J. Math. Biol. (2007) 55:223–247
DOI 10.1007/s00285-007-0081-y

Mathematical Biology

Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates

Richard J. Boys · Philip R. Giles

Received: 13 December 2005 / Revised: 12 February 2007 / Published online: 15 March 2007

Abstract Stochastic compartmental models of the SEIR type are often used to make inferences on epidemic processes from partially observed data in which only removal times are available. For many epidemics, the assumption of constant removal rates is not plausible. We develop methods for models in which these rates are a time-dependent step function. A reversible jump MCMC algorithm is described that permits Bayesian inferences to be made on model parameters, particularly those associated with the step function. The method is applied to two datasets on outbreaks of smallpox and a respiratory disease. The analyses highlight the importance of allowing for time dependence by contrasting the predictive distributions for the removal times and comparing them with the observed data.

Keywords MCMC methods · Predictive fit · Reversible jump · Stochastic epidemic models

Mathematics Subject Classification (2000) 62M05 · 62F15 · 62P10 · 92D25 · 92D30

1 Introduction

Statistical analyses of infectious disease data frequently assume a stochastic model containing Susceptible-Exposed-Infective-Removed (SEIR) compartments. This class of compartmental model provides a natural description of many epidemics as individuals will often pass through several stages of disease development before being removed

R. J. Boys (✉) · P. R. Giles
School of Mathematics and Statistics, University of Newcastle upon Tyne, Newcastle upon Tyne, NE1 7RU, UK
e-mail: richard.boys@ncl.ac.uk

from further participation in disease spread. Incorporating stochasticity into this type of model is readily achieved by specifying the probabilities of transfer between the compartments of the model. These probabilities can often be formulated in terms of a few key parameters which have natural interpretation in terms of the rates at which individuals contract the disease and are subsequently removed from further participation. If the values of these parameters can be determined using data collected from an epidemic outbreak, then steps may be taken to alter these rates in any future outbreak using public health measures (for example, through inoculation) and hence control disease spread during further outbreaks.

Inference on the parameters of compartmental models is, however, a non-trivial statistical problem. The main complication arises from the partial observation of data from epidemic outbreaks. For example, while the times at which individuals are diagnosed with the disease (and hence prevented from further infecting susceptible individuals) are often available from sources such as medical records, the times at which individuals contract the pathogen are usually much more difficult to ascertain. Recent research, for example, in [10,21,22], has focused on the implementation of Markov chain Monte Carlo (MCMC) techniques to help overcome the missing data problem and enable parameter estimation to be performed. In particular, analyses can be performed from a Bayesian perspective, which allows (prior) information on the epidemic (possibly from previous outbreaks) to be incorporated into the current analysis. This approach also yields full probabilistic information on the parameters of interest. Moreover, inferences on functions of model parameters are easily obtained given MCMC output from an analysis of epidemic data. Epidemiological investigators regularly require information on key indicators of disease spread such as the basic reproduction number (see Sect. 3.2), which may be expressed as functions of more fundamental model parameters such as the infection and removal rates in many compartmental models.

An assumption that is usually made with regard to the spread of an infectious disease is that the time between an individual becoming infective and their removal (the infectious period) is independent and identically distributed according to some well known distribution. The most common choice is the exponential distribution as this invariably results in the formulation of models that are simpler to study analytically than those with more complex infectious period distributions. However, more realistic and flexible choices for the infectious period can produce models with very different properties to the exponential distribution model; see, for example, [2,16] and [17].

An alternative choice for modelling the infectious period is the gamma distribution [1,17]. This distribution is used by [20] to analyse data from the 1967 smallpox outbreak in Abakaliki, Nigeria. Their analysis reveals that the length of time spent in the infectious category during this outbreak has two distinct components. Some individuals are shown to have a shorter infectious period of around 14 days, while others remain infectious for approximately 18 days. There are two possible explanations for this finding. The first is that there is a biological interpretation as to why some individuals remain infectious for around 4 days longer than others. The second reason is that individuals who contract the disease during the early part of the outbreak remain infectious for considerably longer as few resources are allocated to isolate these infectives (perhaps due to a slow response to the seriousness of the ensuing epidemic). Once it is known that a serious outbreak of smallpox is in progress, further resources

are made available to identify and isolate infectious individuals who therefore remain infectious for a shorter period of time (by approximately 4 days according to the analysis of [20]). The possibility of different length infectious periods during the early and later stages of this smallpox outbreak is identified in [3]: the author assumes that the first two infective individuals are infective for a fixed period of 14 days, and when each of the subsequently infected individuals become infective they remain so for only 7 days.

However, despite evidence of time-inhomogeneity in the period that individuals remain infectious (and hence in the rate at which infectious individuals are removed from further participation in the outbreak), there have been few previous attempts to explicitly capture this feature of epidemics in a compartmental model. One reason for this may be due to the complexity of such a model and the inherent difficulty in making inferences on model parameters given partial observation of the data. In this paper, we develop a stochastic compartmental model for disease outbreak that includes time-inhomogeneity in the rate at which infectious individuals are removed from the epidemic. The model assumes that the infectious period is exponentially distributed for all individuals, however, the rate parameter of the distribution is a step function in time. The number of change-points in the step function, the position of the changes and the rates of removal in each step are all assumed to be unknown and are therefore objects for inference. The model for the infection process allows for heterogeneity of susceptibility, and the possibility of a latent period for the disease. Inferences for model parameters, and functions of the model parameters, are made using a Bayesian approach which incorporates prior information. We develop a reversible jump MCMC methodology [12] that enables calibration of the model for outbreaks of disease where the data are only partially observed. An analysis of the Abakaliki smallpox outbreak discussed previously shows that the model is able to capture changes in the removal rate and highlight where such changes occur.

The contribution of the paper to the literature of stochastic epidemic modelling is threefold. We develop an extension of existing stochastic compartmental models that enables more realistic modelling of disease outbreaks. We provide a methodology that allows the fitting of this model to partially observed data, and results in more accurate inferences being made on key epidemiological parameters such as the basic reproduction number. Furthermore, much of the work in this area describes procedures for making inferences about model parameters without assessing the appropriateness of the model, for example, through associated measures of model fit. We describe how to assess the adequacy of models by studying their predictive fit, that is, by comparing the predictive distribution (assuming the model) with the observed data. By providing evidence that the model outlined here better describes a range of disease outbreaks, public health measures based on the estimates of the model parameters may be better suited to preventing future epidemics.

The paper is arranged as follows. In Sect. 2, we describe the SEIR model, the observed data and the form of prior information which together allow inferences to be made about key parameters of the epidemic process. An overview of the MCMC algorithm is given in Sect. 3 (with more details in the Appendix). The algorithm is tested on simulated data in Sect. 4 and then used to analyse data from the Abakaliki smallpox outbreak and an outbreak of a respiratory disease in Sects. 5 and 6, respec-

tively. The paper concludes in Sect. 7 with some general remarks on the additional insights gained by using our model and on the importance of using predictive measures of overall model adequacy.

2 A multitype SEIR model

2.1 Model

We adopt a Susceptible-Exposed-Infective-Removed (SEIR) model in which the susceptible category is partitioned into m sub-categories. Each member of the population of individuals susceptible to a disease belongs to one of the m groups, and each group is considered to have a potentially different susceptibility to the disease in question. Group membership is assumed known and might be affected by factors such as age or sex, although within each grouping, susceptibility is assumed to be homogeneous. Initially, there are N_i individuals in group i, $i = 1, 2, \ldots, m$, and the epidemic begins once an individual from one of the groups contracts the disease. Having contracted the disease, the individual enters a latent period, where they show no symptoms of the disease and are unable to infect susceptible individuals, before becoming infective. Models which include independent parameterisations of the latent and infectious period distributions are not identifiable given the partial information available in the data considered here, unless we have very informative prior information about one of these distributions. Therefore, following the example of [20], we assume a latent period of fixed length c, with c chosen appropriately according to known properties of the disease. The epidemic then proceeds according to the following transition probabilities: in a (small) interval $[t, t + dt)$, the process evolves according to

$$\begin{aligned}\Pr\left\{\big(S_i(t+dt), E_i(t+dt)\big) = \big(S_i(t)-1, E_i(t)+1\big)\right\} &= \beta_i S_i(t) I(t) dt + o(dt)\\ \Pr\left\{\big(I_i(t+dt), R_i(t+dt)\big) = \big(I_i(t)-1, R_i(t)+1\big)\right\} &= \gamma(t) I_i(t) dt + o(dt)\end{aligned}$$

for $i = 1, 2, \ldots, m$, where $(S_i(t), E_i(t), I_i(t), R_i(t))$ denotes the number of individuals in group i at time t that are susceptible, exposed, infected, and removed, with $I(t) = \sum_{i=1}^{m} I_i(t)$. The model assumes homogeneous mixing within the population. Each infected individual makes infective contact with a member of susceptibility group i at rate β_i. Infectives are assumed to have the same propensity to infect regardless of the susceptibility group from which they originate. Note that each infected individual must pass through the latent period before becoming infective. Since the infection times are unobserved, the first sign that the epidemic is in progress is at the time of the first removal. For this reason, the time of the first removal is set to be zero and all infection and removal times are set relative to this reference time, in the same manner as [21]. The group to which the first infective belongs is also unknown. The epidemic is observed until there are no infected individuals remaining in the population and this occurs at (relative) time T.

The time-dependence of the removal rate is modelled as a step function, with k steps at times $\boldsymbol{s} = (s_1, s_2, \ldots, s_k)$. The function $\gamma(t)$ takes the value γ_j when $t \in [s_j, s_{j+1})$ for $j = 0, 1, \ldots, k$, where for convenience $s_0 \equiv -\infty$ and $s_{k+1} \equiv T$. This formula-

tion is equivalent to a model in which the distribution of the infectious period has an exponential intensity γ_j within interval $[s_j, s_{j+1})$. The end of the epidemic is defined to be the first time at which no infectives or exposed individuals remain in the population. Note also that the population is assumed to be closed, and removed individuals play no further part in the epidemic.

2.2 Data

Information on the epidemic process is assumed to consist only of the initial numbers of susceptibles in each group, $\boldsymbol{N} = (N_1, N_2, \ldots, N_m)$, along with the removal time of each infective individual. Let r_{ij} $(i = 1, 2, \ldots, m, j = 1, 2, \ldots, n_i)$ denote the time of removal j in group i, and n_i the total number of group i removals. The infection times and the group from which each infective originates are unknown and are therefore treated as parameters of the model. We write τ_{ij} for the time of infection j in group i, and so this individual becomes infective at time $\tau_{ij} + c$ (after the end of their latent period). The group from which the initial infective originates is denoted by i_{min}. Let $\boldsymbol{r} = (r_{ij})$ denote the matrix of removal times and $\boldsymbol{\tau} = (\tau_{ij})$ the matrix of infection times, but excluding the time of the first infection in the whole population, $\tau_{i_{min},1}$. Since the complete epidemic is observed, the total number of infections of individuals in group i is also equal to n_i. The total number of infections and the total number of removals are both equal to $n = \sum_{i=1}^{m} n_i$.

2.3 Complete data likelihood

The parameters for which we would like to make inferences are the infection rates $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$, and the parameters specifying the removal rate step function, namely the number of steps, k, the step positions $\boldsymbol{s} = (s_1, s_2, \ldots, s_k)$, and the removal rates $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_k)$.

Assuming that both infection and removal times data are fully observed, the likelihood function is

$$\begin{aligned}
&\pi(\boldsymbol{\tau}, \boldsymbol{r} | \boldsymbol{\beta}, \boldsymbol{\gamma}, k, \boldsymbol{s}, i_{min}, \tau_{i_{min},1}) \\
&\quad = \left\{ \prod_{i=1, i \neq i_{min}}^{m} \beta_i S_i(\tau_{i1}^-) I(\tau_{i1}^-) \right\} \left\{ \prod_{i=1}^{m} \prod_{l=2}^{n_i} \beta_i S_i(\tau_{il}^-) I(\tau_{il}^-) \right\} \left\{ \prod_{i=1}^{m} \prod_{l=1}^{n_i} \gamma(r_{il}^-) I_i(r_{il}^-) \right\} \\
&\qquad \times \exp\left\{ -\sum_{i=1}^{m} \left(\beta_i \int_{\tau_{i_{min},1}}^{T} S_i(t) I(t) dt \right) - \int_{\tau_{i_{min},1}}^{T} \gamma(t) I(t) dt \right\}
\end{aligned}$$

where $S_i(t^-)$ denotes the number of susceptible individuals in group i immediately prior to time t, and similarly for $\gamma(t^-)$ and $I(t^-)$.

2.4 Prior model

We assume independent prior distributions for the infection rates, the group and time of the first infection, and the step function for the removal rates. Following other

authors such as [21], we take independent gamma priors for the infection rates β_i and the removal rates γ_j (given the number of steps):

$$\begin{aligned}\beta_i &\sim \Gamma(g_{\beta_i}, h_{\beta_i}), \quad i = 1, 2, \ldots, m \\ \gamma_j|k &\sim \Gamma(g_\gamma, h_\gamma), \quad j = 0, 1, \ldots, k.\end{aligned}$$

Such a choice provides reasonable flexibility for describing prior beliefs and has the added attraction of facilitating a conjugate update in the MCMC scheme.

The prior model for the step function is similar to that described by [12]. The number of steps k in $[0, T]$ is assumed to follow a Poisson distribution with mean λ truncated above at k_{max}, with probability function

$$\pi(k) \propto \frac{\lambda^k}{k!}, \qquad k = 0, 1, \ldots, k_{\max},$$

and k_{max} chosen to reflect prior beliefs about the maximum number of change-points in the removal rate. Prior knowledge about the number of change-points is closely related to that of the removal rates and so care must be taken to ensure a coherent choice of parameter values for the joint prior distribution of $(\boldsymbol{\gamma}, k)$. For example, knowledge of a constant step function corresponds to strong prior information on both $k\,(\pi(k = 0) \simeq 1)$ and the $\gamma_j|k\,(g_\gamma \ll h_\gamma^2)$. In general, an informative prior distribution for k is not consistent with an uninformative prior specification for the $\gamma_j|k$ and leads to the posterior distribution assigning all probability to the simplest $(k = 0)$ model. Conditional on k, the joint distribution of the positions of the change-points $\boldsymbol{s}$ is taken to be that of every ath order statistic from a sample of size $a(k+1)-1$ points sampled uniformly over $[0, T]$. Note that it is assumed *a priori* that no jumps in the step function occur before time zero. This choice of prior distribution results in the scaled distance between consecutive change-points $(s_{j+1} - s_j)/T$ following a $Beta(a, ak)$ distribution. Thus the value of a can be chosen to reflect prior beliefs about change-points in the removal rate function, with large values of a essentially putting step positions on a fixed grid. Small values of a can result in the MCMC algorithm accepting intervals containing no data (removal times), and therefore leads to estimates based solely on prior opinion. Thus care must be taken to choose a appropriately and its impact on the posterior distribution assessed.

As the first infection time occurs before time zero (relative to the time of the first removal), we adopt a uniform $U(-d, 0)$ distribution for $\tau_{i_{min},1}$ $(d > 0)$, with the lower endpoint chosen to be a time prior to which it is believed the epidemic cannot have begun. This prior is of the same form as that used by [13] in their analysis of a multitype epidemic. Also, in the absence of further information, we assume a discrete uniform distribution for the susceptibility group from which the initial infective originates, that is, $i_{min} \sim U\{1, 2, \ldots, m\}$.

3 Posterior inference

Information about the unknown parameters (and unobserved data) in the complete data likelihood and prior distribution are combined using Bayes theorem to give the

posterior distribution

$$\pi\left(\boldsymbol{\beta},\boldsymbol{\gamma},k,\boldsymbol{s},i_{min},\tau_{i_{min},1},\boldsymbol{\tau}\middle|\boldsymbol{r}\right) \propto \pi\left(\boldsymbol{\tau},\boldsymbol{r}\middle|\boldsymbol{\beta},\boldsymbol{\gamma},k,\boldsymbol{s},i_{min},\tau_{i_{min},1}\right) \times\pi\left(\boldsymbol{\beta},\boldsymbol{\gamma},k,\boldsymbol{s},i_{min},\tau_{i_{min},1}\right).$$

Posterior inference can be conducted by simulating from a Markov chain for which the equilibrium distribution is the posterior distribution of interest. There are several standard methodologies available for constructing such a Markov chain, and the process is generally known as Markov chain Monte Carlo (MCMC) [8].

An MCMC scheme begins with some suitable choice of parameter values to initialise the Markov chain. Simulation of the Markov chain proceeds by successively updating parameter values according to a standard methodology. This is most simply achieved when the conditional distributions of each parameter (given the other parameters of the model) are known standard distributions. Such an algorithm is known as a Gibbs sampler. If this is not the case then parameter values are simulated using a slightly more complex procedure known as a Metropolis-Hastings sampler. Here updates are proposed from a standard distribution (known as the proposal distribution) which is hopefully similar to the actual non-standard conditional distribution. The proposed update is in turn accepted with a probability that ensures that the equilibrium distribution of the Markov chain is the one required, that is, the posterior distribution. Further, if the proposed update results in a change to the dimensionality of the Markov chain, we use a technique known as a reversible jump update in which the acceptance probability also takes account of the dimension change [12]. We note that other dimension changing moves are possible [23] but are not considered any further here.

In order to simulate from the posterior distribution for the epidemic model given above, we construct an MCMC algorithm that employs both Gibbs sampler updates and, where appropriate, Metropolis-Hastings updates. The method uses techniques similar to those of [10,21] and [22] to explore the space of the infection rates and hidden infection times, and those outlined by [12] for the step function. Mixing over uncertainty for the step function is complicated by the need to use reversible jumps which propose changes to the number of steps k.

3.1 MCMC scheme

The MCMC scheme begins with an initial choice of parameters and hidden infection times, though particular care must be taken to ensure that the infection times are feasible (and have positive density). A simple example of an infeasible sequence of infection times would be one in which the infection of a susceptible individual occurred whilst there were no infectious individuals in the population. Further, the only point at which no individuals lie in the exposed or infective categories is at the end of the epidemic.

The scheme proceeds by cycling through the following steps. Note that the technical details of the algorithm, including the conditional distributions and the acceptance probabilities for the Metropolis-Hastings moves, are reserved for the Appendix.

Update infection times: Propose an update to the sequence of hidden infection times by reallocating a randomly (uniformly) chosen infection time to a new time sampled from a $U(-d, T)$ proposal distribution. The associated time at which the individual becomes infective is moved to a new time $\tau' + c$. The move is rejected immediately if the proposed sequence of infections and removals is infeasible. Otherwise the move is accepted with the appropriate Metropolis-Hastings acceptance probability.

Update infection parameters: Sample a new value for each infection rate parameter β_i $(i = 1, 2, \ldots, m)$ from its conditional posterior distribution given all other states of the chain.

Update removal rate step function: The proposal of a new realisation of the removal rate step function is more complicated and involves different moves for each aspect of the step function. At each iteration, one of the following choices is made:

(i) *Update removal rates*: with probability q_1, simulate a new value for γ_j $(j = 0, 1, \ldots, k)$ from the conditional distribution of each removal rate given $k, \boldsymbol{s}$ and the other states of the chain.

(ii) *Update existing step positions*: with probability q_2, select one of the step positions s_j uniformly from the k existing steps and propose moving s_j to a new position s_j' sampled from a $U(s_{j-1}, s_{j+1})$ distribution. Accept the proposed position with the appropriate Metropolis-Hastings acceptance probability.

(iii) *Increase number of steps*: with probability q_3, propose an additional step at a position s' sampled uniformly over $(0, T)$. Note that this is a reversible jump move which induces an enlargement of the parameter space as additional parameters are required to specify the position of the step and the removal rates between the old and new change-points.

(iv) *Decrease number of steps*: with probability $q_4 = 1 - q_1 - q_2 - q_3$, propose a removal of a step position, s_j say, chosen uniformly from the k existing steps. The move also requires a change in the dimensionality of the parameter space, and is the reverse of the proposal to increase the number of steps.

3.2 Inference for the basic reproduction number

In addition to being able to make inferences about model parameters such as the infection rates, the unobserved infection times and the parameters describing the removal rate step function, it is straightforward to make inferences about other parameters of epidemiological interest such as the basic reproduction number, R_0. This is usually defined to be the expected number of new cases caused by a typical infective in a large and entirely susceptible population during the early stages of an epidemic; see, for example, [14]. Heterogeneity in infectivity across the population has been studied by [18] using a model which allows individuals to have different reproductive numbers, each of which is a draw from a distribution with mean R_0. However, we focus on the more basic task of estimating the population level reproduction number R_0. This is important since a major epidemic will occur with positive probability if $R_0 > 1$, and will typically die out quickly if $R_0 \leq 1$; see [14]. The basic reproduction number is of particular importance to epidemiological investigators since inferences

on R_0 can be used to guide vaccination policies aimed at reducing the number of susceptible individuals in a population to a level such that R_0 is brought below the unity threshold [6].

After approximating the early stages of the epidemic to a Poisson process with rate $\sum_{i=1}^{m} \beta_i N_i$, and noting that a typical infective remains so for a mean period of $1/\gamma_0$ during this period, the basic reproduction number for our model can be shown to be

$$R_0 = \frac{1}{\gamma_0} \sum_{i=1}^{m} \beta_i N_i.$$

The analyses of data in following sections focus particularly on the possibility of improved estimation of this important parameter given the model developed in this paper compared to less sophisticated models.

4 Analysis of simulated data

We now validate our MCMC algorithm by demonstrating its performance on simulated data. The data were generated from a population comprising a single ($m = 1$) homogeneous susceptibility group in which there were $N_1 = 100$ initially susceptible individuals. The length of the latent period was specified to be $c = 10$, and the infection rate was chosen to be $\beta_1 = 0.0012$. The removal rate during the early part of the epidemic was set at $\gamma_0 = 0.08$, increasing to $\gamma_1 = 0.4$ after the change-point at s_1. The time of the change was chosen to coincide with that of the 10th removal, which occurred at $s_1 = 59.8$. A total of $n_1 = 27$ infections and subsequent removals were observed before the end of the simulated epidemic at time $T = 100.4$.

The analysis assumed that the data consisted of the removal times of infected individuals, given in Table 1, along with the number of heterogeneous susceptibility groups ($m = 1$), the initial population size ($N_1 = 100$), and the length of the latent period ($c = 10$). The infection times and the times at which individuals entered the infective category were assumed to be unobserved, as were those parameters specifying the removal rate step function and the rate of infection.

Prior specifications were made on the unknown parameters as follows. The prior mean infection rate was chosen to be 0.001, which is of the same order of magnitude as the infection rate specified in the simulation. A relatively large value of 1.0 was chosen for the prior variance so that the analysis used fairly uninformative prior knowledge for the infection rate β_1. A similar strategy was employed for choosing the parameters of the prior distribution for the removal rates $\gamma_j|k$: the prior mean and

Table 1 Simulated data

Removal times	0	3.9	22.3	29.8	30.8	40.2	42.8	47.7	55.5
	59.8	61.0	61.2	61.3	61.4	62.0	63.3	64.8	68.1
	69.6	70.5	71.6	76.3	80.9	81.0	87.0	89.6	100.4

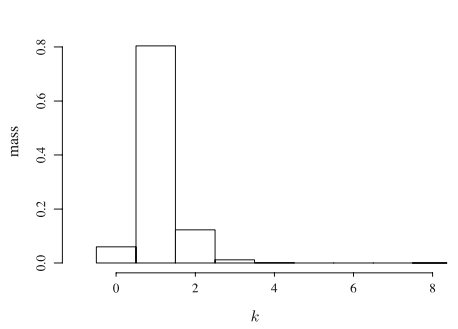
variance were both chosen to be 0.1. The parameter choices for the prior distribution on the removal rate step function were made so that the prior information was sufficiently uninformative to allow inferences to be based primarily on the information in the data, yet sufficiently coherent to ensure that the MCMC scheme explored the full space of the posterior distribution and did not allocate all the posterior probability to the model with $k = 0$ removal rate change-points. Specifically, the maximum number of changes in the removal rate was set to be $k_{max} = 8$, with the number of steps k assumed to follow the truncated Poisson distribution with rate $\lambda = 1$. The parameter of the prior distribution for the step locations $\boldsymbol{s}$ was set at $a = 2$. Finally, the time of the initial infection was assumed to have occurred no earlier than $d = 50$ time units before the first removal.

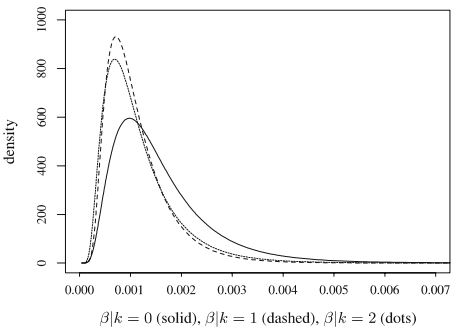
The analysis was performed by implementing the MCMC algorithm described in Sect. 3, with the move probabilities q_i for the proposals to update the removal rate step function each chosen to be equal to 0.25 ($i = 1, 2, 3, 4$). A sample of 100,000 iterations was required before convergence seemed to have occurred, and this was subsequently confirmed by using diagnostics such as the Heidelberger and Welch [15] and Geweke [9] tests. A further 100 million iterations with a thin of 10,000 produced a sample of 10,000 (almost uncorrelated) values from the posterior distribution. The MCMC output is displayed in the form of marginal and conditional posterior distributions in Fig. 1, with corresponding summary measures in Table 2.

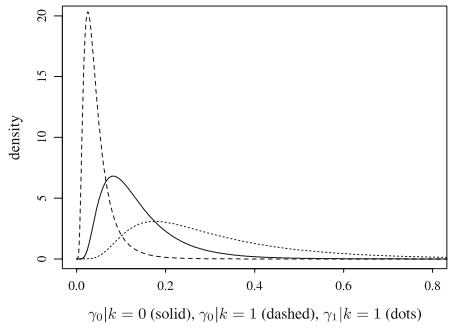
The posterior distribution of k shows that the majority of the posterior mass lies with the model for which there is one change-point in the removal rate. The posterior distribution of $s_1|k = 1$ is centred around the correct (but assumed unknown) value of the change-point ($s_1 = 59.8$). Further, the summary measures for the removal rates given a single change-point highlight the more rapid rate of removal after s_1 compared to that earlier in the simulated epidemic, and are fairly close the theoretical values underpinning the simulated data ($\gamma_0 = 0.08$, $\gamma_1 = 0.4$). The conditional posterior distribution of the infection rate, $\beta_1|k = 1$, is also located near to its theoretical value ($\beta_1 = 0.0012$).

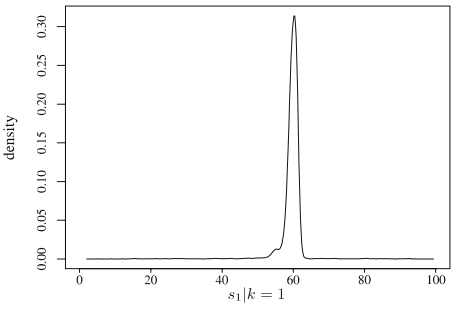
Significantly smaller amounts of posterior mass are allocated to the models for which there are $k = 0$ or $k = 2$ removal rate changes. Note that these relatively implausible models give misleading inferences on the removal process, though inferences for the infection rate are fairly similar to those in the "true" model ($k = 1$). In particular, as expected, the posterior distribution for the removal rate $\gamma_0|k = 0$ is located between those of $\gamma_0|k = 1$ and $\gamma_1|k = 1$, as $\gamma_0|k = 0$ is essentially an average of the slower early rate and the more rapid later rate in the epidemic. For the $k = 2$ change-points model, the conditional posterior distribution of $s_2|k = 2$ is similar to that of $s_1|k = 1$, however, that of $s_1|k = 2$ has a large variance and is close to its prior distribution (as shown by the dotted line on the lower right plot of Fig. 1).

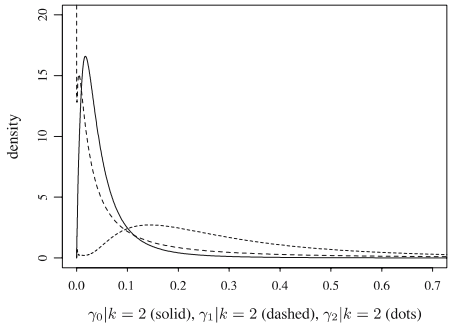
The figure (bottom plot) and Table 2 also highlight the effect of mis-specifying the number of change-points in the removal rate. For example, assuming a time-homogeneous removal rate ($k = 0$) clearly underestimates the basic reproduction number R_0 as this model is unable to capture the slower rate of removal during the early stages of the epidemic. Correctly including a single change-point leads to doubling the estimate for R_0 but also increases its standard deviation four-fold. Recall that initial predictions

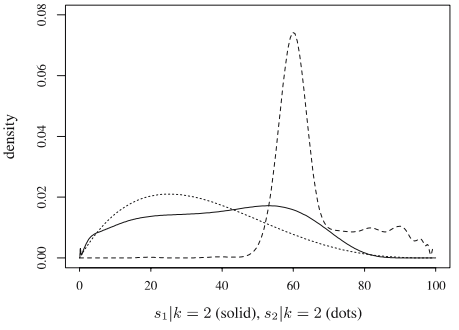












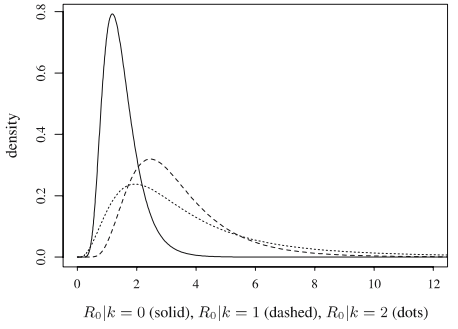


Fig. 1 Selected marginal posterior distributions for the simulated data analysis

Table 2 Posterior means (and standard deviations) for selected parameters

Parameter			
k	1.09 (0.48)		
β_1	0.0011 (0.0005)		
R_0	2.78 (3.33)		
k	0	1	2
$\beta_1 \mid k$	0.0014 (0.0006)	0.0011 (0.0005)	0.0011 (0.0005)
$\gamma_0 \mid k$	0.127 (0.053)	0.045 (0.028)	0.050 (0.044)
$\gamma_1 \mid k$		0.300 (0.168)	0.125 (0.224)
$\gamma_2 \mid k$			0.310 (0.232)
$R_0 \mid k$	1.19 (0.34)	2.77 (1.22)	3.41 (5.23)

for whether or not the epidemic will die out are made using $Pr(R_0 \leq 1)$. In our simulation model the true value for R_0 is 1.5. All three models ($k = 0, 1, 2$) correctly indicate that the epidemic will not die out quickly: $\widehat{\Pr}(R_0 \leq 1 | k, \boldsymbol{r}) = 0.311$ ($k = 0$), 0.009 ($k = 1$), 0.044 ($k = 2$). However, the time-homogeneous model clearly underplays the strength of evidence for a serious outbreak of the epidemic.

Finally the robustness of these inferences was assessed by observing how sensitive the inferences were to small changes in the parameters of the prior distribution. This is not straightforward since marginal likelihood functions are difficult to interpret due to the complex posterior dependence structure induced by the missing data. Instead we have chosen to simply rerun the analysis many times and to study the effect of small changes in the prior specification. Overall, we found that such changes resulted in only very modest changes to the posterior distribution; further details can be found in [11]. We also found that the posterior mass for the incorrect models ($k \neq 1$) is largely due to the prior distribution rather than any particularly supportive evidence in the data.

In summary, the analysis in this section illustrates how our MCMC methodology can be used to make reliable and accurate inferences about the key parameter values of this SEIR epidemic process despite the complications due to partial observation of the epidemic and the use of fairly uninformative prior beliefs on the nature of the heterogeneity in the removal rate process.

5 Analysis of the Abakaliki smallpox data

The Abakaliki smallpox data are much studied by many previous authors; see, for example, [3,7,20,21]. The data have a simple structure and so are ideal for illustrating the methodology we have developed for making inferences on the parameters of our more complex model. The analysis does not intend to provide new insight into this outbreak in particular or the smallpox disease in general (a fuller analysis based on more detailed data from this outbreak is given in [7]). Rather, the analysis serves to show that improved inferences can be made regarding the spread of this disease by considering the model proposed in this paper. For example, as already stated in Sect 3.2, a key epidemiological parameter of interest is the basic reproduction number, R_0. By adopting a model which explicitly models changes in the removal rate, we aim

Table 3 Abakaliki smallpox data

Day	0	13	20	22	25	26	30	35	38	40	42	47
No. of removals	1	1	1	1	3	1	1	1	1	2	2	1
Day	50	51	55	56	57	58	60	61	66	71	76	
No. of removals	1	1	2	1	1	1	2	1	2	1	1	

to demonstrate that more accurate inferences may be made on the rate of removal during the early stages of the epidemic, and hence can lead to improved inference on R_0. This in turn can lead to an improved understanding of the transmission potential of smallpox during the early stages of the epidemic by providing a more accurate estimate of $\Pr(R_0 \leq 1)$, the probability that the epidemic dies out quickly.

The Abakaliki smallpox data are reported here as the days on which the removal of individuals actually took place, with the first day set to be time zero (see Table 3). Inspection of the data reveals large intervening periods between the first three removal times and thereafter, removals occurring far more frequently. This observation is consistent with the choice of two different fixed length infectious periods used by [3] for the first two removed individuals and for the remaining infectives. The analysis in [3] uses a single-type ($m = 1$) model, implying homogeneous susceptibility of all susceptible individuals, with no latent period ($c = 0$) and an initial susceptible population size of $N_1 = 120$. In our analysis, we also assume the single-type model and the same initial population size, but use a duration of $c = 13$ days as this is generally accepted to be the approximate length of the latent period for the smallpox virus [20].

The parameters of the prior distribution were chosen as follows. Infections were thought to occur at a rate of around 1 in 1,000 susceptible-infective contacts and so this was translated into fairly weak prior information by setting the prior mean and standard deviation for the (single) infection rate β_1 as 0.001 and 0.1, respectively. Prior information suggesting a mean infectious period of around ten days was included by choosing the prior mean and standard deviation for the $\gamma_j|k$ to both be 0.1. We sought basic step functions for the removal rate by taking a step location rate of $\lambda = 1$ and a maximum permitted number of steps of $k_{max} = 8$. This choice assigns most of the prior probability to the cases $k = 0$ and $k = 1$, with mean and standard deviation both approximately equal to 1, and $\pi(k = 0) = \pi(k = 1)$. Note that this choice, together with that for the $\gamma_j|k$, is sufficiently informative to ensure that the posterior distribution does not assign all probability to the simplest model. Also, taking $a = 2$ for the prior on the step positions incorporates fairly weak information whilst downweighting both small and large distances between consecutive step positions for all choices of k, and resulted in no small steps appearing in the posterior distribution. Finally, the first infection was assumed to have taken place no more than 50 days before the first removal ($d = 50$).

We report the results of a typical run of the the reversible jump MCMC algorithm described in Sect. 3. Convergence of the sampler was assessed by using standard diagnostics. Equal move probabilities ($q_1 = q_2 = q_3 = q_4 = 0.25$) for the step function proposals gave acceptable mixing rates. The algorithm required 100,000 iterations before convergence was apparent, and then a posterior sample of size 10,000

obtained by running the chain for a further 100 million iterations with a thin of 10,000. Figure 2 shows several marginal and conditional posterior distributions obtained from this MCMC output, and their summary measures are given in Table 4.

The posterior distribution of k shows that the majority of the posterior mass lies away from the time-homogeneous model ($k = 0$) and suggests that there is at least one change in the removal rate. Conditioning on the posterior modal number of change-points ($k = 1$), the posterior densities for γ_0 and γ_1 show a clear increase in the removal rate after the change-point, with $\widehat{\Pr}(\gamma_0 < \gamma_1 | k = 1, \boldsymbol{r}) \simeq 0.981$ (*s.e.* 0.002). The posterior density for the location of this change-point (s_1) has two large modes, the largest of these occurring at approximately $t = 20$ days after the first removal, that is, the time at which the third removal takes place. This is consistent with the hypothesis that the initial cases are removed far more slowly than those removed once the epidemic is known to be present in the population. It also agrees with [3], although that analysis assumed that only the first two removals had a longer infectious period rather than the first three removals found here. Interestingly, our analysis assuming $k = 1$ change-point has a further mode in the posterior density for s_1, taking place at approximately 55 days after the first removal. This finding is consistent with the locations of the change-points in the two change-point model. The $k = 2$ model is well supported by the posterior distribution and separates the bimodal distribution for $s_1 | k = 1$ into two (mainly) unimodal distributions for s_1 and s_2 (given $k = 2$). The figure also shows that, in this two change-point model, the removal rates increase over time, with $\widehat{\Pr}(\gamma_0 < \gamma_1 < \gamma_2 | k = 2, \boldsymbol{r}) \simeq 0.764$ (*s.e.* 0.008). This latter finding suggests that an as yet unknown factor at approximately day 55 caused removals to occur more quickly, resulting in the end of the epidemic on day 76. These conclusions corroborate those of [20], who detect bimodality in the posterior distribution of the mean infectious period and suggest that the mode representing the longer infectious period of approximately 18 days was attributable to the initial cases of infection. Our analysis also gives some indication of where the switch to a shorter infectious period occurs. O'Neill and Becker [20] follow up their analysis by deleting the first eight removals from the full data and refitting the model. As a result, the mode corresponding to the infectious period of the initial infectives at around 18 days in the posterior distribution for the mean infectious period disappears. The primary mode at approximately day 14 remains but a new mode appears at around day 8. Our analysis suggests that this new mode corresponds to a group of individuals with a particular short infectious period of around 4–5 days, these individuals being those removed after around 55 days.

Turning to inferences for the infection rate β_1, Fig. 2 shows its posterior distribution in the zero, one and two change-point models for the removal rate. The (unconditional) distribution obtained after taking account of the uncertainty regarding the number of change-points is very similar to that in the $k = 1$ model and therefore is not shown. We note that the posterior mean for $\beta_1 | k = 1$ is close to the maximum likelihood estimate ($\hat{\beta} = 0.00131$) obtained by [3] when assuming that the first two cases are infective for 14 days and the later cases for only 7 days. Its posterior standard deviation is larger than the standard error of the m.l.e. quoted by [3] (s.e.$(\hat{\beta}) = 0.00024$), mainly due to allowing for uncertainty on the infection times.

As noted earlier in the paper, inferences on the basic reproduction number are key to gaining an understanding of the spread of the disease and hence in developing

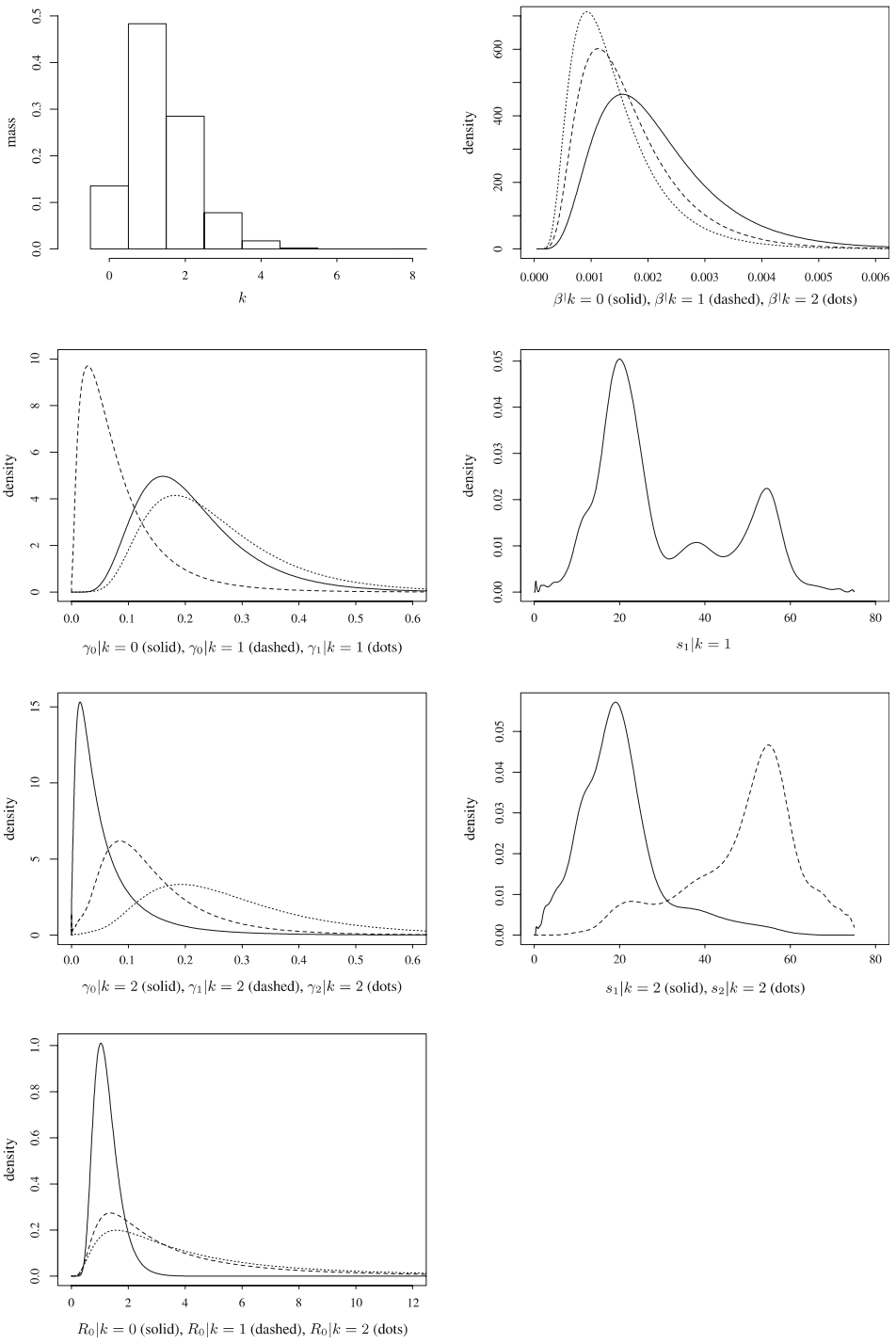


Fig. 2 Selected marginal posterior distributions for the smallpox data analysis

Table 4 Posterior means (and standard deviations) for selected parameters

Parameter			
k	1.37 (0.90)		
β_1	0.0016 (0.0007)		
R_0	4.02 (5.29)		
k	0	1	2
$\beta_1 \vert k$	0.0021 (0.0008)	0.0016 (0.0007)	0.0014 (0.0006)
$\gamma_0 \vert k$	0.206 (0.072)	0.077 (0.052)	0.055 (0.045)
$\gamma_1 \vert k$		0.246 (0.098)	0.136 (0.074)
$\gamma_2 \vert k$			0.272 (0.120)
$R_0 \vert k$	1.24 (0.34)	3.95 (4.94)	4.99 (5.99)

methods to tackle further epidemic spread. The lower plot in Fig. 2 demonstrates that posterior inferences for the basic reproduction number are sensitive to the number of removal rate change-points assumed in the model. For example, the model with a time-homogeneous removal rate is unable to capture the slower rate of removal during the early stages of the epidemic. Therefore inferences on γ_0 are dominated by the more rapid removal of later cases, resulting in this parameter being overestimated and, consequently, R_0 being underestimated. In contrast, models which allow for changes in the removal rate do capture the slower diagnosis of the earliest cases and enable more accurate inferences on γ_0 and hence on R_0. The (conditional) posterior distributions of $R_0|k$ for different k in Fig. 2 highlight the sensitivity of inferences about R_0, and hence about whether the epidemic will die out, on the number of change-points in the removal rate: $\widehat{\Pr}(R_0 \leq 1|k, \boldsymbol{r}) = 0.249$ ($k = 0$), 0.025 ($k = 1$), 0.015 ($k = 2$).

The sensitivity of the posterior distribution to the choice of prior distribution was assessed by using repeated runs of the analysis to study the effect of small changes in the prior specification. We found that such changes resulted in only very modest changes to the posterior distribution; further details can be found in [11].

The inferential comments above are only appropriate if they relate to a model which adequately describes the observed epidemic. We have compared the predictive distributions of the removal times for both the full model and the time-homogeneous model ($k = 0$) with the observed data. In this context, the predictive distribution is determined as the posterior average of realisations of the epidemic process, where the average is taken over the posterior distribution, that is, the predictive density for removal time r_{ij} is

$$f(r_{ij}|\boldsymbol{r}) = E_{\boldsymbol{\theta}|\boldsymbol{r}}\left[f(r_{ij}|\boldsymbol{\theta})\right]$$

where $\boldsymbol{\theta}$ represents all model unknowns (parameters and unobserved data) and $f(r_{ij}|\boldsymbol{\theta})$ denotes the density of the removal time induced by the epidemic model. Therefore, the predictive distributions can be obtained by repeatedly simulating epidemic realisations for randomly sampled iterates from the MCMC output $\boldsymbol{\theta}_i$. Figure 3 shows box and whisker plots of the predictive distributions for every fourth removal time (relative to the first removal time). The left plot of each pair is the distribution for the removal time given the simpler (time-homogeneous removal rate) model, and that on the right,

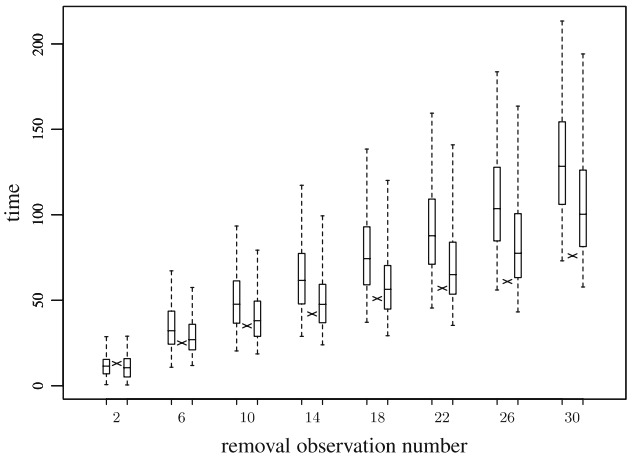


Fig. 3 Predictive distributions for smallpox removal times

the distribution for the more complex (time-inhomogeneous) model. The observed removal time is represented by the cross between the plots. The whiskers extend to the upper and lower $2\frac{1}{2}$ percentiles, and the box gives the median and inter-quartile range. The figure demonstrates that both models give a reasonable description of the removal times early in the epidemic. Later in the epidemic, the observed removals fall well into the lower tails of the predictive distribution for the simpler model. This model fails to capture the change in removal rate and therefore over-predicts the removal times. However, our larger and more flexible model removes much of this lack-of-predictive fit by explicitly modelling such changes. It also impacts inferences for the infection rate and thereby has the capacity to permit more accurate modelling of the unobserved infection times. This analysis shows clearly the benefit of allowing change-points in the removal rate on the fit of the model to the data.

6 Analysis of respiratory disease data

The second dataset concerns an outbreak of a respiratory disease on the South Atlantic island of Tristan da Cunha [4]. The data are presented in Table 5 and take the form of the number of daily removals in each of three different susceptibility categories (infants $= 1$, children $= 2$ and adults $= 3$). As in the analysis of the smallpox data, the times of infection are not observed. An initial study of the data reveals that there is a large intervening period between the first and second removal times, with more frequent removals occurring thereafter. This may indicate a slower rate of removal during the early stages of the epidemic. Once again the purpose of the analysis performed here is to locate change-points in the removal rate and to provide a model that better describes the transmission of the disease in question. In particular, an improvement in the accuracy of inferences made on key parameters such as the basic reproduction number during the early part of the epidemic are vital in providing estimates of the potential of the outbreak to become a major epidemic.

Table 5 Tristan da Cunha respiratory disease data

Group	Day															
	0	7	9	10	11	12	14	15	16	17	18	19	20	21	28	29
Infants	0	0	0	3	1	3	1	0	0	1	0	0	0	0	0	0
Children	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0
Adults	1	1	1	0	2	3	1	4	1	1	3	2	1	2	1	1

A Bayesian analysis of these data has been performed by [13] assuming a multitype model with three categories, containing initial numbers of susceptibles $N_1 = 25$, $N_2 = 36$ and $N_3 = 193$. As the disease is not known to have any significant latent period, they assume a negligible latent period by taking $c = 0$. We also adopt this model and investigate the extent to which the removal rate is time-inhomogeneous. Again we incorporate weak prior beliefs for the infection rate parameters β_i $(i = 1, 2, 3)$ by taking prior mean and standard deviations of 0.001 and 0.1, respectively. Weak prior beliefs are also included for the removal rate parameters $\gamma_j|k$: mean 0.1 and standard deviation 0.3. This choice of standard deviation is different to that used in the previous analyses, and appears to give a more coherent specification of the joint prior distribution of $(\boldsymbol{\gamma}, k)$ in this case. Similar arguments to those used in the smallpox analysis led to the same choice of the maximum number of change-points ($k_{max} = 8$), and of the other removal step function parameters ($\lambda = 1$, $a = 2$, $d = 50$).

The MCMC algorithm was implemented with $q_i = 0.25$ $(i = 1, 2, 3, 4)$ for a burn-in of 100,000 iterations, followed by a further 100 million iterations with a thin of 10,000, giving a sample of 10,000 iterates. Figure 4 gives posterior densities for the main model parameters, with summary statistics included in Table 6.

The posterior distribution for k shows strong support for the inclusion of one removal rate change in the model, with some support for two change-points. Conditional on there being $k = 1$ change-point, the position of the change seems to occur at a time s^* which is approximately on day 9 or day 10 of the epidemic. This is in contrast with the smallpox data, for which the distribution of $s_1|k = 1$ showed much larger variation. Note that the bi-modality of this distribution is due to the discrete nature of these data and disappears when the removal times are perturbed from their integer values. As in the previous analyses, given $k = 1$, the posterior distributions of γ_0 and γ_1 highlight a significant increase in the removal rate for cases diagnosed later in the epidemic compared to the rate at which the initial cases are removed. Turning to the model with two change-points in the removal rate, the marginal posterior distributions for the two change-points $(s_1 < s_2)$ both indicate a change-point at time s^*. Further, a study of the bivariate distribution of (s_1, s_2) reveals that this apparent contradiction is due to uncertainty about the location of a change-point in addition to that around time s^*. In particular, the conditional posterior distributions for $s_1|s_2 = s^*$ and $s_2|s_1 = s^*$ both strongly resemble their respective prior distributions which in turn explains the low posterior probability attached to this more complex model.

The marginal posterior distributions of the infection rate parameters β_i $(i = 1, 2, 3)$ in Fig. 4 show a consistent change in location between those obtained assuming a

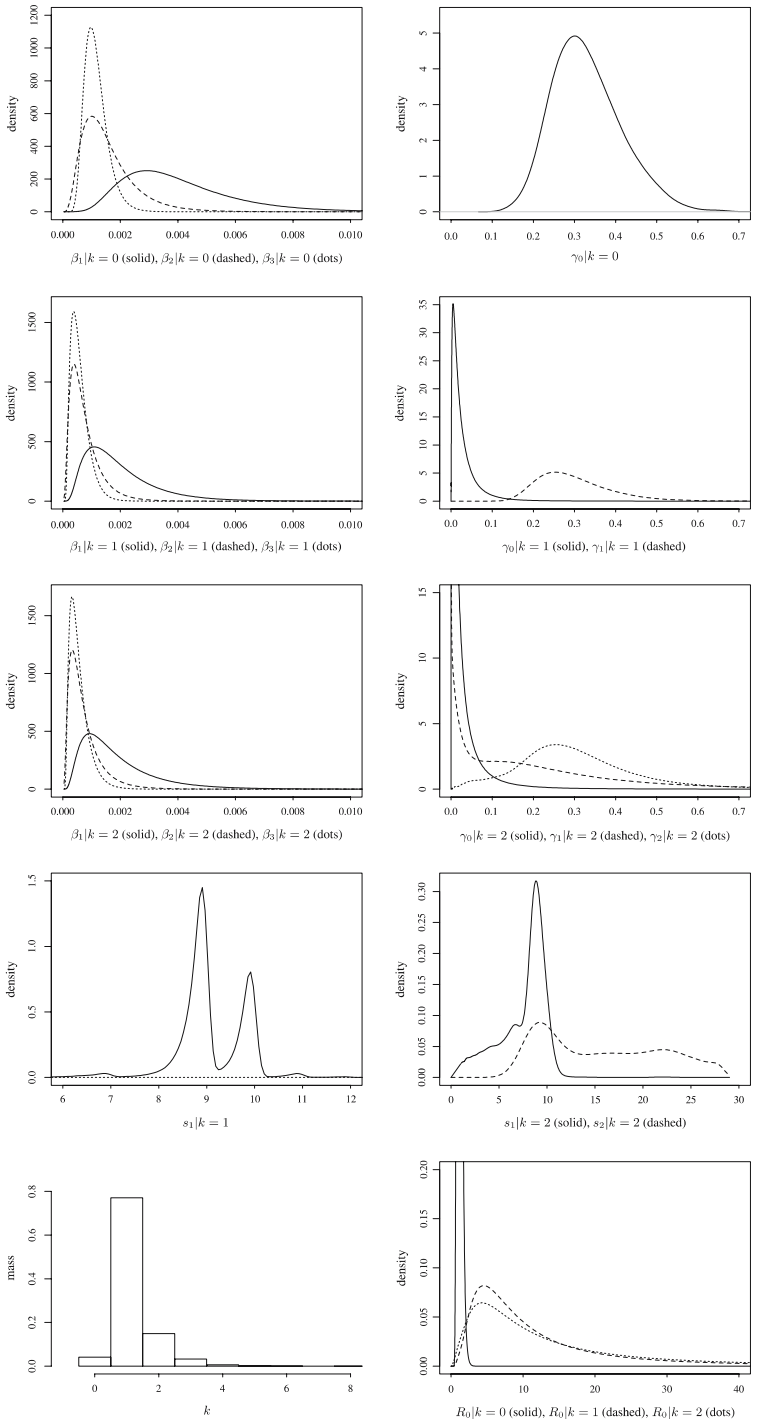


Fig. 4 Marginal posterior distributions for the respiratory disease data analysis

Table 6 Posterior means (standard deviations) for selected parameters

Parameter			
k	1.20 (0.60)		
β_1	0.0020 (0.0012)		
β_2	0.0008 (0.0005)		
β_3	0.0006 (0.0003)		
R_0	18.61 (144.56)		
k	0	1	2
$\beta_1 \vert k$	0.0039 (0.0015)	0.0019 (0.0011)	0.0018 (0.0011)
$\beta_2 \vert k$	0.0015 (0.0007)	0.0007 (0.0005)	0.0007 (0.0005)
$\beta_3 \vert k$	0.0011 (0.0003)	0.0006 (0.0003)	0.0005 (0.0003)
$\gamma_0 \vert k$	0.3304 (0.0830)	0.0304 (0.0396)	0.0294 (0.0514)
$\gamma_1 \vert k$		0.2982 (0.0886)	0.1827 (0.1557)
$\gamma_2 \vert k$			0.3166 (0.1852)
$R_0 \vert k$	1.15 (0.27)	16.49 (93.22)	31.62 (302.72)

time-inhomogeneous ($k \neq 0$) and a time-homogeneous ($k = 0$) removal rate, typically halving the posterior mean. The standard deviations are only slightly smaller for models which include changes in the removal rate. However, all these models indicate that children and adults have similar levels of susceptibility to the disease and that both are at a lower level than that of infants. Further details on statistical procedures for amalgamating susceptibility categories with similar infection rates can be found in [11].

Inferences on the basic reproduction number also depend on the form of the removal rate step function assumed in the model. As with the smallpox data analysis, it appears that the model with constant removal rate suggests that R_0 is smaller than models for which the removal rate is permitted to vary through time. Once again this is most likely due to the inability of the model for which $k = 0$ to capture the slower removal rate during the early stages of disease spread. The overestimation of the removal rate at the beginning of the epidemic results in posterior inferences on R_0 which underestimate its likely magnitude. This is also noticeable in the posterior probability that the basic reproduction number exceeds unity, as $\widehat{\Pr}(R_0 \leq 1 | k = 0, \boldsymbol{r}) = 0.314$, compared to $\widehat{\Pr}(R_0 \leq 1 | k = 1, \boldsymbol{r}) = 0.0017$ and $\widehat{\Pr}(R_0 \leq 1 | k = 2, \boldsymbol{r}) = 0.0061$. The model for which $k = 1$, which has far greater posterior probability than that for which $k = 0$, gives very different inferences on R_0 than the time-homogeneous model. The sensitivity of the posterior distribution to the choice of prior distribution was very similar to that described in the previous analyses; see [11] for full details of the sensitivity analysis.

An assessment of the improvement in fit of our model to these data, compared to the simpler time-homogeneous removal rate model, can once again be made by considering predictive distributions of the removal times. Epidemics with exactly 40 removals were simulated using parameters sampled from the joint posterior distribution of both the simple ($k = 0$) and more complex (variable k) models. The predictive distributions for the second to the eighth removals are included in Fig. 5. Box and whisker plots representing the predictive distributions of removal times for the simple model are on the left of each pair, and the cross gives the observed data point.For the simple model,

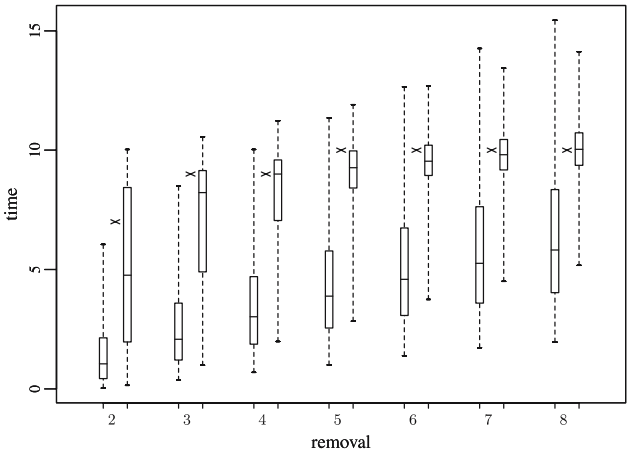


Fig. 5 Predictive distributions for respiratory disease removal times

inference on the removal rate is dominated by the later removal times, which occur at a faster rate than the first three removals. This explains why the observed times of the second and third removals occur later than the most likely times predicted by this simple model. The simple model provides a reasonable fit to the data after the occurrence of the initial removals. By including a time-dependent removal rate, the model gives a far better fit to the initial data points, and the predictive distribution of the subsequent removals has a smaller variance. The predictive distributions for later removals are not included here, but the more complex model continues to provide a better fit to these data than the simple model.

7 Conclusions

The models considered in this paper represent generalisations of the multitype SEIR compartmental model to include a set of models for which the removal rate is time-inhomogeneous. An MCMC algorithm similar to that of, amongst others, [21] and the reversible jump MCMC algorithm of [12] are jointly implemented to fit the models to simulated data and two real life epidemic outbreaks. Although only the removal times of the infective individuals are available for analysis, this information is sufficient to be able to determine the posterior distributions for the number and positions of the removal rate changes.

For both the smallpox data and the respiratory disease data, it appears that the initial cases are removed at a slower rate than later cases. This is most likely due to misdiagnoses of the initial cases as the epidemic is not known to be present within the population. The smallpox data also shows a possible removal rate change later in the course of the epidemic. The construction of predictive distributions for the removal times for both models demonstrates the improved fit of the models which include time-inhomogeneous removal rates. The improvement in fit might be more marked if

these methods were applied to analyse data from much longer epidemics, for which time-dependence of model parameters assumes greater importance [19].

Fitting models with time-dependent removal rate parameters also results in changes to the posterior distribution of the time-homogeneous infection rates. Moreover, large changes are exhibited in the posterior distributions of the basic reproduction number when assuming models for which the removal rate may vary through time. This information may be be vital for epidemiologists determining strategies for tackling the spread of a disease. Basing such strategies on the inferences made on R_0 clearly demands that the model accurately reflects the dynamics of disease spread, and takes into account changes in the dynamics through time. Time variation in the removal rate is likely to be exhibited in a wide range of epidemics, and the techniques outlined here give a method for determining such variation. The infection rates of the epidemic are also candidates for modelling as time-inhomogeneous parameters since the rate of transmission of the disease might change, for example, once knowledge of the epidemic is made public [5]. We have not investigated a model with time-dependent infection rates as the removal times data alone are not sufficient to determine the number and location of the changes in the infection rates. The reversible jump techniques considered here could be adapted to fit models with time dependency in both infection and removal rate parameters given more informative data.

Acknowledgments PRG gratefully acknowledges receipt of a research studentship from the Engineering and Physical Sciences Research Council.

A Appendix

Further details of the MCMC scheme outlined in Section 3.1 are included here. Recall that the MCMC scheme begins with some initial choice of parameters and hidden infection times, although the initial choice of parameters must be made such that the selection is feasible (and has positive density). The scheme then proceeds by cycling through the following steps.

Update infection times: Propose an update to the sequence of hidden infection times by reallocating a randomly (uniformly) chosen infection time to a new time τ', sampled from a $U(-d, T)$ distribution. Shift the associated time at which the individual becomes infective to a new time $\tau' + c$. Note that this will also update i_{min} if the group of the randomly selected infective is different to i_{min} and $\tau' < \tau_{i_{min},1}$. Accept the proposed move with probability $\min\{1, A\}$, where

$$A = \frac{\pi\left(\boldsymbol{\tau}', \boldsymbol{r} \middle| \boldsymbol{\beta}, \boldsymbol{\gamma}, k, \boldsymbol{s}, i'_{min}, \tau'_{i'_{min},1}\right)}{\pi\left(\boldsymbol{\tau}, \boldsymbol{r} \middle| \boldsymbol{\beta}, \boldsymbol{\gamma}, k, \boldsymbol{s}, i_{min}, \tau_{i_{min},1}\right)}.$$

Update infection parameters: Sample a new value for each infection rate parameter β_i $(i = 1, 2, \ldots, m)$ from its conditional posterior distribution given all other states

of the chain, that is

$$\beta_i|\cdot \sim \Gamma\left(g_{\beta_i} + n_i - \delta_{i,i_{\min}}\,,\; h_{\beta_i} + \int_{\tau_{imin},1}^{T} S_i(t)I(t)dt\right), \quad i = 1,2,\ldots,m$$

where δ_{ij} is Kronecker's delta function.

Update removal rate step function: The proposal of a new realisation of the removal rate step function is more complicated and involves different moves for each aspect of the step function. At each iteration, one of the following choices is made:

(i) *Update removal rates*: with probability q_1, simulate a new value for each removal rate γ_j ($j = 0,1,\ldots,k$), conditional on k, $\boldsymbol{s}$ and the other states of the chain, using

$$\gamma_j|k,\boldsymbol{s},\cdot \sim \Gamma\left(g_\gamma + w_j\,,\; h_\gamma + \int_{s_j}^{s_{j+1}} I(t)dt\right), \quad j = 0,1,\ldots,k$$

where w_j is the total number of removals occurring in (s_j, s_{j+1}).

(ii) *Update existing step positions*: with probability q_2, select one of the step positions s_j uniformly from the k existing steps. Propose moving s_j to a new position s_j' sampled from a $U(s_{j-1}, s_{j+1})$ distribution. Accept the proposed position with probability $\min\{1, A\}$, where

$$A = \frac{\pi\left(\boldsymbol{\tau}, \boldsymbol{r}\middle|\boldsymbol{\beta}, \boldsymbol{\gamma}, k, \boldsymbol{s'}, i_{min}, \tau_{i_{min},1}\right)}{\pi\left(\boldsymbol{\tau}, \boldsymbol{r}\middle|\boldsymbol{\beta}, \boldsymbol{\gamma}, k, \boldsymbol{s}, i_{min}, \tau_{i_{min},1}\right)} \times \left[\frac{(s_{j+1} - s_j')(s_j' - s_{j-1})}{(s_{j+1} - s_j)(s_j - s_{j-1})}\right]^{a-1}.$$

(iii) *Increase number of steps*: with probability q_3, propose an additional step at a position s' sampled uniformly over $(0, T)$. Suppose $s' \in (s_j, s_{j+1})$. The current removal rate γ_j over (s_j, s_{j+1}) must be split into removal rates γ_j' over the interval (s_j, s') and γ_{j+1}' over (s', s_{j+1}). Note that adding a new step induces an enlargement of the parameter subspace by two new parameters, s' and an additional removal rate. A proposed division of γ_j into $(\gamma_j', \gamma_{j+1}')$ is achieved by using a stochastic innovation and preserving a weighted geometric mean, that is, such that

$$(s' - s_j)\log\gamma_j' + (s_{j+1} - s')\log\gamma_{j+1}' = (s_{j+1} - s_j)\log\gamma_j$$

and

$$\frac{\gamma_{j+1}'}{\gamma_j'} = \frac{1-u}{u}$$

where u is sampled from a $U(0,1)$ distribution. The move is accepted with probability $\min\{1, A\}$, where

$$A = \frac{\pi(\boldsymbol{\tau}, \boldsymbol{r}|\boldsymbol{\beta}, \boldsymbol{\gamma}', k', \boldsymbol{s}', i_{min}, \tau_{i_{min},1})}{\pi(\boldsymbol{\tau}, \boldsymbol{r}|\boldsymbol{\beta}, \boldsymbol{\gamma}, k, \boldsymbol{s}, i_{min}, \tau_{i_{min},1})} \times \frac{\pi(k+1)}{\pi(k)}$$
$$\times \frac{\big[a(k+2)-1\big]!}{T^a\big[a(k+1)-1\big]!(a-1)!} \times \left[\frac{(s'-s_j)(s_{j+1}-s')}{(s_{j+1}-s_j)}\right]^{a-1}$$
$$\times \frac{h_\gamma^{g_\gamma}}{\Gamma(g_\gamma)}\left(\frac{\gamma_j'\gamma_{j+1}'}{\gamma_j}\right)^{g_\gamma-1} \exp\big\{-h_\gamma(\gamma_j' + \gamma_{j+1}' - \gamma_j)\big\}$$
$$\times \frac{q_4 T}{q_3(k+1)} \times \frac{(\gamma_j' + \gamma_{j+1}')^2}{\gamma_j}.$$

Note that a proposed increase in the number of steps beyond k_{max} cannot be accepted as $\pi(k = k_{max} + 1) = 0$.

(iv) *Decrease number of steps*: with probability $q_4 = 1 - q_1 - q_2 - q_3$, propose a removal of a step position, s_j say, chosen uniformly from the k existing steps. The removal rates (γ_{j-1}, γ_j) are replaced by a new removal rate γ_{j-1}' that satisfies the weighted geometric mean condition

$$(s_j - s_{j-1})\log\gamma_{j-1} + (s_{j+1} - s_j)\log\gamma_j = (s_{j+1} - s_{j-1})\log\gamma_{j-1}'.$$

The proposed move is accepted with probability $\min\{1, A\}$, where

$$A = \frac{\pi(\boldsymbol{\tau}, \boldsymbol{r}|\boldsymbol{\beta}, \boldsymbol{\gamma}', k', \boldsymbol{s}', i_{min}, \tau_{i_{min},1})}{\pi(\boldsymbol{\tau}, \boldsymbol{r}|\boldsymbol{\beta}, \boldsymbol{\gamma}, k, \boldsymbol{s}, i_{min}, \tau_{i_{min},1})} \times \frac{\pi(k-1)}{\pi(k)}$$
$$\times \frac{T^a\,(ak-1)!(a-1)!}{\big[a(k+1)-1\big]!} \times \left[\frac{(s_{j+1}-s_{j-1})}{(s_{j+1}-s_j)(s_j-s_{j-1})}\right]^{a-1}$$
$$\times \frac{\Gamma(g_\gamma)}{h_\gamma^{g_\gamma}}\left(\frac{\gamma_{j-1}'}{\gamma_{j-1}\gamma_j}\right)^{g_\gamma-1} \exp\big\{-h_\gamma(\gamma_{j-1}' - \gamma_{j-1} - \gamma_j)\big\}$$
$$\times \frac{q_3 k}{q_4 T} \times \frac{\gamma_{j-1}'}{(\gamma_{j-1} + \gamma_j)^2}.$$

Note that a proposed reduction in the number of steps below zero cannot be accepted as $\pi(k = -1) = 0$.

References

1. Anderson, D., Watson, R.: On the spread of a disease with gamma distributed latent and infectious periods. Biometrika **67**, 191–198 (1980)
2. Bailey, N.: Some stochastic models for small epidemics in large populations. Appl. Statist. **13**, 9–19 (1964)
3. Becker, N.G.: Analysis of data from a single epidemic. Aust. J. Statist. **25**, 191–197 (1983)

4. Becker, N.G., Hopper, J.L.: Assessing the heterogeneity of disease spread through a community. Amer. J. Epidem. **117**, 362–374 (1983)
5. Becker, N.G., Yip, P.S.F.: Analysis of variations in an infection rate. Austral. J. Statist. **31**, 42–52 (1989)
6. Dietz, K.: The estimation of the basic reproduction number for infectious diseases. Statist. Meth. Med. Res. **2**, 23–41 (1993)
7. Eichner, M., Dietz, K.: Transmission potential of smallpox: estimates based on detailed data from an outbreak. Am. J. Epidem. **158**, 110–117 (2003)
8. Gamerman, D., Lopes, H.F.: Markov Chain Monte Carlo, 2nd edn. Chapman & Hall, London (2006)
9. Geweke, J.: Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Smith, A.F., Dawid, A.P., Berger, J.O. (eds.) Bayesian Statistics 4, pp. 169–193. Oxford University Press, New York (1992)
10. Gibson, G.J., Renshaw, E.: Estimating parameters in stochastic compartmental models using Markov chain methods. IMA J. Math. Appl. Med. Biol. **15**, 19–40 (1998)
11. Giles, P.R.: Bayesian inference for stochastic epidemic models. Ph.D. thesis, School of Mathematics & Statistics, Newcastle University, UK (2005)
12. Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**, 711–732 (1995)
13. Hayakawa, Y., O'Neill, P.D., Upton, D., Yip, P.S.F.: Bayesian inference for a stochastic epidemic model with uncertain numbers of susceptibles of several types. Aust. N. Z. J. Statist. **45**, 491–502 (2003)
14. Heesterbeek, J.A.P., Dietz, K.: The concept of R_0 in epidemic theory. Statist. Neerl. **50**, 89–110 (1996)
15. Heidelberger, P., Welch, P.D.: Simulation run length control in the presence of an initial transient. Oper. Res. **31**, 1109–1144 (1983)
16. Keeling, M., Grenfell, B.T.: Effect of variability in infection period on the persistence and spatial spread of infectious diseases. Math. Biosci. **147**, 207–226 (1998)
17. Lloyd, A.: Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. Theor. Popul. Biol. **60**, 59–71 (2001)
18. Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., Getz, W.M.: Superspreading and the effect of individual variation on disease emergence. Nature **438**, 355–359 (2005)
19. O'Neill, P.D.: Epidemic models featuring behaviour change. Adv. Appl. Prob. **27**, 960–979 (1995)
20. O'Neill, P.D., Becker, N.G.: Inference for an epidemic when susceptibility varies. Biostatistics **2**, 99–108 (2001)
21. O'Neill, P.D., Roberts, G.O.: Bayesian inference for partially observed stochastic epidemics. J. R. Statist. Soc. A **62**, 121–129 (1999)
22. Renshaw, E., Gibson, G.J.: Can Markov chain Monte Carlo be usefully applied to stochastic processes with hidden birth times? Inverse Probl. **14**, 1581–1606 (1998)
23. Stephens, M.: Bayesian analysis of mixtures with an unknown number of components - an alternative to reversible jump methods. Ann. Statist. **28**, 40–74 (2000)