Waleed Almutiry[1] / Rob Deardon[2]

Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease using Approximate Bayesian Computation

[1] Department of Mathematics, College of Science and Arts, Qassim University,Ar Rass, Qassim, Saudi Arabia, E-mail: wkmtierie@qu.edu.sa
[2] Department of Mathematics and Statistics and Department of Production Animal Health, University of Calgary, Calgary, Alberta, Canada, E-mail: robert.deardon@ucalgary.ca

Abstract:
Infectious disease transmission between individuals in a heterogeneous population is often best modelled through a contact network. However, such contact network data are often unobserved. Such missing data can be accounted for in a Bayesian data augmented framework using Markov chain Monte Carlo (MCMC). Unfortunately, fitting models in such a framework can be highly computationally intensive. We investigate the fitting of network-based infectious disease models with completely unknown contact networks using approximate Bayesian computation population Monte Carlo (ABC-PMC) methods. This is done in the context of both simulated data, and data from the UK 2001 foot-and-mouth disease epidemic. We show that ABC-PMC is able to obtain reasonable approximations of the underlying infectious disease model with huge savings in computation time when compared to a full Bayesian MCMC analysis.

Keywords: contact networks, epidemic models, approximate Bayesian computation, population Monte Carlo, Markov chain Monte Carlo

DOI: 10.1515/ijb-2017-0092

Received: November 15, 2017; **Revised:** September 10, 2019; **Accepted:** November 19, 2019

1 Introduction

The catastrophic threat of infectious disease outbreaks in human, animal, and plant populations has increased the demand for statistical modelling in order to understand the mechanisms behind, and dynamics of, disease transmission. For example, the 2001 UK foot-and-mouth disease (FMD) epidemic resulted in over four million animals destroyed, and brought severe economic consequences for the country as a whole [1, 2]. Gaining an understanding of the transmission dynamics quickly in an outbreak situation can be extremely important, as it can enable stakeholders to better make quick and reliable decisions regarding the control of the spread of the disease.

Infectious disease dynamics are a function of complex interactions between infectious and susceptible individuals. Often, these are modelled via spatial mechanisms; in the case of foot-and-mouth disease (FMD) for example [1, 3, 4]. However, this can often be a proxy for network information that may be unavailable (e. g. movement of animals, individuals and vehicles), and the underlying process of disease transmission can often better explained via a contact network or networks (which may or may not be spatial in nature).

We consider a contact network to be a graph of nodes representing individuals (or group of individuals) and edges (connections) between individuals (or groups) in the population. The edges represent possible transmission paths of a disease. Such networks can be social, or based on trade or movement. Incorporating such networks in disease transmission modelling can enhance the prediction of the epidemic pattern and, thus, aid in controlling the spread of a disease [5, 6]. However, the underlying contact networks for epidemics are often not observed in real life, especially for large population sizes, since collecting accurate and complete contact data can prove difficult, time-consuming and/or expensive [5].

Incomplete data is generally a problem in disease modelling as the infection and recovery event times are rarely observed and so need to be inferred. Such missing data can be accounted for in a Bayesian data augmentation framework using Markov chain Monte Carlo (DA-MCMC), though this often comes at high computational

Waleed Almutiry is the corresponding author.
© 2020 Walter de Gruyter GmbH, Berlin/Boston

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

cost. This problem may be even worse if we have uncertainty regarding the contact network which we also wish to infer.

Many, if not most, published studies involving network estimation in an infectious disease setting seem to focus on modelling the transmission, rather than contact, network (although, Welch et al. [7] an excellent review of methods for estimating contact networks from different types of data). Britton and O'Neill [8] describe modelling the transmission network as a Bernoulli random graph within the context of susceptible-infectious-removed ($\mathcal{SIR}$) homogeneous stochastic epidemic models. They showed that parameters of the network and epidemic process can be successfully estimated under a partially observed disease system (known removal times) using MCMC methods. Groendyke et al. [9] followed their general approach in the context of using a susceptible-exposed-infectious-removed ($\mathcal{SEIR}$) stochastic epidemic model to analyze the 1861 measles outbreak data in Hagelloch, Germany. An extension of this work was also undertaken by Groendyke et al. [10], who used exponential-family random graph models (ERGMs) to describe the pattern of contacts between individuals. Their ERGMs allow for incorporating covariates in the transmission network model. They showed that ERGMs network model better fit the 1861 measles data than a Bernoulli network model. Finally, Sainudiin and Welch [11] developed a biparametric Beta-splitting family model for generating susceptible-infectious ($\mathcal{SI}$) transmission networks without the explicit modelling of the underlying contact network. This is done via maximum likelihood estimation for a number of classes of underlying contact network.

However, our approach focuses not on estimating the transmission network, but incorporating the contact network as a covariate in the model to allow for heterogeneity in the population. When this contact network is unknown, we require an estimate of the whole contact network (i. e. all the connections between individuals) along with all the missing event times. Fitting such models in an MCMC framework can be extremely computationally intensive, due to numerous likelihood function evaluations required.

Deardon et al. [1] introduced a class of discrete-time so-called individual-level models (ILMs) (also known as individual-based models) that can be used to make inferences about the spread of a disease at the individual level in a population. The key feature of these models is the flexibility of incorporating various risk factors at the individual level. Both spatial distance between susceptible and infectious individuals and contact network information can easily be incorporated into these ILMs as risk factors.

One of the most useful features of ILMs is that we can use them to make predictions at a fine level. For example, on a given day of an FMD epidemic we can fit an ILM and predict which farms at highest risk in the near future allowing our model to be used to inform control policies (e. g. [12]). However, to do this fast inference techniques are required.

While ILMs are flexible and intuitive, inference can be computationally problematic, especially for large data sets. This is worse in high-dimensional problems such as when a large amount of missing information needs to be imputed. Although this can be improved using MCMC algorithms such as Hamiltonian Monte Carlo [13], the computational intractability problem still exists in very high-dimensional problems.

Several approaches have been taken to speed up the computation time via likelihood approximation methods when fitting spatial ILMs. Deardon et al. [1] used a kernel linearization technique to do this when fitting $\mathcal{SEIR}$ discrete time ILMs to the UK 2001 FMD outbreak data. They linearized the power-law spatial kernel to allow the splitting of a major component of the likelihood into two parts: one which is computationally expensive but, being independent of the parameters, needs to be calculated only once; and one which is quicker to compute, but depends upon the parameters and therefore needs to be recalculated every MCMC iteration. Other examples for similar settings are given by Malik et al. [14], who developed a data sampling-based likelihood approximation method, and Pokharel and Deardon [15], who used Gaussian process emulation to approximate the ILM likelihood function.

An alternative approach is to use Approximate Bayesian Computation (ABC) methodology, a suite of simulation-based techniques devised for fitting models that have an analytically or computationally intractable likelihood functions [16–20]. Under such an approach, likelihood evaluation is avoided completely, being approximated by comparing observed data to data simulated from the model.

Several studies have used ABC methods to analyze epidemic data. For example, McKinley et al. [21] applied various ABC approaches for fitting a stochastic model to Ebola Haemorrhagic Fever data in the Democratic Republic of Congo. Numminen et al. [22] utilized a population Monte Carlo ABC approach to study the transmission dynamics of Streptococcus Pneumoniae from strain prevalence data. Finally, Walker et al. [23] used so-called partial rejection control ABC for fitting small world network models to study Severe Acute Respiratory Syndrome (SARS) spread in Hong Kong.

The main contribution of this paper is to show the potential for using population Monte Carlo ABC methods (ABC-PMC) for fitting continuous-time network-based ILMs to data when an underlying undirected, binary, spatially-informed contact network is unknown. This is done both in terms of successfully estimating the model parameters, and also predicting epidemic curves. To achieve this, we consider two objectives. First, as a proof of concept, we consider the performance of ABC-PMC in comparison with a full Bayesian MCMC analysis

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

when fitting relatively simple ILMs to simulated data sets with small population sizes. This is done under the assumption the we have some prior knowledge about global characteristics of the network. Second, we use ABC-PMC methods for analyzing larger simulated and real data sets; here, data from the 2001 UK FMD epidemic. In each analysis, we explore the performance of ABC-PMC under a set of summary statistics, seeking sets of summary statistics that produce sound results.

The remainder of the paper is laid out as follows. In Section 2 we introduce the continuous-time individual level models and contact networks considered, along with a description of how epidemic simulation is carried out. In Section 3, both ABC-PMC and MCMC parameter estimation methods considered are described. Section 4 details the epidemic simulation studies carried out, along with a description of the data from the UK 2001 FMD epidemic used to assess the performance of the ABC-PMC methods. The results are described in Section 5, with a final discussion presented in Section 6.

2 Methodology

2.1 General continuous-time ILMs

The general framework of disease transmission model introduced here is a modification of the framework of the individual-level models (ILMs) of Deardon et al. [1]. The key modification here is that we now set the models in continuous-, rather than discrete-, time.

The model framework is defined as follows. First, let λ_{jt} denote the infectivity rate of the susceptible individual j at time t:

$$\lambda_{jt} = \left[\Omega_S(j) \sum_{i \in \mathscr{I}(t)} \Omega_T(i)\kappa(i,j)\right] + \epsilon(j,t) \tag{1}$$

where, $\mathscr{I}(t)$ is the set of infectious individuals at time t; $\Omega_S(j)$ and $\Omega_T(i)$ are functions of potential risk factors associated with susceptible individual j contracting, and infectious individual i transmitting, the disease, respectively; $\kappa(i,j)$ is an infection kernel, a function of risk factors shared between pairs of infected and susceptible individuals; "random appearing" infections may be introduced by the spark term $\varepsilon(j,t)$ (e. g. infection of a susceptible individual by an infectious individual from outside the observed population). As is often the case, we shall assume $\varepsilon(j,t) = \varepsilon$ is a fixed constant.

The infection kernel $\kappa(i,j)$ is a key component of the model that allows for spatial or contact network-based mechanisms. In spatial models $\kappa(i,j)$ is typically power-law function of Euclidean distance. However, for many disease systems, an infection kernel which is a function of a contact network, series of such networks, or indeed, a function of both of distance and network(s), is more appropriate.

Here, ILMs are further placed within the context of an $\mathscr{SIR}$ compartmental framework, although extensions to other frameworks, such as $\mathscr{SEIR}$ compartmental framework, could be straightforwardly made. In an $\mathscr{SIR}$ framework, each individual i in the population is assumed to be in one of the three states at each time point: $\mathscr{S}$ (susceptible), $\mathscr{I}$ (infectious) or $\mathscr{R}$ (removed). Individuals are assumed to be in the susceptible state when they can potentially contract the disease, but are not yet infected. Individuals become infectious as soon as they are infected with the disease and are then able to infect other susceptible individuals during their infectious period. Following their infectious period, an individual transitions to the removed state. Individuals within the removed state no longer have a role in spreading the disease to other individuals, and can no longer be infected. The removed state may represent individuals who have died from the disease, recovered from the disease with an acquired immunity, or have been quarantined in some way.

2.1.1 Likelihood function

We now present the likelihood function for this continuous-time class of ILMs. First, we denote the number of individuals that become infected during the course of the epidemic as m. We consider data sets (infection and removal event times for each individual, as well as any relevant covariates) for complete epidemics only here; that is, the last individual recorded in our data set has been through the infectious period and entered the removed state, and that this individual is the last individual infected in the population. This assumption is also very easy to relax, but fits the scenarios we consider here.

We label the m infected individuals $i = 1, 2, \ldots, m$ corresponding to their infection (I_i) and removal (R_i) times; and the $N - m$ individuals who remain uninfected, we label $i = m+1, m+2, \ldots, N$ with $I_i = R_i = \infty$. We then

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

denote infection and removal times sets for the population as $\boldsymbol{I} = \{I_1, ..., I_m\}$ and $\boldsymbol{R} = \{R_1, ..., R_m\}$, respectively. Then the infectious period of an infected individual i is defined as $\gamma_i = R_i - I_i$. The infectious period is generally assumed to follow some non-negative distribution, typically an exponential or gamma distribution [3].

The likelihood can then be constructed as the product of two parts detailing infection and removal, respectively. Note that, each susceptible individual j that becomes infected would have specific infectious pressure from an infected individual i prior to be infected if and only if $I_i < I_j < R_i$. The likelihood is given as follows:

$$\begin{aligned}
L(\boldsymbol{I}, \boldsymbol{R}|\boldsymbol{\theta}) &= \prod_{j \neq k} \left(\epsilon + \sum_{i: I_i < I_j \leq R_i} \Omega_S(j)\Omega_T(i)\kappa(i,j) \right) \\
&\times \exp\left\{ -\int_{I_k}^{T_{max}} \left(\sum_{i \in \mathcal{S}(u)} \epsilon + \sum_{i \in \mathcal{I}(u)} \sum_{j \in \mathcal{S}(u)} \Omega_S(j)\Omega_T(i)\kappa(i,j)(u - I_i) \right) du \right\} \prod_{i=1}^{m} f(\gamma_i; \delta) \\
&= \prod_{j \neq k} \left(\epsilon + \sum_{i: I_i < I_j \leq R_i} \Omega_S(j)\Omega_T(i)\kappa(i,j) \right) \\
&\times \exp\left\{ -\sum_{i=1}^{m} \left(\sum_{j=1}^{N} ((R_i \wedge I_j) - (I_i \wedge I_j))\Omega_S(j)\Omega_T(i)\kappa(i,j) \right) \right\} \\
&\times \exp\left(-\epsilon \sum_{i=1}^{N} [(T_{max} \wedge I_i) - I_k] \right) \prod_{i=1}^{m} f(\gamma_i; \delta)
\end{aligned}$$

where the wedge symbol $\wedge$ denotes the minimum operator; $\boldsymbol{\theta}$ is the vector of unknown parameters; $f(\cdot\,; \delta)$ indicates the density of the infectious periods; T_{obs} indicates the time of the last infected individual being removed; and k is the label of the initial infective individual with I_k being its corresponding infection time. The integral term which represents the total person-to-person infectious pressure through the epidemic, can be simplified by the product of the two exponential terms in the lower equation [3, 8, 24].

2.1.2 The Network-Based ILM

All the ILMs considered here are network-based, with epidemics spreading between individuals through a single, binary, undirected time-homogeneous contact network. Specifically, we assume $\kappa = c_{ij}$, where

$$c_{ij} = \begin{cases} 1 & \text{if a connection exist between individuals } i \text{ and } j. \\ 0 & \text{otherwise.} \end{cases}$$

We now introduce two specific forms of network-based ILMs considered in this paper.

Simple network-based ILM

We consider a simple network-based ILM for use in a simulated data setting as proof of concept. This simple model contains a single binary susceptibility covariate with no transmissibility covariates (i. e. $\Omega_T(i) = 1$) or spark term (i. e. $\varepsilon = 0$). Specifically, the infectivity rate is given by

$$\lambda_{jt} = \left[(\alpha_0 + \alpha_1 z_j) \sum_{i \in \mathcal{I}(t)} c_{ij} \right] + \epsilon \tag{2}$$

where: α_0 is a baseline infectivity parameter; z_j is a binary covariate that could represent, for example, the vaccination status of individuals; and α_1 is the covariate effect. Since we assume $\varepsilon = 0$, the infectivity rate for an individual with no connections to infectious individuals at time t will be zero.

FMD network-based ILM

Our second network-based ILM is designed for use with data from the outbreak of foot-and-mouth disease (FMD) in the UK in 2001. It is a heavily simplified version of that introduced by Deardon et al. [1] for the same purpose. We consider the individuals in the population to be the farms, rather than animals themselves, and the binary contact network is assumed to represent connections between farms through which infection could potentially occur. Such mechanisms could include animal movements, human or vehicular movement (since

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

the disease can be carried by vectors), contact between farm animals on contiguous farms, pathogen carried in the air and/or wildlife vectors carrying the disease between farms.

The spread of FMD was mainly mostly confined to sheep and cattle farms [2, 25]. Therefore, we consider only sheep and/or cattle farms, and include the number of sheep and cattle on the farm as susceptibility and transmissibility covariates in our model. Specifically, we assume the rate of infectivity of farm j at time t to be

$$\lambda_{jt} = \left[(\alpha_s n_j^s + \alpha_c n_j^c) \sum_{i \in \mathcal{F}(t)} (\phi_s n_i^s + \phi_c n_i^c) c_{ij} \right] + \epsilon \tag{3}$$

where: n_j^s and n_j^c represent the number of sheep and cattle at farm j; and α_s and α_c are the susceptibility parameters, and ϕ_s and ϕ_c are the transmissibility parameters, for sheep and cattle, respectively. We arbitrarily fix α_s in order to avoid identifiability issues [1, 14] so that other susceptibility and transmissibility parameters are interpreted with respect to this baseline value. Here, the constant spark term ε, as well as explaining infections not well explained by the contact network and/or covariates, also allows for infection coming from outside the observed population.

2.2 Contact networks

Networks considered here are undirected with $c_{ij} = c_{ji}$ for $i \neq j$; $i, j = 1, ..., N$, so that each contact network is defined by $\binom{N}{2}$ elements. Here, we confine ourselves to considering only contact networks that are spatial in nature, with connections more likely to be present between two individuals close together than far apart. Specifically, we use a generalized version of the power-law contact networks of Bifolchi et al. [26], in which the probability of a connection between individual i and j is given by:

$$P(c_{ij} = 1) = 1 - \exp(-\nu d_{ij}^{-\beta}), \quad \beta, \nu > 0 \tag{4}$$

where: d_{ij} is the Euclidean distance between individual i and j; β is the spatial parameter; and ν is the scale parameter. In many cases we consider here we set $\nu = 1$ so that the spatial parameter β plays the key role in structuring the contact network. The increase of β (for a given value of ν) leads to more sparse networks, while small values of β tend to produce intense networks with a large number of connections between individuals. Figure 1 shows two simulated contact networks, one relatively intense, and the other relatively sparse, on a population size of 25 individuals randomly distributed in space.

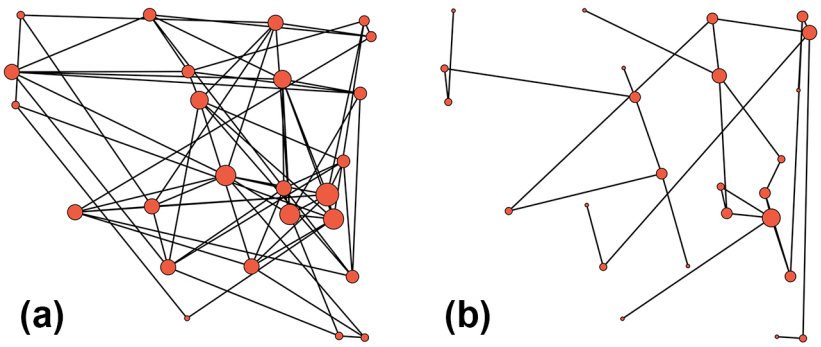


Figure 1: Two samples of spatial contact networks. Red dots represent nodes with different sizes that are corresponding to their degree (number of edges connected).
(a) Intense contact network, $\nu = 1$, $\beta = 0.9$; (b) Sparse contact network, $\nu = 1$, $\beta = 1.8$.

2.3 Epidemic simulation

We consider simulation of epidemics from the models described above in the following way. Each infected individual has an infection life history defined by their time of infection and the length of time spent in the infectious state. We are assuming that infection events follow a non-homogenous Poisson process, so that the time to the next infected individual, given that the last infection occurred at time t, is assumed to follow $W_i \sim \text{Exp}(\lambda_{it})$, where W_i represents the "waiting time" for susceptible individual i becoming infected.

Then, with a given infectious period distribution, an epidemic is simulated across a given contact network, starting with initial infected individual k at time I_k. This is done in the following way. First, we compute the

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

rate of infection λ_{it} for each susceptible individual i at time t. Then we generate $W_i \sim \text{Exp}(\lambda_{it})$. Next we choose the individual with minimum W as the next infected individual and assign $I_{t+1} = I_t + W$ as the infection time for this individual. The infectious period γ_i for this newly-infected individual is generated from $f(\gamma_i; \delta)$ and $R_{t+1} = I_{t+1} + \gamma_i$ assigned as the removal time for this individual. These steps are then repeated until no infectives remain in the population.

3 Model fitting

In most infectious disease data sets, salient information we would wish to observe is missing; for example, infection times and infectious period are usually unknown. It is typical to consider these missing data as latent variables to be inferred. If, additionally, the contact network is unknown, the number of these latent variables becomes even larger, increasing as population size increases.

The estimation of the model parameters, along with the latent variables, is typically carried out within a Bayesian framework using data-augmented MCMC to generate samples from the joint posterior of the model and latent parameters. However, this approach can prove computationally intractable if the population size is large, and, thus, the number of latent variables is large.

In this paper, infection times and the underlying contact network are assumed unknown. However, unlike many situations, we assume the removal times are known. This follows our UK 2001 FMD epidemic motivating case study in which removal times were recorded as the date on which animals on infected farms were culled.

3.1 Approximate Bayesian computation

In an attempt to avoid the computational intractability often associated with an MCMC approach in high dimensional problems, we consider the use of (increasingly popular) approximate Bayesian computation (ABC) methods for fitting network-based ILMs to data. In such methods, the likelihood function is approximated by simulating data sets from the model and comparing them to the observed data in some way.

Early variants of ABC were based upon rejection sampling [19]. However, more efficient sequential ABC methods have since been developed, examples of which are Partial Rejection Control (PRC) [20], Sequential Monte Carlo (SMC) [27] and Population Monte Carlo (PMC) [16]. For the purpose of this study, we will focus on the ABC-PMC method.

3.1.1 ABC-PMC algorithm

Beaumont et al. [16] introduced an adaptive ABC-PMC algorithm that was inspired by the Population Monte Carlo algorithm of Cappé et al. [28]. The ABC-PMC algorithm is shown in Algorithm (1) in the Web Supplementary Material.

The ABC-PMC algorithm is based on sampling $\boldsymbol{\theta}$, called particles, from a sequence of approximate posteriors (importance densities) $q_t(\theta)$ with decreasing tolerance levelsξ. At each generation of the algorithm, new particles are sampled from the previous generation. This is typically by sampling from the importance density as given by:

$$q_t(\theta) = \begin{cases} \pi(\theta) & \text{if } t = 1, \\ \sum_{i=1}^{N} w_i^{(t-1)} K(\theta_i^{(t-1)}, \theta) / \sum_{j=1}^{N} w_j^{(t-1)} & \text{otherwise.} \end{cases}$$

Sampling from $q_t(\theta)$ is done by sampling from the prior distribution $\pi(\theta)$ in the first generation, and from the particle set $\theta_i^{(t-1)}, i = 1, \ldots, N$ with probability $w_i^{(t-1)}$ for the subsequent generations. The sampled particles are then perturbed by a proposal distribution $K(\cdot|\cdot)$, that is adapted at each generation t based on the accepted particles of generation $t-1$. Typically a multivariate normal distribution is used as the proposal kernel with an adaptive scale Σ set to be twice the weighted covariance of $\theta_j^{(t-1)}, j = 1, \ldots, N$ [16, 29]. The sampled particles are then accepted or rejected as in Algorithm (1) in the Web supplementary materials, and weights are then assigned to the accepted particles. The ABC-PMC algorithm is stopped at generation t when there is no significant difference between the distribution of particles between generations t and $(t-1)$.

Sufficient summary statistics are usually unavailable for models with intractable likelihoods, and summary statistics (S in Algorithm (1); see the Web supplementary materials) are generally chosen in a way that balances the need for high levels of information with low dimensionality, in order to gain as good an approximation

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

to the likelihood function as possible [30]. However, the difficulty of defining those summary statistics can be increased when important information regarding the epidemic process is missing, since this missing information must be modelled too (e. g. in epidemic models a delay between the unobserved infection time and the observed reporting time), and we require summary statistics that capture all aspects of the model. Of course, this difficulty can be alleviated if we have strong prior information about some aspects of the model we can incorporate into our analysis (e. g. through the prior distribution).

Considering a k dimensional summary statistics $S = (s_1, s_2, \dots, s_k)$, the discrepancy between summary statistics of the simulated $S^{(sim)}$ and observed $S^{(obs)}$ data are usually assessed by the L^1-norm or L^2-norm (Euclidean distance). Here, the Euclidean distance used:

$$\rho(S^{(sim)}, S^{(obs)}) = \left[\sum_{i=1}^{k} (s_i^{(sim)} - s_i^{(obs)})^2 \right]^{1/2}$$

The sequence of tolerances $\xi^{(t)}$ is adaptively determined in the algorithm. The algorithm starts with initial tolerance level $\xi^{(1)} = \infty$ at $t = 1$, resulting in the acceptance of all the sampled parameters from the prior distributions. Then, the tolerance level of the next generation is set adaptively, by considering the $\boldsymbol{\alpha}$ quantile of $\rho_1^{(t)}, \rho_2^{(t)}, \dots, \rho_N^{(t)}$, where these are the $\rho(S^{(sim)}, S^{(obs)})$ distances from the accepted simulations at time t [22, 31] ($\boldsymbol{\alpha} = 0.4$ is assumed throughout the paper).

As we are assuming unknown contact networks in our analyses, generating a contact network is essential for using ABC-PMC to fit the network-based ILM. Therefore, for each epidemic simulated within the ABC-PMC algorithm a contact network is also generated from eq. (4). Therefore, we augment our parameter vector to estimate the unknown parameters of the contact network model as part of our ABC analyses.

The specific summary statistics used in our analyses are described in Section 4.

3.2 Data augmentation MCMC

In order to assess performance of the ABC-PMC algorithm, the simple network-based ILM is fitted using full Bayesian data-augmented MCMC (DA-MCMC) methodology. Specifically, the Metropolis Hastings algorithm is used to update both the model parameters and latent variables representing the unobserved infection times and contact network.

As discussed in Section 1, we assume that we have global information about the contact network. Specifically here, we assume we know the total number of connections of our network, ϕ_{tot}, where $\phi_{tot} = \sum_{j=1}^{N} \phi_i = \sum_{j=1}^{N} \sum_{i=1}^{N} c_{ij}$. This information can be estimated in a number of ways. For example, an estimate of ϕ_{tot} can be extracted by studying individuals' contacts in small regions via surveys, and then generalized to represent the whole population. For example, see Eames et al. [32]. Then, given independent gamma distribution priors $\pi(\alpha_0)$, $\pi(\alpha_1)$, $\pi(\varepsilon)$ and $\pi(\delta)$, for α_0, α_1, ε, and δ, respectively, the posterior density up to proportionality for the simple network-based ILM is given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}, \boldsymbol{Z}) \propto L(\boldsymbol{R}, \boldsymbol{I}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{\theta}) \pi(\alpha_0) \pi(\alpha_1) \pi(\epsilon) \pi(\delta) \pi(\phi_{tot}).$$

The conditional distribution of the infectious period rate parameter δ can be shown to be a gamma distribution:

$$\delta|\alpha_0, \alpha_1, \epsilon, \boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}, \boldsymbol{Z} \sim \Gamma(m + a_\delta, M + b_\delta)$$

where, $M = \sum_{i=1}^{m} (R_i - I_i)$ and a_δ and b_δ are the rate and shape parameters of the prior distribution of δ, respectively. Therefore, a Gibbs update (i. e. sampling from the conditional posterior distribution) is used to update δ. An independence sampler is used to update infection times, proposing new infection times I_i^* from the infectious period distribution, specifically $\gamma_i \sim \text{Exp}(\delta)$. Then, the new infection time is just the difference between the fixed known removal time and the new infectious period of the ith individual. Each infection time/infectious period is updated in this way in turn. Each element of the contact network ($c_{ij} = c_{ji}$) is also updated in turn. As c_{ij} has only two states, 0 (no connection) or 1 (connection), these parameters are updated by proposing a move from the current state to the only other possible state with probability 1. α_0, α_1 and ε are updated in turn, using Gaussian random-walk updates, tuned to achieve good mixing properties.

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

4 Epidemic data and summary statistics

4.1 Simple network-based data

Here, we consider the objective of comparing the performances of the ABC-PMC and MCMC approaches for small, simulated data sets. We compare the two approaches within two scenarios: one in which the contact networks are fairly dense, and one in which they are fairly sparse. The data sets considered here consist of a population of size $N = 25$ in which individuals are distributed uniformly at random across a square area of size 10×10 units. Two contact network scenarios are considered. In one scenario $\beta = 0.9$ and $\nu = 1$ (intense contact network scenario), and in the other $\beta = 1.8$ and $\nu = 1$ (sparse contact network scenario). Ten contact networks were generated for each scenario. Then, an epidemic was simulated across each of the 20 contact networks using the simple network-based ILM, resulting in 10 data sets for analysis under each scenario. Epidemics were generated with model parameters, $\alpha_0 = 0.8$, $\alpha_1 = 0.5$, $\delta = 2.0$ and $\varepsilon = 0$. It was assumed here that our population is isolated, hence the choice of $\varepsilon = 0$. Other sets of model parameters were considered, and produced results in line with those shown in Section 5 (results not shown).

ABC-PMC was implemented for each epidemic data set using various sets of summary statistics to find which provide good approximations to results under the MCMC analysis. For brevity, three sets of such statistics are included here, some of low dimension and some of high dimension (shown in Table 1). Following McKinley et al. [21], we based our choice of summary statistics primarily on the removal and infection epidemic curves (i. e. numbers removed and infected over time, respectively). We found the epidemic curves to be crucial for capturing information about the epidemic dynamics. However, we were also interested to see if information about these dynamics could be captured by summary statistics of the epidemic curves such as the length of the epidemic, final number of infections, etc. Thus, we tested a variety of summary statistics sets with varying dimension.

Table 1: The sets of summary statistics used, where $P_{(\cdot)}$ denote the time at the peak of the epidemic curve (R removal and I infection times), n_I is the number of infected individuals, L_{epi} is the length of epidemic, $\mu_{(\cdot)}$ and $\sigma_{(\cdot)}$ are the mean and standard deviation of the infection or removal times, $\mathcal{I}_0^T$ is the number of infection events over equal time intervals; and $\chi^2_{s,c}$ is the chi-squared statistic as described in Section 4.2.1.

Simple Network-based data			Simulated FMD data		Real FMD data	
ABC-simple (1)	**ABC-simple (2)**	**ABC-simple (3)**	**ABC-sim (1)**	**ABC-sim (2)**	**ABC-real (1)**	**ABC-real (2)**
P_R	P_R	P_R	μ_R	μ_R	P_R	P_R
L_{epi}	L_{epi}	σ_R	σ_R	σ_R	L_{epi}	σ_R
σ_R	$\mathcal{I}_0^T$	P_I	μ_I	$\mathcal{I}_0^T$	$\mathcal{I}_0^T$	P_I
n_I		σ_I	σ_I	$\chi^2_{s,c}$	n_I	σ_I
		n_I	n_I		$\chi^2_{s,c}$	$\chi^2_{s,c}$
			$\chi^2_{s,c}$			n_I

We start with a low dimensional set of summary statistics, ABC-simple(1), based mainly on the removal times. This set includes scalar summary statistics that fully represent the removal time curve. As the removal times of most of the data sets have right-skewed distribution, we use the time at the peak of the removal times density rather than the mean. The time at the peak of the removal times curve is computed through estimating the kernel density of the removal times using Gaussian kernel smoothing, with bandwidth $h = 1.06\ \hat{\sigma}_R\ n^{-1/5}$, where $\hat{\sigma}_R$ is the standard deviation of the removal times [33].

Information regarding the infection time curves are also included in the second and the third sets. We considered a high dimensional set of summary statistics (ABC-simple(2)) that incorporates the numbers of infection events occurring within a set of equal time intervals of 0.2 time units ($\mathcal{I}_0^T$) from $t = 0$ to $t = T$ instead of the standard deviation of the removal times (σ_R) and the number of infected individual (n_I). In the third group of summary statistics (ABC-simple(3)), we consider scalar summaries of $\mathcal{I}_0^T$, thus, forming a representation of ABC-simple(2) in lower dimensions.

The DA-MCMC method was also used to carry out a full "gold standard" Bayesian analysis of each epidemic data set, sampling from the joint posterior of the model parameters, infection times, infectious periods and the upper triangular (since the network is undirected) contact matrix. The degree distribution of the network is incorporated into the analysis via an observation model for the total number of connections (ϕ_{tot}), as described in Section 3.2. For this observation model, we use a normal distribution with mean equal to the true number of connections and variance 10.

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

Although our intention here is to fit a model with $\varepsilon = 0$, for the MCMC-based analysis ε was included as a free parameter to aid with MCMC mixing. This done to avoid high rejection rates resulting from many MCMC moves being proposed with zero likelihood (e. g. if a proposed contact network change results in an infected individual now having no contacts with anybody infectious individuals under the model). However, we fix $\varepsilon = 0$ under ABC-PMC since the above issue is not exist with ABC-PMC.

For all data sets, independent prior distributions were assigned to all model parameters. The marginal prior distributions for the common model parameters under both ABC-PMC and DA-MCMC were as follows: $\alpha_0 \sim \Gamma(2, 2)$, $\alpha_1 \sim \text{Exp}(2)$, and $\delta \sim \Gamma(4, 2)$. Under the ABC-PMC, the scale parameter of the contact network is fixed at $\nu = 1$, and a uniform prior assigned to the spatial parameter β such that $\beta \sim \text{U}(0, 5)$. Also, we assigned a non-negative, half-normal prior distribution with a variance of 100 and a mode of 0 for the spark term ε under the DA-MCMC. The ABC-PMC was run sampling 500 particles at each generation. The DA-MCMC was run for 150,000 iterations with a burn-in period of 10,000 iterations. The convergence of the both the ABC-PMC and MCMC algorithms was diagnosed visually, monitoring the sample output chains.

4.2 The UK 2001 foot-and-mouth disease

Here, we consider the performance of ABC-PMC on a more complex network-based ILM in a larger population. Specifically, we consider modelling data from, or based upon, the UK 2001 foot-and-mouth disease epidemic.

The full UK 2001 FMD data set consists of X and Y ordnance survey locations for the farmhouse of the farms, and the number of cattle and sheep on each farm recorded at the census carried out previous to the FMD outbreak. All farms on which FMD was detected had an infection time estimated (by veterinarians/veterinary epidemiologists on the ground), as well as a removal time recorded (the date on which animals were culled as part of the control policy). A further set of farms on which the presence of the infection was not confirmed, but which had their animals culled as part of the "high-risk farm" control policy, are also recorded along with the date of the cull. It is assumed that no farms which had animals infected with FMD went undetected.

Foot and mouth disease has been extensively modelled, in many cases assuming an $\mathcal{SEIR}$ framework (e. g. [1–3]). Our model is simplified in comparison, primarily since we follow Malik et al. [14] and Deeth et al. [34] and assume the simpler $\mathcal{SIR}$ compartmental framework (extensions to other frameworks could be made relatively easily). We consider subsets of farms in the county of Cumbria, one of the counties most highly affected by the 2001 FMD outbreak.

4.2.1 Simulated FMD data

First, we consider simulating epidemics from the FMD network-based ILM of eq. (3) through a population consisting of 674 farms, those being all the farms within a region of 25 × 30km; see Web Figure 1. The data consist of the number of cattle and sheep on each farm, and an XY coordinate.

Simulated data sets were produced by first simulating a contact network from eq. (4) with spatial parameter $\beta = 4$ and $\nu = 1$. The value of $\beta = 4$ was chosen to produce relatively sparse networks, as the network structure of most real contact networks is typically sparse in nature [5]. Then, an epidemic was simulated through the contact network. The infectious period for each individual was simulated from a gamma distribution with shape parameter 10 and rate δ. A number of scenarios were tested, but here we show results for parameters $\alpha_s = 0.0001$, $\alpha_c = 0.01$, $\phi_s = 0.001$, $\phi_c = 0.005$ and $\delta = 1.5$. Conclusions from other scenarios were similar to these shown here.

Two scenarios regarding the spark term were considered. First, in Scenario 1, ten epidemics were simulated from the ILM with $\varepsilon = 0$, thus, assuming that no infections come from outside the observed population. Second, in Scenario 2, ten epidemics were simulated from an ILM with $\varepsilon = 0.001$, allowing for infections from outside the observed population.

The FMD network-based ILM was fitted to the simulated FMD epidemic via ABC-PMC with two different sets of summary statistics (see Table 1). We found that stratifying the number of infected farms (n_I) by farm size and type helped to inform the contact network parameter. The summary statistic used was the following chi-squared statistic: $\chi^2_{s,c} = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$ where O_i and E_i are the observed and expected number of infected farms in the ith stratified class.[1] The four classes are defined by the number of sheep and cattle upon the farms. First, let m_c and m_s denote the median numbers of cattle and sheep on farms in the observed population, respectively. Then, large cattle farms are defined as having more than m_c cattle, but fewer than m_s sheep; large sheep farms are defined as having more than m_s sheep, but fewer than m_c cattle; large mixed farms are defined as having more than m_c cattle and m_s sheep; and the final class, small mixed farms, are defined as having fewer than m_c cattle and m_s sheep. This $\chi^2_{s,c}$ is then calculated for both the observed and ABC-simulated data. We present

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

results for a set of six summary statistics (ABC-sim(1)), as well as a higher dimensional second set (ABC-sim(2)) containing $\mathcal{F}_0^T$.

For all analyses, we assume the scale parameter of the contact network is fixed at $\nu = 1$ and estimate unknown spatial parameter β. Independent prior distributions were assigned to the model parameters, as follows: α_c, ϕ_s, ϕ_c and $\varepsilon \sim \text{Exp}(100)$, $\delta \sim \Gamma(3, 2)$ and $\beta \sim \text{U}(0, 8)$. Since the model of Equation (3) as given is non-identifiable, following Deardon et al. [1] we fix α_s. For convenience, here we set it equal to its true value of 1×10^{-4}. The ABC-PMC was run with a sample size of 500 particles at each generation.

Note that, although the population size of 674 individual farms is still relatively small, a data augmented MCMC analysis to estimate the joint posterior of the model parameters and contact network proves computationally intractable and so could not be carried out.

4.2.2 Real FMD data

Finally, we consider fitting the FMD network-based ILM to real data, consisting of a subset of 1795 farms observed during the 2001 UK FMD epidemic within a region of 58.4 × 52.5km; see Web Figure 1. This data set includes 456 farms which were infected during the epidemic. We treat the cull dates of animals on these farms as the removal times. Connections between farms in the contact network represents any potential infection route (e. g. contact between animals and/or farm workers, transfer of infective material by vectors such as wildlife or vehicles, etc.). We fit our model to data from the date of animal movement ban introducted on 23 February, conditioning on the infection status of farms at that date. That is, all farms deemed infectious on 23 February we consider to be "initial infections". We assume known infection times (estimated by vets on the ground) and removal times when fitting models to this data set. Since simulation of the "high risk" farm culling procedure implemented in 2001 is extremely difficult, we do not model this aspect of the epidemic here. Instead, we follow Jewell et al. [3], and assume all these "high risk" farms are susceptible throughout the epidemic.

In this work, we once again present our investigation of the performance of ABC-PMC using two sets of summary statistics. Table 1 shows the two chosen sets, one of high, and one of low, dimension. The high dimensional set of summary statistics (ABC-real(1)) consists of P_R, L_{epi}, n_I, $\chi^2_{s,c}$ and $\mathcal{F}_0^T$ at equal time intervals of 5 days. The lower dimensional group of summary statistics (ABC-real(2)) is similar to ABC-real(1), but with the infection time counts being replaced by P_I and σ_I. The inclusion of the stratification procedure for the number of infected farms $\chi^2_{s,c}$ by their farm size was inspired by the findings of Jewell et al. [3] and Deardon et al. [1], both who found a strong effect of farm size and type on both susceptibility and infectivity.

The ABC-PMC was performed with fixed scale parameter $\nu = 1$ for both sets of summary statistics. Independent prior distributions are assigned to the model parameters as follows: $\beta \sim \text{Exp}(0.3)$, α_c, ϕ_s and $\phi_c \sim \text{Exp}(100)$ and $\delta, \varepsilon \sim \text{Exp}(10)$. Once again, to solve the issue of non-identifiability, we fixed α_s (arbitrarily) to 1×10^{-7}. Additionally, analyses were carried out in which ν was treated as unknown. We show such results for ABC-real(1). In this case independent, marginal prior distributions were assumed for the contact network parameters such that $\beta, \nu \sim \text{Exp}(1)$. The other model parameters have the same prior distributions as above.

5 Results

In this section, we present the results of our analyses. The independence of sampling in the ABC-PMC algorithm at each generation allows the computation to be run in parallel to achieve lower computation times. Parallel computation was implemented using a distributed-memory architecture by the Message Passing Interface Standard (MPI), and running on multiprocessors servers. This was done for ABC-PMC for the FMD network-based ILMs. Note that the used servers have restricted limit of time for running jobs to 168 hours. Computations for the simple network-based ILMs were performed on an Apple iMac with i5-core Intel 2.9 GHz processors with 16 GB of RAM without parallelization. All coding was done in Fortran 95.

5.1 Simple network-based ILM data

Figure 2 shows the (approximate) posterior means of the ILM parameters with 95 % credible (percentile) intervals, under both the full Bayesian MCMC and ABC-PMC analyses, and under both intense and sparse contact network scenarios. The results shown for the ABC-PMC analysis are for the ABC-simple(3) summary statistics.

Under the ABC-PMC analysis, we observe good approximate posterior estimates of the model parameters that tended to be close to their true values. Under both sparse and intense networks, we appeared to be able

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

to estimate the contact network spatial parameter quite well. However, we observe better estimates of this parameter under the intense network data sets, with systematic underestimation and a larger posterior variance under the sparse networks. These results show good performance of the ABC-PMC under the chosen summary statistics ABC-simple(3).

The performances under the other two sets of summary statistics are displayed in Web Figure 2 in the supplementary materials. Their performances were similar to ABC-simple(3), with true values of parameters being included within 95 % credible intervals throughout. However, we found that including more information from the infection times in the summary statistics, (ABC-simple(2) and ABC-simple(3)), appeared to slightly enhance estimation of the parameters with tighter credible intervals, especially for α_0 and α_1. The approximate posterior estimates of α_0 and α_1 were closer to their true values under ABC-simple(2) and ABC-simple(3) than ABC-simple(1), where overestimation of both parameters was observed. Thus, the inclusion of the infection times in the summary statistics appeared to enhance the estimation of these parameters. This can be seen most clearly for α_0 with approximate posterior means closer to its true value, and with less variation under ABC-simple(2) and ABC-simple(3) for both intense and sparse networks. The infectious period δ estimates were similar under each of the three sets of summary statistics.

Under all of the three ABC-simple analyses on both type of network data sets, we have found clear correlation between the estimated of the contact network that represented by its spatial parameter β and the baseline infectivity parameter α_0, and to a lesser extent the infectious period rate δ. Web Figure 6 in the supplementary materials shows the pairwise posteriors of the model parameters for two data sets (one intense and one sparse network data set) based on the ABC-PMC analyses using ABC-simple(3) summary statistics. Strong correlations between β and α_0 were observed under both type of networks. A much less pronounced correlation is also observed between β and the infectious period δ of around - 0.3 under both networks. Correlations between other pairs of parameters are close to negligible.

Under the full Bayesian MCMC analysis, more accurate estimation of α_0 and α_1 with tighter credible intervals under both types of networks was achieved than under ABC-PMC. However, the infectious period rate δ had wider credible intervals than under ABC-PMC with a lot of variation in performance between data sets. The estimated degree distributions under both types of networks were quite close to the observed ones; see Web Figure 3. This indicated the benefits of incorporating the observational model of the total number of connections in the update of each c_{ij}. The analyses were also carried out without the observation model and large bias estimate of the degree distribution were detected. In both types of network, the estimated degree distribution tended to be overestimated resulting in very intense contact network.

Both MCMC and ABC methods generally produced good estimates of the model parameters under both types of network. However, as expected, approximate posterior variances were larger under ABC-PMC than the estimated posterior variances under the full Bayesian MCMC analyses.

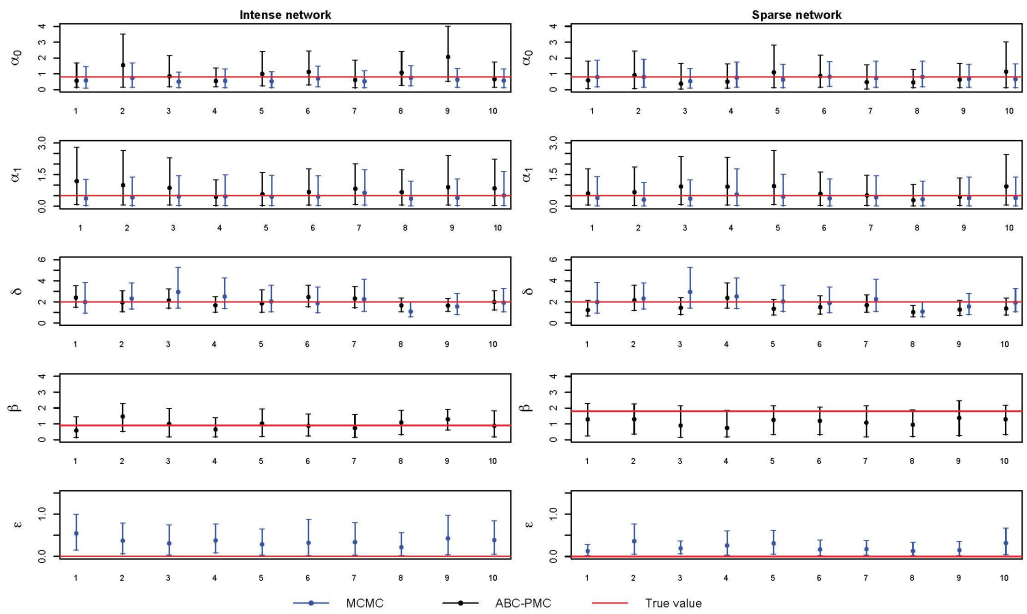


Figure 2: (Approximate) posterior means for the simple network-based ILM with 95 % credible intervals, for intense (left) and sparse (right) contact network epidemics, under both the full Bayesian MCMC and ABC-PMC analyses, respectively. Horizontal dotted lines show the true values of the parameter; $\alpha_0 = 0.8$, $\alpha_1 = 0.5$, $\delta = 2$, $\beta = 0.9$ (intense network) and 1.8 (sparse network), and $\varepsilon = 0$.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

5.2 Simulated FMD data

We examine the performance of ABC-PMC with low and high dimensional summary statistics. Figure 3 shows the resulting approximate posterior means and 95 % credible intervals of applying the ABC-PMC method on the simulated FMD data sets. Results, in general, indicate good estimation of the model parameters for most of the epidemics. Both low and high dimensional summary statistics provided similar estimates of the model parameters, with the true values of the model parameters contained in the 95 % credible intervals in most instances. The approximate posterior variance was noticeably higher under the high dimensional summary statistics ABC-sim(2), however. Model parameters estimates were reasonably good under both scenarios considered; Scenario 1 in which no spark term was used in either the fitted or epidemic data generating model, and Scenario 2 in which $\varepsilon = 0.001$ was used in the generating model and was estimated as part of the ABC analysis.

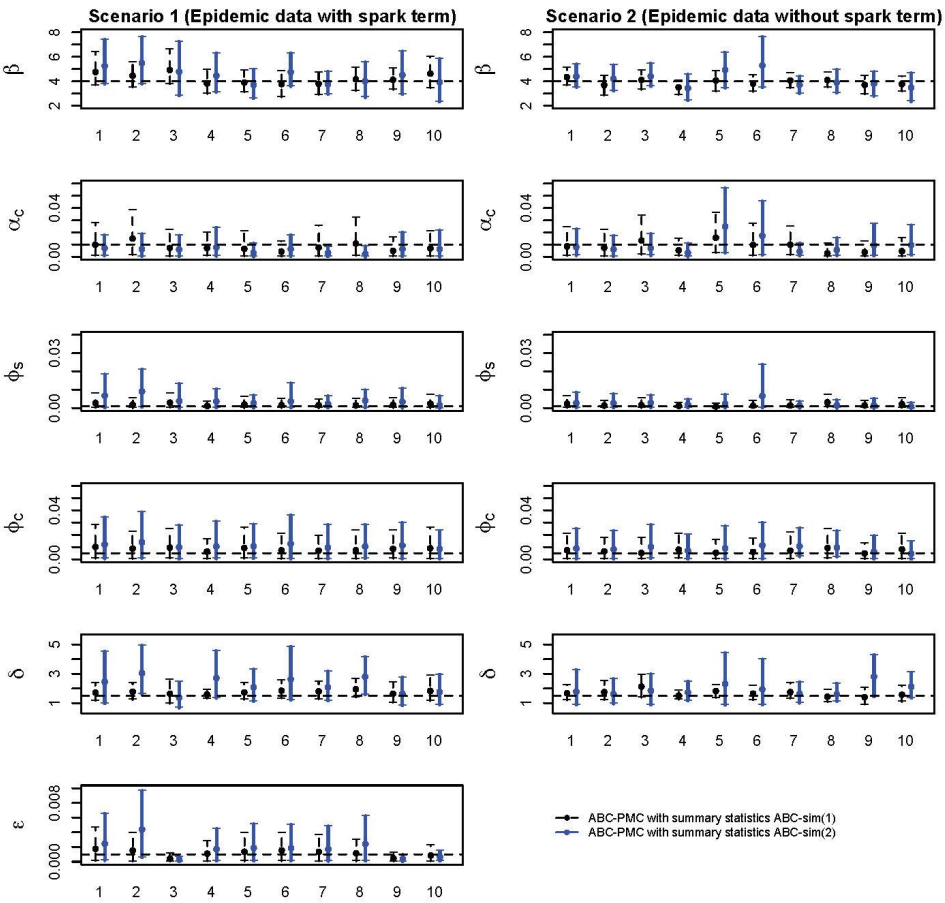


Figure 3: Posterior means and 95 % credible intervals for fitting the FMD network-based ILM to the simulated FMD epidemics with the spark term (left, scenario 1) and without spark term (right, scenario 2). The black dotted lines represent results of applying ABC-PMC with summary statistics ABC-sim(1) and blue solid lines of applying ABC-PMC with summary statistics ABC-sim(2); as described in Section 4.2.1. Horizontal dotted lines are the true values of the parameters; $\beta = 4, \alpha_c = 0.01, \phi_s = 0.001, \phi_c = 0.005, \delta = 1.5$ and $\varepsilon = 0.001$.

We also achieved good estimation of the latent information, both infection times and contact networks. The approximated posterior predictive distributions of the infection time density of some representative simulated FMD epidemics are displayed in Web Figure 4 for both groups of summary statistics. As observed with the parameter estimates, we observe much greater variation under the ABC-sim(2) analysis (high dimensional summary statistics) than under ABC-sim(1) (low dimensional summary statistics).

5.3 Real FMD data

Approximate posterior means and 95 % credible intervals of the model parameters for the real FMD epidemic data analyses are shown in Table 2. Although there is reasonable agreement between some parameters (e. g. the sparks term and infectious period rate parameters), with substantially overlapping credible intervals, there are

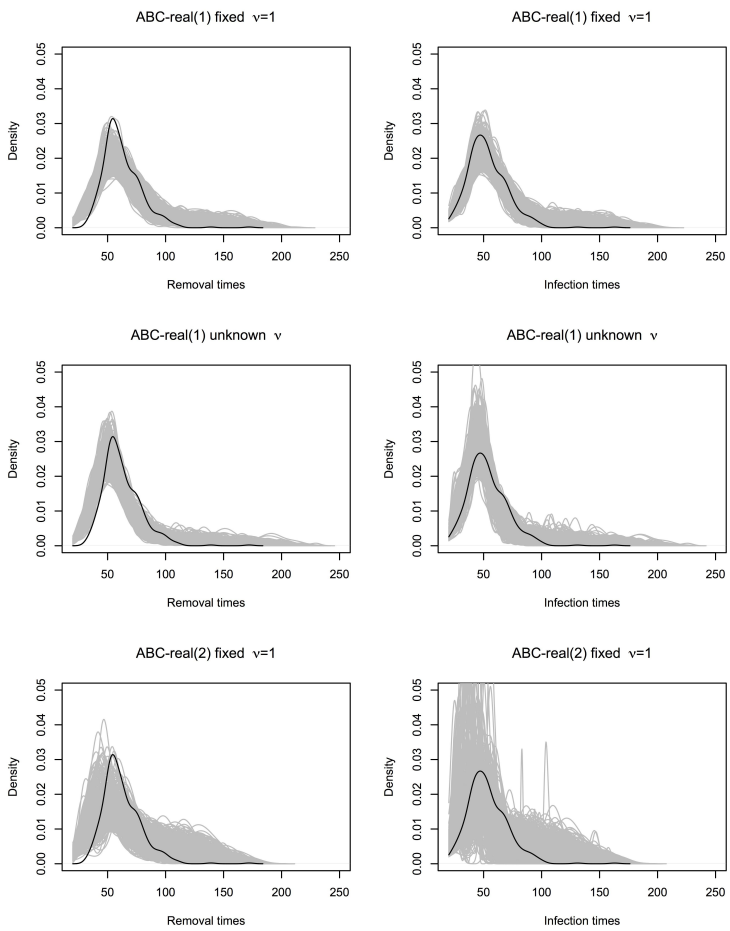
Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

also some noticeable differences in the three analyses. This is perhaps most obvious with the spatial network parameter, β, with a much larger estimate of this parameter for the summary statistics ABC-real(2). These larger estimates of β imply a very sparse underlying network is being inferred in which connections between farms occur predominantly over very short distances (e. g. less than 1 KM).

Table 2: Posterior means and 95 % credible intervals for parameters for the FMD network-based ILM fitted to the real FMD epidemic data.

Parameters	Units	ABC-real(1)		ABC-real(2)
		ν known	ν unknown	ν known
β		3.012 (1.253, 5.871)	2.840 (0.914, 5.645)	12.731 (4.932, 23.799)
ν		1.0	0.622 (0.053, 1.970)	1.0
α_c	10^{-4}	2.354 (0.077, 11.165)	7.664 (0.274, 34.901)	45.206 (3.334, 143.498)
ϕ_s	10^{-3}	3.273 (0.035, 13.922)	5.822 (0.125, 22.476)	11.242 (0.265, 34.315)
ϕ_c	10^{-2}	1.042 (0.086, 2.880)	0.912 (0.039, 2.518)	0.795 (0.047, 2.243)
δ	10^{-1}	1.788 (0.804, 2.969)	1.717 (1.097, 2.289)	1.475 (0.575, 3.112)
ε	10^{-4}	2.092 (0.071, 5.769)	3.206 (0.188, 8.964)	1.365 (0.123, 5.231)

Figure 4 shows realizations from the ABC-approximated posterior predictive epidemic curves (infection times and removal times) from the three ABC-PMC analyses. We can see that prediction of these curves was better under the two analyses of ABC-real(1) than ABC-real(2), with much lower variance. This would imply that the estimates under ABC-real(1) are more reliable.



Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

Figure 4: The approximated posterior predictive distribution of the infection and removal time curves for the real FMD epidemic data analyses.

Web Figure 5 shows the intermediate posterior distribution in sequential generations of the ABC-PMC, using ABC-real(1) with fixed ν. We can see that the analysis is informative, with the final approximate posterior distributions appearing quite different to the prior distributions. The prior distributions chosen were relatively informative. Similar analyses were carried out with less informative priors, and the approximate posterior distributions found to converge to approximately the same distributions as seen here. However, this was accompanied with a slower convergence rate. (This is discussed further Section 5.4).

Since we are using a relatively small subset of the 2001 UK FMD epidemic data set, a direct comparison of our results to previously published studies (e. g. [1–3]), which were based on the whole of the UK or much larger subsets, is difficult. However, the approximate posterior means of the parameter values based on the ABC-real(1) analysis show that individual cattle were both more likely to transmit the disease, and were more susceptible to the disease, than sheep. These results are qualitatively similar to those reported in Deardon et al. [1], Jewell et al. [3], and Diggle [35].

Moreover, the estimate of the network spatial parameters under ABC-real(1) are also similar to the spatial infection kernels estimated in those studies. Figure 5 shows the posterior mean and 95 % credible intervals of the probability of a connection within the farm network under the fitted model assuming both known and fixed, and unknown, ν. These results imply a very small chance of connections within the network between farms that were 4 KM apart or greater, which is in agreement with these other published works. This fact, along with reasonably good posterior prediction of the epidemic curves provide some evidence that results we observe are reasonable. Whether treating the scale parameter of the contact network as unknown, or fixing $\nu = 1$, similar posterior estimates of the model parameters results were observed. Of course, fixing $\nu = 1$ forces $p(c_{ij} = 1) = 1 - e^{-1}$ when $d_{ij} = 1$ and leads to tight credible intervals for d_{ij} near to 1 as can be seen in the magnified part of Figure 5. However, this does not appear to have much effect on the model fit as the proportion of distances (≤ 1 KM) between farms is very small. On the other hand, varying ν added more flexibility to the probability of connections within farms in short distances, but with larger posterior variation.

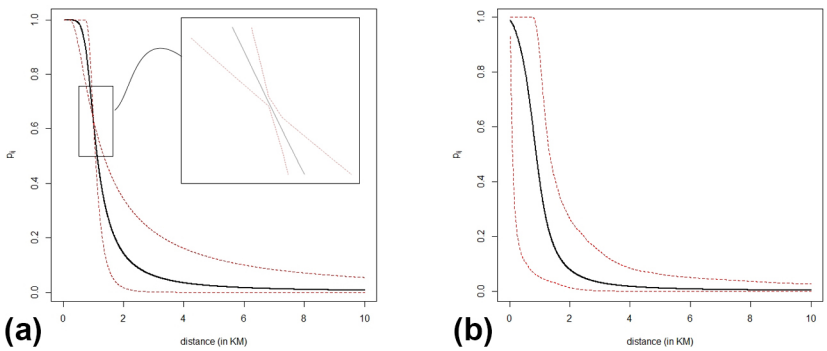


Figure 5: Posterior mean and 95 % credible intervals of the probability of connections p_{ij} between farms in the inferred network for the real FMD data using summary statistics ABC-real(1) based on the posterior mean of the contact network parameters.
(a) Fixed $\nu = 1$, (b) Unknown ν.

5.4 Computation time

Table 3 shows the computation time using unparalleled ABC-PMC and DA-MCMC algorithms to fit the network-based ILM to a single simple network-based data set, and a simulated FMD epidemic data. The computation time is for obtaining 150,000 DA-MCMC samples and 500 particles at each generation of ABC-PMC (15 generations). The computational time of running DA-MCMC becomes infeasible as the population size increases and is approximated from running a single MCMC iteration. This rough approximation was used since the number of missing connections increases drastically with increases in population size, and as a result, updating each connection via DA-MCMC requires evaluating the likelihood a huge number of times at each iteration.

Table 3: The computational time (in minutes) of using both DA-MCMC and ABC-SMC on fitting network-based ILMs for a simple network-based data (N = 25) and a simulated FMD epidemic data (N = 674).

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Population size	Number of c_{ij}	Computation time in minutes	
		DA-MCMC	ABC-PMC
N = 25	300	14.54	44.41
N = 674	226801	$\approx 6.22 \times 10^7$	9807.95

Using DA-MCMC took approximately three times less than the ABC-PMC with small data contains N=25 individuals. However, as the population size increases, ABC-PMC becomes increasingly more efficient than DA-MCMC. Using ABC-PMC with the simulated FMD epidemic contains N=674 individuals took 9807.95 minutes while DA-MCMC had infeasible computation time. The computation time of applied parallel computing on the same simulated FMD epidemic was 629.67 minutes with 12 processors. Parallel computing was also performed for fitting the model to real FMD epidemic for both groups of summary statistics, with 168 hours and 64 processors.

6 Discussion

The main focus of this paper has been fitting network-based ILMs of disease transmission in situations where there is uncertainty both about infection times and the underlying content network. Our results suggest that ABC-PMC approaches to inference in such scenarios can produce good estimates of model parameters and fitted models with good predictive qualities, and also obtain good estimations of the underlying unknown contact network. However, this area of research is a work in progress, and the positive results we have achieved here do come with some caveats that require further exploration.

The first caveat pertains to the selection of summary statistics used to compare simulated and observed data within the ABC algorithm. We have explored the use of numerous potential combinations of summary statistics for fitting our models, only a small subset of which have been included here. In many situations that summary statistics considered here might be hard to be obtain. For example, we may have underreporting and have to include this mechanism in the model. However, for epidemics such as the UK FMD outbreak of 2001 having such summary statistics is entirely plausible. Selecting an "optimal" set of summary statistics for a given analysis is not easy. For example, in some situations it appeared that including a discretized version of the epidemic curve produced more satisfactory results than a more parsimonious set of summary statistics of the epidemic curve, and sometimes vice versa. It is still not clear how we would go about predicting which would work better in advance. This would suggest that either multiple ABC-PMC analyses should be carried out (preferably) in parallel when analyzing epidemic data, or some approach that allows for the comparison of different sets of summary statistics to be carried out as part of the analysis [36].

In some situations, normalizing the summary statistics is important and necessary to avoid the domination of the most variable summaries. The potential for this is high when the summary statistics have different magnitudes or scales. Therefore, by scaling the summary statistics, each summary will have the same effect on the total level of similarities [37]. However, results shown in this paper are for analyses carried out with unnormalized summary statistics throughout. We here performed the ABC-PMC both with and without re-scaling the summary statistics. However, we did not see any noticeable effect on either the posterior estimates or computation time. In our case, all the summary statistics are bounded below by zero (can take any positive numbers) but have different magnitudes. We used different scaling methods, such as, median absolute deviation (MAD), mean, and the adaptive scaling method employed by Prangle [37].

The second caveat pertains prior choice for model parameters. Simulation-based approaches such as ABC tend to perform much better when informative priors are used. This is less of a problem for ABC-PMC approaches than, say, rejection-based ABC approaches, since the sequential learning inherent in the ABC-PMC algorithm can successfully hone in on the areas of higher approximate posterior mass with even quite uninformative priors. However, the computational ramifications of using less informative priors can be large, exhibited by both a slower decrease in the tolerance level ξ between generations, and a higher rejection rate within generations. This highlights the care which must be taken when selecting priors given limited computational resources.

Of course, even with this caveat, the huge reduction in computational burden resulting from the use of ABC-PMC as opposed to DA-MCMC, is extremely impressive. In fact, as we have seen here, ABC-PMC approaches can be used to deal with otherwise computationally intractable analyses. Furthermore, it is possible to envisage ways of increasing this improvement even more. For example, one issue with individual-level epidemic models is the amount of variation observed between simulations from the same model (i. e. with the same parameters). This variation means that rejection rates within the ABC algorithm can be very high. One possible approach to dealing with this problem is the technique of so-called "Lazy ABC" introduced by Prangle [38], in which

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

simulated data sets which differ from the observed data substantially are rejected early in the simulation, rather than generating an entire data set (here, epidemic).

There are also aspects of the model that would be interesting to explore in further work. We have confined ourselves to considering static, binary, spatial undirected networks. Of course, in many situations the underlying contact network may not be spatial in nature, and may be directional and/or weighted. For example, consider livestock diseases in which – often long distance – animal movement between farms and/or markets may be drive disease spread through a dynamic, weighted, non-symmetric network.

Another avenue for further work regards the performance of ABC methods when the fitted model is misspecified. The simple rejection ABC method has been found more robust than full Bayesian MCMC approach for misspecified models in other contexts [39, 40]. Whether this is the case for our scenarios, especially with the presence of large amount of important missing information is something the authors wish to explore further.

We tended in this work to mainly investigate the possibility of getting good approximated estimates of the model parameters as well as the missing information. Thus, investigating the performance of ABC-PMC method for such misspecified models would be an interesting further work.

In conclusion, methods such as ABC-PMC do seem to offer a plausible way of carrying out inference for complex infectious disease models in the presence of multiple layers of data uncertainty. Of course, in scenarios where Monte Carlo methods such as data augmented MCMC are plausible, they are to be preferred providing, as they do, a full, coherent Bayesian analysis. However, in infectious disease epidemiology it is very easy to find systems in which the size of the population and/or level of uncertainty in the data precludes such approaches in anything like an acceptable timeframe. We thus conclude that further exploration into the use of ABC methods for fitting infectious disease models to data is certainly warranted.

Notes

1 The observed is the number in each stratified class, and the expected is calculated as the product of the total number of the stratified class's row and column divided by the total number of classes.

References

[1] Deardon R, Brooks SP, Grenfell BT, Keeling MJ, Tildesley MJ, Savill NJ, et al. Inference for individual-level models of infectious diseases in large populations. Stat Sin. 2010;20:239.
[2] Keeling MJ, Woolhouse ME, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, et al. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. Science. 2001;294:813–7.
[3] Jewell CP, Kypraios T, Neal P, Roberts GO, et al. Bayesian analysis for emerging infectious diseases. Bayesian Anal. 2009;4:465–96.
[4] Ster IC, Ferguson NM. Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. PLoS One. 2007;2:e502.
[5] Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, et al. Networks and the epidemiology of infectious disease. Interdiscip Perspect Infect Dis. 2011:1–28.
[6] Keeling MJ, Eames KT. Networks and epidemic models. J R Soc Interface. 2005;2:295–307.
[7] Welch D, Bansal S, Hunter DR. Statistical inference to advance network models in epidemiology. Epidemics. 2011;3:38–45.
[8] Britton T, O'Neill PD. Bayesian inference for stochastic epidemics in populations with random social structure. Scand J Stat. 2002;29:375–90.
[9] Groendyke C, Welch D, Hunter DR. Bayesian inference for contact networks given epidemic data. Scand J Stat. 2011;38:600–16.
[10] Groendyke C, Welch D, Hunter DR. A network-based analysis of the 1861 hagelloch measles data. Biometrics. 2012;68:755–65.
[11] Sainudiin R, Welch D. The transmission process: a combinatorial stochastic process for the evolution of transmission trees over networks. J Theor Biol. 2016;410:137–70.
[12] Tildesley MJ, Savill NJ, Shaw DJ, Deardon R, Brooks SP, Woolhouse ME, et al. Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. Nature. 2006;440:83.
[13] Neal RM. MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo. 2011;2:2.
[14] Malik R, Deardon R, Kwong GP. Parameterizing spatial models of infectious disease transmission that incorporate infection time uncertainty using sampling-based likelihood approximations. PloS One. 2016;11:e0146253.
[15] Pokharel G, Deardon R. Gaussian process emulators for spatial individual-level models of infectious disease. Can J Stat. 2016;44:480–501.
[16] Beaumont MA, Cornuet JM, Marin JM, Robert CP. Adaptive approximate Bayesian computation. Biometrika. 2009;96:983–90.
[17] Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162:2025–35.
[18] Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods. Proc Nat Acad Sci. 2003;100:15324–8.
[19] Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol. 1999;16:1791–8.
[20] Sisson SA, Fan Y, Tanaka MM. Sequential Monte Carlo without likelihoods. Proc Nat Acad Sci. 2007;104:1760–65.
[21] McKinley T, Cook AR, Deardon R. Inference in epidemic models without likelihoods. Int J Biostat. 2009;5:1–40.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

[22] Numminen E, Cheng L, Gyllenberg M, Corander J. Estimating the transmission dynamics of streptococcus pneumoniae from strain prevalence data. Biometrics. 69: 748–57.
[23] Walker DM, Allingham D, Lee HW, Small M. Parameter inference in small world network disease models with Approximate Bayesian Computation methods. Phys A: Stat Mech Appl. 2010;389:540–48.
[24] Neal P, Roberts G. A case study in non-centering for data augmentation: stochastic epidemics. Stat Comput. 2005;15:315–27.
[25] Donaldson A, Alexandersen S. Relative resistance of pigs to infection by natural aerosols of fmd virus. Vet Rec. 2001;148:600–2.
[26] Bifolchi N, Deardon R, Feng Z. Spatial approximations of network-based individual level infectious disease models. Spatial Spatio-temporal Epidemiol. 2013;6:59–70.
[27] Del Moral P, Doucet A, Jasra A. An adaptive sequential Monte Carlo method for approximate Bayesian computation. Stat Comput. 2012;22:1009–20.
[28] Cappé O, Guillin A, Marin JM, Robert CP. Population Monte Carlo. J Comput Graph Stat. 2004;13:907–29.
[29] Filippi S, Barnes CP, Cornebise J, Stumpf MP. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. Stat Appl Genet Mol Biol. 2013;12:87–107.
[30] Kypraios T, Neal P, Prangle D. A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. Math Biosci. 2017;287:42–53.
[31] Drovandi CC, Pettitt AN. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. Biometrics. 2011;67:225–33.
[32] Eames K, Bansal S, Frost S, Riley S. Six challenges in measuring contact networks for use in modelling. Epidemics. 2015;10:72–7.
[33] Silverman BW. Density estimation for statistics and data analysis. London: Chapman & Hall/CRC, 1986.
[34] Deeth LE, Deardon R, et al. Latent conditional individual-level models for infectious disease modeling. Int J Biostat. 2013;9:75–93.
[35] Diggle PJ. Spatio-temporal point processes, partial likelihood, foot and mouth disease. Stat Methods Med Res. 2006;15:325–36.
[36] Joyce P, Marjoram P. Approximately sufficient statistics and Bayesian computation. Stat Appl Genet Mol Biol. 2008;7:1–18.
[37] Prangle D. Adapting the ABC distance function. Bayesian Anal. 2017;12:289–309.
[38] Prangle D. Lazy ABC. Stat Comput. 2016;26:171–85.
[39] Chaudhuri S, Ghosh S, Nott DJ, Pham KC. An easy-to-use empirical likelihood ABC method. arXiv preprint arXiv:1810.01675. 2018.
[40] Frazier DT, Robert CP, Rousseau J. Model misspecification in ABC: consequences and diagnostics. arXiv preprint arXiv:1708.01974. 2017.

Supplementary Material: The online supplementary material contains the algorithm of ABC-PMC, Web Figures referenced in Sections 4.2.1, 4.2.2, 5.1,5.2 and 5.3. (DOI:https://doi.org/10.1515/ijb-2017-0092).

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd