JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2008) **5**, 885–897
doi:10.1098/rsif.2007.1292
Published online 3 January 2008

Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London

Simon Cauchemez* and Neil M. Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London W2 1PG, UK

We present a new statistical approach to analyse epidemic time-series data. A major difficulty for inference is that (i) the latent transmission process is partially observed and (ii) observed quantities are further aggregated temporally. We develop a data augmentation strategy to tackle these problems and introduce a diffusion process that mimicks the susceptible–infectious–removed (SIR) epidemic process, but that is more tractable analytically. While methods based on discrete-time models require epidemic and data collection processes to have similar time scales, our approach, based on a continuous-time model, is free of such constraint. Using simulated data, we found that all parameters of the SIR model, including the generation time, were estimated accurately if the observation interval was less than 2.5 times the generation time of the disease. Previous discrete-time TSIR models have been unable to estimate generation times, given that they assume the generation time is equal to the observation interval. However, we were unable to estimate the generation time of measles accurately from historical data. This indicates that simple models assuming homogenous mixing (even with age structure) of the type which are standard in mathematical epidemiology miss key features of epidemics in large populations.

Keywords: susceptible–infectious–removed model; Cox–Ingersoll–Ross model; Markov chain Monte Carlo methods; epidemics; Bayesian inference; measles

1. INTRODUCTION

Over the last century, mathematical epidemiology has played a critical role in our understanding of the spread of communicable diseases in human populations. Continuous-time mechanistic models constitute the backbone of the discipline. One of the most studied models is the susceptible–infectious–removed (SIR) model, in which individuals are successively susceptible to infection, infectious and removed (may no longer transmit the disease). Denoting S_t and I_t, the numbers of susceptibles and infectives in the population at time t, respectively, new infections occur at rate $\beta S_t I_t$; recoveries at rate γI_t (Soper 1929; Bailey 1975; Anderson & May 1991).

Likelihood-based estimation of the parameters of such a model would be relatively straightforward if the times of infection and removal were observed for each case (Becker & Britton 1999); but this detail of data is rarely obtained in practice. In general, the underlying transmission process is partially observed (e.g. times of infection are observed, but not the times of removal; some cases are not reported), and observed quantities may be further aggregated (e.g. times of detection are aggregated weekly, monthly…). In this context, calculation of the likelihood quickly becomes intractable since it requires to integrate over all unobserved quantities.

Other concepts have therefore been used in an attempt to develop easier methods of estimation. For example, Becker (1989) and Becker & Hasofer (1997) rested on martingale methods to estimate transmission parameters when observations consist of the initial state of the epidemic, plus (i) the final state of the epidemic or (ii) the whole removal process. The approach provided simple but nevertheless efficient estimators of key quantities and approximate confidence regions for the parameters. However, it would be difficult to extend it to more complex situations, such as the one we are interested in, where (i) times of detection are temporally aggregated, (ii) the initial state of the system is unknown, and (iii) we must account for under-reporting, seasonality (and possibly long-term variations) in transmission rates. It seems that only likelihood-based methods can provide an integrated framework to deal simultaneously with these issues.

*Author for correspondence (s.cauchemez@imperial.ac.uk).

Received 5 November 2007
Accepted 29 November 2007

Over the last decade, data augmentation methods have been extensively used to tackle the missing data problem that makes likelihood-based estimation so tedious. The idea is to augment the observed data with the pieces of information required to write easily the likelihood; here the times of infection/removal. In a Bayesian setting, the joint posterior distribution of parameters and augmented data is then explored by Markov chain Monte Carlo (MCMC) sampling (Gilks *et al.* 1996). Using reversible jump MCMC sampling (Green 1995), the methodology has been extended to the situation where the exact amount of missing data is unknown, for example owing to under-reporting (Gibson & Renshaw 1998; Auranen *et al.* 2000; Cauchemez *et al.* 2006). Although the method is flexible and allows investigation of complex models, it is essentially limited by the size of the augmented data, which increases with the number of cases. Consequently, the approach has been used only for the data collected in small communities such as households (Auranen *et al.* 2000; O'Neill *et al.* 2000; Cauchemez *et al.* 2004) or schools (Cauchemez *et al.* 2006), when the number of cases does not exceed a few thousands. Computation times would become prohibitive when dealing with larger datasets, such as those collected by surveillance systems, for which the number of cases can easily reach tens of thousands.

For large epidemics in large populations, there is therefore no option but to find approximations of the SIR model, which are analytically tractable. Consider, for example, epidemic time-series data. These data typically provide counts of cases reported daily, weekly or monthly on a local or national ground. For inference, a natural choice is to approximate continuous-time models by discrete-time models (Finkenstadt & Grenfell 2000; Morton & Finkenstadt 2005). In these latter models, each time period is made of one generation of cases; generation of period k is simply the offspring of the generation of period $k-1$. However, an important constraint is that one observation period must effectively capture one generation of cases. This may be achieved only if the generation time of the disease (delay between infection of a case and infection of their typical secondary case) is equal to the length T of observation periods, or is a multiple of T. In the latter case, the data must be further aggregated, which may lead to an additional loss of information. There are therefore a variety of situations where discrete-time models cannot be used, since few generations may occur during a single observation period.

In this paper, we design a statistical framework to estimate the continuous-time SIR model from time-series data, when (i) the number of cases is too large to augment the data with the times of infection/removal of each case and (ii) epidemic and data collection processes have different time scales, so that the use of discrete-time models is excluded. To tackle the problem of temporal aggregation (and missing data), the data are augmented with the latent state $\{I_{kT}, S_{kT}\}$ at the beginning of each observation period k (=time interval $]kT,(k+1)T]$). The main difficulty is then to define the relationship between $\{I_{kT}, S_{kT}\}$, $\{I_{(k+1)T}, S_{(k+1)T}\}$ and the observation (number of infections reported for period k). This is achieved by introducing a diffusion process that mimics the SIR process, for which exact solution is readily available. The diffusion process is the Cox–Ingersoll–Ross process, which is commonly used in finance to model interest rates (Cox *et al.* 1985). The method is applied to measles time series in London in the pre-vaccination era (1948–1964).

2. MATERIAL AND METHODS

2.1. The SIR epidemic model

2.1.1. The SIR model. The SIR epidemic model is a continuous-time Markovian model that describes the spread of a communicable disease in a population. Denoting $\{S_t, I_t, R_t\}$, the numbers of susceptibles, infectives and removed cases at time t, respectively, and H_t the σ-algebra generated by the history $\{S_u, I_u, R_u; 0 \le u \le t\}$, the model is defined by the following equations:

$$\begin{cases} P(\mathrm{d}S_t = 1 | H_t) = \nu(t)\mathrm{d}t + o(\mathrm{d}t), \\ P(\mathrm{d}S_t = -1, \mathrm{d}I_t = 1 | H_t) = \beta(t) S_t I_t \,\mathrm{d}t + o(\mathrm{d}t), \\ P(\mathrm{d}I_t = -1, \mathrm{d}R_t = 1 | H_t) = \gamma I_t \,\mathrm{d}t + o(\mathrm{d}t), \end{cases} \quad (2.1)$$

where ν is the birth rate; β is the transmission rate; and $1/\gamma$ is the mean infectious period. In this formulation, we neglect the mortality due to disease. We also neglect the number of individuals who leave the susceptible population owing to death or migration.

In practice, this continuous-time process is only partially observed, and observed quantities are further aggregated. Surveillance data typically consist of numbers $\{U_k\}_{k=0,\ldots,K}$ of new infections occurring during periods of length T, i.e. U_k is the number of times in interval $]kT,(k+1)T]$ when I_t increases by $+1$.

When epidemic and data collection processes have the same time scale ($1/\gamma \approx T$), a discrete approximation of the model may be considered, where the expected number of cases $E(U_k)$ for period k is proportional to the number of cases U_{k-1} for period $k-1$ (Finkenstadt & Grenfell 2000; Finkenstadt *et al.* 2002). However, such relationship no longer holds when $1/\gamma \neq T$ since few generations of cases may then occur during a single observation period. There is then no option but to come back to the continuous-time model.

Figure 1 shows two possible trajectories for the number of infectives I_t consistent with four new infections occurring during period k ($U_k=4$). The larger rate of recovery in figure 1*a* implies that, although the same number of new infections is observed in figure 1*a*,*b* trajectories of I_t are very different. Owing to stochastic fluctuations, important differences could be observed between trajectories, even if the recovery rate was the same.

Let us first assume that we observe $\{I_{kT}, S_{kT}\}_{k=0,\ldots,K}$, the numbers of infectives and susceptibles at the beginning of each observation period (in practice, this is not the case; these terms will be considered as nuisance parameters of the final inference framework). The main issue for inference is to determine the probability $P(I_{(k+1)T}, U_k | I_{kT}, S_{kT})$, the joint probability of the number of infectives at the beginning of period $k+1$

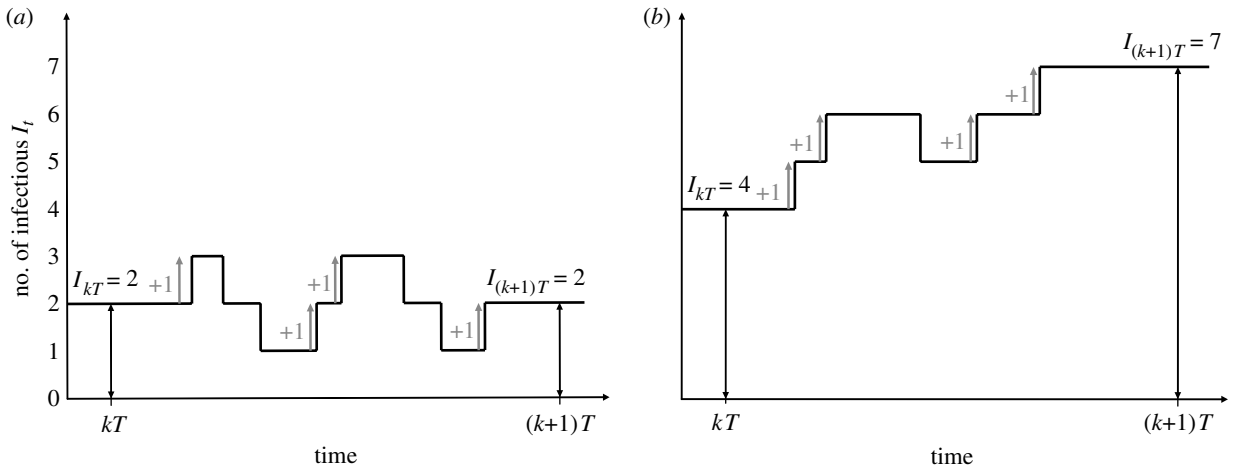

Figure 1. Examples of trajectories for the number of infectives I_t consistent with $U_k=4$, where U_k is the number of new infections occurring during time period $]kT,(k+1)T]$: (*a*) $I_{kT}=2$, $I_{(k+1)T}=2$ and (*b*) $I_{kT}=4$, $I_{(k+1)T}=7$.

and the number of new infections for period k given I_{kT}, S_{kT}.

$$P(I_{(k+1)T}, U_k|I_{kT}, S_{kT}) = P(I_{(k+1)T}|I_{kT}, S_{kT})P(U_k|I_{(k+1)T}, I_{kT}, S_{kT}). \quad (2.2)$$

In the TSIR framework, which relies on the relatively strong assumption that the number of infectives I_t during time period k is constant and equal to the number of new infections (so $U_k=I_{(k+1)T}$), only the first term of equation (2.2) is needed (the second term is equal to 1 for $U_k=I_{(k+1)T}$, 0 otherwise). This simplification does not hold here. Consider, for example, the situation in figure 1*a*; there is an infinity of values for U_k consistent with $I_{(k+1)T}=2$ and $I_{kT}=2$ (any $U_k\geq 0$ is consistent $I_{(k+1)T}=2$ and $I_{kT}=2$).

Without loss of generality, let us assume $k=0$.

2.1.2. Approximation: Cox–Ingersoll–Ross diffusion process. Here, we need to make two additional assumptions. We assume that length T of observation periods is small enough to neglect within-period changes in the number of susceptibles

$$H1 : \forall t \in [0, T[\quad S_t \approx \overline{S}_0,$$

and that the transmission parameter is constant during an observation period

$$H2 : \forall t \in [0, T[\quad \beta(t) \approx \beta_0.$$

The value of $\overline{S}_0$ is discussed in appendix C.

Under assumptions $H1$ and $H2$, the number of infectives I_t is a birth and death process over time period $[0,T]$, with birth rate $\beta_0\overline{S}_0 I_t$ and death rate γI_t at time t. When there is one infective at the beginning of the period ($I_0=1$), the number I_T of infectives at the end of the period follows a geometric distribution (Bailey 1964; Renshaw 1991). If $I_0>1$, the distribution of I_T is still available, but the resulting expression is 'really too messy to be of much practical use' as I_0 increases (Bailey 1964; Renshaw 1991). Considering the development of a population of initial size I_0 as being equivalent to the development of I_0 separate populations each of initial size 1, it can be shown that the distribution of I_T is negative binomial; however, it is unclear how large I_0 may become before the 'populations' can no longer be assumed to develop independently of each other (Renshaw 1991). Here, we introduce an alternative approximation where the assumption of independence is not required.

In the models described above, the number of infectives is a discrete variable. Here, we intend to model the *effective number of infectives* as a continuous variable I'_t. A natural candidate is the diffusion process that mimics the epidemic SIR process under $H1$ and $H2$, i.e. with trend and volatility

$$\begin{cases} E\{I'_{t+\mathrm{d}t}|I'_t\} = (\beta_0\overline{S}_0-\gamma)I'_t\,\mathrm{d}t, \\ \mathrm{var}\{I'_{t+\mathrm{d}t}|I'_t\} = (\beta_0\overline{S}_0+\gamma)I'_t\,\mathrm{d}t, \end{cases}$$

which is the solution of the stochastic differential equation

$$\mathrm{d}I'_t = r_0 I'_t\,\mathrm{d}t + \sigma_0\sqrt{I'_t}\,\mathrm{d}W, \quad (2.3)$$

where $r_0=\beta_0\overline{S}_0-\gamma$, $\sigma_0^2=\beta_0\overline{S}_0+\gamma$ and W is the Brownian motion. Equation (2.3) characterizes the Cox–Ingersoll–Ross process, which is used to model interest rates on financial markets (Cox *et al.* 1985). Interestingly, the exact solution of equation (2.3) is readily available (Cox *et al.* 1985). It has a non-central χ^2 distribution with zero d.f. (Siegel 1979; see appendix A).

Instead of equation (2.2), inference will therefore rely on

$$P(I'_T, U_0|I'_0, S_0) = P(I'_T|I'_0, S_0)P(U_0|I'_T, I'_0, S_0). \quad (2.4)$$

For $I'_0>0$, the first term of equation (2.4) is (appendix A)

$$P(I'_T|I'_0, S_0) = \begin{cases} \exp(-u_0) & \text{if } I'_T=0, \\ 2c_0 f_{2u_0}(2c_0 I'_T) & \text{otherwise,} \end{cases} \quad (2.5)$$

where $c_0=2r_0(\mathrm{e}^{r_0T}-1)/\sigma_0$; $u_0=c_0(\mathrm{e}^{r_0T}-1)I'_0$; and f is defined in appendix A. The mean and variance of $I'_T|I'_0,S_0$ are $\mathrm{e}^{r_0T}I'_0$ and $(\sigma_0)^2\mathrm{e}^{r_0T}(\mathrm{e}^{r_0T}-1)I'_0/r_0$, respectively.

The second term $P(U_0|I'_T, I'_0, S_0)$ of equation (2.4) can only be approximated. We use the fact that, given the expected number of new infections

$\Lambda_0 = \beta_0 \overline{S_0} \int_0^T I'_t \, \mathrm{d}t$, the number U_0 of infections occurring in $[0,T]$ is Poisson distributed with mean Λ_0. Assuming that $\Lambda_0 | I'_T, I'_0, S_0$ may be approximated by a gamma distribution with mean M_0 and variance V_0 to be determined, a first approximation of the distribution $P(U_0 | I'_T, I'_0, S_0)$ is

$$P(U_0 | I'_T, I'_0, S_0) = \int_0^\infty P(U_0 | \Lambda_0) P(\Lambda_0 | I'_T, I'_0, S_0) \mathrm{d}\Lambda_0, \tag{2.6}$$

which characterizes the negative binomial distribution with mean M_0 and variance $M_0 + V_0$ (Poisson distribution with gamma distributed parameter). In this first approximation, it is assumed that, given Λ_0, the number of infections in $[0,T]$ is independent of (I'_T, I'_0). In practice, however, there is the additional constraint that if $I'_T - I'_0 > 0$, the number U_0 of persons infected in $[0,T]$ must be $\geq I'_T - I'_0$. We therefore condition the negative binomial distribution by $U_0 \geq \max(0, I'_T - I'_0)$.

Eventually, to determine the mean M_0 and variance V_0 of $\Lambda_0 | I'_T, I'_0, S_0$, we rely on the linear model

$$\Lambda_0 = x_0 + y_0 I'_T + \epsilon_0, \tag{2.7}$$

where ϵ_0 is the error. Denoting $\{\tilde{x}_0, \tilde{y}_0\}$, the scalars that minimize $E[(\Lambda_0 - x_0 - y_0 I'_T)^2]$ (minimization is performed analytically and there is no need to use a minimization routine; see appendix B), and $\tilde{v}_0$ the variance of $\Lambda_0 - \tilde{x}_0 - \tilde{y}_0 I'_T$, we approximate the mean M_0 by $\tilde{x}_0 + \tilde{y}_0 I'_T$, and the variance V_0 by $\tilde{v}_0$. We derive $\tilde{x}_0, \tilde{y}_0, \tilde{v}_0$ from the Laplace transform of $(I'_T, \int_0^T I'_t \, \mathrm{d}t) | I'_0, S_0$ (appendix B).

2.1.3. Approximation when the hypothesis of mass action is violated. When the hypothesis of mass action is violated, different authors have suggested to use the force of infection $\beta_0 \overline{S_0} (I_t)^{1-\epsilon}$ with ϵ close to 0 in general (Finkenstadt & Grenfell 2000; Morton & Finkenstadt 2005). Unfortunately, the results of §2.1.2 apply only for $\epsilon = 0$ since the exact solution of equation (2.3) is not available otherwise. However, the force of infection can be linearized for ϵ close to 0. Denoting $\bar{I}'_0$, the expectation of the average number of infectives over period $[0,T]$ given I'_0, S_0

$$\bar{I}'_0 = E\left(\frac{1}{T} \int_0^T I'_t \, \mathrm{d}t | I'_0, S_0 \right),$$

the force of infection may be approximated by the following term, linear in I'_t:

$$\beta_0 \overline{S_0} (I'_t)^{1-\epsilon} \approx \beta_0 \overline{S_0} \bar{I}'^{-\epsilon}_0 I'_t. \tag{2.8}$$

We can use equation (2.8) in the model described in §2.1.2 for inference. An approximation of $\bar{I}'_0$ is given in appendix C.

2.2. Statistical framework

2.2.1. Observed and augmented data. We consider the situation where data consist of the time series $\{U^*_k\}_{k=0,\ldots,K}$ of numbers of reported cases, plus birth rates $\{B_k\}_{k=0,\ldots,K}$. Unobserved variables required for model specification are

(i) the total number of cases U_k (i.e. reported + unreported cases) during observation period $k = 0, \ldots, K$,
(ii) the number of infectives I'_{kT} at the beginning of observation period $k = 0, \ldots, K$, and
(iii) the number of susceptibles S_0 at the beginning of the first period.

Given S_0, $\{U_k, B_k\}_{k=0,\ldots,K}$, the number of susceptibles at the beginning of period k, is given by the deterministic relationship

$$S_{kT} = S_0 + \sum_{i=0}^{k-1} (B_i - U_i). \tag{2.9}$$

In the statistical framework, observations $\{U^*_k, B_k\}_{k=0,\ldots,K}$ are augmented with $\{(U_k, I'_{kT})_{k=0,\ldots,K}, S_0\}$, which may be considered as nuisance parameters of the model.

2.2.2. Joint distribution. Denoting Θ, the parameters of the model, the joint distribution of observations, augmented data and parameters are

$$P\left(I'_{(K+1)T}, \left\{ I'_{kT}, U_k, U^*_k \right\}_{k=0,\ldots,K}, S_0, \Theta | \{B_k\}_{k=0,\ldots,K} \right)$$
$$= \prod_{k=0,\ldots,K} \{ P(U^*_k | U_k, \Theta) P(I'_{(k+1)T}, U_k | I'_{kT}, S_0,$$
$$\{U_i, B_i\}_{i=0,\ldots,k-1}, \Theta) \} P(S_0, I'_0 | \Theta) P(\Theta). \tag{2.10}$$

The first term on the r.h.s. of equation (2.10) characterizes the reporting process. Here, we assume that the number of reported cases follows a binomial distribution

$$U^*_k | U_k \sim Bin(U_k, \rho), \tag{2.11}$$

where ρ is the proportion of cases which are reported.

The second term has been discussed in §2.2.1

$$P\left(I'_{(k+1)T}, U_k | I'_{kT}, S_0, \{U_i, B_i\}_{i=0,\ldots,k-1}, \Theta \right)$$
$$= P\left(I'_{(k+1)T}, U_k | I'_{kT}, S_{kT} = S_0 + \sum_{i=0}^{k-1} (B_i - U_i), \Theta \right).$$

The third term models the state of the system at the beginning of the follow-up. Denoting M, the size of the population, we assume that I'_0 and S_0 are uniformly distributed in $[0,M]$ and $\{0,1,\ldots,M\}$, respectively. For London in the period 1948–1964, we specified $M = 10\,000\,000$.

The last term gives our priors for model parameters. For parameters defined on $]0,\infty[$, we specify an exponentially distributed prior with mean 10^{10}. This distribution is flat on the range of values possible for the parameters. The reporting rate ρ has a uniform prior distribution $U[0,1]$.

2.2.3. MCMC sampling. The joint posterior distribution of augmented data and parameters was explored by Metropolis–Hastings MCMC sampling (Gilks *et al.* 1996). The following steps were sequentially applied.

(i) Update the parameters Θ: parameters defined on $]0, +\infty[$ were updated by a random walk on the log scale. Power parameter ϵ was updated by a random walk on the real line.
(ii) Update the numbers of infections U_k, for $k=0,\ldots,K$: an independence sampler was used to update U_k (Gilks *et al.* 1996). New candidate $\tilde{U}_k$ was drawn as follows: $\tilde{U}_k = U_k^* + X_k$, where X_k is drawn from the negative binomial distribution $(U_k^* + a, ((\rho + b)/(1 + a)))$ with $a = b = 10^{-5}$. This choice is motivated by the fact that, if U_k is Poisson distributed with gamma (a, b)-distributed parameter, the distribution of $U_k | U_k^*$ is the proposal used here.
(iii) Update the numbers of susceptibles S_0 at the beginning of the follow-up: a random walk was used.
(iv) Update the numbers of infectives I'_{kT} at the beginning of period $k=0,\ldots,K$: a random walk on the log scale was used.

To reduce correlation between the transmission rate β and the initial number of susceptibles S_0, it was useful to reparametrize the transmission rate $\beta = \beta^*/\overline{S}$, with $\overline{S} = \sum_{k=0}^{K-1} S_{kT}/K$.

The standard deviations of the proposals were tuned to obtain an acceptance rate of 20–40%. We performed 4 000 000 iterations for each run of the MCMC algorithm. The first 400 000 were discarded as the burn-in period. The output was then recorded on every 200 iterations to constitute a sample from the posterior distribution. One MCMC run took roughly 20 hours on a desktop. Convergence of the MCMC was visually assessed.

2.3. Applications

2.3.1. Simulation study. We first considered the situation where transmission rates vary every two weeks with a period of 1 year, and where data are collected every two weeks too ($T=14$ days). Epidemics were simulated from the continuous-time SIR model with mean infectious period $1/\gamma$ equal to 7, 14 and 21 days. We also simulated epidemics where the hypothesis of mass action was violated ($\epsilon=3$, 5, 7%). Eventually, we simulated epidemics from the susceptible–exposed (infected but not infectious)–infectious–removed (SEIR) model, with constant or exponentially distributed latent period (time period during which the subject is infected but not infectious, mean $L=2$, 3.5, 7, 10 days) and exponentially distributed infectious period, with mean $I=7$ days.

To assess how the method could cope with temporal aggregation in the data, we also simulated an epidemic for 20 years from the continuous-time SIR model with mean infectious period $1/\gamma=14$ days. In the simulation, two seasons with high and low transmissibility, respectively, were defined for each calendar year. We then estimated parameters of the SIR model for different levels of temporal aggregation. It is not always possible to split seasons in equally sized observation periods. For example, for a target duration of observation period of eight weeks, each season (26 weeks) is split in three observation periods with size 8, 8 and 10, respectively (average duration: 8.7 days). We investigated scenarios where the average duration of observation periods was 1, 2, 2.9, 4.3, 5.2, 6.5, 8.7 and 13 weeks. Note that the length T of observation periods is not constant in a dataset; the method can account for these variations in T.

For all scenarios, birth rate in the simulations was $B=2152$ births per month, which is the average birth rate in London between 1944 and 1964. Simulation values for the parameters were defined, based on their posterior mean given historical data on measles transmission in London, and under appropriate constraints (e.g. mean infectious period equal to 7 days for the scenario $1/\gamma=7$ days). For some scenarios (e.g. SIR model with $1/\gamma=14$ days), simulation values were simply equal to their posterior mean. For other scenarios (e.g. $\epsilon=-7\%$), parameters were adjusted so that the average number of cases per year was roughly consistent with the one observed in London between 1948 and 1964.

2.3.2. Comparison with existing methods: measles epidemics in London. We also analysed the time series of the number of measles cases, collected bi-weekly in London between 1948 and 1964 (http://www.zoo.ufl.edu/bolker/measdata.html; Finkenstadt & Grenfell 2000; Morton & Finkenstadt 2005).

Previous studies have shown that, for the pre-vaccination era in the UK, models in which parameters have only seasonal variations fail to exhibit the same cyclical pattern as observed epidemics (Finkenstadt & Grenfell 2000; Morton & Finkenstadt 2005). This result, which was also observed with the method presented here, is probably due to changes in the structure of the population (through changes in birth rates) that modify transmission parameters themselves. The problem has been previously tackled by the use of local regressions, leading to the estimation of a sequence of reporting rates $\{\rho_k\}_{k=0,\ldots,K}$ (Finkenstadt & Grenfell 2000). Here, we use an alternative approach, where the person-to-person transmission rate is inversely proportional to the size of the core group—group of individuals who contribute the most to the chain of transmission—(De Jong *et al.* 1995). For measles, we assumed that the core group consists of children with age below 4 years (i.e. below 104 bi-weeklies), so that the size of the core group for observation period k is

$$N_k = \sum_{j=0}^{103} B_{k-j},$$

and the contribution to the force of infection of an infective is $\beta_k \overline{S}_k / N_k$ during this period, where β_k varies bi-weekly, with a period of 1 year.

For $\epsilon=0$, the effective reproduction number R_t (average number of persons infected by a typical case at time t) is simply $(\beta_k/\gamma).(\overline{S}_k/N_k)$ for time t in period k. When the assumption of mass action is violated ($\epsilon \neq 0$), the effective reproduction number also depends on the number of infectives in the population $R_t = (\beta_k/\gamma).(\overline{S}_k/N_k).(I_t)^{\epsilon}$. However, this quantity can easily be computed from the output of our algorithm for the sequence of times $\{kT\}_{k=0,\ldots,T}$.

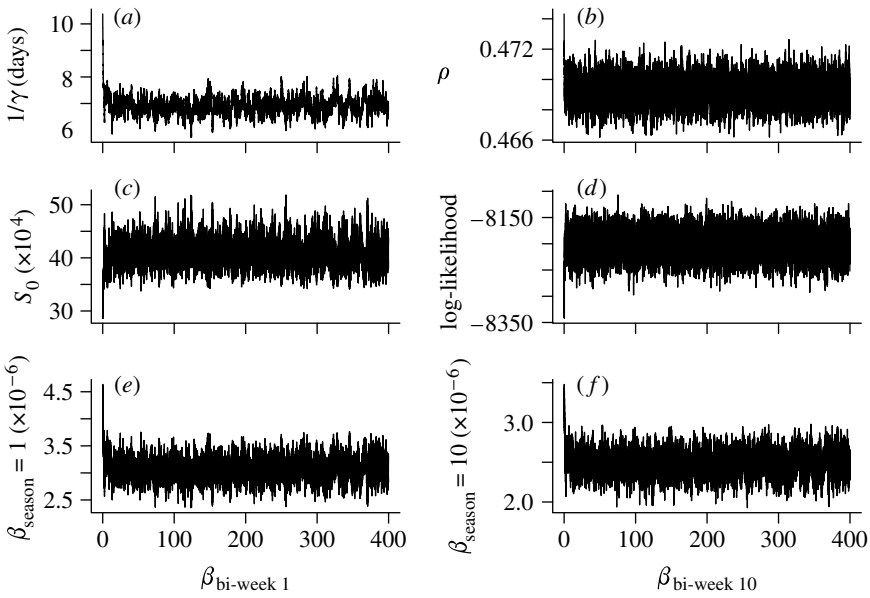

Figure 2. Convergence of the MCMC algorithm for the SIR epidemic simulated with $1/\gamma=7$ days. (*a*) Recovery rate, $1/\gamma$; (*b*) reporting rate, ρ; (*c*) initial number of susceptibles, S_0; (*d*) log-likelihood; (*e*) transmission rate for bi-week 1; (*f*) transmission rate for bi-week 10.

Table 1. Epidemics simulated from the SIR model. (Simulation values, posterior mean and 95% credible interval of the mean infectious period $1/\gamma$, the initial number of susceptibles S_0 and the reporting rate ρ. The mean (s.d.) of the relative error of β/γ is also given. The duration of observation periods is 14 days.)

	$1/\gamma$ (days)		S_0 ($\times 10^3$)		ρ (%)		error β/γ (%)
data	simul	estimate	simul	estimate	simul	estimate	mean (s.d.)
simul 7	7	6.90 [6.30,7.55]	415	408 [365,457]	47.00	46.93 [46.76,47.10]	2.97 (1.09)
simul 14	14	13.89 [12.72,15.09]	160	159 [146,174]	47.00	46.96 [46.84,47.08]	1.08 (1.33)
simul 21	21	20.66 [19.46,21.84]	94	96 [91,101]	47.00	46.93 [46.83,47.04]	−0.87 (1.27)

Two runs of the MCMC algorithm were performed: (i) all the parameters of the model are estimated, including the mean infectious period $1/\gamma$ and (ii) the mean infectious period of measles is known and equal to $1/\gamma=14$ days.

For this dataset, the TSIR method is expected to be applicable since it is reasonable to assume that the generation time for measles is approximately two weeks. For comparison purpose, we also estimated our model with the TSIR method (Finkenstadt & Grenfell 2000; Morton & Finkenstadt 2005). Under TSIR assumptions (see above), equation (2.2) simplifies to $P(U_k|U_{k-1}, S_{kT})$, where the number U_k of new infections occurring during period k has a negative binomial distribution with mean $E_k = r_k S_{kT}(U_{k-1})^{1-\epsilon}$, variance E_k/q and density

$$\frac{\Gamma(U_k+q)}{\Gamma(q)U_k!}(E_k/q)^{U_k}(1+E_k/q)^{-U_k-q}.$$

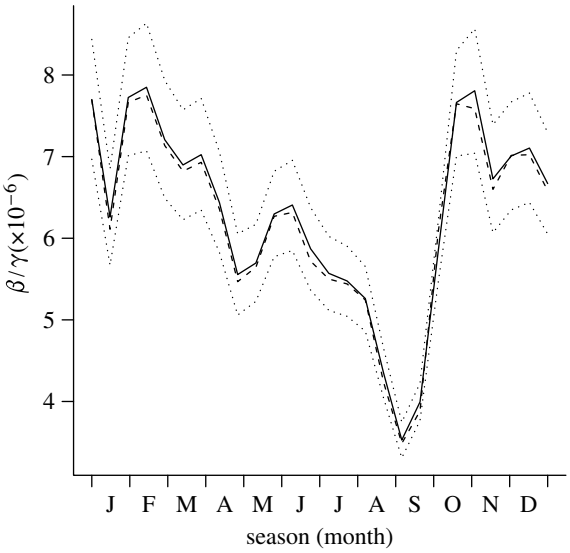

Figure 3. Posterior mean (solid line), 95% credible interval (dotted lines) and simulation value (dashed line) of the ratio β/γ for the SIR epidemic simulated with $1/\gamma=14$ days. β is the transmission rate (seasonal variations with period=1 year) and $1/\gamma$ is the mean infectious period.

3. RESULTS

3.1. Simulation study

3.1.1. SIR model. Figure 2 shows convergence of the MCMC algorithm for the SIR epidemic simulation with $1/\gamma=7$ days. Convergence is quickly obtained.

Table 1 gives simulation values, posterior mean and 95% credible interval of the mean infectious period $1/\gamma$, the initial number of susceptibles S_0 and the reporting rate ρ. The mean (s.d.) of the relative error for β/γ is also given. For the three simulated datasets: posterior means are close to simulation values; simulation values are always within the 95% credible interval; and the relative error of β/γ is small (less than 3%). Figure 3

Table 2. Epidemics simulated when the hypothesis of mass action is violated ($\epsilon \neq 0$). (Simulation values, posterior mean and 95% credible interval of the power ϵ, the mean infectious period $1/\gamma$, the initial number of susceptibles S_0 and the reporting rate ρ. The mean (s.d.) of the relative error of β/γ is also given. The force of infection is $\beta S_t I_t'^{1-\epsilon}$. The duration of observation periods is 14 days.)

ϵ (%)		$1/\gamma$ (days)		S_0 ($\times 10^3$)		ρ (%)		error β/γ (%)
simul	estimate	simul	estimate	simul	estimate	simul	estimate	mean (s.d.)
0.00	0.28 [−0.10,0.68]	14.00	13.50 [12.38,14.72]	160	163 [150,179]	47.00	46.96 [46.84,47.08]	0.37 (1.35)
3.00	3.27 [2.83,3.72]	14.00	13.79 [12.48,15.15]	180	182 [169,197]	47.00	46.92 [46.80,47.03]	1.82 (1.06)
5.00	6.70 [5.00,8.42]	14.00	14.89 [13.21,16.84]	180	197 [170,233]	47.00	46.94 [46.82,47.06]	3.57 (3.14)
7.00	9.42 [6.98,11.70]	14.00	14.72 [12.93,16.66]	180	154 [123,196]	47.00	47.04 [46.92,47.15]	30.12 (1.98)

Table 3. Epidemics simulated from the SEIR model. (Simulation values, posterior mean and 95% credible interval of the mean infectious period $1/\gamma$, the initial number of susceptibles S_0 and the reporting rate ρ. The mean (s.d.) of the relative error of β/γ is also given. In the simulations, the latent period L is constant or exponentially distributed; the infectious period is exponentially distributed with mean $I=7$ days. The duration of observation periods is 14 days.)

data		$1/\gamma$ (days)	S_0 ($\times 10^3$)		ρ (%)		error β/γ (%)
L	$L+I$	estimate	simul	estimate	simul	estimate	mean (s.d.)
cst							
2	9	9.23 [8.38,10.10]	415	416 [377,460]	47.00	46.89 [46.70,47.07]	1.00 (3.01)
3.5	10.5	10.80 [9.82,11.87]	415	402 [364,442]	47.00	47.01 [46.85,47.18]	3.69 (4.92)
7	14	13.82 [12.52,15.11]	415	414 [375,460]	47.00	46.90 [46.72,47.08]	0.02 (8.12)
10	17	17.11 [15.10,19.14]	415	415 [370,467]	47.00	47.09 [46.92,47,27]	−0.14 (9.33)
exp							
2	9	8.83 [8.01,9.74]	415	403 [361,446]	47.00	47.09 [46.90,47.27]	3.34 (3.01)
3.5	10.5	10.80 [9.72,11.97]	415	396 [351,448]	47.00	47.03 [46.83,47.21]	5.22 (4.32)
7	14	13.49 [12.15,14.84]	415	420 [377,470]	47.00	47.08 [46.88,47.27]	−1.23 (6.57)
10	17	15.30 [13.78,16.67]	415	436 [397,485]	47.00	46.98 [46.79,47.18]	−4.54 (7.57)

shows the seasonal variations of the ratio β/γ for the epidemic simulated with $1/\gamma=14$ days.

3.1.2. Estimation when the hypothesis of mass action is violated. Table 2 gives the simulation values, posterior mean and 95% credible interval of ϵ, $1/\gamma$, S_0 and ρ when the hypothesis of mass action is violated. Power ϵ is correctly estimated when the simulation value is less than 5%; it is overestimated otherwise. Estimates of other parameters remain satisfying for $\epsilon \leq 5\%$. For $\epsilon=7\%$, both the ratio β/γ and the power coefficient ϵ are overestimated by 30%.

3.1.3. SEIR model. Table 3 gives the simulation values, posterior mean and 95% credible interval of $1/\gamma$, S_0 and ρ when the epidemic is generated from an SEIR model. In this context, the estimate of $1/\gamma$ corresponds to the generation time, i.e. the sum $L+I$ of the latent and infectious period, rather than to the effective infectious period I. Estimates for other parameters and the relative error of β/γ remain satisfying.

3.1.4. Accuracy of estimates according to the level of temporal aggregation. Table 4 gives the posterior mean, 95% credible interval and relative error of parameters according to the level of coarseness in the data. We find that, so far as the average duration of observation periods is less than or equal to 5.2 weeks ($=2.5\times$generation time of the disease), relative errors remain small for all parameters. For larger degrees of temporal aggregation in the data, important biases are observed.

3.2. Measles epidemics in London in the pre-vaccination era

3.2.1. Posterior distribution. Table 5 gives the posterior mean and 95% credible interval of $1/\gamma$, S_0, ρ and ϵ. Figure 4 shows the seasonality, the trend of the transmission rate and the effective reproduction number when $1/\gamma=14$ days. Minimum transmission is obtained during holidays, at the end of August.

When all parameters of the model are estimated, including the mean infectious period, we find that the mean infectious period is very short (3–4 days), the number of susceptibles at the beginning of the follow-up is roughly 400 000, half of the cases are reported and the hypothesis of mass action is violated ($\epsilon>0$), although the estimate of ϵ is close to zero (95% credible interval: 0.46 and 1.30%). When the mean infectious period $1/\gamma$ is assumed to be known ($=14$ days), the number of

Table 4. Robustness of the estimates according to the average duration $\overline{T}$ of the observation period. (Simulation values, posterior mean (95% credible interval; relative error %) for the reporting rate ρ, the mean infectious period $1/\gamma$, the initial number of susceptibles S_0 and infectives I_0 and the transmission rates. Epidemics are simulated from the SIR model.)

	ρ (%)		$1/\gamma$ (days)		S_0 ($\times 10^3$)		I_0 ($\times 10^3$)		β_1($\times 10^{-7}$)		β_2($\times 10^{-7}$)	
simulation value	47		14		160 000		900		5.1		3.9	
estimates												
$\overline{T}=1$	46.91	[46.79,47.03] (−0.2)	13.83	[13.27,14.38] (−3.9)	166	[159,174] (−0.8)	893	[807,983] (−0.8)	4.98	[4.83,5.12] (−2.4)	3.81	[3.68,3.94] (−2.3)
$\overline{T}=2$	46.92	[46.80,47.03] (−0.2)	13.95	[13.38,14.50] (−2.7)	164	[157,172] (−1.0)	891	[805,982] (−1.0)	5.00	[4.85,5.14] (−2.0)	3.81	[3.69,3.94] (−2.2)
$\overline{T}=2.9$	46.92	[46.80,47.03] (−0.2)	13.89	[13.29,14.46] (−2.6)	164	[157,172] (−2.2)	880	[791,974] (−2.2)	5.03	[4.88,5.17] (−1.4)	3.84	[3.71,3.97] (−1.6)
$\overline{T}=4.3$	46.95	[46.83,47.07] (−0.1)	13.25	[12.63,13.86] (−14.7)	184	[175,192] (−6.5)	841	[751,938] (−6.5)	4.73	[4.59,4.88] (−7.2)	3.57	[3.44,3.70] (−8.5)
$\overline{T}=5.2$	46.96	[46.83,47.08] (−0.1)	13.05	[12.38,13.68] (−14.9)	184	[176,193] (−6.3)	843	[747,944] (−6.3)	4.80	[4.65,4.94] (−6.0)	3.63	[3.49,3.77] (−7.0)
$\overline{T}=6.5$	46.98	[46.77,47.18] (0.0)	6.16	[4.81,8.18] (−143.9)	390	[310,472] (−52.3)	429	[307,575] (−52.3)	4.53	[4.23,4.85] (−11.2)	3.93	[3.52,4.31] (−0.7)
$\overline{T}=8.7$	47.03	[46.61,47.46] (−0.1)	3.36	[2.91,3.84] (−411.5)	818	[713,946] (−73.8)	236	[142,345] (−73.8)	3.81	[3.49,4.14] (−25.3)	3.50	[3.20,3.83] (−10.2)
$\overline{T}=13$	46.84	[45.79,47.89] (−0.3)	2.33	[2.03,2.66] (−978.5)	1726	[1450,2067] (−82.7)	156	[55,293] (−82.7)	2.59	[2.26,2.94] (−49.3)	2.42	[2.10,2.76] (−38.0)

susceptibles at the beginning of the follow-up is halved and the estimate of ϵ doubles (posterior mean 2.09 instead of 0.88).

Our estimates are similar to those obtained with the TSIR approach (see table 5 and figure 4*a*). The main difference is obtained for the power coefficient ϵ which is twice larger for the TSIR method.

3.2.2. Model checking. We simulated epidemics from the model, with parameters equal to their posterior means. Simulations started at the beginning of year 1948 and birth rates for period 1944–1964 were an input of the simulations. For the posterior distribution with $1/\gamma$ estimated from the data (3.38 days), simulated epidemics faded out. For $1/\gamma=14$ days fixed, figure 5 compares the observed and expected (average of 40 realizations) time series. The model captures the biannual pattern of the epidemics, although predicted incidence for inter-epidemic years is slightly more important than the observed one (figure 5*a*,*c*). The model also captures the magnitude of biannual epidemics, except for two of the three very large epidemics for which the incidence is underestimated (years 1955 and 1961; figure 5*a*). Predicted trend in the number of susceptibles is relatively close to the observed one (figure 5*b*).

4. DISCUSSION

We have presented a method to estimate continuous-time epidemic models from time-series data. Compared with existing methods based on discrete-time models, the approach can be used when epidemic and data collection processes have different time scales or when data are collected at irregular intervals.

A diffusion process for which an exact solution is readily available was introduced to approximate the SIR epidemic process. In large populations, modelling the number of infectives with a diffusion process does not raise major concerns. Quite obviously, this choice would be much more questionable for data collected in small communities or households.

We proposed a simple approach to capture changes in the transmission rate due to modifications in the structure of the population. The basic idea is that an individual's number of contacts is fixed, so that each infective does not contact 10% more people if the population grows 10%. This leads to assuming that the person-to-person transmission rate is inversely proportional to the size of the core group (Anderson & May 1991; De Jong *et al.* 1995). The core group for measles is clearly the group of young children, although defining a clear cut-off appears to be relatively arbitrary. Here, we specified the cut-off at 4 years because the birth rate at a delay time of 4 years has been found to have a positive effect on the number of cases and a negative effect on the fade-out probability (Finkenstadt & Grenfell 1998). This is consistent with an increase in the person-to-person transmission rate when an important number of children leave the group of children under 4 years old (core group). Under the assumption that the core group was the group of 4–6-year-old children (early school

Table 5. Transmission parameters for measles epidemics in London in the pre-vaccination era (1948–1964). (Posterior mean and 95% credible interval of the mean infectious period $1/\gamma$, the initial number of susceptibles S_0, the reporting rate ρ and the power ϵ. Estimates obtained with the TSIR approach are also given (parameter $q=26.17$ [22.19,30.51]). The duration of observation periods is 14 days.)

	$1/\gamma$ (days)	S_0 ($\times 10^3$)	ρ (%)	ϵ (%)
$1/\gamma$ estimated	3.34 [3.12,3.56]	428 [410,448]	47.50 [47.25,47.75]	0.95 [0.54,1.38]
$1/\gamma$ fixed	14 —	220 [215,225]	48.25 [48.10,48.41]	2.25 [1.84,2.66]
TSIR	—	246 [219,277]	48.04 [47.22,48.86]	4.63 [3.07,6.22]

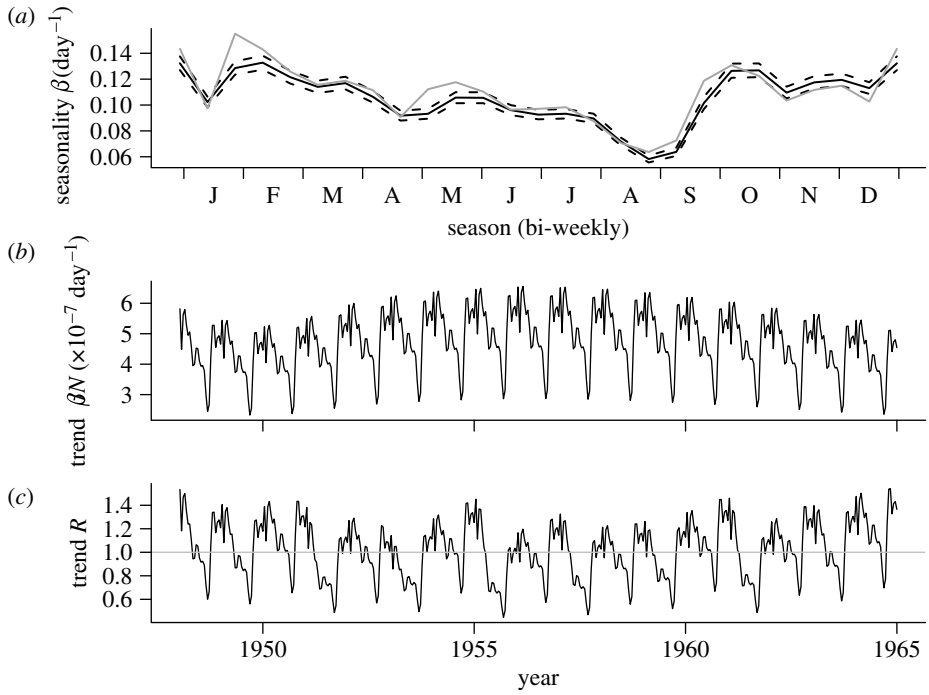

Figure 4. Seasonality and trend of the transmission rate and the effective reproduction number for measles epidemics in London, estimated under the assumption that the mean infectious period is equal to 14 days. (*a*) Seasonality of the transmission rate (solid line, posterior mean; dashed line, 95% credible interval; grey line, posterior mean of the daily transmission rate $r_k/14$ estimated with the TSIR approach). (*b*) Trend of the transmission rate β/N. (*c*) Trend of the effective reproduction number (R). The transmission rate for period k is β_k/N_k where N_k is the size of the core group (children with age below 4 years). The formula for the effective reproduction number is given in the main text. We correct for the fact that the measles latent period is 8 days (Anderson & May 1991).

years) and that there was a 4-year delay between birth and introduction to the susceptible compartment, the fit was similar to that for our baseline scenario. When the core group was the group of 0–9 years old, there was no improvement of the fit compared with the situation where it is assumed that core group size and therefore transmission rates remain constant over time (in both cases, the expected period of epidemic cycles was 1 year). Further research could try to determine which core group gives optimal fit in a more systematic way.

In a context where the TSIR approach is expected to be applicable (generation time $\approx$ duration of the observation period), we found that our estimates were similar to those obtained with TSIR. The main difference was observed for the power coefficient ϵ, which was larger for the TSIR model than our diffusion method. One possible explanation is that the relatively crude way TSIR deals with temporal aggregation in the data leads to the overestimation of the 'gap' between mass action and historical contact patterns. Apart from this effect, our results validate the use of the TSIR approach when generation time $\approx$ duration of observation period.

When the observation interval and generation time are not approximately equal, more refined statistical models are needed, such as the one we presented here. Using simulated data, we found that our approach provided accurate estimates for all transmission parameters (including the mean infectious period) so long as the observation interval was less than or equal to 2.5-fold more than the generation time of the disease. This suggests that, using weekly surveillance data, our approach could be used to study most of the more common respiratory diseases, even those with very short generation time (for influenza, for example, two recent estimates of the generation time obtained from independent datasets are 2.6 days (Ferguson *et al.* 2005) and 2.85 days (Wallinga & Lipsitch 2007), respectively)). The relevance of the approach for rare diseases is more difficult to assess since the reporting rate may then vary dramatically with time. When the effective reproduction number is high, the

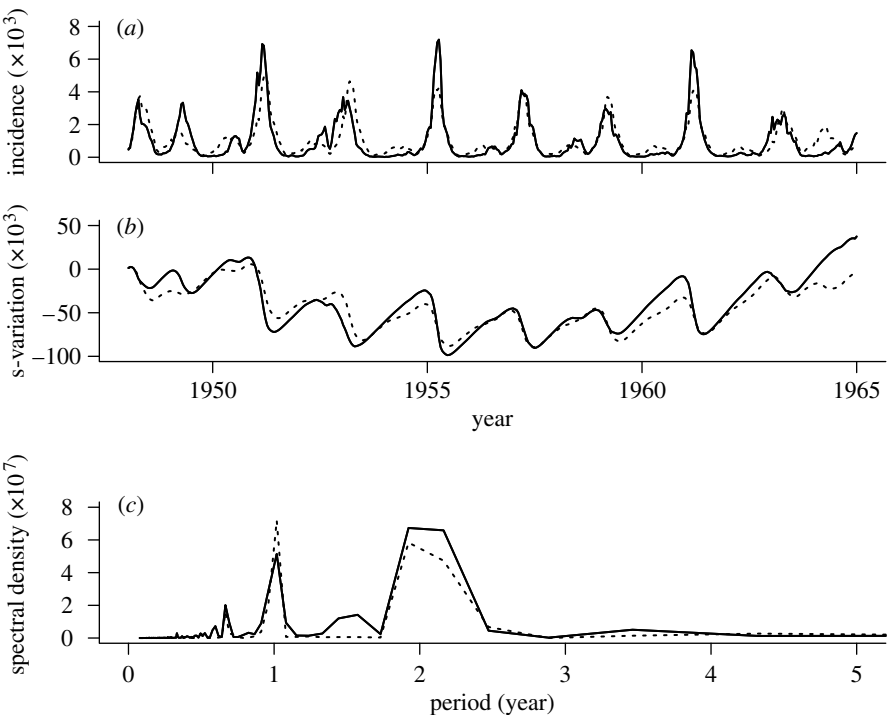

Figure 5. Model checking. (*a*) Number of cases of measles reported in London between 1948 and 1964. (*b*) Variations in the number of susceptibles. (*c*) Spectral density. The solid line is the observed curve and the dashed line is the predicted curve (average of 40 epidemics simulated from the model, with parameters equal to their posterior mean). The mean infectious period is assumed to be equal to 14 days.

method might not be able to cope with the same level of temporal aggregation as for measles, since the assumption that the number of susceptibles is roughly constant during one observation period might start to break down. When we used an SEIR model to generate the simulated data, we found that the estimate of $1/\gamma$ from the SIR diffusion model used to fit the data corresponded to the generation time of the SEIR model. When the assumption of mass action was violated, estimates remained accurate so far as ϵ was relatively small.

Although our approach provided accurate estimates of the generation time for simulated data, estimation of the generation time for measles from historical data was, disappointingly, unsuccessful. There were two contradictory observations regarding the quality of the fit for estimated γ ($1/\gamma = 3.34$ days) and fixed γ ($1/\gamma = 14$ days). First, likelihood comparison suggests that the model with estimated γ (log-lik $= -7011$) has a better fit than the model with fixed γ (log-lik $= -9226$). However, epidemics simulated with $1/\gamma = 3.34$ days fade out quickly, while those simulated with $1/\gamma = 14$ days gave a good fit. A possible explanation is that the long- and short-term predictions contradict each other, perhaps because more refined modelling is required to capture structural changes in contact patterns. It is also possible that the high removal rate we estimated for compartment I is due to an overestimated flow into this compartment. This might for example happen if the reporting rate is positively correlated with the number of cases, which is a plausible phenomenon. However, estimates of $1/\gamma$ were unchanged when we defined different reporting rates for high/low incidence periods.

Our inability to estimate the generation time of measles therefore has an intriguing and novel interpretation. It suggests that standard models assuming homogenous mixing (even within an age-defined core group) miss key features of epidemics in large populations. Possible ways to relax the assumption of homogenous mixing models are of course well known (Anderson & May 1991)—for example allowing for heterogeneity in susceptibility/infectiousness, the age structure of the population, spatial substructuring or network structure. Determining which of these elements are needed to improve parameter estimates would provide an important insight into which are the most important determinants of epidemic patterns in large populations. This is the topic of ongoing research.

Assuming that the generation time was known, the method we developed provided estimates of other parameters consistent with estimates from the TSIR model (when the TSIR is applicable) and the model captured both the magnitude and biannual pattern of measles epidemics. The most relevant epidemiological model for measles is the SEIR model with a latent period of 6–9 days and an infectious period of 6–7 days (Anderson & May 1991). It is therefore relatively unlikely that measles cases are effectively infectious for 14 days. We nonetheless assumed $1/\gamma = 14$ days because we found that, when our approach was applied to data simulated from the SEIR model, the estimate of $1/\gamma$ corresponded to the generation time of the disease (sum of the latent and infectious periods).

In general, it is not possible to identify both the reporting rate and the transmission parameters. For surveillance data on influenza for example, by appropriate re-scaling of the transmission parameters, it is

probable that data would be consistent with a wide range of reporting rates. The fact that here we were able to obtain a precise estimate of the reporting rate is therefore relatively intriguing. Our ability to estimate the reporting rate for measles (as our inability to estimate it for influenza) is due to the fact that measles, as opposed to influenza, confers permanent immunity. In a context where the system is roughly stationary (i.e. no major change in epidemic patterns over time and relatively similar birth rates over time), we expect that the number of susceptible individuals in the population is roughly stationary too. This is possible only if the flow of births is compensated by the flow of infections. In this context, a very intuitive estimate of the reporting rate is the coefficient a of the linear regression: (cumulated number of infections) $= a$ (cumulated number of births) $+ b$, with $a = 0.478$. For diseases like influenza that do not confer permanent immunity, strong assumptions on the reporting process or the history of immunity are needed to estimate transmission parameters (Finkenstadt *et al.* 2005).

Standard results on the Cox–Ingersoll–Ross model (Cox *et al.* 1985) provided the distribution $I'_T|I'_0, S_0$. However, more work was required to relate I'_T and I'_0 to the observation, i.e. the number U_0 of infections in $[0, T]$. No exact solution could be obtained for $P(U_0|I'_T, I'_0, S_0)$; we could derive only an approximated form of the distribution. We used that, given the expected number of new infections $\Lambda_0 = \beta_0 \overline{S_0} \int_0^T I'_t \, \mathrm{d}t$, the number U_0 of infections should be Poisson distributed with mean Λ_0. The main issue was then to find a good predictor of Λ_0 given I'_t, I'_0, S_0. The analysis of the Laplace transform (appendix C) suggests that the linear predictor $\tilde{x}_0 + \tilde{y}_0 I'_t$ used here has satisfying properties

(i) Λ_0 and I'_T are highly correlated

$$\mathrm{cor}(\Lambda_0, I'_T|I'_0, S_0) = \frac{\sqrt{3}}{2} + O(r_0 T) \approx 0.866 + O(r_0 T),$$

and

(ii) The predictor explains a large part of the variance of the expected number of new infections

$$\frac{\mathrm{var}(\Lambda_0 - \tilde{x}_0 - \tilde{y}_0 I'_T|I'_0, S_0)}{\mathrm{var}(\Lambda_0|I'_0, S_0)} = \frac{1}{4} + O(r_0 T).$$

We therefore assumed that $\Lambda_0|I'_T, I'_0, S_0$ was gamma distributed with mean $\tilde{x}_0 + \tilde{y}_0 I'_T$ and variance $\mathrm{var}(\Lambda_0 - \tilde{x}_0 - \tilde{y}_0 I'_T)$. The choice of the gamma distribution had no theoretical foundation, but simplified the computation since the number of infections U_0 had then a negative binomial distribution.

In the pre-vaccination era, in large cities like London, measles was endemic and it is not necessary to model introduction of cases (Morton & Finkenstadt 2005). In smaller towns, fade outs were common. The statistical framework could be extended to take into account the introduction of cases in this context. Under the assumption that cases are introduced at the beginning of each observation period, stochastic differential equation (2.3) would still apply.

In the standard SIR model, it is assumed that the duration of infectiousness is exponentially distributed. This is motivated by mathematical tractability (under this assumption, the system is Markovian) rather than biological realism. Data augmentation techniques can cope with more realistic distributions for the duration of infectiousness (Cauchemez *et al.* 2004), but those techniques are available for relatively small datasets only. For time-series data, designing estimation methods that do not rely on the Markovian assumption is the subject of further research.

We thank the MRC, European Union FP6 SARSTRANS and INFTRANS projects, and the NIGMS MIDAS initiative for research funding.

APPENDIX A. NON-CENTRAL χ^2 SOLUTION OF SDE

Consider the solution of the stochastic differential equation (2.3)

$$\mathrm{d}I'_t = r_0 I'_t \, \mathrm{d}t + \sigma_0 \sqrt{I'_t} \, \mathrm{d}W.$$

Denoting $c_0 = 2r_0(\mathrm{e}^{r_0 T} - 1)/\sigma_0$ and $u_0 = c_0(\mathrm{e}^{r_0 T} - 1)$ I'_0, $2c_0 I'_T|I'_0, S_0$ has a non-central χ^2 distribution with zero d.f. and with non-centrality parameter $2u_0$. Described by Siegel (1979), the non-central χ^2 distribution with zero d.f. has a mass at 0, which corresponds here to the probability of extinction of the outbreak. If $Y \sim \chi^2_0(\lambda)$ where λ is the non-centrality parameter,

$$P(Y = 0) = \exp(-\lambda/2).$$

The positive part of the distribution has a density f_λ in the sense that, if $Y \sim \chi^2_0(\lambda)$ and $0 \le a < b$, then

$$P(a < Y < b) = \int_a^b f_\lambda(y) \mathrm{d}y,$$

$$f_\lambda(y) = 0.5 \left(\frac{\lambda}{y}\right)^{0.5} \exp(0.5(\lambda + y)) \boldsymbol{I}'_1\left\{\sqrt{\lambda y}\right\},$$

where $\boldsymbol{I}'_1\{\cdot\}$ is the modified Bessel function of the first kind. Note that f_λ is not a proper density since $\int_0^\infty f_\lambda(y)\mathrm{d}y = 1 - \exp(-\lambda/2) < 1$. The distribution has mean λ and variance 4λ, and may be approximated by the normal distribution $N(\lambda, 4\lambda)$ when λ is large.

APPENDIX B. ANALYSIS OF THE LAPLACE TRANSFORM

The Laplace transform of $(I'_T, \int_0^T I'_t \, \mathrm{d}t)|I'_0, S_0$ is (Ben-Ameur *et al.* 2006)

$$F(\nu, \omega) = E\left\{\exp\left(-\nu I'_T - \omega \int_0^T I'_t \, \mathrm{d}t\right)|I'_0, S_0\right\}$$

$$= \exp(-Y(\nu, \omega) I'_0),$$

where

$$\xi(\omega) = \sqrt{r_0^2 + 2\omega\sigma_0^2},$$

$$Y(\nu, \omega) = \frac{\nu(\xi(\omega) - r_0 + \mathrm{e}^{\xi(\omega)T}(\xi(\omega) + r_0)) + 2\omega(\mathrm{e}^{\xi(\omega)T} - 1)}{(\nu\sigma_0^2 + \xi(\omega) - r_0)(\mathrm{e}^{\xi(\omega)T} - 1) + 2\xi(\omega)}.$$

Define $(\tilde{u}, \tilde{v})$ the scalars that minimize the function

$$L(u,v) = E\left[\left(\int_0^T I'_t\,\mathrm{d}t - u - vI'_T\right)^2 | I'_0, S_0\right].$$

The solution can be obtained analytically

$$\begin{cases} \dfrac{\partial L}{\partial u}(\tilde{u},\tilde{v}) = 0 \\ \dfrac{\partial L}{\partial v}(\tilde{u},\tilde{v}) = 0 \end{cases} \Leftrightarrow \begin{cases} E\left[\int_0^T I'_t\,\mathrm{d}t | I'_0, S_0\right] \\ \quad -\tilde{u} - \tilde{v}E[I'_T | I'_0, S_0] = 0, \\ E\left[I'_T.\int_0^T I'_t \mathrm{d}t | I'_0, S_0\right] \\ \quad -\tilde{u} - \tilde{v}E\left[(I'_T)^2 | I'_0, S_0\right] = 0, \end{cases}$$

$$\begin{cases} \tilde{v} = \dfrac{\mathrm{cov}\left(I'_T, \int_0^T I'_t\,\mathrm{d}t | I'_0, S_0\right)}{\mathrm{var}(I'_T | I'_0, \mathrm{S}_0)}, \\ \tilde{u} = E\left(\int_0^T I'_t\,\mathrm{d}t | I'_0, S_0\right) - \tilde{v}E(I'_T | I'_0, S_0). \end{cases}$$

Standard results on Laplace transforms give

$$\begin{cases} \mathrm{cov}\left(I'_T, \int_0^T I'_t\,\mathrm{d}t | I'_0, S_0\right) = \partial^2 F/\partial\nu\,\partial\omega|_{(0,0)}, \\ \mathrm{var}(I'_T | I'_0, S_0) = \partial^2 F/\partial^2\nu|_{(0,0)}, \\ E\left(\int_0^T I'_t\,\mathrm{d}t | I'_0, S_0\right) = -\left.\dfrac{\partial F}{\partial\omega}\right|_{(0,0)}, \\ E(I'_T | I'_0, S_0) = -\left.\dfrac{\partial F}{\partial\nu}\right|_{(0,0)}. \end{cases}$$

Eventually, $(\tilde{u}, \tilde{v})$ are equal to

$$\begin{cases} \tilde{v} = \left.\dfrac{\partial^2 F/\partial\nu\partial\omega}{\partial^2 F/\partial^2\nu}\right|_{(0,0)} = \dfrac{1}{r_0} - \dfrac{T}{\mathrm{e}^{r_0 T} - 1}, \\ \tilde{u} = -\left.\dfrac{\partial F}{\partial\omega}\right|_{(0,0)} + \tilde{v}\left.\dfrac{\partial F}{\partial\nu}\right|_{(0,0)} = \dfrac{I'_0(\mathrm{e}^{r_0 T} - 1)}{r_0} - \tilde{v}I'_0\mathrm{e}^{r_0 T}. \end{cases}$$

Denoting $Z_T = \int_0^T I'_t\,\mathrm{d}t - \tilde{u} - \tilde{v}I'_T$, the residual, the Laplace transform of $(I'_T, Z_T)|I'_0, S_0$ is

$$\begin{aligned} G(\nu,\omega) &= E\{\exp(-\nu I'_T - \omega Z_T)|I'_0, S_0\} \\ &= E\left\{\exp\left(-vI'_T - \omega\left(\int_0^T I'_t\mathrm{d}t - \tilde{u} - \tilde{v}I'_T\right)|I'_0, S_0\right)\right\} \\ &= F(v - \omega\tilde{v}, \omega)\exp(\omega\tilde{u}). \end{aligned}$$

It is easy to check that the residual has mean 0 and variance

$$\tilde{z} = \frac{\partial^2 G}{\partial^2\omega} = \frac{I'_0\sigma_0^2(1 + \mathrm{e}^{2r_0 T} - \mathrm{e}^{r_0 T}(2 + r_0^2 T^2))}{r_0^3(\mathrm{e}^{r_0 T} - 1)}.$$

Scalars $\{\tilde{x}_0, \tilde{y}_0, \tilde{v}_0\}$ are then straightforward to calculate from $\{\tilde{u}, \tilde{v}, \tilde{z}\}$ and the definition of $\Lambda_0 = \beta_0 S_0 \int_0^T I'_t\,\mathrm{d}t$

$$\begin{cases} \tilde{y}_0 = \beta_0\overline{S}_0(1/r_0 - T/(\mathrm{e}^{r_0 T} - 1)), \\ \tilde{x}_0 = \beta_0\overline{S}_0 I'_0(\mathrm{e}^{r_0 T} - 1)/r_0 - I'_0\mathrm{e}^{r_0 T}\tilde{y}_0, \\ \tilde{v}_0 = (\beta_0\overline{S}_0)^2\dfrac{I'_0\sigma_0^2(1 + \mathrm{e}^{2r_0 T} - \mathrm{e}^{r_0 T}(2 + r_0^2 T^2))}{r_0^3(\mathrm{e}^{r_0 T} - 1)}. \end{cases} \quad (\mathrm{B}\,1)$$

APPENDIX C. AVERAGE NUMBER OF SUSCEPTIBLES AND INFECTIVES

We need to approximate the average numbers of susceptibles $\overline{S}_0$ and infectives $\bar{I}'_0$ over time period $[0, T]$.

Assume first that there is no birth in the population. A natural choice is to specify $\overline{S}_0$ equal to the number S_0 of susceptibles at the beginning of the period. In this case, the expected number of infections in $[0, T[$ is (from the Laplace transform of appendix B)

$$\overline{U_0} = \beta_0 S_0 E\left(\int_0^T I'_t\,\mathrm{d}t | I'_0\right) = \beta_0 S_0 I'_0\frac{(\mathrm{e}^{(\beta_0 S_0 - \gamma)T} - 1)}{\beta_0 S_0 - \gamma},$$

and the average number of susceptibles in the interval can be roughly approximated by

$$\begin{aligned} &S_0 - \overline{U_0}/2 \\ &\quad = S_0\left(1 - \beta_0\frac{(\exp(\beta_0 S_0 T - \gamma T) - 1)}{2\beta_0 S_0 - 2\gamma}I'_0\right). \end{aligned} \quad (\mathrm{C}\,1)$$

This value underestimates the average number of susceptibles when births occur. To correct for this bias, we simply assume that half of the births occur at the beginning of the period and that there is no birth during the period. Equation (C 1) becomes

$$\begin{aligned} &\overline{S}_0 = (S_0 + 0.5B_0) \\ &\left(1 - \beta_0\frac{(\exp(\beta_0(S_0 + 0.5B_0)T - \gamma T) - 1)}{2\beta_0(S_0 + 0.5B_0) - 2\gamma}I'_0\right). \end{aligned} \quad (\mathrm{C}\,2)$$

We use this last value in our inference framework.

The same type of iterative approach can be used when the hypothesis of mass action is violated

$$\begin{cases} \bar{I}'_0 = \dfrac{I'_0}{T}\dfrac{\exp\left(\beta(S_0 + 0.5B_0)I_0'^{-\epsilon}T - \gamma T\right) - 1}{\beta(S_0 + 0.5B_0)I_0'^{-\epsilon} - \gamma}, \\ \overline{S}_0 = (S_0 + 0.5B_0) \\ \left(1 - \beta_0\bar{I}_0'^{-\epsilon}\dfrac{\exp\left(\beta_0\bar{I}_0'^{-\epsilon}(S_0 + 0.5B_0)T - \gamma T\right) - 1}{2\beta_0\bar{I}_0'^{-\epsilon}(S_0 + 0.5B_0) - 2\gamma}\bar{I}'_0\right). \end{cases} \quad (\mathrm{C}\,3)$$

REFERENCES

Anderson, R. M. & May, R. M. 1991 *Infectious diseases of humans: dynamics and control*. Oxford, UK: Oxford University Press.

Auranen, K., Arjas, E., Leino, T. & Takala, A. K. 2000 Transmission of pneumococcal carriage in families: a

latent markov process model for binary longitudinal data. *J. Am. Stat. Assoc.* **95**, 1044–1053. (doi:10.2307/2669741)

Bailey, N. T. J. 1964 *The elements of stochastic processes: with applications to the natural sciences*. New York, NY: Wiley.

Bailey, N. T. J. 1975 *The mathematical theory of infectious diseases and its applications*. London, UK: Griffin.

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

Becker, N. G. & Britton, T. 1999 Statistical studies of infectious disease incidence. *J. R. Stat. Soc. B* **61**, 287–307. (doi:10.1111/1467-9868.00177)

Becker, N. G. & Hasofer, A. M. 1997 Estimation of epidemics with incomplete observations. *J. R. Stat. Soc. B* **59**, 415–429. (doi:10.1111/1467-9868.00076)

Ben-Ameur, H., Breton, M., Karoui, L. & L'Ecuyer, P. 2006 A dynamic programming approach for pricing options embedded in bonds. *J. Econ. Dynam. Control* **31**, 2212–2233. (doi:10.1016/j.jedc.2006.06.007)

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A. J. & Boelle, P. Y. 2004 A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Stat. Med.* **23**, 3469–3487. (doi:10.1002/sim.1912)

Cauchemez, S., Temime, L., Guillemot, D., Varon, E., Valleron, A. J., Thomas, G. & Bolle, P. Y. 2006 Investigating heterogeneity in pneumococcal transmission: a Bayesian MCMC approach applied to a follow-up of schools. *J. Am. Stat. Assoc.* **101**, 946–958. (doi:10.1198/016214506000000230)

Cox, J. C., Ingersoll, J. E. & Ross, S. A. 1985 A theory of the term structure of interest rates. *Econometrica* **53**, 385–408. (doi:10.2307/1911242)

De Jong, M. C. M., Diekmann, O. & Heesterbeek, H. 1995 How does transmission of infection depend on population size? In *Epidemic models: their structure and relation to data* (ed. D. Mollison), pp. 84–94. Cambridge, UK: Cambridge University Press.

Ferguson, N. M., Cummings, D. A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S. & Burke, D. S. 2005 Strategies for containing an emerging influenza pandemic in southeast Asia. *Nature* **437**, 209–214. (doi:10.1038/nature04017)

Finkenstadt, B. & Grenfell, B. 1998 Empirical determinants of measles metapopulation dynamics in England and Wales. *Proc. R. Soc. B* **265**, 211–220. (doi:10.1098/rspb.1998.0284)

Finkenstadt, B. F. & Grenfell, B. T. 2000 Time series modelling of childhood diseases: a dynamical systems approach. *J. R. Stat. Soc. C* **49**, 187–205. (doi:10.1111/1467-9876.00187)

Finkenstadt, B. F., Bjornstad, O. N. & Grenfell, B. T. 2002 A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. *Biostatistics* **3**, 493–510. (doi:10.1093/biostatistics/3.4.493)

Finkenstadt, B. F., Morton, A. & Rand, D. A. 2005 Modelling antigenic drift in weekly flu incidence. *Stat. Med.* **24**, 3447–3461. (doi:10.1002/sim.2196)

Gibson, G. J. & Renshaw, E. 1998 Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA. J. Math. Appl. Med. Biol.* **15**, 19–40. (doi:10.1093/imammb/15.1.19)

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. 1996 *Markov Chain Monte Carlo in practice*. London, UK: Chapman and Hall.

Green, P. J. 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. (doi:10.1093/biomet/82.4.711)

Morton, A. & Finkenstadt, B. F. 2005 Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *J. R. Stat. Soc. C* **54**, 575–594. (doi:10.1111/j.1467-9876.2005.05366.x)

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M. & Mollison, D. 2000 Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J. R. Stat. Soc. C* **49**, 517–542. (doi:10.1111/1467-9876.00210)

Renshaw, E. 1991 *Modelling biological populations in space and time*. Cambridge, UK: Cambridge University Press.

Siegel, A. F. 1979 The noncentral chi-squared distribution with zero degrees of freedom and testing for uniformity. *Biometrika* **66**, 381–386. (doi:10.1093/biomet/66.2.381)

Soper, H. E. 1929 The interpretation of periodicity in disease prevalence. *J. R. Stat. Soc.* **92**, 34–73. (doi:10.2307/2341437)

Wallinga, J. & Lipsitch, M. 2007 How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B* **274**, 599–604. (doi:10.1098/rspb.2006.3754)