J. R. Statist. Soc. B (1999)
61, *Part* 2, *pp.* 309–330

On the time to extinction in recurrent epidemics

Ingemar Nåsell

Royal Institute of Technology, Stockholm, Sweden

[Received October 1997. Final revision July 1998]

Summary. An approximation is derived for the expected time to extinction in a stochastic model for recurrent epidemics. Numerical illustrations indicate that the approximation is crude but that it has the correct order of magnitude. The quasi-stationary distribution plays an important role in the derivation. Approximations for the critical community size and for the persistence threshold are derived. Comments are made on the classical study by Bartlett (1956–1960).

Keywords: Critical community size; Persistence threshold; Quasi-stationary distribution; Stochastic fade-out; Time to extinction

1. Introduction

A recurrence of epidemic outbreaks is known to happen for a number of diseases that confer immunity. Recurrence can be explained by the combined influence of epidemic and demographic forces. Stochastic models that account for these two forces in a closed population predict that the infection will eventually become extinct. The time to extinction is an important measure of the persistence of the infection. The aim of the present paper is to study the time to extinction for two closely related stochastic models for recurrent epidemics.

The mathematical problem of determining the time to extinction has proved to be surprisingly difficult. Classical results in this area are contained in a series of papers by Bartlett (1956, 1957, 1960a, b, c). A description of Bartlett's contributions is given by Bailey (1975). One of our findings is that Bartlett's approximation for the expected time to extinction gives a serious underestimate in an important region of the parameter space.

The two models that we analyse can somewhat imprecisely be described as endemic SIR models. Here, the letters S, I and R refer to the successive states of the individuals: S for susceptible, I for infected and R for recovered and immune. The impreciseness in this description is related to the fact that the endemicity is limited in time with the duration measured by the time to extinction.

Both of the models that we deal with have been used previously in both deterministic and stochastic versions for studying recurrent epidemics. The two models treat the epidemic events of infection and recovery in identical ways, but they differ in their treatment of demographic forces. Both allow for an inflow of susceptible individuals at a constant immigration rate, but only one of the models, treated in Section 2, allows the individuals to die, whereas the model of Section 3 allows all individuals to live for ever. The model of Section 3 is therefore less realistic. Heuristically one can, however, argue that this difference between the

Address for correspondence: Ingemar Nåsell, Department of Mathematics, Royal Institute of Technology, S-10044 Stockholm, Sweden.
E-mail: ingemar@math.kth.se

 1369–7412/99/61309

hypotheses of the two models should have small consequences in the interesting case where the rate of recovery from infection is much larger than the death-rate. The model in Section 2 has the additional desirable feature that its deterministic version has a threshold, whereas the model in Section 3 does not. Our main reason for analysing the model of Section 3 is that this model was used in Bartlett's work, and in the work of several later researchers.

The stochastic versions of both models are bivariate birth–death processes with absorbing states. The concept of quasi-stationarity is of importance for such models. This concept was introduced for continuous time Markov chains by Darroch and Seneta (1967). Quasi-stationarity in this sense in the epidemic context was first used by Kryscio and Lefèvre (1989). They were concerned with a stochastic logistic model that also can be interpreted as an SIS model. The SIS model is used for infections that give no immunity; in this model an individual who recovers from an infection is assumed to be susceptible to reinfection. The same model has been analysed further by Nåsell (1996). He extended the results of Kryscio and Lefèvre (1989) by the systematic use of asymptotic approximations. Such approximations are conceptually useful in a search for qualitative features in stochastic models.

Our results support the claim by Bartlett (1956) that recurrence and extinction can be understood through the study of a fully stochastic model. The same claim is further emphasized by Dietz (1995).

2. The Martini model

In the Martini model, all individuals are subject to an immigration–death process, with the death-rate independent of the state of infection. Furthermore, all newly immigrated individuals are assumed to be susceptible. The infection of susceptible individuals and recovery of infected individuals are treated in the same way as in the classical SIR model.

The deterministic version of this model has previously been studied by Martini (1921), Lotka (1923, 1956), Hethcote (1974, 1976) and Anderson and May (1991). The corresponding deterministic model for a host vector situation has been studied by Dietz (1975). A slight extension of the model is denoted an endemic SEIR model. Here, the letter E refers to exposed individuals. In this model, each individual spends some time in the exposed state after receiving infection but before becoming infected. A deterministic SEIR model has been studied by Olsen and Schaffer (1990) and Sugihara and May (1990). The stochastic version of this model, also incorporating age structure, has been studied by simulation by Schenzle (1984) and Keeling and Grenfell (1997). The stochastic version of the model has been studied mathematically by van Herwaarden and Grasman (1995).

2.1. Model formulation

We use two state variables, namely the number of susceptible individuals $S(t)$ and the number of infected individuals $I(t)$ at time t. They take values in the state space $\{(m, n)$: $m = 0, 1, 2, \ldots, n = 0, 1, 2, \ldots\}$. The joint distribution at time t is denoted

$$p_{mn}(t) = P\{S(t) = m, I(t) = n\}.$$

We use this notation even when m and/or n are negative, with the convention that $p_{mn}(t)$ is then equal to 0.

The model accounts for four basic events, i.e. infection of a susceptible individual, recovery of an infected individual, immigration of a susceptible individual and death of an individual

Table 1. Hypothesized transition rates for the stochastic version of the Martini model

Event	*Transition*	*Transition rate*
Immigration of susceptible individual	$(m, n) \to (m+1, n)$	$\lambda_1(m, n) = \mu N$
Death of susceptible individual	$(m, n) \to (m-1, n)$	$\mu_1(m, n) = \mu m$
Infection of susceptible individual	$(m, n) \to (m-1, n+1)$	$\nu_2(m, n) = \beta mn/N$
Death or recovery of infected individual	$(m, n) \to (m, n-1)$	$\mu_2(m, n) = (\mu+\gamma)n$

(who can be susceptible or infected). Four parameters are used at the outset, namely the total population size N, the death-rate per individual μ, the contact rate β and the recovery rate per infected individual γ. All four parameters are assumed to be strictly positive. The hypotheses for the model are set out by specifying the transition rates as shown in Table 1.

All the states $(m, 0)$ communicate with each other, but not with any state (m, n) with $n \geqslant 1$. The states $(m, 0)$ are absorbing and the states (m, n), $n \geqslant 1$, are transient. The quasi-stationary distribution is supported on the set of transient states, and the (trivial) stationary distribution on the set of absorbing states.

The Kolmogorov forward equations for the model can be written as

$$p'_{mn}(t) = \lambda_1(m-1, n)\, p_{m-1,n}(t) + \mu_1(m+1, n)\, p_{m+1,n}(t) + \nu_2(m+1, n-1)\, p_{m+1,n-1}(t) + \mu_2(m, n+1)\, p_{m,n+1}(t) - \kappa(m, n)\, p_{mn}(t), \tag{2.1}$$

where

$$\kappa(m, n) = \lambda_1(m, n) + \mu_1(m, n) + \nu_2(m, n) + \mu_2(m, n).$$

2.2. The deterministic model and reparameterization

We consider the deterministic approximation of the stochastic model. This is useful both for the reparameterization that it suggests and for its clear exhibition of the bifurcation phenomenon that corresponds to a threshold for the deterministic model.

The deterministic model can be expressed by the following pair of differential equations:

$$S' = \mu N - \frac{\beta}{N} SI - \mu S,$$
$$I' = \frac{\beta}{N} SI - (\gamma + \mu) I.$$

We reparameterize the model by introducing R_0 to denote the basic reproduction ratio, defined as the ratio of the contact rate β to the sum of the recovery rate γ and the death-rate μ. Furthermore, we let α denote the ratio of the sum of the recovery rate γ and the death-rate μ to the death-rate μ. This means that α is the ratio of the average life length to the average duration of infection. In our numerical illustrations we shall use two values of α, namely $\alpha = 500$ and $\alpha = 2500$. These values are typical for polio and measles respectively, with an average life length of 50 years and an average duration of infection of 1 month and 1 week respectively. The derived parameters are

$$R_0 = \beta/(\gamma + \mu),$$
$$\alpha = (\gamma + \mu)/\mu. \tag{2.2}$$

The system of differential equations above has critical points at $(N, 0)$ and at $(N/R_0, N(R_0 - 1)/\alpha R_0)$. The first of these critical points corresponds to the absence of infection, and the second corresponds to an endemic infection level provided that $R_0 > 1$. Stability analysis can be used to show that bifurcation occurs for $R_0 = 1$. We express this by saying that the deterministic model has a threshold at $R_0 = 1$. This threshold result is perhaps the most important result to emerge from the deterministic analysis. We shall see below that the threshold values of R_0 based on the stochastic model are substantially larger than 1 for realistic values of the parameters N and α. This means that the deterministic threshold is a poor approximation to the stochastic threshold.

2.3. The quasi-stationary distribution

We derive first a differential equation that will be used both later in this subsection and in the discussion of time to extinction in the next subsection. Put $n = 0$ in the Kolmogorov forward equations (2.1) and insert the explicit expressions for the transition rates given above. Also, use the reparameterization given by equations (2.2). Furthermore, introduce

$$p_{\cdot n}(t) = \sum_{m=0}^{\infty} p_{mn}(t) = P\{I(t) = n\}$$

to denote the marginal distribution of the number of infected individuals at time t. By summing the forward equations for $n = 0$ over all m-values we obtain

$$p'_{\cdot 0}(t) = \mu\alpha\, p_{\cdot 1}(t). \tag{2.3}$$

Next we derive a system of equations for the quasi-stationary distribution. The state probabilities conditioned on not being absorbed are denoted $q_{mn}(t)$ and given by

$$\begin{aligned} q_{mn}(t) &= P\{S(t) = m,\, I(t) = n | I(t) \neq 0\} \\ &= \frac{p_{mn}(t)}{1 - p_{\cdot 0}(t)}, \qquad m = 0, 1, \ldots, \quad n = 1, 2, \ldots . \end{aligned} \tag{2.4}$$

We introduce $q_{\cdot n}(t) = \Sigma_{m=0}^{\infty} q_{mn}(t)$ to denote the marginal distribution for the number of infected individuals at time t, conditioned on not having reached any state in the absorbing set. By differentiating the expression for $q_{mn}(t)$ in equation (2.4) and applying equation (2.3) we obtain

$$q'_{mn}(t) = \frac{p'_{mn}(t)}{1 - p_{\cdot 0}(t)} + \mu\alpha\, q_{\cdot 1}(t)\, \frac{p_{mn}(t)}{1 - p_{\cdot 0}(t)}. \tag{2.5}$$

By applying the forward equations for $p_{mn}(t)$ in equation (2.1) we obtain the following system of differential equations for the conditional state probabilities $q_{mn}(t)$:

$$\begin{aligned} q'_{mn}(t) = {} & \lambda_1(m-1, n)\, q_{m-1,n}(t) + \mu_1(m+1, n)\, q_{m+1,n}(t) + \nu_2(m+1, n-1)\, q_{m+1,n-1}(t) \\ & + \mu_2(m, n+1)\, q_{m,n+1}(t) - \kappa(m, n)\, q_{mn}(t) + \mu\alpha\, q_{\cdot 1}(t)\, q_{mn}(t), \\ & \qquad m = 0, 1, \ldots, \quad n = 1, 2, \ldots . \end{aligned} \tag{2.6}$$

The quasi-stationary distribution $\{q_{mn}\}$ is the stationary solution of this system of equations.

We proceed to derive a recursion relationship for the marginal distribution $q_{\cdot n}$ of the number of infected individuals in quasi-stationarity. Accordingly, we put $q'_{mn}(t) = 0$ in

equation (2.6) and insert the explicit expressions for the transition rates, using the reparameterization (2.2). We then add the resulting equations over all m-values and introduce

$$e_S(n) = \sum_{m=1}^{\infty} m q_{mn}/q_{\cdot n}$$

to denote the conditional expectation of S, given $I = n$, in quasi-stationarity. By furthermore putting

$$f(n) = n q_{\cdot n} - (R_0/N)(n-1) q_{\cdot n-1}\, e_S(n-1), \qquad n = 1, 2, \ldots,$$

we find that

$$f(n+1) = f(n) - q_{\cdot 1}\, q_{\cdot n}, \qquad n = 1, 2, \ldots.$$

We can easily solve this recursion relationship since $f(1) = q_{\cdot 1}$. The result is

$$f(n+1) = \left(1 - \sum_{k=1}^{n} q_{\cdot k}\right) q_{\cdot 1}, \qquad n = 0, 1, \ldots.$$

It follows that the marginal distribution $\{q_{\cdot n}\}$ of infected individuals in quasi-stationarity satisfies the recursion relationship

$$q_{\cdot n+1} = \frac{R_0}{N} \frac{n}{n+1} e_S(n)\, q_{\cdot n} + \frac{1 - \sum_{k=1}^{n} q_{\cdot k}}{n+1} q_{\cdot 1}, \qquad n = 1, 2, \ldots. \tag{2.7}$$

This relationship cannot be used to solve for the marginal distribution in quasi-stationarity, since the conditional expectation $e_S(n)$ is unknown. We include it here as a possibly useful basis for future work.

2.4. The distribution of the time to extinction

The time to extinction is an important random variable that is of great interest because of its epidemiological interpretation. It was used to define thresholds for stochastic endemic models by Nåsell (1995).

Two initial distributions are especially interesting. One is the quasi-stationary distribution and the other corresponds to one infected individual. The thresholds for both of these cases occur at those values of the basic reproduction ratio R_0 for which the expected time to extinction equals some fixed value. The corresponding function of N gives the persistence threshold when the initial distribution equals the quasi-stationary distribution, and the invasion threshold when initially we have one infected individual. In the following we shall be concerned with the persistence threshold only.

The distribution of the time to extinction τ can be determined if we can solve the Kolmogorov forward equations (2.1) for $p_{mn}(t)$. The reason is that the event that τ exceeds t is equal to the event that the number of infected individuals at time t is positive. By considering the complementary events we obtain

$$P\{\tau \leqslant t\} = P\{I(t) = 0\} = p_{\cdot 0}(t).$$

Thus the cumulative distribution function of the time to extinction at time t equals the marginal probability that the number of infected individuals at time t equals 0.

The distribution of the time to extinction is especially simple when the initial distribution is

equal to the quasi-stationary distribution. To be explicit, let us denote the time to extinction from quasi-stationarity by τ_Q. We show that τ_Q has an exponential distribution and that its expected value is equal to

$$E(\tau_Q) = 1/\mu\alpha\, q_{\cdot 1}. \tag{2.8}$$

To derive this result, we put $q'_{mn}(t) = 0$ in equation (2.5). Thus we are led to the initial value problems

$$p'_{mn} = -\mu\alpha\, q_{\cdot 1} p_{mn}, \qquad p_{mn}(0) = q_{mn}, \qquad m = 0, 1, \ldots, \quad n = 1, 2, \ldots,$$

with solutions

$$p_{mn}(t) = q_{mn} \exp(-\mu\alpha\, q_{\cdot 1} t), \qquad m = 0, 1, \ldots, \quad n = 1, 2 \ldots.$$

By adding these expressions for $p_{mn}(t)$ over all m we obtain

$$p_{\cdot n}(t) = q_{\cdot n} \exp(-\mu\alpha\, q_{\cdot 1} t), \qquad n = 1, 2, \ldots.$$

The differential equation for $p_{\cdot 0}(t)$ in equation (2.3) can now be solved since the right-hand side of this equation is known from above. Recalling that we have the initial value $p_{\cdot 0}(0) = 0$ we obtain $p_{\cdot 0}(t) = 1 - \exp(-\mu\alpha\, q_{\cdot 1} t)$. This establishes the claim that τ_Q has an exponential distribution with expected value given by equation (2.8).

2.5. Diffusion approximation

In this subsection we derive a diffusion approximation for the process formulated in Section 2.1 with $R_0 > 1$. The diffusion approximation has continuous state space, in contrast with the discrete state space of the original process. The main result is that the quasi-stationary distribution is approximated by a bivariate normal distribution if N is sufficiently large. The method is based on work by Kurtz (1971) and Barbour (1972, 1976). A text-book treatment is given by Gardiner (1985).

We rescale the state space by putting $x_1 = S/N$ and $x_2 = I/N$ and consider the bivariate process $x(t) = (x_1(t),\ x_2(t))$. The critical point of the rescaled deterministic model that corresponds to an endemic infection level for $R_0 > 1$ is denoted $\hat{x} = (\hat{x}_1,\ \hat{x}_2) = (1/R_0,\ (R_0 - 1)/\alpha R_0)$.

The changes in the scaled state variables x_1 and x_2 during the time interval from t to $t + \Delta t$ are denoted Δx_1 and Δx_2: $\Delta x_i = x_i(t + \Delta t) - x_i(t)$, $i = 1, 2$. From the hypotheses for the original process we determine the mean and the covariance of the vector with components Δx_1 and Δx_2. We begin with the mean:

$$E\begin{pmatrix} \Delta x_1 \\ \Delta x_2 \end{pmatrix} = \mu\begin{pmatrix} 1 - \alpha R_0 x_1 x_2 - x_1 \\ \alpha R_0 x_1 x_2 - \alpha x_2 \end{pmatrix} \Delta t + o(\Delta t) = b(x)\, \Delta t + o(\Delta t).$$

The Jacobian matrix of the vector $b(x)$ with respect to x is denoted $B(x)$:

$$B(x) = \frac{\partial b(x)}{\partial x} = \mu\begin{pmatrix} -\alpha R_0 x_2 - 1 & -\alpha R_0 x_1 \\ \alpha R_0 x_2 & \alpha R_0 x_1 - \alpha \end{pmatrix}.$$

We approximate the matrix $B(x)$ by evaluating it at the deterministic critical point $\hat{x} = (1/R_0,\ (R_0 - 1)/\alpha R_0)$:

$$B(\hat{x}) = \mu \begin{pmatrix} -R_0 & -\alpha \\ R_0 - 1 & 0 \end{pmatrix}.$$

Next we determine the covariance matrix of the vector of changes in the state variables during the time interval $(t, t + \Delta t)$:

$$\operatorname{cov}\begin{pmatrix} \Delta x_1 \\ \Delta x_2 \end{pmatrix} = \frac{\mu}{N} \begin{pmatrix} 1 + \alpha R_0 x_1 x_2 + x_1 & -\alpha R_0 x_1 x_2 \\ -\alpha R_0 x_1 x_2 & \alpha R_0 x_1 x_2 + \alpha x_2 \end{pmatrix} \Delta t + o(\Delta t)$$

$$= \frac{S(x)}{N} \Delta t + o(\Delta t).$$

The matrix $S(x)$ is approximated by evaluating it at the critical point $\hat{x}$ corresponding to the endemic infection level:

$$S(\hat{x}) = \frac{\mu}{R_0} \begin{pmatrix} 2R_0 & -(R_0 - 1) \\ -(R_0 - 1) & 2(R_0 - 1) \end{pmatrix}.$$

The process $N^{1/2}\{x(t) - \hat{x}\}$ is approximated for large N by a bivariate Ornstein–Uhlenbeck process with local drift matrix $B(\hat{x})$ and local covariance matrix $S(\hat{x})$. Its stationary distribution approximates the quasi-stationary distribution. It is approximately bivariate normal with mean 0 and covariance matrix Σ, where the matrix Σ is determined from the matrices $B(\hat{x})$ and $S(\hat{x})$ through the relationship

$$B(\hat{x})\Sigma + \Sigma B^{\mathrm{T}}(\hat{x}) = -S(\hat{x}),$$

where the superscript T is used to denote transpose. By solving this equation with the above expressions for $B(\hat{x})$ and $S(\hat{x})$ we obtain

$$\Sigma = \frac{1}{R_0^2} \begin{pmatrix} \alpha + R_0 & -R_0 \\ -R_0 & R_0 - 1 + R_0^2/\alpha \end{pmatrix}.$$

2.6. Approximation for the quasi-stationary distribution

We conclude from the above result that the marginal distributions of the number of susceptible and infected individuals in quasi-stationarity are approximately normal with means μ_S and μ_I respectively and standard deviations σ_S and σ_I respectively, where

$$\begin{aligned} \mu_S &= \frac{N}{R_0}, \\ \sigma_S &= \frac{\sqrt{N}}{R_0} \sqrt{(\alpha + R_0)}, \end{aligned} \tag{2.9}$$

$$\begin{aligned} \mu_I &= N \frac{R_0 - 1}{\alpha R_0}, \\ \sigma_I &= \frac{\sqrt{N}}{R_0} \sqrt{(R_0 - 1 + R_0^2/\alpha)}. \end{aligned} \tag{2.10}$$

Furthermore, the covariance is approximated by

$$\sigma_{SI} = -\frac{N}{R_0}. \tag{2.11}$$

The model results are of particular interest when α, the ratio of the average life length to the average length of infection, is large. The expressions for σ_S and σ_I can then be simplified. We obtain

$$\sigma_S \approx \tilde{\sigma}_S = \frac{\sqrt{N}}{R_0}\sqrt{\alpha}$$

and

$$\sigma_I \approx \tilde{\sigma}_I = \frac{\sqrt{N}}{R_0}\sqrt{(R_0 - 1)},$$

given that

$$\alpha \gg \frac{R_0^2}{R_0 - 1}.$$

The coefficients of variation of the distributions of susceptible and infected individuals are therefore approximated by the expressions

$$\mathrm{CV}(S) \approx \frac{\tilde{\sigma}_S}{\mu_S} = \frac{\sqrt{\alpha}}{\sqrt{N}}$$

and

$$\mathrm{CV}(I) \approx \frac{\tilde{\sigma}_I}{\mu_I} = \frac{\alpha}{\sqrt{N}\sqrt{(R_0 - 1)}},$$

given that

$$\alpha \gg \frac{R_0^2}{R_0 - 1}.$$

These relationships show that the coefficient of variation of infected individuals is larger than the coefficient of variation of susceptible individuals under our assumed inequality $\alpha \gg R_0^2/(R_0 - 1)$, since then α is much larger than $R_0 - 1$.

The approximating expression for the coefficient of variation of I coincides with an expression given by Schenzle and Dietz (1987) if we adopt our notation for the derived parameters. They dealt with measles and assumed an average length of infection of $1/(\gamma + \mu) = 12$ days and an average life length of $1/\mu = 70$ years. Hence, $\alpha = 70 \times 365/12 = 2129$. Furthermore, they put $R_0 = 15$. With a population size as large as $N = 10^6$ they found from this formula that the coefficient of variation of I is equal to 0.57. This result prompted them to the insightful comment that we can expect large variations in the prevalence of measles even in very large communities.

The asymptotic analysis of the SIS model in Nåsell (1996) identifies three regions in parameter space with qualitatively different behaviours of the quasi-stationary distribution. The identification of the regions is based on the reparameterization $\rho = N^{1/2}(R_0 - 1)$. Asymptotic analysis is carried out by requiring $R_0 = O(1)$ with $R_0 > 1$ in the first of the regions, $\rho = O(1)$ in the second region and $R_0 = O(1)$ with $R_0 < 1$ in the third region, all as $N \to \infty$. The three parameter regions for the SIS model are described by saying that R_0 is distinctly above or below the deterministic threshold value 1 in the first or the third region respectively and that R_0 lies in the transition region close to the deterministic threshold value

1 in the second region. The quasi-stationary distribution is approximated by a normal distribution in the first region and by a geometric distribution in the third region, whereas the behaviour is more complicated in the transition region.

The relationship between R_0 and ρ allows us to judge which parameter region any given community belongs to. The formal requirement $\rho = O(1)$ can for practical purposes be replaced by $|\rho| < 3$. (This choice of interval for ρ is related to the fact that a normally distributed random variable takes values within ± 3 standard deviations of its mean with high probability.) Thus, any community described by the SIS model lies in the transition region if $|\rho| = |N^{1/2}(R_0 - 1)| < 3$. The transition region for the SIS model is narrow. For example, if $N = 10\,000$, then the community lies in the transition region only if $0.97 < R_0 < 1.03$.

Our investigation will be limited to $R_0 > 1$. This is no limitation for the goal of our study, since the persistence threshold always occurs for $R_0 > 1$. In this parameter range there are two qualitatively different behaviours of the quasi-stationary distribution, and two corresponding parameter regions. We shall refer to the two regions as 'R_0 is distinctly large' and 'R_0 is in the transition region' respectively. This means that we avoid the wording used for the SIS model, where the transition region is described as being close to the deterministic threshold value 1. The reason is that the transition region for the model studied in this section is very wide; R_0 can belong to the transition region without having its value close to 1.

The reparameterization that is necessary to describe the two parameter regions is found by putting the inverse of our approximation for the coefficient of variation of I, $\tilde{\sigma}_I/\mu_I$, equal to ρ. This gives $\sqrt{N}\sqrt{(R_0 - 1)/\alpha} = \rho$. We are thus led to the relationship

$$R_0 = 1 + \rho^2/M,$$

where $M = N/\alpha^2$. Note that the new parameter M accounts for the combined influence of the two parameters N and α on this rescaling of R_0.

We show that the transition region for the model in this section is wide. As for the SIS model, it is determined by the requirement $|\rho| < 3$, which translates into $M(R_0 - 1) < 9$. With $\alpha = 500$ and $N = 10\,000$ we find that the transition region contains the interval $1 < R_0 < 226$.

We recall that the diffusion approximation led to the conclusion that the marginal distribution of the number of infected individuals I in quasi-stationarity is approximately normal. We modify the approximating normal distribution by truncation at 0.5 to achieve consistency with the fact that I is positive. Thus we find that this distribution can be approximated by the expression

$$q_{\cdot n} \approx \frac{1}{\sigma_I} \frac{\varphi\{(n - \mu_I)/\sigma_I\}}{\Phi\{(\mu_I - 0.5)/\sigma_I\}}, \tag{2.12}$$

where Φ and φ denote the normal cumulative distribution function and the normal density function respectively.

Numerical evaluations were undertaken to illustrate the approximations that we present. Following the ideas of Eichner *et al.* (1996) we used Monte Carlo simulations. The programming of the simulations was done in Fortran, whereas some of the analysis and all of the graphing utilized MATLAB®. Both programs were run on a personal computer. The random number generator used for the simulations was the minimal standard generator of Park and Miller with Bays–Durham shuffle; see Press *et al.* (1992).

The simulations were used to study the process starting from the quasi-stationary distribution. All simulations started from the deterministic equilibrium of the model and were allowed to run for an initial time period before data collection began. The initial time period

was longer for smaller values of ρ; it was judged to be sufficiently long if the observed distribution of time to extinction was close to an exponential distribution. The marginal distributions of S and I in quasi-stationarity, the two conditional expectations of either as a function of the other and the distribution of the time to extinction are examples of data that were gathered during the simulations.

Two different sets of simulations were run. In both cases we used $R_0 = 10$ and $\alpha = 500$. These parameter values are close to a set of values used in simulation studies of the persistence of polio by Eichner *et al.* (1996). Two different population sizes were used, namely $N = 800000$ and $N = 50000$. The corresponding values of ρ were 5.37 and 1.34 respectively. In the first of these cases, R_0 is distinctly large, whereas in the second case it is in the transition region.

Comparisons between the simulation results for the marginal distributions of the number of susceptible individuals S and the number of infected individuals I in quasi-stationarity show the following: the distributions of S agree reasonably well with the normal distributions predicted by the diffusion approximation in both cases. The distribution of I is close to approximation (2.12) when R_0 is distinctly large, at least in the body of the distribution, but it deviates somewhat more from this approximation when R_0 is in the transition region. An illustration of the latter case is given in Fig. 1.

2.7. Approximation for the time to extinction

We present two approximations for the expected time to extinction from quasi-stationarity, $E(\tau_Q)$. The first, denoted $E_1(\tau_Q)$, is based on our approximation of the quasi-stationary

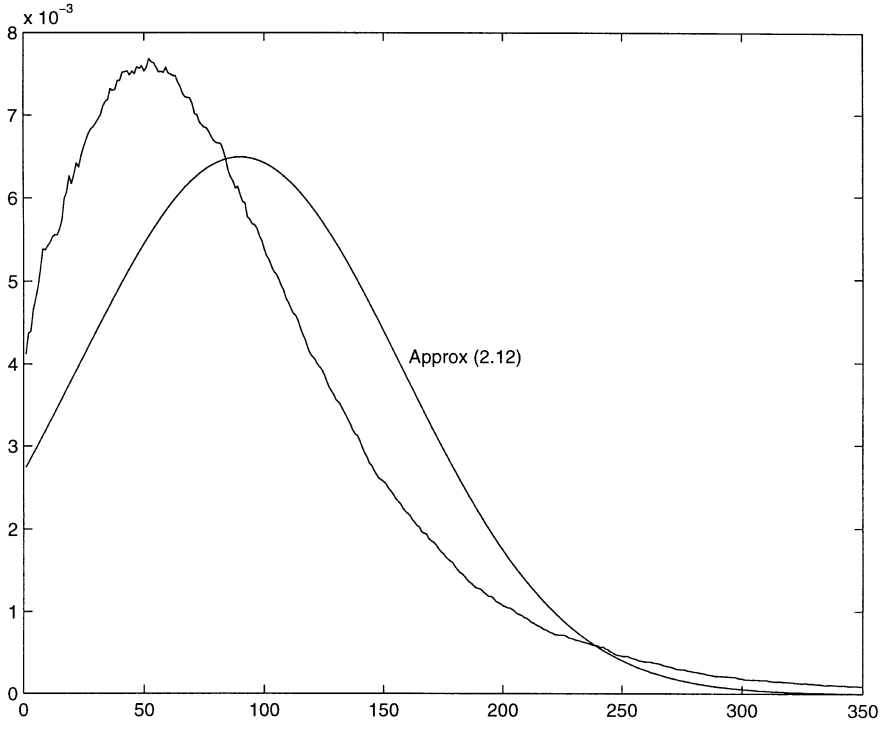

Fig. 1. Simulated marginal distribution of the number of infected individuals in quasi-stationarity, together with approximation (2.12) ($N = 50000$, $R_0 = 10$, $\alpha = 500$ and hence $\rho = 1.34$)

distribution, whereas the second, $E_2(\tau_Q)$, is a further approximation to the first. By inserting the approximation to $q_{\cdot 1}$ from expression (2.12) into the relationship (2.8) that expresses $E(\tau_Q)$ in terms of $q_{\cdot 1}$, we find that

$$E(\tau_Q) \approx E_1(\tau_Q) = \frac{\sigma_I}{\mu\alpha} \frac{\Phi\{(\mu_I - 0.5)/\sigma_I\}}{\varphi\{(\mu_I - 1)/\sigma_I\}}. \tag{2.13}$$

A further approximation can be derived by first approximating σ_I by $\tilde{\sigma}_I$ and then noting that the arguments of Φ and φ can be approximated by $\mu_I/\tilde{\sigma}_I = \rho$. The result is

$$E(\tau_Q) \approx E_2(\tau_Q) = \frac{\rho}{\mu R_0} \frac{\Phi(\rho)}{\varphi(\rho)}. \tag{2.14}$$

The two approximations are based on different parameter sets. The first depends on the four parameters μ, N, R_0 and α, whereas the second depends on the three parameters μ, M and ρ (since $R_0 = 1 + \rho^2/M$). The parameter μ plays the same role in both cases, since both of the approximations are proportional to $1/\mu$. This means that we can put $\mu = 1$ in the numerical evaluations and interpret the results in terms of a new unit of time which equals the average life length $1/\mu$.

Fig. 2 gives a numerical comparison between the simulation results and the two approximations given above. The two approximations are seen to be quite close. Fig. 2 supports our claim that either of the two approximation formulae gives the correct order of magnitude, but

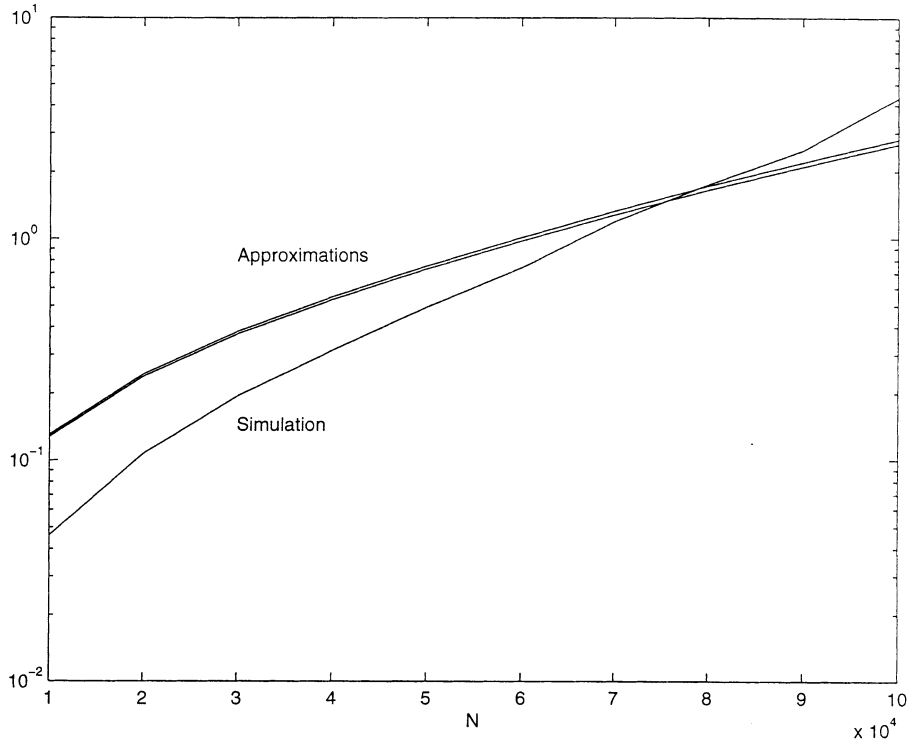

Fig. 2. Expected time to extinction from quasi-stationarity as a function of N: simulation results compared with approximations (2.13) and (2.14) ($R_0 = 10$, $\alpha = 500$, $\mu = 1$ and hence ρ ranges from 0.60 when $N = 10000$ to 1.90 when $N = 100000$)

it also shows that there is a definite need for improvement in the approximations, for both small and large values of N.

The time to extinction has also been analysed recently by van Herwaarden and Grasman (1995), who worked with the time to extinction starting from the endemic equilibrium, whereas our results are based on the assumption that we start from the quasi-stationary distribution. On the basis of the results in Nåsell (1996) we would expect that these two times to extinction are close if R_0 is distinctly large, and that they deviate somewhat when R_0 is in the transition region.

van Herwaarden and Grasman (1995) give a formula for their approximation for the expected time to extinction from the endemic equilibrium. The formula takes the form

$$C(N) \sim \frac{Q_1}{\sqrt{N}} \exp(Q_0 N)$$

when R_0 is distinctly large. We note that in the same parameter region the second of our approximations for the expected time to extinction from the quasi-stationary distribution can be written

$$E_2(\tau_Q) = \frac{\sqrt{\{2\pi(R_0-1)N\}}}{\mu\alpha R_0} \exp\left\{\frac{(R_0-1)N}{2\alpha^2}\right\}. \tag{2.15}$$

A comparison between these two expressions shows that they are not of the same order as $N \to \infty$.

The quantities Q_1 and Q_0 in the formula for $C(N)$ are functions of the parameters R_0, α and μ. No expression is given for Q_1 in van Herwaarden and Grasman (1995). To determine Q_0 they described a numerical procedure that requires the use of a shooting method to determine a particular solution to an autonomous system of four ordinary differential equations. Comparisons with simulation results show that the procedure gives good agreement for the parameter values $R_0 = 2$, $\alpha = 2$ and $\mu = 0.2$. However, the numerical procedure appears to break down in the interesting case where the ratio α between the life length and the length of the infectious period is large; we found it impossible to determine the particular solution to the system of differential equations in a test case with $R_0 = 10$ and $\alpha = 500$.

2.8. *Approximation for the critical community size and for the persistence threshold*

The expected time to extinction from quasi-stationarity, $E(\tau_Q)$, is an increasing function of the population size N. The critical community size N_C can be defined as that value of N for which $E(\tau_Q)$ takes some fixed value. We study the situation where $E(\tau_Q)$ takes the fixed value K/μ. The critical community size N_C is therefore determined implicitly by the relationship

$$E(\tau_Q) = K/\mu. \tag{2.16}$$

We conclude that N_C is a function of the three parameters R_0, α and K. Equivalently, the persistence threshold R_{0P} is that value of R_0 for which $E(\tau_Q)$ takes the fixed value K/μ. The persistence threshold R_{0P} is a function of N, α and K. The critical community size N_C and the persistence threshold R_{0P} are related in the sense that when α and K are fixed then the functions $N_C(R_0)$ and $R_{0P}(N)$ are the inverses of each other.

We derive two approximations for the critical community size (or, equivalently, for the persistence threshold) by inserting our two approximations for $E(\tau_Q)$ in the defining relationship (2.16). By using the approximation $E(\tau_Q) \approx E_1(\tau_Q)$ we obtain

$$\frac{\sigma_I}{\alpha}\frac{\Phi\{(\mu_I - 0.5)/\sigma_I\}}{\varphi\{(\mu_I - 1)/\sigma_I\}} = K. \tag{2.17}$$

Similarly, by replacing $E(\tau_Q)$ by the approximation $E_2(\tau_Q)$ we obtain

$$\frac{\rho}{1 + \rho^2/M}\frac{\Phi(\rho)}{\varphi(\rho)} = K. \tag{2.18}$$

The second of these relationships defines implicitly an approximation to the critical value ρ_C of ρ. This approximation is denoted ρ_{C2}. It is a function of M and K. The functional values $\rho_{C2}(M, K)$ can be determined numerically. We use the reparameterization in terms of ρ and M also in the study of the first of these two relationships. This relationship determines implicitly an approximation ρ_{C1} to the critical value ρ_C of ρ. This approximation depends on three parameters, namely M, K and α, just as the critical value ρ_C does itself. As above, the functional values $\rho_{C1}(M, K, \alpha)$ can be determined numerically.

The persistence threshold R_{0P} is related to the critical value of ρ through the relationship

$$R_{0P}(N, K, \alpha) = 1 + \frac{\alpha^2}{N}\rho_C^2\left(\frac{N}{\alpha^2}, K, \alpha\right).$$

By inserting the two approximations for ρ_C into the right-hand side of this expression we are led to two corresponding approximations for the persistence threshold. Finally, two corresponding approximations for the critical community size N_C are found from the fact

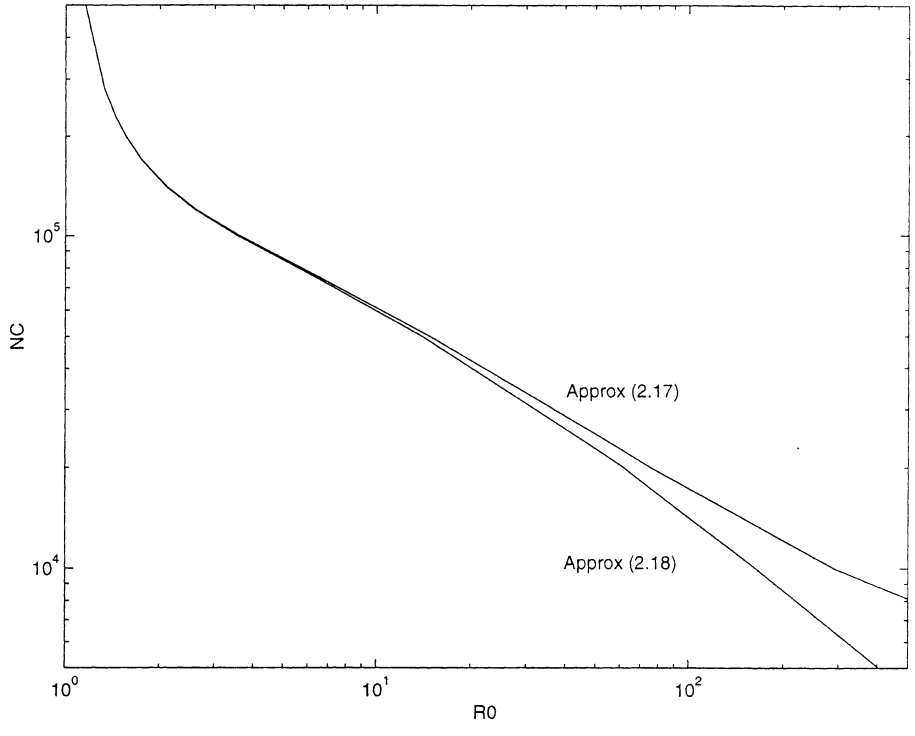

Fig. 3. Two approximations to the critical community size N_C as a function of R_0 for $K = 1$ and $\alpha = 500$

that the function $N_C(R_0)$ is the inverse of the function $R_{0P}(N)$ for fixed values of K and α. Fig. 3 gives the results of numerical evaluations of the two approximations for the critical community size with $\alpha = 500$ and $K = 1$. The two approximations are shown as functions of R_0 over the interval $1 < R_0 < 500$. They deviate appreciably only for large values of R_0.

3. The Bartlett model

The model in this section differs from the model of the previous section only in the sense that no deaths of individuals are assumed to occur.

A deterministic version of this model was first formulated by Hamer (1906) and Soper (1929). Bartlett (1956, 1957, 1960a, b, c) was the first to realize that an understanding of both recurrence and extinction could be provided through stochastic modelling. Further work on the stochastic version of the model has been published by Stirzaker (1975), Ridler-Rowe (1967), Pollett and Stewart (1994) and Stewart and Bebbington (1996).

The analysis of the model runs closely parallel to the analysis of the Martini model of the previous section.

3.1. Model formulation

The hypotheses of the Bartlett model are set down by specifying the transition rates in Table 2. We equip the transition rates with a subscript B (for Bartlett) to distinguish them from the transition rates of the preceding section.

3.2. The deterministic model and reparameterization

The deterministic version of the model leads to the differential equations

$$S' = \mu N - \frac{\beta}{N} SI,$$
$$I' = \frac{\beta}{N} SI - \gamma I.$$

We reparameterize by introducing R_B and α_B as follows:

$$\begin{aligned} R_B &= \beta/\gamma, \\ \alpha_B &= \gamma/\mu. \end{aligned} \tag{3.1}$$

The system of differential equations has one critical point only, with co-ordinates $S = N/R_B$ and $I = N/\alpha_B$. It is asymptotically stable for all positive values of the parameters. Hence the deterministic version of the Bartlett model has no threshold. This means that there is no parameter that can be referred to as the basic reproduction ratio. The parameter R_B is a substitute for R_0. Even though the deterministic model has no threshold, we expect the stochastic version of the model to have threshold behaviour for finite values of N.

Table 2. Hypothesized transition rates for the Bartlett model

Event	*Transition*	*Transition rate*
Immigration of susceptible individual	$(m, n) \to (m+1, n)$	$\lambda_{1B}(m, n) = \mu N$
Infection of susceptible individual	$(m, n) \to (m-1, n+1)$	$\nu_{2B}(m, n) = \beta mn/N$
Recovery of infected individual	$(m, n) \to (m, n-1)$	$\mu_{2B}(m, n) = \gamma n$

3.3. The quasi-stationary distribution

The main result in this section is that the marginal distribution $q_{\cdot n}$ of the number of infected individuals in quasi-stationarity satisfies the recursion relationship

$$q_{\cdot n+1} = \frac{R_B}{N}\frac{n}{n+1}\, e_S(n) q_{\cdot n} + \frac{1 - \sum_{k=1}^{n} q_{\cdot k}}{n+1}\, q_{\cdot 1}, \qquad n = 1, 2, \ldots . \tag{3.2}$$

As mentioned in Section 1 the concept of quasi-stationarity for continuous time Markov chains was introduced by Darroch and Seneta (1967). The concept was therefore unknown to Bartlett at the time when he carried out his studies. Bartlett used a closely related concept, namely the stationary distribution (let us call it r_{mn}) of the model modified with a small rate of immigration of infected individuals. He derived an approximate recursion relationship for the marginal distribution $r_{\cdot n}$ of infected individuals in this stationary distribution. It takes the form

$$r_{\cdot n+1} \approx \frac{R_B}{N}\frac{n}{n+1}\, e_S(n) r_{\cdot n}, \qquad n = 1, 2, \ldots,$$

and is clearly different from the recursion relationship (3.2) for the marginal distribution $q_{\cdot n}$.

3.4. The distribution of the time to extinction

Our main result in this subsection is that the time to extinction from quasi-stationarity τ_Q has an exponential distribution with expectation determined by

$$E(\tau_Q) = 1/\mu\alpha_B q_{\cdot 1}. \tag{3.3}$$

Bartlett's counterpart to the random variable 'time to extinction' is 'recurrence time to zero infectives'. He derived an approximate relationship between the mean recurrence time to zero infectives, R, and the probability $r_{\cdot 1}$. It takes the form

$$R \approx 1/\mu\alpha_B r_{\cdot 1}.$$

This relationship can be seen to be a close counterpart to our relationship (3.3).

3.5. Diffusion approximation

The main result in this section is that the covariance matrix Σ of the stationary distribution of the approximating Ornstein–Uhlenbeck process is

$$\Sigma = \frac{1}{R_B^2}\begin{pmatrix} \alpha_B + R_B & -R_B \\ -R_B & R_B + R_B^2/\alpha_B \end{pmatrix}.$$

This agrees with the results reported by Barbour (1976) and by Pollett and Stewart (1994).

The methods and results for the diffusion approximation to discrete state space Markov chains were not developed at the time when Bartlett published his results. It is therefore of considerable interest to note that one part of the study undertaken by Bartlett was directed at a linearization of the bivariate stochastic model about the deterministic critical point. This study allowed Bartlett to conclude that the covariance matrix of the process can be approximated with exactly the same covariance matrix Σ that we have given above as a result of applying the standard diffusion approximation. Bartlett did not draw the conclusion that the bivariate quasi-stationary distribution is approximately normal.

3.6. Approximation for the quasi-stationary distribution

We conclude that the marginal distributions of the number of susceptible and infected individuals in quasi-stationarity are approximately normal with means μ_{SB} and μ_{IB} respectively and standard deviations σ_{SB} and σ_{IB} respectively, where

$$\mu_{SB} = \frac{N}{R_B},$$

$$\sigma_{SB} = \frac{\sqrt{N}}{R_B}\sqrt{(\alpha_B + R_B)},$$

$$\mu_{IB} = \frac{N}{\alpha_B},$$

$$\sigma_{IB} = \frac{\sqrt{N}}{R_B}\sqrt{(R_B + R_B^2/\alpha_B)}.$$

Furthermore, the covariance is approximated by

$$\sigma_{SI} = -\frac{N}{R_B}. \tag{3.4}$$

We obtain the following approximations for the standard deviations σ_{SB} and $\sigma_{IB}$when α_B is much larger than R_B:

$$\sigma_{SB} \approx \tilde{\sigma}_{SB} = \frac{\sqrt{N}}{R_B}\sqrt{\alpha_B}$$

and

$$\sigma_{IB} \approx \tilde{\sigma}_{IB} = \sqrt{\left(\frac{N}{R_B}\right)},$$

when

$$\alpha_B \gg R_B.$$

The coefficients of variation of the distributions of infected and susceptible individuals are then approximated as

$$\mathrm{CV}(S) \approx \frac{\tilde{\sigma}_{SB}}{\mu_{SB}} = \frac{\sqrt{\alpha_B}}{\sqrt{N}}$$

and

$$\mathrm{CV}(I) \approx \frac{\tilde{\sigma}_{IB}}{\mu_{IB}} = \frac{\alpha_B}{\sqrt{(R_B N)}},$$

when

$$\alpha_B \gg R_B.$$

These relationships show that the coefficient of variation for infected individuals is larger than the coefficient of variation for susceptible individuals when the assumed inequality $\alpha_B \gg R_B$ holds. As was the case for the main model, we truncate the approximation of the

marginal distribution of infected individuals in quasi-stationarity. The resulting approximation can be expressed as

$$q_{\cdot n} \approx \frac{1}{\sigma_{IB}} \frac{\varphi\{(n-\mu_{IB})/\sigma_{IB}\}}{\Phi\{(\mu_{IB}-0.5)/\sigma_{IB}\}}. \tag{3.5}$$

We rescale the parameters of the Bartlett model by introducing ρ_B and M_B to satisfy the relationships

$$R_B = \rho_B^2/M_B,$$
$$M_B = N/\alpha_B^2.$$

Bartlett derived an approximation for the marginal stationary distribution $r_{\cdot n}$. The result can be expressed in the form

$$r_{\cdot n} \approx \frac{N}{n\alpha_B} \frac{1}{\sigma_{IB}} \varphi\left(\frac{n-\mu_B}{\sigma_{IB}}\right), \tag{3.6}$$

where

$$\mu_B = \mu_{IB} + \frac{\alpha_B}{R_B} + 1.$$

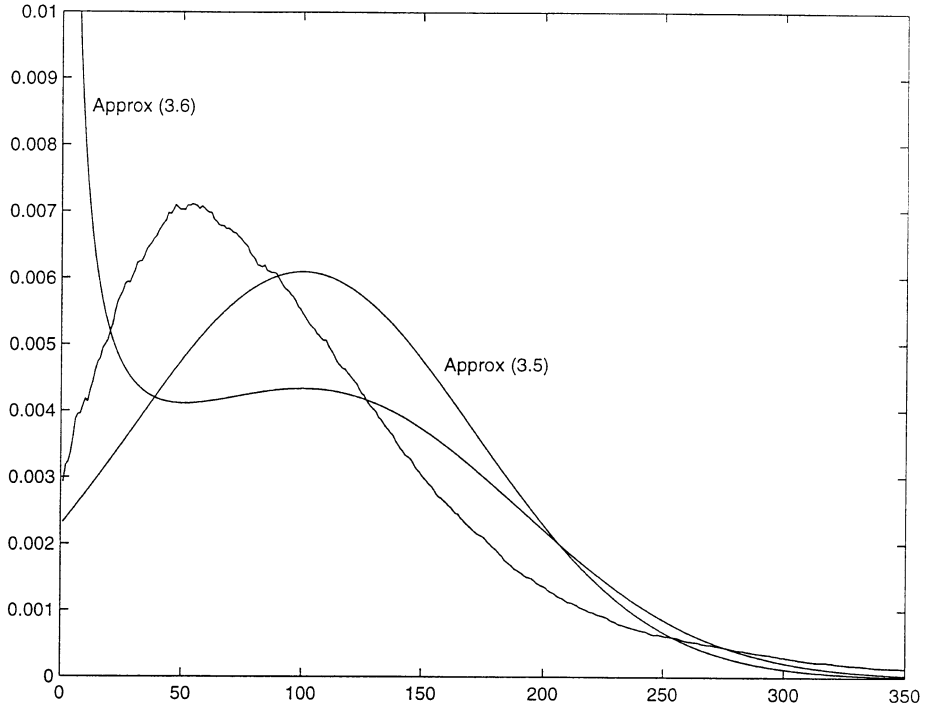

Fig. 4. Marginal distribution of the number of infected individuals in quasi-stationarity in Bartlett's model determined by simulation, together with approximations (3.5) and (3.6) ($N = 50000$, $R_B = 10$, $\alpha_B = 500$ and hence $\rho_B = 1.41$)

A simulation study of the Bartlett model shows that the marginal distribution of the number of infected individuals in quasi-stationarity is rather well approximated by both approximations (3.5) and (3.6) in the body of the distribution when R_B is distinctly large. The situation is different in the transition region. A comparison of the simulation results with the two approximations when R_B lies in the transition region is shown in Fig. 4. Here, the Bartlett approximation is seen to deviate rather strongly from the simulation result in the important left-hand tail of the distribution.

The reason for the disagreement between the Bartlett approximation and the simulation result lies in Bartlett's choice of approximation for the quasi-stationary distribution. Bartlett modified his process by allowing a small rate of immigration of infected individuals. Several alternative modifications are available. One would be to allow for one individual in the population to be immortal and permanently infected. Another would be to remove the possibility of transitions from the state $(m, 1)$ to the state $(m, 0)$. All three of the modified processes are free from absorbing states, so they have stationary distributions. Heuristic arguments lead to the conjecture that all three of these stationary distributions are good approximations for the quasi-stationary distribution. This conjecture is correct when R_B is distinctly large, but not in the transition region. Here, subtle differences appear that can only be discovered by careful analysis.

Numerical methods for the evaluation of joint quasi-stationary distributions are discussed by Pollett and Stewart (1994) and by Stewart and Bebbington (1996). It is not clear whether these methods can be used for the large values of N and α_B that are used for example in Fig. 4.

3.7. Approximation for the time to extinction

Similarly to the case for the main model in the previous section we derive two approximations for the expected time to extinction from quasi-stationarity. They can be written as

$$E(\tau_Q) \approx E_1(\tau_Q) = \frac{\sigma_{IB}}{\mu\alpha_B} \frac{\Phi\{(\mu_{IB} - 0.5)/\sigma_{IB}\}}{\varphi\{(\mu_{IB} - 1)/\sigma_{IB}\}}, \tag{3.7}$$

and

$$E(\tau_Q) \approx E_2(\tau_Q) = \frac{\rho_B}{\mu R_B} \frac{\Phi(\rho_B)}{\varphi(\rho_B)}. \tag{3.8}$$

Bartlett derived an approximation for the expected recurrence time to zero infectives, R. His result can be expressed as

$$R \approx \frac{\sigma_{IB}}{\mu N} \frac{1}{\varphi\{(\mu_B - 1)/\sigma_{IB}\}}. \tag{3.9}$$

A different approximation for the time to extinction is presented in Stirzaker (1975). He approximated the time dependence of the scaled realization of the process as a perturbation of a periodic solution $A\cos(\omega t + \psi)$. The time to extinction is studied from the initial amplitude A_0. For $A_0 = 0$ this initial condition is equivalent to starting in the deterministic equilibrium.

The expected time to extinction from the initial amplitude A_0 is approximated by the expression

$$M_1(A_0) = \frac{1}{n} \int_{A_0}^{1} \int_{0}^{t} \exp\left\{ -\frac{m(s^2 - t^2)}{2n} \right\} \mathrm{d}s\, \mathrm{d}t,$$

where, in our notation,

$$m = \frac{1}{2}\,\mu R_B,$$

$$n = \frac{\mu\alpha_B}{2N}\,(\alpha_B + R_B).$$

It follows that $M_1(A_0)$ can be expressed in the form

$$M_1(A_0) = \frac{2\rho_{B1}}{\mu R_B}\int_{A_0}^{1} \frac{\Phi(\rho_{B1}t) - 0.5}{\phi(\rho_{B1}t)}\,\mathrm{d}t,$$

where

$$\rho_{B1} = \sqrt{\left(\frac{m}{n}\right)} = \frac{\mu_{IB}}{\sigma_{IB}} = \frac{\sqrt{(R_B N)}}{\sqrt{\{\alpha_B(\alpha_B + R_B)\}}}$$

is the inverse of the coefficient of variation of I, without introducing any approximation for large α. It is of interest to compare this result with our approximation $E_2(\tau_Q)$. We note therefore first that

$$\int_0^1 \frac{\Phi(\rho y) - 0.5}{\phi(\rho y)}\,\mathrm{d}y \sim \frac{1}{2\rho^2\,\phi(\rho)}, \qquad \rho \to \infty.$$

Thus, Stirzaker's expression for $M_1(0)$ can be approximated as follows when R_B is distinctly large:

$$M_1(0) \sim \frac{1}{\mu R_B \rho_{B1}\,\phi(\rho_{B1})}, \qquad \rho \to \infty. \tag{3.10}$$

Since $\rho_{B1} = O(\sqrt{N})$, we find from this result that the N-dependence of $M_1(0)$ has the same form as the N-dependence for $C(N)$ given by van Herwaarden and Grasman (1995) in Section 2.7, and that this N-dependence differs from that of our approximation $E_2(\tau_Q)$.

A different analysis of the same model was undertaken by Ridler-Rowe (1967). It follows from his results that the expected time to extinction, starting at the deterministic equilibrium $(N/R_B,\ N/\alpha_B)$, is given by

$$\tau(m, n) \sim \frac{1}{\mu\alpha_B}\log\left(\frac{N}{R_B} + \frac{N}{\alpha_B}\right), \qquad N \to \infty. \tag{3.11}$$

This approximation deviates substantially in form from the several approximations given above; it behaves like $O\{\log(N)\}$ for all positive values of R_B and α_B, whereas all the other approximations above for the expected time to extinction grow exponentially with N for R_B sufficiently large. The explanation for this difference in behaviour is that Ridler-Rowe's model is based on different hypotheses. On p. 22 of Ridler-Rowe (1967), he uses the (implicit) assumption that his parameter $\lambda = \beta/N$ is independent of N. This is equivalent to the unrealistic assumption that the contact rate β is proportional to the population size N.

The five approximations for the expected time to extinction that are available for the Bartlett model are compared with simulations in three cases in Table 3. In all three cases we use the parameter values $R_B = 10$, $\alpha_B = 500$ and $\mu = 1$, whereas the population size N takes the three values 20000, 50000 and 100000. All three cases are in the transition region, with ρ-values 0.89, 1.41 and 2 respectively. The table entries may be compared with the results

Table 3. Expected time to extinction from quasi-stationarity for the Bartlett model determined by simulation and compared with five explicit approximations†

N	*Simulation*	$E_1(\tau_Q)$	$E_2(\tau_Q)$	*Bartlett approximation*	*Stirzaker approximation*	*Ridler-Rowe approximation*
20000	0.12	0.27	0.27	0.041	0.090	0.015
50000	0.67	0.86	0.89	0.032	0.28	0.017
100000	5.7	3.4	3.6	0.054	0.87	0.018

†The approximations $E_1(\tau_Q)$ and $E_2(\tau_Q)$ are given in equations (3.7) and (3.8) respectively. The Bartlett approximation is given in equation (3.9), the Stirzaker approximation in equation (3.10) and the Ridler-Rowe approximation in equation (3.11). The parameters are $R_B = 10$, $\alpha_B = 500$ and $\mu = 1$.

in Fig. 2 for the closely related model of Section 2. Table 3 shows that none of the five approximations is very accurate. Both of the approximations $E_1(\tau_Q)$ and $E_2(\tau_Q)$ overestimate the time to extinction for $N = 20000$ and $N = 50000$ and underestimate it for $N = 100000$, but they are all of the same order of magnitude as the simulation results. The remaining three approximations give underestimates in all three cases, and several of them are low by one or two orders of magnitude.

4. Concluding comments

Both stochastic and deterministic formulations of endemic SIR models have been studied in the literature. It is well known that the deterministic version of any stochastic model is actually an approximation to it. The question therefore arises whether this approximation is acceptable. The deterministic versions of the fully stochastic endemic SIR models cannot be used to study extinction, since this is a purely stochastic phenomenon. Furthermore, our results show (for the Martini model) that the persistence threshold is orders of magnitude larger than the deterministic threshold $R_0 = 1$ for realistic parameter values. We are led to conclude that the deterministic version of the model gives an unacceptable approximation to the stochastic model.

A deterministic endemic SIR (or SEIR) model with a periodic contact rate shows chaotic behaviour if the variation in the contact rate is sufficiently large. This fact forms the basis for a number of papers that argue that chaos is important for understanding recurrent epidemics. Examples are given by Sugihara and May (1990), Olsen and Schaffer (1990) and Ellner *et al.* (1995). The chaos argument is, however, completely at odds with our result that the deterministic model is an unacceptable approximation to the stochastic model.

The models of Sections 2 and 3 omit several factors that are known to be of importance. Schenzle (1984) described a stochastic model that accounts for age-dependent exposure and seasonality in transmission. The model was analysed by simulation. Keeling and Grenfell (1997) extended the Schenzle model to account for waiting times that are not exponentially distributed. It is a challenge to analyse mathematically the effect of any of these factors on the model properties.

Our comparisons with simulation results show that our approximation to the expected time to extinction from quasi-stationarity is crude. An improvement requires improved approximation for the marginal distribution of infected individuals in quasi-stationarity. When R_0 is distinctly large, an improvement is needed in the left-hand tail of the distribution, and when R_0 lies in the transition region the improvement is needed in the body of the distribution.

The analysis that we have presented is based on the concept of quasi-stationarity, the diffusion approximation result, the result that scaling can be used to study the transition region and the tools of modern computer software and hardware. It is noteworthy that all four components of this basis have been developed since the time when Bartlett undertook his study. Our brief review of Bartlett's work has shown that his approach is rather similar to ours. We can only admire his ability to discover the structure of the problem at a time when the basis for today's analysis was just not available. Bartlett's work on this problem remains impressive even though his approximation for the expected time to extinction fails when R_B lies in the transition region.

Acknowledgements

The ideas in this paper were discussed at the Royal Statistical Society Epidemics Workshop 1997 on the Isle of Skye. I thank Håkan Andersson, Tom Britton, Odo Diekmann, Klaus Dietz, Alun Lloyd and Anders Martin-Löf for good contributions to the discussion. I also thank the referee for careful reading of the manuscript and for valuable comments.

References

Anderson, R. M. and May, R. M. (1991) *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Bailey, N. T. J. (1975) *The Mathematical Theory of Infectious Diseases and Its Applications*. London: Griffin.

Barbour, A. D. (1972) The principle of the diffusion of arbitrary constants. *J. Appl. Probab.*, **9**, 519–541.

———(1976) Quasi-stationary distributions in Markov population processes. *Adv. Appl. Probab.*, **8**, 296–314.

Bartlett, M. S. (1956) Deterministic and stochastic models for recurrent epidemics. In *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, vol. 4, pp. 81–109. Berkeley: University of California Press.

———(1957) Measles periodicity and community size (with discussion). *J. R. Statist. Soc.* A, **120**, 48–70.

———(1960a) *Stochastic Population Models in Ecology and Epidemiology*. London; Methuen.

———(1960b) Some stochastic models in ecology and epidemiology. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (eds I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann), pp. 89–96. Stanford: Stanford University Press.

———(1960c) The critical community size for measles in the United States. *J. R. Statist. Soc.* A, **123**, 37–44.

Darroch, J. N. and Seneta, E. (1967) On quasi-stationary distributions in absorbing continuous-time finite Markov chains. *J. Appl. Probab.*, **4**, 192–196.

Dietz, K. (1975) Transmission and control of arbovirus diseases. In *Epidemiology* (eds D. Ludwig and K. L. Cooke), pp. 104–121. Philadelphia: Society for Industrial and Applied Mathematics.

———(1995) Some problems in the theory of infectious disease transmission and control. In *Epidemic Models: Their Structure and Relation to Data* (ed. D. Mollison), pp. 3–16. Cambridge: Cambridge University Press.

Eichner, M., Hadeler, K. P. and Dietz, K. (1996) Stochastic models for the eradication of poliomyelitis: minimum population size for polio virus persistence. In *Models for Infectious Human Diseases, Their Structure and Relation to Data* (eds V. Isham and G. Medley), pp. 315–327. Cambridge: Cambridge University Press.

Ellner, S., Gallant, A. R. and Theiler, J. (1995) Detecting nonlinearity and chaos in epidemic data. In *Epidemic Models: Their Structure and Relation to Data* (ed. D. Mollison), pp. 229–247. Cambridge: Cambridge University Press.

Gardiner, C. W. (1985) *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Berlin: Springer.

Hamer, W. H. (1906) Epidemic disease in England — the evidence of variability and of persistence of type. *Lancet*, 733–739.

van Herwaarden, O. A. and Grasman, J. (1995) Stochastic epidemics: major outbreaks and the duration of the endemic period. *J. Math. Biol.*, **33**, 581–601.

Hethcote, H. W. (1974) Asymptotic behavior and stability in epidemic models. *Lect. Notes Biomath.*, **2**.

———(1976) Qualitative analysis of communicable disease models. *Math. Biosci.*, **28**, 335–356.

Keeling, M. J. and Grenfell, B. T. (1997) Disease extinction and community size: modelling the persistence of measles. *Science*, **275**, 65–67.

Kryscio, R. J. and Lefèvre, C. (1989) On the extinction of the SIS stochastic logistic epidemic. *J. Appl. Probab.*, **26**, 685–694.

Kurtz, T. G. (1971) Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.*, **8**, 344–356.
Lotka, A. J. (1923) Martini's equations for the epidemiology of immunising diseases. *Nature*, **111**, 633–634.
———(1956) *Elements of Mathematical Biology*. New York: Dover Publications.
Martini, E. (1921) *Berechnungen und Beobachtungen zur Epidemiologie und Bekämpfung der Malaria*. Hamburg: Gente.
Nåsell, I. (1995) The threshold concept in stochastic epidemic and endemic models. In *Epidemic Models: Their Structure and Relation to Data* (ed. D. Mollison), pp. 71–83. Cambridge: Cambridge University Press.
———(1996) The quasi-stationary distribution of the closed endemic SIS model. *Adv. Appl. Probab.*, **28**, 895–932.
Olsen, L. F. and Schaffer, W. M. (1990) Chaos versus noisy periodicity: hypotheses for childhood epidemics. *Science*, **249**, 499–504.
Pollett, P. K. and Stewart, D. E. (1994) An efficient procedure for computing quasi-stationary distributions of Markov chains with sparse transition structure. *Adv. Appl. Probab.*, **26**, 68–79.
Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992) *Numerical Recipes in Fortran: the Art of Scientific Computing*, 2nd edn. Cambridge: Cambridge University Press.
Ridler-Rowe, C. J. (1967) On a stochastic model of an epidemic. *J. Appl. Probab.*, **4**, 19–31.
Schenzle, D. (1984) An age-structured model of pre- and post-vaccination measles transmission. *IMA J. Math. Appl. Biol. Med.*, **1**, 171–178.
Schenzle, D. and Dietz, K. (1987) Critical population sizes for endemic virus transmission. In *Räumliche Persistenz und Diffusion von Krankheiten*. Heidelberg: Heidelberger Geographische Arbeiten.
Soper, H. E. (1929) The interpretation of periodicity in disease prevalence (with discussion). *J. R. Statist. Soc.*, **92**, 34–73.
Stewart, D. E. and Bebbington, M. S. (1996) An iterative aggregation-disaggregation procedure for modelling the long-term behaviour of continuous-time evanescent random processes. *J. Statist. Computn Simuln*, **56**, 77–95.
Stirzaker, D. R. (1975) A perturbation method for the stochastic recurrent epidemic. *J. Inst. Math. Applic.*, **15**, 135–150.
Sugihara, G. and May, R. M. (1990) Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, **344**, 734–741.