Biostatistics (2004), **5**, 2, *pp.* 249–261
Printed in Great Britain

Statistical inference and model selection for the 1861 Hagelloch measles epidemic

PETER J. NEAL[†], GARETH O. ROBERTS
Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK
P.Neal-2@umist.ac.uk

SUMMARY

A stochastic epidemic model is proposed which incorporates heterogeneity in the spread of a disease through a population. In particular, three factors are considered: the spatial location of an individual's home and the household and school class to which the individual belongs. The model is applied to an extremely informative measles data set and the model is compared with nested models, which incorporate some, but not all, of the aforementioned factors. A reversible jump Markov chain Monte Carlo algorithm is then introduced which assists in selecting the most appropriate model to fit the data.

Keywords: Model choice; Reversible jump MCMC; Stochastic epidemics.

1. INTRODUCTION

The statistical analysis of temporal infectious disease data is often complicated by a lack of complete data. Therefore it is crucial to make the most of 'complete' data sets in order to try to understand infection paths, and to devise effective control strategies (for example, vaccination policies). One such data set where the data are far more 'complete' than usual is the Hagelloch data set of Pfeilsticker (1863).

This data set concerns a severe outbreak of measles in an isolated German village (Hagelloch) in the winter of 1861. The last previous measles outbreak in Hagelloch occurred in the winter of 1847, and this suggests that only those children born after 1847 are susceptible to measles. This conjecture is supported by the data since 185 out of the 197 children under the age of 14 were infected, whilst only three (all of whom were aged either 14 or 15 and had not previously contracted measles) out of the remaining 380 village inhabitants are infected. Pfeilsticker also provides limited information about the 12 children under the age of 14 not infected during the course of the epidemic. Seven were infants under the age of twelve months and were therefore presumably carrying placental immunity. Three children were kept totally isolated during the course of the epidemic, one child was aged two years old (and therefore did not attend the school) and the final child was an immigrant who had previously had measles. We shall therefore assume that the total susceptible population became infected. This is a reasonable assumption which is necessitated by the lack of data about the uninfected members of the population (the vast majority of whom were almost certainly immune to infection). For the same reason we shall ignore the interactions in the population between those infected and the remaining two-thirds of the population. This assumption is not as unreasonable as at first it may seem since the epidemic is restricted to children, of whom over 90% are infected. Therefore we restrict the population of interest to the children within the village and ignore the adults.

[†]To whom correspondence should be addressed.

The data set has previously been studied by Lawson and Leimich (2000). Their approach is based on methods developed in spatial epidemiology and the model is fitted using a proportional hazards approximation. We take a very different approach using a stochastic epidemic model (see, for example, O'Neill and Roberts (1999)). The specific scientific question of interest is the significance of the spatial effect in the transmission mechanism. However, the methodology developed here can be used more generally to explore various important features of the transmission of infection.

Usually, temporal infectious disease data only give information about the appearance of first visible symptoms (e.g. the appearance of a rash for measles). The Hagelloch data set, by contrast, is extremely rich. For each infected individual the following information was obtained by Pfeilsticker: name, age (in years), sex, date of first sign of symptoms, date rash appeared, class of child at the village school, date of death (where appropriate), most likely source of infection, number of days between first sign of symptoms in infector (most likely source of infection) and infected, complications due to other diseases, location of the individual's home, number of cases within family, maximum temperature and day of maximum fever (days after rash appears).

By exploiting this unusually rich data-set, we wish to investigate which factors are the most important in the spread of the disease. The period when an individual is infectious is of particular importance. In accordance with usual practice for measles, we assume that each individual is infectious from some time before the first sign of symptoms to some time after the appearance of the measles rash. (Typically the first symptoms of measles are Koplik spots which appear on the inside of the cheek.) Further, we assume that an individual is equally infectious throughout their infectious period. This is a questionable assumption since it is well known that an individual's infectivity varies during their infectious period. However, we make the assumption to avoid the model becoming over-complicated. For ease of exposition, initially we do not include latent periods in the model. The latent period for measles is approximately a week in length, the time from infection until the individual becomes infectious. As we shall see later the results obtained with latent periods included are very similar to those obtained without latent periods. We consider both fixed length and unknown (imputed) infectious periods.

The three factors we consider in the model beyond the infection times (appearance of symptoms and rash) are the household and school class (if any) to which an individual belongs and the distance between the different households. We have chosen to omit the other factors for the following reasons. The day of maximum fever (and the corresponding maximum temperature) should provide a good indication of how long an individual is infectious after the appearance of the rash. However, this is complicated by more than 50% missing data and the data which are known are often distorted by the presence of other diseases. The age and sex of an individual are important in determining their mixing patterns with the rest of the population. With regard to age, a general grouping is provided by school class, which suffices. The data suggests that characteristics of the disease such as susceptibility, infectiousness, severity of disease, etc, do not depend upon the sex of an individual. Therefore we chose to omit the sex of an individual from the model. Our analysis of the data does not require us to assign a particular individual as the infector of a given individual.

2. MODEL

Let n and m denote the total size of the population and the eventual number infected, respectively. Label the individuals who are infected as $i = 1, 2, \ldots, m$ and those who remain susceptible throughout the course of the epidemic as $i = m + 1, m + 2, \ldots, n$. (Note that for the Hagelloch data, we assume that $m = n = 188$.) For individual i, let S_i and Q_i denote the dates upon which the symptoms and the rash first appear, respectively. Let I_i denote the time at which individual i becomes infected and let R_i denote the end of individual i's infectious period, i.e. when individual i is removed. Note that S_i and Q_i

are known while I_i and R_i are unknown and need to be imputed. We assume that there exists a constant $d_0 \in \mathbb{N}$ such that $R_i = \min\{D_i, Q_i + d_0\}$ where D_i denotes the date of individual i's death. For $S_i - I_i$ we consider two models. In the first model we assume that there exists $d_1 \in \mathbb{N}$ such that $S_i - I_i = d_1$. Therefore the dates, $\mathbf{I} = (I_1, I_2, \ldots, I_n)$, upon which the individuals are infected are assumed known given the data. In the second model we assume that the infection times are unknown, and therefore we treat $\mathbf{I}$ as parameters in the model. Furthermore, we shall assume that $S_i - I_i \sim \text{Gam}(\omega, \delta)$ where ω and δ are unknown parameters. For $k = m + 1, m + 2, \ldots, n$, set $I_k = \infty$.

We assume that individual, i say, exerts a constant infection rate throughout the course of their infectious period (i.e. over the time interval $(I_i, R_i]$). We assume that, while infectious, individual i makes infectious contacts with individual j, say, at rate α_{ij} per day where α_{ij} depends upon the relationship between individuals i and j. In other words, the probability that individual i fails to make an infectious contact with individual j on any given day is $\exp(-\alpha_{ij})$. In particular, we assume that

$$\alpha_{ij} = \beta_H 1_{\{\rho(i,j)=0\}} + \beta_C^1 1_{\{L_i=L_j=1\}} + \beta_C^2 1_{\{L_i=L_j=2\}} + \beta_G \exp(-\theta\rho(i, j)) \tag{2.1}$$

where β_H, β_C^1, β_C^2 β_G and θ are non-negative parameters of interest, $\rho(i, j)$ denotes the distance between the households of individuals i and j and L_i denotes the school classroom (either 1 or 2) to which individual i belongs and $L_i = 0$ if individual i does not attend the school. Therefore β_H, β_C^1 and β_C^2 denote the within-household, within-classroom 1 and within-classroom 2 infection rates, respectively. (Preliminary analysis of the data suggest a different within-classroom infection rate for each of the two classrooms.) Also β_G denotes the global infection rate whilst θ governs the extent to which distance between individuals reduces the global transmission rate. Therefore the infection rate between two individuals can be split into three categories: household, spatial (global) and classroom. The household and classroom effects are self-explanatory, in that we assume there is increased contact between two individuals if they share the same house or classroom. In particular, we assume that there is a 'nugget' household effect on top of the spatial effect. This seems a reasonable assumption owing to the relatively cramped living conditions (by modern standards) in the middle of the nineteenth century. The spatial component needs a bit more consideration. The choice of $\beta_G \exp(-\theta\rho(i, j))$ to model the spatial component in the disease spread is somewhat arbitrary, but does seem qualitatively reasonable. By choosing an exponential decay for the spatial infection, we allow global/spatial infection throughout the whole population, whilst maintaining a spatial element to the spread of the disease. In Section 3, we shall consider two alternative models for the spatial component, and we will show that all three different spatial components considered produce similar qualitative results.

Clearly, θ is highly dependent on the distance measure used. The households were plotted by Oesterle on a square reference grid approximately 100 × 100 units (Oesterle, 1992). (The length of each unit is 2.5 m.) We rescale so that the reference grid is approximately a unit square and give a plot of the households in Figure 1. In particular, we rescale so that each unit is 250 metres in length.

Let $g(x)$ $(x \geqslant 0)$ be the probability density function for $S - I$ and let γ denote the parameters of the density $g(\cdot)$. The parameters γ may either be known or unknown. Then

$$f(\mathbf{I}, \mathbf{R}|\gamma, \theta, \boldsymbol{\beta}, \mathbf{Z}, \mathbf{P}, I_\kappa) = \prod_{j \neq \kappa} \left\{ \sum_{i \in Y_j} \alpha_{ij} \right\} \exp(-A) \prod_{i=1}^{m} g(S_i - I_i) \tag{2.2}$$

where κ denotes the initial infective (i.e. $I_\kappa = \min_{1 \leqslant j \leqslant m}\{I_j\}$), $Y_j = \{k : I_k < I_j \leqslant R_k\}$, $\mathbf{Z} = (\mathbf{D}, \mathbf{L}, \mathbf{S}, \mathbf{Q})$ (the infectious periods data), $\boldsymbol{\beta} = (\beta_H, \beta_C^1, \beta_C^2, \beta_G)$, $A = \sum_{j=1}^{m} \sum_{k=1}^{n} \alpha_{jk}\{(R_j \wedge I_k) - (I_j \wedge I_k)\}$ and P_i $(1 \leqslant i \leqslant n)$ denotes the spatial position of individual i's household. Therefore Y_j $(1 \leqslant j \leqslant m)$ denotes the set of individuals who are infectious when individual j becomes infected

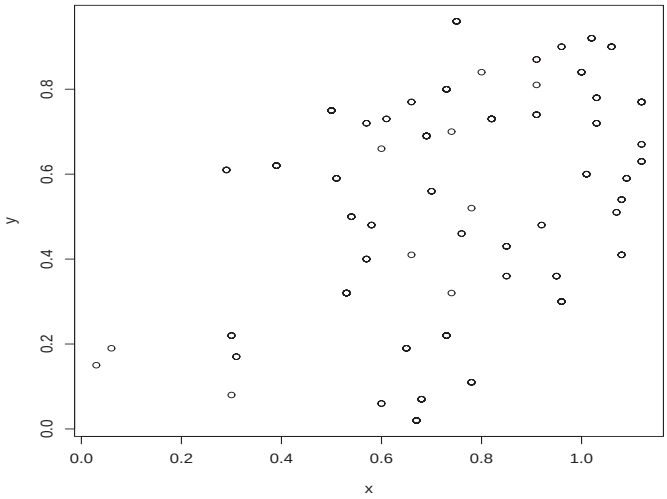


Fig. 1. Location of the households in Hagelloch containing at least one infected individual with locations on the approximate unit square.

and A denotes the total amount of person-to-person infectious pressure observed during the course of the epidemic.

To each parameter β_H, β_C^1, β_C^2, β_G and θ we assign independent Gamma priors, namely, $\pi(\zeta) \sim \text{Gam}(\nu_\zeta, \lambda_\zeta)$ where $\zeta = \beta_H, \beta_C^1, \beta_C^2, \beta_G, \theta$. Suppose that γ comprises k parameters $\gamma_1, \gamma_2, \ldots, \gamma_k$. Then we assign Gamma priors, namely, $\pi(\gamma_i) \sim \text{Gam}(\nu_i^\gamma, \lambda_i^\gamma)$ $(1 \leqslant i \leqslant k)$. Also, if $\mathbf{I}$ is unknown, we assign a prior to I_κ, in particular $\pi(I_\kappa) \propto \exp(\epsilon I_\kappa)$ for some $\epsilon > 0$. (Note that $I_\kappa < Q_1$.) In each iteration we update each of the parameters γ and the infection times $\mathbf{I}$. At each iteration we update 18 of the infection times, updating one infection time at a time. (Updating approximately 10% of the infection times in each iteration was found to be close to optimal in terms of computing time and convergence of the MCMC algorithm.) The rationale for single site rather than block updating is given in Neal and Roberts (2003).

We adopted the following MCMC scheme to simulate from the joint distribution $f(\mathbf{I}, \mathbf{R}, \gamma, \theta, \boldsymbol{\beta}, \mathbf{Z}, \mathbf{P})$.

For each of the parameters θ, β_H, β_G, β_C^1 and β_C^2, we updated the parameter using random walk Metropolis with a Gaussian proposal; for example, $\theta' \sim N(\theta, \sigma)$, where σ is 'tuned' to give an overall acceptance rate of approximately 0.4 (see, for example, Gelman *et al.* (1996)). This simple MCMC method is easy to implement and we found it gave adequate results.

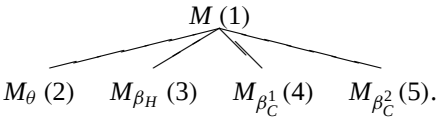
The γ parameters are updated using a Gibbs step where appropriate, or random walk Metropolis with a Gaussian proposal otherwise. In particular, when the infectious periods are assumed to be Gamma distributed with parameters ω and δ, i.e. $\text{Gam}(\omega, \delta)$, we update δ using a Gibbs step while updating ω requires random walk Metropolis.

Updating $\mathbf{I}$. Suppose that we are updating I_j. We propose a new value I_j' by sampling $S_j - I_j'$ from its conditional posterior density. Let κ' denote the initial infective under the proposed change. Then we accept the move with probability p_{acc}, where

$$p_{acc} = \min\left\{1, \frac{\prod_{j \neq \kappa'}\{\sum_{i \in Y_j'} \alpha_{ij}\}}{\prod_{j \neq \kappa}\{\sum_{i \in Y_j} \alpha_{ij}\}} \exp(-(A' - A)) \exp(\epsilon(I_{\kappa'}' - I_\kappa))\right\}.$$

We call the model described above the full model, denoted M. However, we are interested in model

selection, and in particular, wish to establish which of the parameters θ, β_H, β_C^1 and β_C^2 are important. (Note that we require $\beta_G > 0$ for the model to be valid.) Therefore for each of the parameters θ, β_H, β_C^1 and β_C^2, we used reversible jump (RJ) MCMC (Green (1995)) to move between M and sub-models for which one of the parameters is set equal to 0. Let M_X denote the model with parameter X set equal to 0, where $X = \theta$, β_H, β_C^1, β_C^2. We also number the models one through five with models 1, 2, 3, 4 and 5 corresponding to models M, M_θ, M_{β_H}, $M_{\beta_C^1}$ and $M_{\beta_C^2}$, respectively. The models are nested as shown below.



The algorithm was then implemented as above except that a model switching step was added. We now describe how the reversible jump moves were implemented.

Firstly, we consider a move from M to M_θ (the latter corresponding to homogeneous spatial mixing). We leave the parameters β_H, β_C^1 and β_C^2 as they are. We propose $\beta_G' = \beta_G \exp(-\tau_1\theta)$ for some $\tau_1 \geqslant 0$. For the reverse move, we require an auxiliary random variable U. Let $U \sim \text{Exp}(\tau_2)$ for some $\tau_2 > 0$. Then set $\theta = u$ and $\beta_G' = \beta_G \exp(\tau_1 u)$. Then τ_1 and τ_2 can be chosen so as to optimize mixing between the two models. The Jacobian for the transformation from M to M_θ is $\exp(-\tau_1\theta)$. Therefore,

$$\alpha_{ij}' = \beta_H 1_{\{\rho(i,j)=0\}} + \beta_C^1 1_{\{L(i)=L(j)=1\}} + \beta_C^2 1_{\{L(i)=L(j)=2\}} + \beta_G',$$

and we therefore accept the proposed move with probability

$$\min\left\{1, 4\prod_{j\neq\kappa}\left(\frac{\sum_{i\in Y_j}\alpha_{ij}'}{\sum_{i\in Y_j}\alpha_{ij}}\right)\exp(-(A'-A))\frac{\tau_2\exp(-\tau_2\theta)\Gamma(\nu_\theta)}{\lambda_\theta^{\nu_\theta}\theta^{\nu_\theta-1}\exp(-\lambda_\theta\theta)}\left(\frac{\beta_G'}{\beta_G}\right)^{\nu_{\beta_G}-1}\exp(-\lambda_{\beta_G}(\beta_G'-\beta_G))\frac{\zeta_2}{\zeta_1}\mathrm{e}^{-\tau_1\theta}\right\}, \tag{2.3}$$

where ζ_i $(1 \leqslant i \leqslant 5)$ denotes the prior assigned to model i. The factor 4 in (2.3), corresponds to the fact that if we are currently using M_θ then we will always propose model M, whilst there is probability $\frac{1}{4}$ of proposing a move from M to M_θ.

Secondly, we consider moves between M and M_{β_H}, with the parameters β_G and θ left unchanged during the move. We propose $(\beta_C^1)' = \beta_C^1\frac{\beta_H+\beta_C^1+\beta_C^2}{\beta_C^1+\beta_C^2}$ and $(\beta_C^2)' = \beta_C^2\frac{\beta_H+\beta_C^1+\beta_C^2}{\beta_C^1+\beta_C^2}$. For the reverse move we use an auxiliary random variable $U \sim U(0,1)$. Then we set $(\beta_C^1)' = u\beta_C^1$, $(\beta_C^2)' = u\beta_C^2$ and $\beta_H' = (1-u)(\beta_C^1+\beta_C^2)$. The Jacobian for the transformation from M to M_{β_H} is $\frac{1}{\beta_C^1+\beta_C^2}$. Therefore

$$\alpha_{ij}' = (\beta_C^1)'1_{\{L(i)=L(j)=1\}} + (\beta_C^2)'1_{\{L(i)=L(j)=2\}} + \beta_G\exp(-\theta\rho(i,j)),$$

and we therefore accept the proposed move with probability

$$\min\left\{1, 4\prod_{j\neq\kappa}\left(\frac{\sum_{i\in Y_j}\alpha_{ij}'}{\sum_{i\in Y_j}\alpha_{ij}}\right)\exp(-(A'-A))\prod_{i=1}^{2}\left\{\left(\frac{(\beta_C^i)'}{\beta_C^i}\right)^{\nu_{\beta_C^i}-1}\exp(-\lambda_{\beta_C^i}((\beta_C^i)'-\beta_C^i))\right\}\frac{\Gamma(\nu_{\beta_H})}{\lambda_{\beta_H}^{\nu_{\beta_H}}}\beta_H^{1-\nu_{\beta_H}}\exp(\lambda_{\beta_H}\beta_H)\frac{\zeta_3}{\zeta_1}\frac{1}{\beta_C^1+\beta_C^2}\right\}.$$

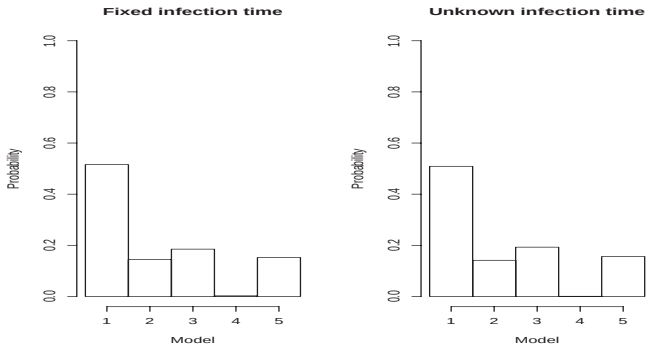


Fig. 2. Histograms for the time spent in each model for fixed and unknown infection times, respectively. The models are labelled as before (i.e. M, M_θ, M_{β_H}, $M_{\beta_C^1}$ and $M_{\beta_C^2}$ correspond to models 1, 2, 3, 4 and 5, respectively).

A similar procedure is used for moves between M and $M_{\beta_C^1}$ (and also M and $M_{\beta_C^2}$). We propose $\beta_H' = \beta_H \frac{\beta_H+\beta_C^1+\beta_C^2}{\beta_H+\beta_C^2}$ and $(\beta_C^2)' = \beta_C^2 \frac{\beta_H+\beta_C^1+\beta_C^2}{\beta_H+\beta_C^2}$. For the reverse move we use $U \sim U(0,1)$ and set $\beta_H' = u\beta_H$, $(\beta_C^2)' = u\beta_C^2$ and $(\beta_C^1)' = (1-u)(\beta_H + \beta_C^2)$.

3. RESULTS

3.1 *Hagelloch data set*

We began by running the algorithm for the Hagelloch data set, firstly with fixed infection times and then with unknown (imputed) infection times. The results for fixed and unknown infection times are very similar especially with regards to model selection as demonstrated in Figures 2 and 3 below. In both cases we obtained samples of size 20 000 taken after every five iterations with a burn-in period of 1000 iterations. For each of the parameters ω, δ and I_κ we assigned an Exp(1) prior when necessary. We set $\pi(\theta) \sim \text{Exp}(0.1)$ and $\pi(\beta) \sim \text{Exp}(10)$ for $\beta = \beta_H, \beta_G, \beta_C^1, \beta_C^2$. We assign a uniform prior for the models, i.e. $\zeta_i = \frac{1}{5}$ $(1 \leqslant i \leqslant 5)$. We take $d_0 = 3$, that is, an individual is infectious for three days after the appearance of the rash (unless they die). Also for fixed infection times we take $d_1 = 1$: that is, an individual becomes infectious a day before the appearance of symptoms (the results are robust to different choices of d_1, for example $d_1 = 5$ produces very similar results). These assumptions on d_0 and d_1 are reasonable for measles. For the model selection step, we take $\tau_1 = 0$ and $\tau_2 = 1$. Thus, when moving between models M and M_θ, β_G is left unchanged and θ (where necessary) is drawn from an Exp(1). The resulting reversible jump step exhibited excellent mixing whilst mixing over the parameter space was more than adequate.

The results give most support for the full model. However, there is also noticeable support for M_{β_H}, $M_{\beta_C^2}$ and M_θ. There is very strong evidence that $\beta_C^1 > 0$: that is, that classroom 1 plays a crucial role in the spread of the disease.

3.2 *Model extensions*

We begin by presenting two alternatives to the exponential spatial component: namely, the Cauchy and discrete spatial components. The MCMC procedure is the same as that presented in Section 2 with the only difference being that (2.1) is replaced by (3.1) and (3.2) for the Cauchy and discrete spatial components,

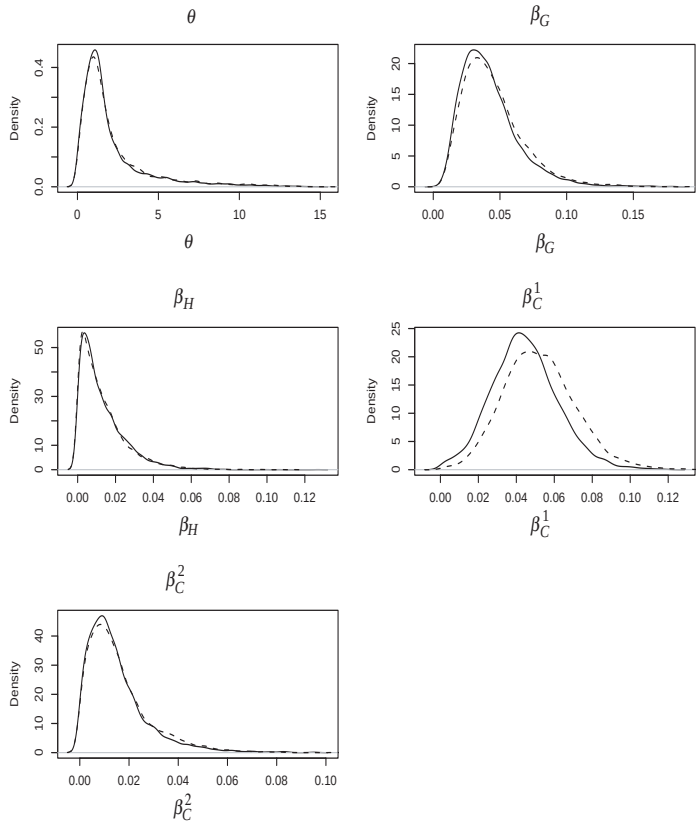


Fig. 3. Density plots for the parameters θ, β_G, β_H, β_C^1 and β_C^2 for both fixed (solid line) and unknown (dashed line) infection times, conditional on the full model.

respectively, where

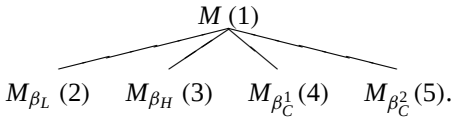
$$\alpha_{ij} = \beta_H 1_{\{\rho(i,j)=0\}} + \beta_C^1 1_{\{L_i=L_j=1\}} + \beta_C^2 1_{\{L_i=L_j=2\}} + \beta_G (1+\theta\rho(i,j)^2)^{-1} \tag{3.1}$$

and

$$\alpha_{ij} = \beta_H 1_{\{\rho(i,j)=0\}} + \beta_C^1 1_{\{L_i=L_j=1\}} + \beta_C^2 1_{\{L_i=L_j=2\}} + \beta_L 1_{\{\rho(i,j)<\theta\}} + \beta_G. \tag{3.2}$$

The Cauchy spatial component is self-explanatory. The discrete spatial component on the other hand requires a bit more explanation.

The discrete spatial component comprises household, local and global infection rates, β_H, β_L and β_G, respectively. We allow θ (the measure of how close two households need to be, to be considered local) to be a random variable and we shall set $\pi(\theta) \sim U(0.1, 0.4)$. Therefore the distance which constitutes local ranges from 25 to 100 m. Note that setting $\beta_L = 0$ and/or $\theta = 0$ eliminates the spatial component. We shall consider the following nested models with M_{β_L} corresponding to no spatial effect:



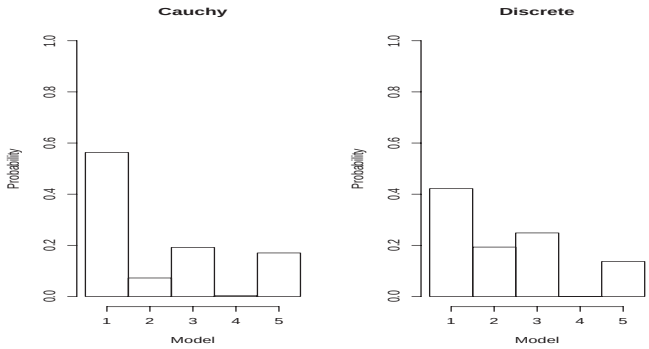


Fig. 4. Histograms for the time spent in each model for Cauchy and discrete spatial components, respectively.

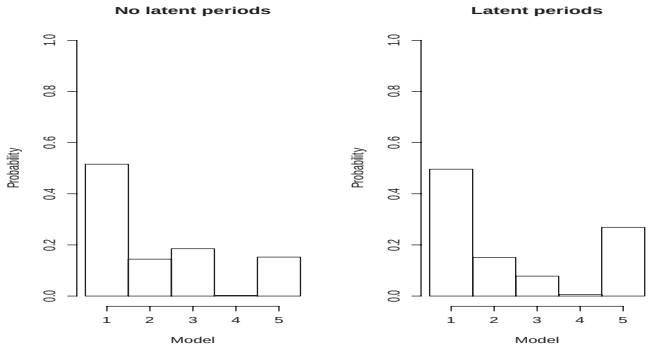


Fig. 5. Histograms for the time spent in each model for the Hagelloch data set without and with latent periods, respectively.

Therefore, once again model 2 represents no spatial effect. The model switching steps between M and each of M_{β_H}, $M_{\beta_C^1}$ and $M_{\beta_C^2}$ are as in Section 2. For a move from M to M_{β_L}, we propose $\beta'_G = \beta_G \frac{\beta_G+\beta_L+\beta_H}{\beta_G+\beta_H}$ and $\beta'_H = \beta_H \frac{\beta_G+\beta_L+\beta_H}{\beta_G+\beta_H}$. For the reverse move we again use $U \sim U(0, 1)$ as an auxiliary random variable setting $\beta'_G = u\beta_G$, $\beta'_H = u\beta_H$ and $\beta'_L = (1-u)(\beta_G + \beta_H)$. The Jacobian for the transformation and, hence, the acceptance probability of the move are then straightforward to calculate.

Obviously, the interpretation of the spatial parameter(s) θ (and β_L) varies from model to model but in terms of model selection and, in particular, the importance of the spatial component, the results are similar as demonstrated in Figure 4. For both models, we set $\pi(\beta) \sim \text{Exp}(10)$ for $\beta = \beta_H, \beta_G, \beta_C^1, \beta_C^2$, ($\beta_L$ for the discrete model) and we set $\pi(\theta) \sim \text{Exp}(0.1)$ for the Cauchy model.

Similar results were obtained with other spatial components we considered. Therefore, since the model selection question is robust to the choice of model for the spatial component, we focus our further analysis on the model of Section 2.

The introduction of a latent period into the original model is very straightforward. As an example, we introduce a fixed length latent period of a week, adjusted where necessary to ensure that the model is consistent with the data, into the model with fixed infectious periods. The results obtained are very similar to those obtained previously. This is demonstrated in Figure 5. Similar results are again obtained if the latent period and/or infectious periods are unknown and need to be imputed.

Table 1. *True model parameters for simulated data sets*

Data set	θ	β_H	β_G	β_C^1	β_C^2
SD1	5	0.05	0.05	0.06	0.04
SD2	0	0.05	0.025	0.075	0.05
SD3	5	0	0.075	0.075	0.05
SD4	5	0.1	0.05	0	0.05
SD5	5	0.075	0.075	0.075	0

3.3 *Comparisons with simulated data*

We now compare the results from the Hagelloch data set with the results obtained from five simulated data sets, thereby allowing us to assess the ability of our methodology to distinguish between models given the observed data set. In each of the simulated data sets we structure the population as in the Hagelloch data set. Then for $X = 1, 2, 3, 4, 5$, we simulate a data set with X as the true model and label the data set SDX. The model parameters were chosen and the simulations run to ensure that the entire susceptible population was infected as in the case of Hagelloch. We used fixed (known) infection times with $d_0 = 3$, $d_1 = 1$ and $Q_i - S_i$ $(1 \leqslant i \leqslant m)$ drawn uniformly at random from $\{1, 2, 3, 4\}$. Simulations where infection times were unknown (imputed) produced similar results.

The parameter values in Table 1 are generally higher than the means of the parameter posterior densities given by Figure 3. The primary reason for this is that the posterior densities in Figure 3 are for the full model M_1, while the simulations are done (with the exception of SD1) using the sub-models M_2–M_5.

The algorithms were again run to obtain samples of size 20 000 taken after five iterations with a burn-in period of 1000 iterations. We assign a uniform prior for the models. We set $\pi(\theta) \sim \text{Exp}(0.1)$ and $\pi(\beta) \sim \text{Exp}(10)$ for $\beta = \beta_G, \beta_H, \beta_C^1, \beta_C^2$. (Therefore we use the same priors as in Section 3.1.) Also, we set $\tau_1 = 0$ and $\tau_2 = 1$ as before. The reversible jump step again exhibits excellent mixing.

Figure 6 shows that the algorithm is very good at detecting the 'correct' model except in the case β_H (SD3). In SD1, more informative priors give increasing support for the full model, since the parameter priors that we have chosen are fairly uninformative and such choice of priors penalize the full model. The results of the simulation study are very promising and give further credence for using the full model for the Hagelloch data set.

One major concern with the algorithm, which we now address, is its apparent inability to detect that $\beta_H = 0$ in SD3. This is due to problems of identification within the model. The absence of a classroom 1 effect (i.e. $\beta_C^1 = 0$) reduces the chance that two individuals in classroom 1 are infected at the same time, especially if they live far apart. Therefore it is fairly easy to detect whether $\beta_C^1 = 0$ or not. The same argument applies for classroom 2. Suppose that we were to assume that $\beta_C^1 = \beta_C^2 = \beta_C$, say. Then it is even easier to detect the absence of a school parameter (i.e. $\beta_C = 0$) since the spread of the epidemic is then purely spatial. (This is particularly noticeable if θ is large.) On the other hand if $\theta = 0$, there is no spatial spread of the disease. The absence of a spatial parameter leads to an overlapping groups model with (uniform) global infection (see, for example, Becker and Dietz (1995)). However, the absence of a household parameter (i.e. $\beta_H = 0$) does not alter the spatial nature of the disease spread. Also, if β_G and θ are large, there is a high within-household infection rate, similar to that found in the full model. In other words, the parameters β_G and θ can often compensate for $\beta_H = 0$ to 'create' a household effect. The classroom mixing is left unaltered from the full model, which implies that the epidemics produced by models M and M_{β_H} are very similar. Therefore unless the priors for either the parameters or the models

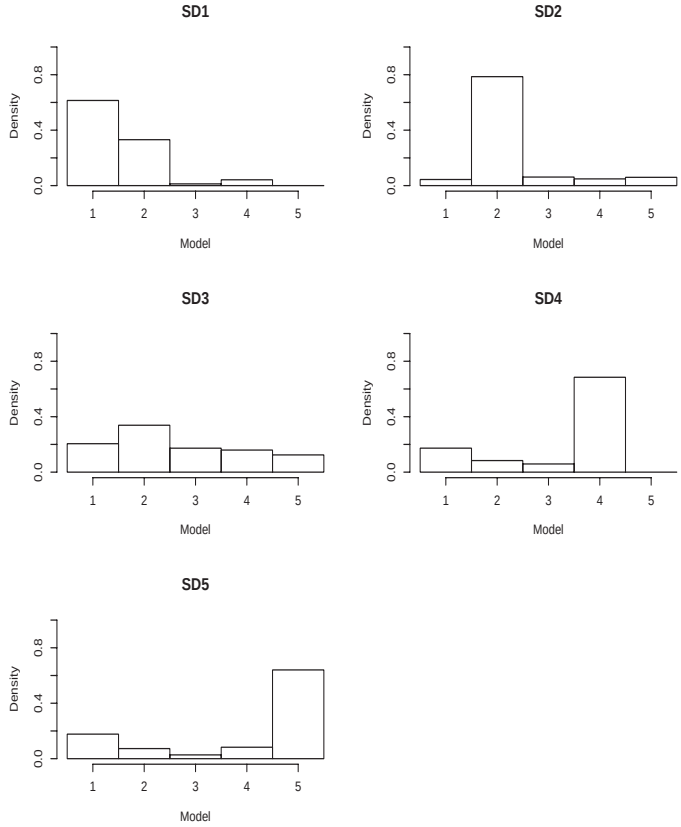


Fig. 6. Histograms for the time spent in each model for SD1, SD2, SD3, SD4 and SD5, respectively.

encourage the algorithm to choose model M_{β_H}, the algorithm will tend to choose the full model M.

We therefore considered a model which completely separates out the household infection from the spatial infection. In particular, we set

$$\alpha_{ij} = \beta_H 1_{\{\rho(i,j)=0\}} + \beta_G 1_{\{\rho(i,j)\neq 0\}} \exp(-\theta\rho(i,j)) + \sum_{k=1}^{2} \beta_C^k 1_{\{L_i=L_j=k\}} \quad (1 \leqslant i, j \leqslant n). \tag{3.3}$$

Then, with α_{ij} given by (3.3), we simulated a data set, SD6, with $\beta_H = 0$, $\beta_G = 0.075$, $\beta_C^1 = 0.075$, $\beta_C^2 = 0.05$ and $\theta = 5$. We ran the algorithm for both the Hagelloch data set and SD6 with α_{ij} of the form given in (3.3). The same priors as before were used. The results are similar to those obtained before in terms of model selection. This is shown to be the case in Figure 7.

By considering SD6, we explain why the above results are similar to those previously obtained. Consider three individuals i, j and k, say, and suppose that i and j belong to the same household and k belongs to a neighbouring household. Thus, i cannot infect j unless they belong to the same class at school. Suppose that in SD6, i infects k and k infects j such that i is still infectious when j is infected. At the height of all the simulated epidemics (SD1–SD6), 65–70% of the population are infectious, and the above situation therefore occurs frequently. That is, even in the absence of household infection, members of the same household will often be infectious at the same time. This is accentuated if members of the

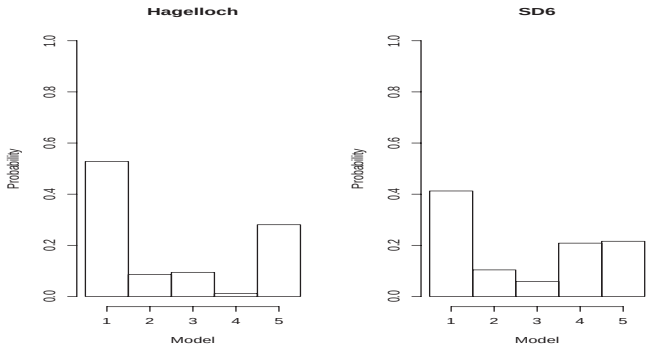


Fig. 7. Histograms for the time spent in each model for the Hagelloch data set and SD6, respectively, with α_{ij} given by (3.3).

household belong to the same class at school. The algorithm will therefore 'assign' the infection of j to i rather than to k. Thus, even in the complete absence of household infection, the model setup and the severity of the epidemic are such that the algorithm will often choose to incorporate household infection.

Unfortunately, there does not seem any way around this problem. All the spatial transmission mechanisms have difficulty differentiating between a true household effect and a high local infection rate. Despite the problems of identifiability associated with the household effect it is important to include it in the model since the Hagelloch data certainly suggest that the households play an important role in the spread of the disease. In the absence of a specific household effect (i.e. $\beta_H = 0$), the spatial effect needs to compensate for this to 'create' a household effect. Therefore if $\beta_H = 0, \theta > 0$ out of necessity and the question of the significance of the spatial effect in the transmission of the measles is redundant.

3.4 *Control strategies*

Clearly, with any disease it is useful to know what effect control strategies have in limiting the size and spread of an epidemic. The most commonly applied control strategies are the immunization of susceptibles and the quarantining of infectives. However, for Hagelloch it would be interesting to study the effect of various control strategies: in particular, closure of the school for the duration of the epidemic. We simulated 5000 epidemics using model parameters from the MCMC output for the full model M. Then, to assess the effect of closing the school, we simulated 5000 epidemics again using the same model parameters except we set $\beta_C^1 = \beta_C^2 = 0$. The results are presented in Figure 8 for epidemics which infect in excess of 165 people. (In both cases over 90% of the epidemics infect at least 165 individuals.) In the uncontrolled case 139 of epidemics failed to infect more than 10 people compared with 225 in the controlled case. The (estimated) mean size of an epidemic is reduced from 177.9 in the uncontrolled case to 169.2 in the controlled case.

The results are not as dramatic as one might hope. While there is definitely a classroom effect in the spread of the epidemic through Hagelloch, the spatial and household infection rates are sufficient to drive the epidemic process in their own right, so that, although the size of the outbreak is not dramatically reduced by closing the school, the nature of the spread of the epidemic is very different being purely spatially driven. The closure of the school accompanied by halving the global infection rate (i.e. setting $\beta_G = 0.5\beta_G$) has a far more dramatic effect on the epidemic. In that case, the probability that the epidemic fails to infect at least 10 people is 0.243 and the mean size of an epidemic is 104.6.

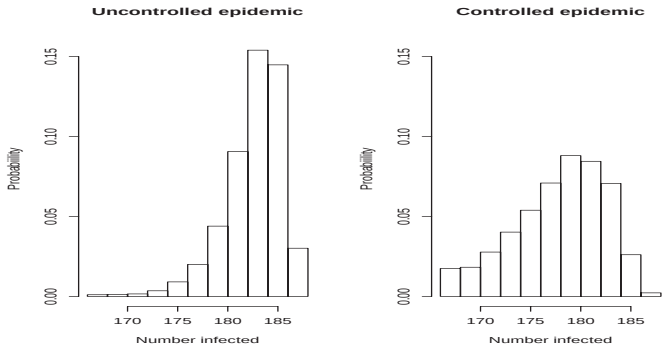


Fig. 8. Histograms of all outbreaks infecting at least 165 individuals from 5000 simulations for both uncontrolled and controlled epidemics, respectively.

4. DISCUSSION

We have considered the question of model selection for an extremely rich epidemic data set. In particular, we have studied which of spatial location, households and classrooms are important in the transmission of measles throughout the population. Our conclusion is that all three effects are important, to varying degrees, in the transmission of the disease. In particular, there is definitely a classroom 1 effect.

The results of the study into model selection are shown to be robust to reasonable changes in the model, such as inclusion of latent periods or unknown (imputed) infectious periods. This is due to the data being very informative about the spread of measles through the population.

The reversible jump step is easy to implement. We note that even a straightforward dimension-swapping step produces excellent mixing. More generally, characteristics of the epidemic model can be used to ensure that the MCMC algorithm moves readily between models.

The epidemic is far more severe than we usually observe, in that the entire susceptible population is assumed to be eventually infected. However, our approach readily extends to the situation where some of the individuals remain susceptible throughout the course of the epidemic. We would then require data on the household location and the school classroom of those individuals who remain susceptible. The likelihood in (2.2) incorporates susceptible individuals, by taking $I_k = \infty$ for $k = m + 1, m + 2, \ldots, n$.

It is therefore clear that the methods introduced and developed in this paper can be applied to a wide range of epidemic models. The complexity of the model that can be analysed will depend on how informative the data are and how easily missing data can be imputed.

As is the case with many variable dimension parameter problems, choosing priors which do not unduly prejudice in favour of one or other model, requires care. The study of simulated data sets is an important tool for understanding the statistical information contained within the data.

ACKNOWLEDGEMENTS

Both authors would like to thank Niels Becker for both bringing the Hagelloch data to the authors' attention and for providing the data set. The work by Heike Oesterle to assemble the data set from the original thesis of Pfeilsticker and other historical records is gratefully acknowledged. The authors would also like to thank Philip O'Neill, the referee and associate editor for their helpful comments on earlier versions of the paper which have greatly improved the presentation of the paper. This research was supported by the UK Engineering and Physical Sciences Research Council, under research grant number GR/M62723.

REFERENCES

BECKER, N. G. AND DIETZ, K. (1995). The effect of household distribution on transmission and control of highly infectious diseases. *Math. Bioscience* **127**, 207–219.

GELMAN, A., ROBERTS, G. O. AND GILKS, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics* **5**, 599–608.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

LAWSON, A. B. AND LEIMICH, P. (2000). Approaches to space-time modelling of infectious disease behaviour. *IMA Journal of Mathematics Applied in Medicine and Biology* **17**, 1–13.

NEAL, P. J. AND ROBERTS, G. O. (2003). Optimal Scaling for MCMC algorithms. Submitted to *Ann. Appl. Prob*.

OESTERLE, H. (1992). Statistiche Reanalyse einer Masernepidemie 1861 in Hagelloch, M.D. Thesis, Eberhard-Karls Universität, Tübingen.

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

PFEILSTICKER, A. (1863). Beiträge zur Pathologie der Masern mit besonderer Berücksichtigung der statistischen Verhältnisse, M.D. Thesis, Eberhard-Karls Universität, Tübingen.

[*Received February 19, 2003; first revision February 27, 2003; second revision July 10, 2003; third revision October 1, 2003; accepted for publication October 15, 2003*]