int_x^\infty p(u \mid \lambda, \gamma) du.$$

Farms in set $\mathcal{B}$, that were infected and culled at the end of their infectious period, contribute to the likelihood function through the total time during which they were infectious. For those in set $\mathcal{C}$, which were infected but culled pre-emptively, we consider their removal time as a censoring time, and compute the probability they would have remained infectious past their culling time. Combining the infection and removal processes gives the augmented likelihood function

$$\begin{aligned}
&\pi(\mathbf{i}, \mathbf{r}^c, \mathcal{B}, \mathcal{C}, \mathcal{D}|\beta, \lambda, \gamma, \omega, i_\omega, \mathbf{r}^p) \\
&= \left(\prod_{j\in\mathcal{B}\cup\mathcal{C}}\prod_{k=1}^{N}\psi_{j,k}\right)\left(\prod_{\substack{j\in\mathcal{B}\cup\mathcal{C}\\ j\neq\omega}}\phi_j\right)\prod_{j\in\mathcal{B}}p(r_j-i_j|\lambda,\gamma)\prod_{j\in\mathcal{C}}S(r_j-i_j|\lambda,\gamma) \\
&= \exp\{-\Psi\}\prod_{\substack{j\in\mathcal{B}\cup\mathcal{C}\\ j\neq\omega}}\left(\sum_{k\in\mathcal{Y}_j}\beta(d_{k,j})\right)\prod_{j\in\mathcal{B}}p(r_j-i_j|\lambda,\gamma)\prod_{j\in\mathcal{C}}S(r_j-i_j|\lambda,\gamma),
\end{aligned} \tag{1}$$

where

$$\begin{aligned}
\Psi = \sum_{j\in\mathcal{B}\cup\mathcal{C}}\Bigg[&\sum_{k\in\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}\beta(d_{j,k})\left((r_j\wedge i_k)-(i_j\wedge i_k)\right) \\
&+\sum_{k\in\mathcal{D}}\beta(d_{j,k})\left((r_j\wedge r_k)-(i_j\wedge r_k)\right)\Bigg].
\end{aligned} \tag{2}$$

Note that the set of culling times determines which farms belong to the set $\mathcal{A}$, which is why $\mathcal{A}$ does not appear explicitly in the left-hand side of Equation (1).

3 | BAYESIAN NONPARAMETRIC INFERENCE

We wish to make Bayesian inference for the unknown model parameters given the observed data of farm locations and culling dates (see Table 1). If a farm was not culled by the end of the outbreak, we assume that it remained susceptible throughout the outbreak. Hence, the observed culling dates determine which farms belong to set $\mathcal{A}$. For a farm that has been pre-emptively culled, its infection status is unknown and therefore we cannot determine from the observed data if such a farm belongs to set $\mathcal{C}$ or $\mathcal{D}$. Also, the infection process defined in our model is not observed directly. Hence, the label of the initially infected farm ω, its infection time i_ω and the infection times of the farms belonging to sets $\mathcal{B}$ or $\mathcal{C}$ are unknown.

We adopt a data augmentation framework (see, e.g. Jewell et al., 2009) in which we include the farms' unknown infection event times and statuses as additional model parameters to the ones which govern the transmission and removal processes. Combining the augmented data likelihood (1) with the joint prior distribution, by using Bayes' theorem, the target posterior density is given by

$$\begin{aligned}
\pi(\beta, \gamma, \omega, i_\omega, \mathbf{i}, \mathcal{C}, \mathcal{D}|\mathcal{B}, \lambda, \mathbf{r}^p, \mathbf{r}^c) \propto\ &\pi(\mathbf{i}, \mathbf{r}^c, \mathcal{B}, \mathcal{C}, \mathcal{D}|\beta, \lambda, \gamma, \omega, i_\omega, \mathbf{r}^p) \\
&\times\pi(\beta)\pi(\lambda)\pi(i_\omega|\omega)\pi(\omega),
\end{aligned}$$

where we have assumed that β, λ and (ω, i_ω) are independent *a priori*.

3.1 | Prior distributions

We now discuss in detail the prior distributions for the infection rate function and the other model parameters.

3.1.1 | The infection rate function

We wish to infer the infection rate function β nonparametrically and to do so we will use a transformed GP as a prior distribution. We follow Rasmussen and Williams (2006) and define a GP as a collection of points, any finite subset of which follow a multivariate Normal distribution. Suppose we wish to model a function f, over a space χ, specifically being interested in the values of the function $f(x_1), \ldots, f(x_n)$ evaluated at the points $\mathbf{x} = \{x_1, \ldots, x_n\}$. We specify the GP prior distribution on f by

$$f \sim \mathcal{GP}(\mu, \Sigma),$$

where μ is the mean function and Σ the covariance matrix, defined using a covariance function k. We build our assumptions about f into the model through the covariance matrix, and to do so we use the squared exponential function. This is given by

$$\Sigma_{i,j} = k(x_i, x_j; \alpha, l), \quad k(x_i, x_j; \alpha, l) = \alpha^2 \exp\left\{ -\frac{(x_i - x_j)^2}{l^2} \right\}.$$

The function k has two hyperparameters, namely α, which controls the overall variance, and l, which controls the length scale. The value of l essentially determines how much the function can change as the input changes. We implicitly assume that f is smooth and differentiable. Many other choices for the kernel function are available (Rasmussen & Williams, 2006), but our choice appears suitable for the application at hand.

The input space of the function β is the space of Euclidean distances. We specifically wish to evaluate β at $\boldsymbol{d}$, the set of pair-wise distances between all farms. As the GP prior distribution gives non-zero probability to negative values and we are modelling a rate which is always positive, we introduce a dummy function g and use the transformation $\beta = \exp\{g\}$. In other words, we are placing a GP prior distribution on $\log \beta$ by specifying that

$$g \sim \mathcal{GP}(0, \Sigma), \quad \Sigma_{ij} = k(d_i, d_j; \alpha, l), \quad \beta = \exp\{g\}$$

where d_i is the Euclidean distance between the i^{th} pair of farms.

A well-known problem arises with GPs when the size of the covariance matrix is large, since this creates computational difficulties with matrix inversion and decomposition; see, for example, Hensman et al. (2013); Csato and Opper (2002); Quinonero-Candela and Rasmussen (2005). For the avian influenza data set, there are over 9 million unique pair-wise distances from which the covariance matrix is constructed. In the MCMC algorithm we will develop, we will require the covariance matrix to be repeatedly decomposed and inverted, which is not feasible in practice with such a large matrix. We therefore approximate the GP prior distribution using a projection method first described in Quinonero-Candela and Rasmussen (2005). We construct a pseudo set of distances, $\overline{\mathbf{d}}$, that is much smaller than the original set $\mathbf{d}$. While $\overline{\mathbf{d}}$ need not be a subset of $\mathbf{d}$ it should provide an adequate representation of $\mathbf{d}$. We then place a GP prior distribution on the pseudo set and draw samples from this distribution. The joint prior distribution of the pseudo function, $\overline{f}$ and the full function f is

$$\begin{pmatrix} \overline{f} \\ f \end{pmatrix} \sim \mathcal{GP}\left(\mathbf{0}, \begin{pmatrix} \Sigma_{\overline{\mathbf{d}},\overline{\mathbf{d}}} & \Sigma_{\mathbf{d},\overline{\mathbf{d}}} \\ \Sigma_{\overline{\mathbf{d}},\mathbf{d}} & \Sigma_{\mathbf{d},\mathbf{d}} \end{pmatrix} \right),$$

where the subscripts on the Σ matrices denote the vectors used to construct them. We can then project the samples onto the full data set by considering the conditional distribution of f given the pseudo function $\overline{f}$. We take f to be the mean of this distribution, which is given by

$$f = \Sigma_{\mathbf{d},\overline{\mathbf{d}}}\Sigma_{\overline{\mathbf{d}}}^{-1}\overline{f},$$

where $\Sigma_{\overline{\mathbf{d}}}$ is the covariance matrix of the approximating prior distribution. Some care is needed to construct the pseudo set $\overline{\mathbf{d}}$, because we must ensure the points are sufficient in number and suitably placed across the entire domain to capture the features of β. Simulation studies in Seymour (2020) suggest that the error introduced by this approximation is small, even when the number of pseudo distances is as small as 10% of the total number of pair-wise distances. This method assumes that the prior distribution over the pseudo data set has the same properties as that over the original data set, which is a reasonable assumption as they are both sets of Euclidean distances.

The value of the length scale parameter l can have a material impact on the results, and it is not obvious how to assign a suitable value. We therefore place a non-informative prior distribution on this parameter, specifically $l \sim \text{Exp}(0.01)$, where $\text{Exp}(a)$ denotes an exponential distribution with mean a^{-1}. Inferring both hyperparameters of the GP (l and α) can be very challenging (Zhang, 2004), and therefore we assume the variance parameter α is known *a priori*. We choose a value such that samples from the prior distribution are over a large enough range to capture the scale of the infection rate.

3.1.2 | Other model parameters

Recall that the infectious period distribution is assumed to be a $\Gamma(\lambda, \gamma)$ distribution. We follow Jewell et al. (2009) and assume that λ is known and place an uninformative prior distribution on γ, specifically $\gamma \sim \text{Exp}(0.01)$.

For the infection times, we place a discrete uniform prior distribution on the label of the initially infected farm ω. We set a time axis by assuming the first culling to be at time zero, so that $r_1 = 0$, and set the prior distribution on the infection time of ω by

$$(i_\omega \,|\, \omega) = -z, \quad z \sim \text{Exp}(0.01).$$

3.2 | Markov chain Monte Carlo algorithm

The density of the full posterior distribution is given by

$$\begin{aligned}
&\pi(\beta, \gamma, \omega, i_\omega, \mathbf{i}, \mathcal{C}, \mathcal{D} | \mathcal{B}, \lambda, \mathbf{r}^c, \mathbf{r}^p) \\
&\propto \exp\{-\Psi\} \prod_{\substack{j=1 \\ j \neq \omega}}^{n} \left(\sum_{k \in \mathcal{Y}_j} \exp\{g(d_{k,j})\} \right) \prod_{j \in \mathcal{B}} p(r_j - i_j | \lambda, \gamma) \prod_{j \in \mathcal{C}} S(r_j - i_j | \lambda, \gamma) \\
&\times \mathcal{GP}(g; \mathbf{0}, \Sigma) \exp\{-0.01l\} \exp\{-0.01\gamma\} \exp\{0.01 i_\omega\}.
\end{aligned}$$

The likelihood function is the same as in Equation (1) and Ψ is given in Equation (2) with β replaced by the inferred value $\exp\{g\}$. The term $\mathcal{GP}(g; \mathbf{0}, \Sigma)$ refers to the finite dimension form of the GP, which is the probability density function of a multivariate Gaussian distribution evaluated at $g(\mathbf{d})$, with the

Algorithm 1 Structure of the MCMC algorithm

1: Initialise the chain with values $g^{(0)}$, $\gamma^{(0)}$, $l^{(0)}$, and $\mathbf{i}^{(0)}$;
Repeat the following steps:
2: Update g using a Metropolis-Hastings step;
3: Update l using a random walk Metropolis-Hastings step;
4: Update γ using a random walk Metropolis-Hastings step;
5: Update ω using a random walk Metropolis-Hastings step;
6: Update i_ω using a random walk Metropolis-Hastings step;
7: Choose one of the following steps with equal probability:

- Update an infection time
- Remove an infection time for a pre-emptively culled farm
- Add an infection time for a pre-emptively culled farm

corresponding mean vector $\mathbf{0}$ and covariance matrix Σ. We cannot sample from the posterior distribution directly so construct an MCMC algorithm, which is shown in Algorithm 1. There are five main steps to the algorithm and these are described in detail below.

3.2.1 | Updating the infection rate

The first step is concerned with sampling the dummy function g, which we do using an underrelaxed proposal mechanism described in Neal (1998). This allows us to update the function as a block while reducing computational complexity. Given the current function g, we propose a new function g' by

$$g'(\mathbf{d}) = \sqrt{1-\delta^2}\, g(\mathbf{d}) + \delta \nu(\mathbf{d}), \quad \nu \sim \mathcal{GP}(\mathbf{0}, \Sigma),$$

where $\delta \in (0, 1]$ is a tuning parameter, $g(\boldsymbol{d})$ is the value of the function g at the current iteration, and $\nu(\boldsymbol{d})$ is a sample drawn from the prior distribution $\mathcal{GP}(\mathbf{0}, \Sigma)$ where Σ denotes the covariance matrix of the GP. The computational advantage of this is that the prior ratio is the inverse of the proposal ratio so the Metropolis–Hastings acceptance probability reduces to the likelihood ratio (see Section 1 of the supplementary material)

$$p_{acc} = \frac{\pi(\mathbf{i}, \mathbf{r}^c, \mathcal{B}, \mathcal{C}, \mathcal{D} \,|\, g', \lambda, \gamma, \omega, i_\omega, \mathbf{r}^p)}{\pi(\mathbf{i}, \mathbf{r}^c, \mathcal{B}, \mathcal{C}, \mathcal{D} \,|\, g, \lambda, \gamma, \omega, i_\omega, \mathbf{r}^p)} \wedge 1.$$

When the projection approximation method is used, we first propose new values for $\overline{g}$ on the input space $\overline{\mathbf{d}}$ using the above proposal, that is, $\overline{g}'(\overline{\mathbf{d}}) = \sqrt{1-\delta^2}\,\overline{g}(\overline{\mathbf{d}}) + \delta \nu(\overline{\mathbf{d}})$, and then project $\overline{g}'$ onto g' which is then used in the Metropolis–Hastings ratio above.

3.2.2 | Updating l

To update the GP prior distribution length scale, l, we use a Gaussian random walk Metropolis algorithm by first proposing a new length scale, l', from $N(l, \sigma_l^2)$ where l is the current value and σ_l^2 is a tuning parameter, and then accept l' with probability

$$p_{acc} = \frac{\mathcal{GP}(g; \mathbf{0}, \Sigma_{l'})}{\mathcal{GP}(g; \mathbf{0}, \Sigma_l)} \frac{\pi(l')}{\pi(l)} \wedge 1.$$

3.2.3 | Updating γ

To update γ, we also use a Gaussian random walk Metropolis algorithm by proposing γ' from the distribution $N(\gamma, \sigma_\gamma^2)$ and accepting this with probability

$$p_{acc} = \frac{\prod_{j\in\mathcal{B}} p(r_j - i_j \mid \lambda, \gamma') \prod_{j\in\mathcal{C}} S(r_j - i_j \mid \lambda, \gamma')}{\prod_{j\in\mathcal{B}} p(r_j - i_j \mid \lambda, \gamma) \prod_{j\in\mathcal{C}} S(r_j - i_j \mid \lambda, \gamma)} \frac{\pi(\gamma')}{\pi(\gamma)} \wedge 1.$$

3.2.4 | Updating infection times

The final step in the algorithm concerns the unobserved infection times. We use a method proposed in O'Neill and Roberts (1999) and then further developed in Jewell et al. (2009). We choose one of three actions with equal probability: (i) propose to move an existing infection time; (ii) propose to add a new infection time; and (iii) propose to delete a previously added infection time.

1. **Updating an infection time** of a farm in sets $\mathcal{B}$ or $\mathcal{C}$ is the simplest of the three procedures. To do this, we randomly choose a farm j that is currently infected and propose a new infection time by $i_j' = r_j - t_j$, where $t_j \sim \Gamma(\lambda, \gamma)$ and γ denotes the current value of the parameter in the chain. We accept i_j' with probability

$$p_{acc} = \frac{p(r_j - i_j \mid \lambda, \gamma)}{p(r_j - i_j' \mid \lambda, \gamma)} \frac{\pi(\mathbf{i} - i_j + i_j', \mathbf{r}^c \mid g, \lambda, \gamma, i_\omega, \mathbf{r}^p, \mathcal{B}, \mathcal{C}, \mathcal{D})}{\pi(\mathbf{i}, \mathbf{r}^c \mid g, \lambda, \gamma, i_\omega, \mathbf{r}^p, \mathcal{B}, \mathcal{C}, \mathcal{D})} \wedge 1,$$

   where $\mathbf{i} - i_j + i_j'$ is the set $\mathbf{i}$ with i_j removed and i_j' included.

2. When **adding an infection time**, first define m to be the number of pre-emptively culled farms. We suppose that at the current iteration of the algorithm, $\tilde{m}$ of the farms which were pre-emptively culled have had infection times added by the algorithm; that is farms belonging in set $\mathcal{C}$. We randomly choose one of the $m - \tilde{m}$ pre-emptively culled farms with no infection time and propose an infection time for it. If $m = \tilde{m}$, we abandon this step. We generate an infection time as above and accept it with probability

$$p_{acc} = \frac{1/(\tilde{m}+1)}{(1/(m-\tilde{m}))p(r_j - i_j' \mid \lambda, \gamma)} \frac{\pi(\mathbf{i} + i_j, \mathbf{r}^c, \mathcal{C}, \mathcal{D} \mid g, \lambda, \gamma, i_\omega, \mathbf{r}^p)}{\pi(\mathbf{i}, \mathbf{r}^c, \mathcal{C}, \mathcal{D} \mid g, \lambda, \gamma, i_\omega, \mathbf{r}^p)} \wedge 1$$

$$= \frac{m - \tilde{m}}{(\tilde{m}+1)p(r_j - i_j' \mid \lambda, \gamma)} \frac{\pi(\mathbf{i} + i_j, \mathbf{r}^c, \mathcal{C}, \mathcal{D} \mid g, \lambda, \gamma, i_\omega, \mathbf{r}^p)}{\pi(\mathbf{i}, \mathbf{r}^c, \mathcal{C}, \mathcal{D} \mid g, \lambda, \gamma, i_\omega, \mathbf{r}^p)} \wedge 1.$$

3. Finally, if we choose to **delete an infection time** for a pre-emptively culled farm, we randomly choose a pre-emptively culled farm j which at the current iteration has an infection time added and we propose to remove its infection time. Should there be no farms with an unknown infection status, which, at the current iteration of the algorithm, have had an infection time added, the step is abandoned. We accept this proposal with probability

$$p_{acc} = \frac{1/(m-(\tilde{m}-1))p(r_j - i_j|\lambda,\gamma)}{1/\tilde{m}} \frac{\pi(\mathbf{i}-i_j, \mathbf{r}^c|\mathrm{g},\lambda,\gamma,i_\omega,\mathbf{r}^p,\mathcal{B},\mathcal{C},\mathcal{D})}{\pi(\mathbf{i},\mathbf{r}^c|\mathrm{g},\lambda,\gamma,i_\omega,\mathbf{r}^p,\mathcal{B},\mathcal{C},\mathcal{D})} \wedge 1$$

$$= \frac{p(r_j - i_j|\lambda,\gamma)\tilde{m}}{m-(\tilde{m}-1)} \frac{\pi(\mathbf{i}-i_j, \mathbf{r}^c,\mathcal{B},\mathcal{C},\mathcal{D}|\mathrm{g},\lambda,\gamma,i_\omega,\mathbf{r}^p)}{\pi(\mathbf{i},\mathbf{r}^c,\mathcal{B},\mathcal{C},\mathcal{D}|\mathrm{g},\lambda,\gamma,i_\omega,\mathbf{r}^p)} \wedge 1.$$

4 | RESULTS

We now present the results of our method applied to two data sets. The first is a simulated data set and the second is the avian influenza data set described in Section 1. We then use the posterior predictive distribution to analyse the impact of various culling strategies for the avian influenza outbreak.

4.1 | Simulation study

We generated the position of 1000 farms uniformly at random on a square with side length 30 km. We then simulated 250 outbreaks of avian influenza using the infection rate function

$$\beta(d) = 0.6\exp\{-2d\}.$$

The infectious period distribution parameters were $\lambda = 4$ and $\gamma = 0.8$. This gives a mean infectious period, which represents the time from infection to culling, of $\lambda/\gamma = 5$ days, suitable for influenza-like diseases among livestock. The simulations also included a deterministic culling strategy such that once a farm was culled following a positive test, all farms within a 1 km radius were pre-emptively culled. Note that although this strategy is inspired by what happened in the actual outbreak, it is somewhat idealized since, as mentioned in Section 2.2, culling strategies may change over time.

We discarded simulated data sets with less than 100 infected farms, since our focus is towards analysing sizeable outbreaks, and any nonparametric modelling approach will struggle with a small data set. This left 175 data sets. Mimicking the data available in the avian influenza outbreak, for each simulated data set we assume that in addition to the coordinates we only observe the culling times and whether a farm has been pre-emptively culled or not. The infectious period shape parameter is assumed to be $\lambda = 4$.

Estimating both the length scale (l) and the variance parameter (α) can be very challenging (Zhang, 2004) and computationally expensive (Chalupka et al., 2013). It is therefore common to treat either parameter, or both of them, fixed and known. Care is indeed needed when specifying a value for the variance parameter α. A very small value yields slow convergence times for the Markov chain. On the other hand, a very large value will lead to poor MCMC mixing. In practice, α needs to chosen such that the prior distribution for β covers a large space, particularly $\beta(0)$, but not so large that the Markov chain mixes poorly. Therefore, we fix the length scale parameter as $l = 3$ and the variance parameter $\alpha = 9$ due to the computation time required to perform inference for both hyperparameters and infection times for pre-emptively culled farms for each of the simulated data sets. We also repeated the analysis for $l = 2$ and $l = 5$, and found that the results described below were essentially unchanged (see Section 2 of the supplementary material).

We fitted the model described in Section 2 by assuming that the infection rate is a function that depends only on the distance between farms and assigned a GP prior distribution as described in Section 3.1.1. Due to the number of farms, we used the GP approximation method, constructing the pseudo set

of distances by taking 256 equally spaced points from zero to the largest distance in the data set. We employed the MCMC algorithm described in Section 3.2 to fit the model to each of the 175 data sets.

Figure 2 shows the median rate compared to the true rate and a 95% credible interval constructed from all 175 posterior medians. The results demonstrate that we can infer the infection rate function well for all pair-wise distances above 0.5 km, but we slightly underestimate the rate between immediate neighbours. This underestimate is caused by there being few farms in each data set that are less than 0.5 km apart. We estimate the median infectious period to be 5.07 days, close to the true value of 5 days, and the 95% credible interval of the 175 estimates contains the true value of 5. This slight overestimation is likely to be caused by the combination of slight underestimation of β at low distances, and the fact that we only considered data sets with at least 100 infections, in which infectious periods may be slightly larger than average.

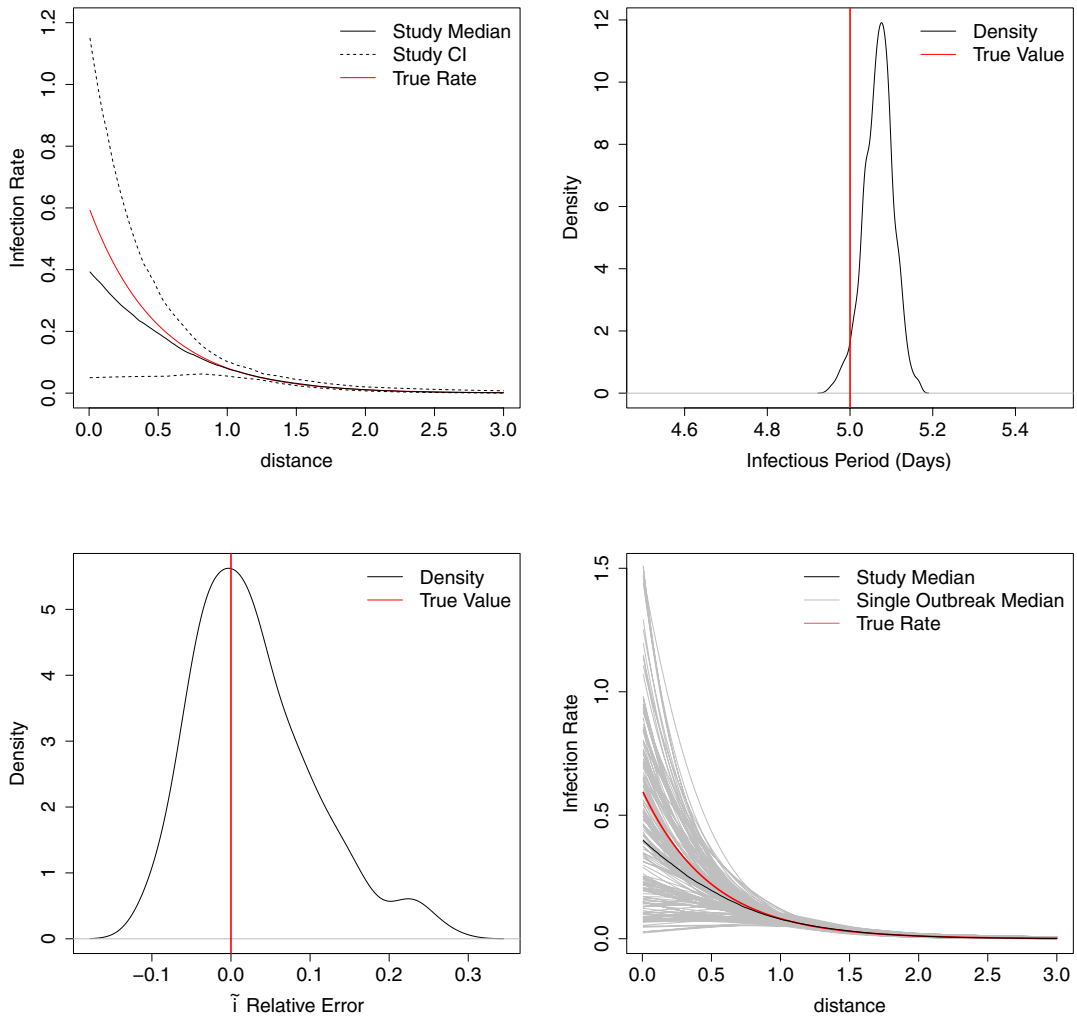

FIGURE 2 Top left: The results of the nonparametric method for the infection rate in the simulated data sets. We report the median and 95% credible interval of all 175 posterior medians. Top right: The distribution of the posterior median estimates for the mean infectious period. Bottom left: The distribution of the relative error in the sum of the infection times. Bottom right: The 175 posterior medians

To assess the results for the infection times across all simulations, we use the relative percentage error in the sum of the infection times for each simulated outbreak, which we denote by $\tilde{i}$, defined as follows. Consider a single simulated data set. Let S denote the sum of all infection times of farms culled either naturally or pre-emptively in the data set. Let $\widehat{S}$ denote the median estimate of S obtained from the MCMC output. Note that $\widehat{S}$ implicitly takes account of which pre-emptively culled farms are imputed to have been infected. Then we define

$$\tilde{i} = \frac{S - \widehat{S}}{S} \times 100\%.$$

The median relative percentage error across all data sets is 1.40%, which demonstrates that our method for inferring infection times gives accurate results. The results for the parameters are shown in Table 3.

4.2 | Avian influenza

We now analyse the avian influenza data described in Section 2.1. Due to the size of the data set, we split the inference into two parts. We first inferred plausible values of the GP prior distribution length scale parameter, l, by fitting our transmission model under the assumptions of a constant infectious period of 7 days and that pre-emptively culled farms were not infected, as in Boender et al. (2007). We obtained a posterior median for l of 2.75 km (95% CI: (2.55, 3.01)). The reason for inferring plausible values for l separately is that estimating l requires decomposing and inverting the covariance matrix inside the MCMC algorithm which is highly computationally intensive and leads to prohibitively long run times. This issue is amplified when the infection times are unknown as well.

We repeated the inference method without assuming that the infection times or the status of the pre-emptively culled farms are known. Based on the results of the method with a fixed infectious period, we fixed $\alpha = 3$ and $l = 3$ km. We employed the GP approximation method for this data set. As we expect the infection rate function to vary considerably over short to medium distances, we included more such distances in the pseudo data set. The pseudo data set was $\overline{\boldsymbol{d}} = \{0, 0.5, 1, \ldots, 19.5, 20, 30, \ldots, 350\}$. We ran the MCMC algorithm for 20,000 iterations, including a burn-in period of 500 iterations. In each iteration of the MCMC algorithm, we proposed updating, adding or deleting 200 infection times. This took 7 days on the University of Nottingham's High Performance Computing facility.

The results for the infection rate are shown in Figure 3, where we see a logistic-type function that decays to zero. From this, we estimate that the probability of a farm infecting another farm which is more than 6 km apart is negligible. From the credible interval, we see that samples from the posterior distribution take a variety of shapes, with functions that have a high infection rate over short distance decaying quickly, and functions that have a lower rate over short distances taking a logistic function form.

TABLE 3 Summary statistics for the 175 posterior median values obtained in the simulation study. The probability interval is from the 2.5% to 97.5% quantiles

Parameter	True value	Median	95% Prob. Int.
γ	0.8	0.787	(0.778, 0.802)
λ/γ	5	5.07	(4.99, 5.13)
$\tilde{i}$	0%	1.40%	(−9.18%, 23.0%)

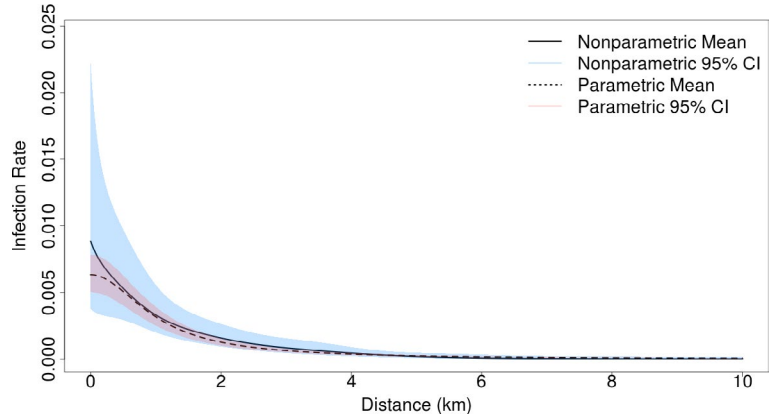

FIGURE 3 The posterior mean for the nonparametric (solid) and parametric (dashed) infection rate functions for the avian influenza data set. The parametric function is kernel 3 in Table 4

TABLE 4 The proposed parametric pair-wise infection rates for the avian influenza data set in Boender et al. (2007)

Rate	Kernel
1	$\beta(d) = \beta_0$
2	$\beta(d) = \beta_0(1 + d)^{-1}$
3	$\beta(d) = \beta_0(1 + d^2)^{-1}$
4	$\beta(d) = \beta_0(1 + d^{\beta_1})^{-1}$
5	$\beta(d) = \beta_0(1 + (d/\beta_2)^{\beta_1})^{-1}$

We compare our results to those in Boender et al. (2007), particularly with a view to comparing estimation of the infection rate function. The authors propose five models, shown in Table 4, which we fitted to the data assuming a fixed infectious period of 7 days. Model 3 was the best of the proposed models according to the deviance information criterion (Spiegelhalter et al., 2002). We refitted model 3 to the data assuming that the infection times are unknown. The results are shown in Figure 3 and one clear difference between the parametric and nonparametric methods is the associated uncertainty. Although the nonparametric method allows for a greater degree of flexibility, it also induces a greater degree of uncertainty. However, we argue that the parametric method may underestimate the uncertainty by imposing stricter assumptions. Despite this, both estimates are of similar shape and scale, and our results broadly agree with existing work. We see a slight difference in the forms of the infection rate function for distances less than 400 m, which is due to there being very few farms that are less than 400 m apart.

Since we assume infection times to be unknown, we infer them via our MCMC algorithm. We estimate the mean infectious period to be 6.4 days, and Figure 4 shows the distribution of median infectious periods by culling status. For farms that were subject to pre-emptive culling, the median infectious period is shorter than for those who were identified as infected. This is expected as pre-emptive culling of infected farms introduces censoring.

We estimate the probability that each pre-emptively culled farm was actually infected, as shown in Figure 4. All of the farms with non-zero probability of infection are located in the two main infection clusters. Our results show that the transmission to the southern cluster cannot be explained by a path

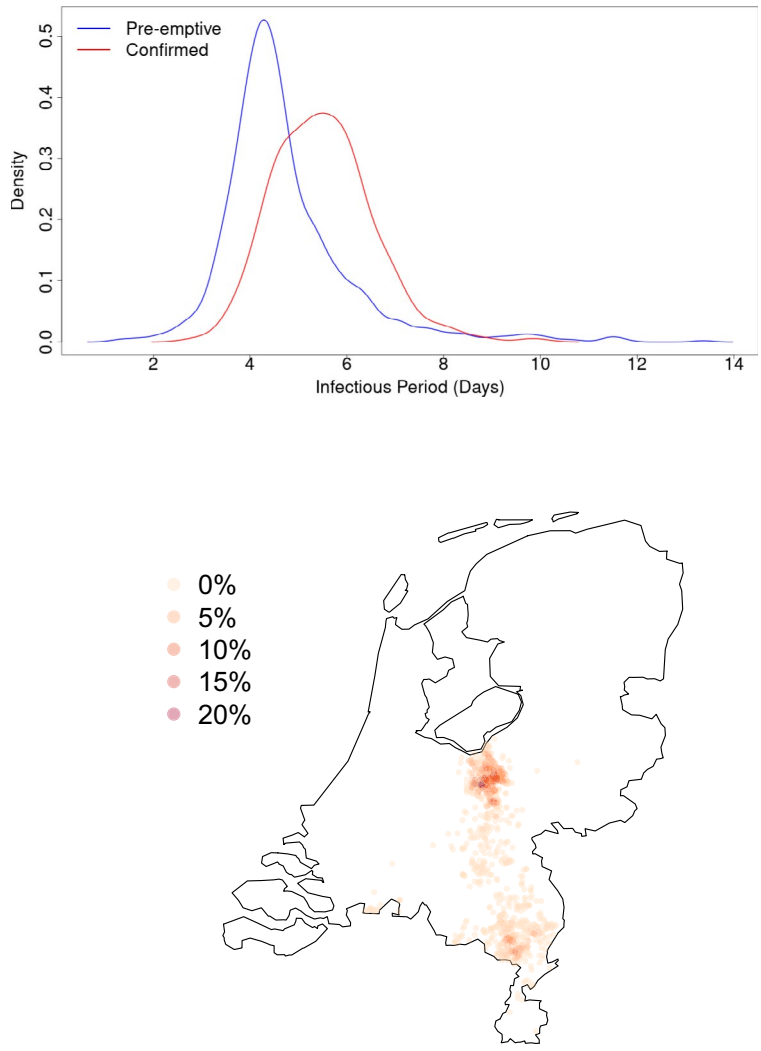

FIGURE 4 Top: Posterior distribution of median infectious periods for farms with confirmed infections and those pre-emptively culled. Bottom: Estimates of probabilities that pre-emptively culled farms were infected. Only farms which were pre-emptively culled are plotted. Each probability is the proportion of iterations in the MCMC algorithm that the pre-emptively culled farm was actually infected

of shorter distance infections that were censored by pre-emptive culling. This is consistent with the hypothesis proposed in Bataille et al. (2011) that this long distance transmission event of avian influenza was the result of a single human-mediated transport of the virus.

4.3 | Culling strategies

We now investigate how to improve the disease control measures by analysing how the culling radius affects the number of infected farms. Culling infected farms has the effect of reducing the time a farm is infectious, and culling susceptible farms means there are fewer farms to be infected. Although this

is an effective measure for controlling the spread of the disease, it can be expensive as farmers are compensated for lost livestock and it can cause negative public attitudes.

To simulate the effect of culling, we sample from the posterior predictive distribution of the infection and culling times. Given the observed culling times, and the posterior distributions of g, γ and ω, we wish to generate new infection times $\mathbf{i}^*$ for all farms, and corresponding culling times $\mathbf{r}^*$. We do this using the posterior predictive distribution, which is given by

$$\pi(\mathbf{i}^*, \mathbf{r}^* \mid \mathbf{r}) = \int\int\int \pi(\mathbf{i}^*, \mathbf{r}^* \mid g, \gamma, \omega, \mathbf{r})\pi(g, \gamma, \omega \mid \mathbf{r})\mathrm{d}g\mathrm{d}\gamma d\omega.$$

To generate samples from this distribution we initiate the outbreak by assuming the initially infected farm ω is the farm that was initially culled in the observed outbreak. To consider the effectiveness of culling strategies, we assume that once an infected farm reaches the end of its infectious period and enters the removed class all farms up to r km away are simultaneously culled and enter the removed class. Culling cannot start immediately as it may take time for the authorities to be notified of the disease and put measures into place, and whereas previous work (Backer et al., 2015) uses a fixed delay after the first detection to initiate the culling measures, we allow for stochasticity in the disease take-off and assume culling takes place once a certain number of farms have been infected. As resources may not be immediately available to the authorities, it may not be possible to cull all farms within r km and we simulate this by fixing a maximum number of farms that can be culled per day. We then increase this number over the course of the outbreak as the authorities have more available resources. The numbers are given in Table 5 and are based on the number of farms we estimate to have been infected in the observed outbreak. Similarly, we assume the authorities will not have sufficient resources to cull all farms within the chosen radius at the start of the outbreak, and we model this by assuming they initially cull farms within a radius half as large.

To investigate the economic consequences of these strategies, we assume each farmer is compensated for their culled livestock. We use additional data from the outbreak which describes the type of poultry on each farm (broiler, duck, turkey and layer) and the number of birds on each farm. The value of the compensation depends on the type of bird culled, the number of birds culled, their age in weeks, and, for turkeys, their gender. We follow Backer et al. (2015) who use the approximate rates shown in Table 6. We acknowledge this method is crude and does not take into account any of the wider economic impacts. However, it allows us to simulate the number of farms that are infected, the number of farms that are culled, and the compensation paid to farmers. These three values can be used to compare the risk to public health, the impact of the poultry industry and the cost to the authorities.

Table 7 shows the results of the culling strategies for radii between 0 km and 5 km. A culling radius of 0 km denotes the authorities taking no action. It is clear that taking any course of action leads to a reduction of the number of infected farms but also an increase in the amount of compensation given. Furthermore, we see that more ambitious strategies show little gain in reducing the median number of farms infected in an outbreak. The effect of culling at larger radii results in a larger number of culled farms and a higher amount of compensation, but does not result in a considerable reduction in the

TABLE 5 The culling strategy as a function of the total number of infected farms

Total number of infected farms (I)	Maximum number of farms culled per day	Proportion of culling radius implemented
$I \leq 33$	0	0
$33 < I \leq 54$	3	$\frac{1}{2}$
$54 < I$	6	1

TABLE 6 Estimates of compensation per bird paid to farmers during the avian influenza outbreak from Backer et al. (2015)

Poultry type	Compensation (€ per bird)
Broiler	0.98
Duck	2.09
Turkey	10.63
Layer	2.05

TABLE 7 Posterior predictive medians (95% probability intervals) for the number of infected and culled farms and the amount of compensation paid

Radius (km)	No. infected farms	No. culled farms	Compensation paid (€ millions)
0	443 (151, 644)	443 (151, 644)	24.8 (8.62, 35.9)
1	297 (110, 535)	489 (215, 709)	27.2(12.2, 38.9)
2	283 (108, 608)	488 (217, 740)	27.5 (12.2, 41.7)
3	283 (112, 582)	517 (242, 775)	29.0 (13.2, 43.1)
4	274 (105, 564)	512 (228, 793)	28.5 (12.3, 43.9)
5	280 (109, 549)	527 (226, 797)	39.2 (12.4, 41.9)

number of infected farms. This is because the maximum number of farms culled per day is quickly reached, even for small culling radii. In the data set, the average density of farms was approximately 2 per km^2, whereas a culling radius of 2 km covers over 12 km^2.

These results are broadly in line with those in Backer et al. (2015), which also suggest that larger culling radii do not result in a considerable reduction in the number of infected farms. However, as we use a much smaller estimate for the maximum number of farms culled per day, we do not observe a large difference between culling radii of 1 km and 2 km.

5 | CONCLUSIONS

We have presented an analysis of an outbreak of avian influenza in poultry farms in the Netherlands using a Bayesian nonparametric approach. Our approach demonstrates that it is possible to model the spatially heterogeneous infection rate for infectious diseases nonparametrically, and that GPs provide a flexible framework for doing so. This nonparametric methodology allows us to reduce the need for strict parametric assumptions, which are often made for mathematical or practical convenience and may have little scientific basis. Our methods also allow us to account for missing data, specifically the unobserved process of infection, without making unrealistic simplifying assumptions.

Although we have focused on an SIR model, in principle our methods can be extended to SEIR models as well, to incorporate a latent period. For our application to avian influenza, transmission experiments suggest that for the A/H7N7 virus in chickens, latent period for an infected animal is between 1 and 2 days (van der Goot et al., 2005). The latent period of an infected farm is often equated to the latent period of the first infected chicken, that is, in this case that would suggest a fairly short latent period of between 1 and 2 days (e.g. Backer et al., 2015). Furthermore, in Ypma et al. (2011)

the authors assumed a latent period of 1 day and subsequently performed a sensitivity analysis, where they compared results across latent periods of lengths 1, 2 and 3 days. Their comparison shows that their estimated kernel parameters were essentially insensitive to the assumed latent period duration. Although we could have considered fitting an SEIR model with a short latent period, we anticipate that this would not have any material impact on our results.

The methods we have described require more time and computational power than the standard parametric methods, especially when employed in conjunction with an MCMC approach to sample from the desired posterior distribution. We have, however, somewhat alleviated these issues by using a GP approximation method which appears to work well in our applications. Simulation studies in Seymour (2020) suggest that our methods work well even in small populations (e.g. $N = 100$), although there needs to be enough transmission in the population leading to a sizeable outbreak. Conversely, in scenarios where fewer data are available, such as small outbreaks or the initial phases of an outbreak, then in common with any nonparametric approach there will be greater uncertainty in parameter estimates. In such situations it might be appropriate to incorporate strong prior information or simply revert to a parametric approach.

For the avian influenza data set, our methodology has allowed us to approach the infection process in a more flexible way than previous methods. Our estimates are in line with previous work, and combining this method with previously developed MCMC techniques and data augmentation allows us to analyse this data set in more detail than has previously been possible, including determining whether pre-emptively culled farms had been infected. The uncertainty around our estimates is larger than that of previous parametric methods, but since we do not assume specific parametric models then our methods are, in some sense, giving a fairer quantification of uncertainty. We are able to use the posterior predictive distribution to analyse the effect of different control strategies which can be used to inform policy in this area.

In this paper, we have focussed on spatial heterogeneity as the key determinant of the infection rate. In reality, it is possible that the number and type of animals on the farms was also important. Given appropriate data, it is natural to build a model which contains such data as covariates. One way of doing this would be to consider each covariate as a separate dimension of the GP. Another possible extension is to consider different covariance functions beyond the squared exponential function, which could be appropriate in some applications. Also, the proposed framework can be extended to analyse the spread of infectious diseases early on in an outbreak. As mentioned above, the lack of data in the initial stages of an outbreak may be problematic. Possible ways to mitigate this by include adding further assumptions to the model, such as monotonicity of the infection rate, as well as employing more informative prior distributions. It is also be of interest to develop methods for model assessment, which is something that we have not considered here. Finally, another avenue for future work is to employ the recently developed methods described in Stockdale et al. (in press) in which the observed data likelihood is approximated without the need for imputing the unknown infection times. This would significantly reduce the computation time needed for our methods, and in conjunction with the GP projection method can make the proposed methodology more applicable for very large data sets.

ACKNOWLEDGEMENTS

We thank Wageningen Bioveterinary Research, The Netherlands Food and Consumer Product Authority and the Dutch Ministry of Agriculture, Nature and Food Quality for sharing anonymized outbreak, culling and denominator data of the Dutch 2003 HPAI epidemic with us.

We are grateful for access to the University of Nottingham High Performance Computing Facility.

We also thank the Associate Editor and the two reviewers for helpful and constructive comments that have improved this article.

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/N50970X/1.

ORCID

Rowland G. Seymour http://orcid.org/0000-0002-8739-3921
Theodore Kypraios http://orcid.org/0000-0002-6190-4762
Philip D. O'Neill http://orcid.org/0000-0001-9084-8184

REFERENCES

Andersson, H. & Britton, T. (2000) *Stochastic epidemic models and their statistical analysis*. Lecture Notes in Statistics. Berlin: Springer.

Backer, J., van Roermund, H., Fischer, E., van Asseldonk, M. & Bergevoet, R. (2015) Controlling highly pathogenic avian influenza outbreaks: an epidemiological and economic model analysis. *Preventive Veterinary Medicine*, 121, 142–150.

Bailey, N.T.J. (1975) *The mathematical theory of infectious diseases and its applications*, 2nd edn. London: Griffin.

Bataille, A., van der Meer, F., Stegeman, A. & Koch, G. (2011) Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. *PLoS Pathogens*, 7, e1002094.

Bavinck, V., Bouma, A., van Boven, M., Bos, M., Stassen, E. & Stegeman, J. (2009) The role of backyard poultry flocks in the epidemic of highly pathogenic avian influenza virus (H7N7) in the Netherlands in 2003. *Preventive Veterinary Medicine*, 88, 247–254.

Becker, N.G. (1989) *Analysis of infectious disease data*. London: Chapman and Hall.

Boender, G.J., Hagenaars, T.J., Bouma, A., Nodelijk, G., Elbers, A.R.W., de Jong, M.C.M.et al. (2007) Risk maps for the spread of highly pathogenic avian influenza in poultry. *PLoS Computational Biology*, 3, e71.

Chalupka, K., Williams, C.K. & Murray, I. (2013) A framework for evaluating approximation methods for Gaussian process regression. *Journal of Machine Learning Research*, 14, 333–350.

Csato, L. & Opper, M. (2002) Sparse online Gaussian processes. *Neural Computation*, 14, 641–668.

Directorate-General for Health and Consumers. (2003) Avian influenza (AI) in the Netherlands, Belgium and Germany – chronology of main events and list of decisions adopted by the commission. *Technical report*, European Commission.

Elbers, A.R.W., Fabri, T.H.F., de Vries, T.S., de Wit, J.J., Pijpers, A. & Koch, G. (2004) The highly pathogenic avian influenza A (H7N7) virus epidemic in the Netherlands in 2003-lessons learned from the first five outbreaks. *Avian Diseases*, 48, 691–705.

Fouchier, R.A.M., Schneeberger, P.M., Rozendaal, F.W., Broekman, J.M., Kemink, S.A.G., Munster, V. et al. (2004) Avian influenza a virus (H7N7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome. *Proceedings of the National Academy of Sciences*, 101, 1356–1361.

van der Goot, J.A., Koch, G., de Jong, M.C.M. & van Boven, M. (2005) Quantification of the effect of vaccination on transmission of avian influenza (H7N7) in chickens. *Proceedings of the National Academy of Sciences*, 102, 18141–18146.

Hensman, J., Fusi, N. & Lawrence, N.D. (2013) Gaussian processes for big data. In *Conference on Uncertainty in Artificial Intelligence*, pp. 282–290.

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

Koopmans, M., Wilbrink, B., Conyn, M., Natrop, G., van der Nat, H., Vennema, H. et al. (2004) Transmission of H7N7 avian influenza a virus to human beings during a large outbreak in commercial poultry farms in the Netherlands. *The Lancet*, 363, 587–593.

Neal, R. (1998) Regression and classification using Gaussian process priors. In: Bernardo, J.M., Berger, J.O., Dawid A.P. & Smith, A.F.M. (Eds.) *Bayesian Statistics 6*. Oxford: Oxford University Press.

O'Neill, P.D. & Roberts, G.O. (1999) Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A*, 162, 121–129.

Quinonero-Candela, J. & Rasmussen, C.E. (2005) A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6, 1939–1959.

Rasmussen, C.E. & Williams, C. (2006) *Gaussian processes for machine learning*. Cambridge: MIT Press.
Seymour, R.G. (2020) Bayesian nonparametric methods for individual-level stochastic epidemic models. Ph.D. thesis, University of Nottingham.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64, 583–639.
Stegeman, A., Bouma, A., Elbers, A.R.W., de Jong, M.C.M., Nodelijk, G., de Klerk, F. et al. (2004) Avian influenza a virus (H7N7) epidemic in the Netherlands in 2003: Course of the epidemic and effectiveness of control measures. *The Journal of Infectious Diseases*, 190, 2088–2095.
Stockdale, J.E., Kypraios, T. & O'Neill, P.D. (in press) Pair-based likelihood approximations for stochastic epidemic models. Biostatistics. Available at https://doi.org/10.1093/biostatistics/kxz053.
Ypma, R.J.F., Bataille, A.M.A., Stegeman, A., Koch, G., Walling, J. & van Ballegooijen, W.M. (2011) Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B*, 279, 444–450.
Zhang, H. (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99, 250–261.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Seymour RG, Kypraios T, O'Neill PD, Hagenaars TJ. A Bayesian nonparametric analysis of the 2003 outbreak of highly pathogenic avian influenza in the Netherlands. *J R Stat Soc Series C*. 2021;00:1–21. https://doi.org/10.1111/rssc.12515