J. Math. Biol. 41, 559–580 (2000)
Digital Object Identifier (DOI):
10.1007/s002850000060

Mathematical Biology

Håkan Andersson · Tom Britton

Stochastic epidemics in dynamic populations: quasi-stationarity and extinction

Received: 14 February 2000 / Revised version: 5 June 2000 /
Published online: 24 November 2000 – © Springer-Verlag 2000

Abstract. Empirical evidence shows that childhood diseases persist in large communities whereas in smaller communities the epidemic goes extinct (and is later reintroduced by immigration). The present paper treats a stochastic model describing the spread of an infectious disease giving life-long immunity, in a community where individuals die and new individuals are born. The time to extinction of the disease starting in quasi-stationarity (conditional on non-extinction) is exponentially distributed. As the population size grows the epidemic process converges to a diffusion process. Properties of the limiting diffusion are used to obtain an approximate expression for τ, the mean-parameter in the exponential distribution of the time to extinction for the finite population. The expression is used to study how τ depends on the community size but also on certain properties of the disease/community: the basic reproduction number and the means and variances of the latency period, infectious period and life-length. Effects of introducing a vaccination program are also discussed as is the notion of the critical community size, defined as the size which distinguishes between the two qualitatively different behaviours.

1. Introduction

For several diseases giving life-long immunity, so called childhood diseases, it has been empirically observed that the disease fades out in small communities but may persist in larger communities (e.g. Anderson & May, 1991, p 83 and Keeling & Grenfell, 1997). It is a classical problem in the mathematical theory of infectious diseases to find good models pointing out this phenomenon. An intuitive explanation of the qualitative difference, already given by Bartlett (1956), goes as follows.

The basic reproduction number, R, can loosely be defined as the expected number of new cases generated by one infectious individual in a large susceptible population. The basic reproduction number is thought to reflect both the social activity

H. Andersson: Department of Mathematics, Stockholm University, 106 91 Stockholm, Sweden.
Present address: Group Financial Risk Control, SwedBank, 105 34 Stockholm, Sweden. e-mail: hakan.b.andersson@foreningssparbanken.se

T. Britton (to whom correspondence should be addressed): Department of Mathematics, Uppsala University, P.O. Box 480, 751 06 Uppsala, Sweden. e-mail: tom.britton@math.uu.se

Key words: Critical community size – Diffusion approximation – Persistence – Quasi-stationary distribution – SIR epidemics – Stochastic fade-out – Vaccination

among individuals and the infectiousness of the disease. If $R > 1$, as is assumed in this paper, then we say that the population is above threshold. Introducing an infective to a susceptible community above threshold may, as is well known, lead to a large outbreak. We expect one of two scenarios.

1. If the community is not too large the epidemic will fade out after a relatively short period of time. This happens because a large proportion becomes infected (and cannot be reinfected) and not enough new susceptible individuals are born into the population to keep the epidemic going. However, by births of new susceptibles and deaths of immune individuals, the proportion of susceptibles will slowly grow as time goes by, eventually taking the population above threshold again. Infectives visiting the community may then again initiate a new large outbreak. In this way the recurrent behaviour of the disease is achieved.

2. On the other hand, in a very large community the susceptible population might be augmented fast enough for the epidemic to be maintained for a long time without any introduction of new infectives to the community. This happens because the population is brought above threshold during the final stages of a large outbreak due to a high enough birth rate of new susceptibles.

The notion of *critical community size*, loosely defined as the population size needed for the epidemic to persist over a given time horizon with a given probability, tries to separate between the two situations above. As opposed to many other problems for infectious diseases, deterministic models are not of much use when aiming to derive expressions for the time to extinction, because extinction is always caused by random fluctuations from the expected (or deterministic) curve. Recently, the stochastic model suggested by Bartlett (1956) has been modified slightly to what is called the SIR epidemic process with demography (van Herwaarden & Grasman, 1995, and Nåsell, 1999). In Section 2 of the present paper we generalise this model in several ways. Life-lengths and the duration of the infectious period are now modelled by Γ-distributions. Further a latency period is introduced which is also modelled by a Γ-distribution. In Section 3 we let the population size tend to infinity and show that the epidemic process converges to a diffusion. The stationary distribution of the limiting diffusion is Gaussian and it is investigated how the mean vector and covariance matrix depend on properties of the disease/community. In Section 4 it is shown that in a finite population the time to extinction is exponentially distributed if the process is started in quasi-stationarity (conditional on non-extinction). An approximate expression for τ, the mean parameter of the exponential distribution, is derived in Section 4.3 (equation 8) where we approximate the quasi-stationary distribution by the stationary distribution of the limiting diffusion (cf. Nåsell, 1999, who uses the same ‘linearising’ approach for a different model). In Section 4.4 it is shown how the introduction of a vaccination program may be incorporated in the model simply by changing certain parameters (equation 9). The section ends with a discussion on how the critical community size depends on the disease/community parameters. The qualitative result of the present paper says that, in relevant parameter regions,

The expected time to extinction τ is increasing in: the community size and the average lengths of the infectious period, the latency period and the life-length, and *decreasing in the proportion vaccinated. Further, if the community size is large enough, then τ is increasing in: the basic reproduction number and the variances of the life-length and the latency period; but not monotone in the variance of the infectious period.*

In Section 5 where τ is computed over relevant parameter regions, it is seen that the most influential parameters, beside the community size, are the average lengths of the infectious period and latency period, and to some extent the basic reproduction number, the variance of the infectious period and the proportion vaccinated. The variances of the life-length and the latency period play a lesser role. In Section 5 we also give some special cases and present simulations to see how well the approximations apply.

2. Description of the model

We start with the population dynamics. Individuals are born into the population according to a Poisson process of *constant* rate μN and live, independently of everything else, for a $\Gamma(i,\ i\mu)$-distributed time. This means that $1/\mu$ is the average life-length, $1/(\mu\sqrt{i})$ is the standard deviation of the life-lengths, and $1/i$ is the squared coefficient of variation. The population size for this model will fluctuate around the 'parameter' N, a parameter which hence should be interpreted as the (average) population size. The reason for choosing size-independent birth rates is to avoid population extinction or explosion – disease extinction is the primary interest of the present paper. In this population we now wish to model the spread of a non-fatal infectious disease giving life-long immunity after recovery. For simplicity we assume that all individuals are homogeneous and mix uniformly. Once an individual gets infected he/she is first latent for a $\Gamma(j,\ j\gamma/r)$-distributed time after which she becomes infectious and remains so for a period having $\Gamma(k,\ k\gamma)$-distribution. This means that the average infection period is $1/\gamma$ long, with a standard deviation $1/(\gamma\sqrt{k})$ and squared coefficient of variation $1/k$. The expected length of the latent period is r/γ (r is the relative length with respect to the infectious period), its standard deviation is $r/(\gamma\sqrt{j})$ and squared coefficient of variation $1/j$. During the infectious period an individual has 'close contact' with any given individual according to a Poisson process of constant rate β/N, so β is (approximately) the contact rate with other individuals. A 'close contact' is defined as a contact resulting in infection if the other individual is susceptible. All contact processes, latent periods, infectious periods, births and deaths are defined mutually independent.

The model defined above will be analysed using a Markovian compartment model relying on the well-known fact that a $\Gamma(n,\ \nu)$-distribution is identical to the distribution of the sum of n independent $\text{Exp}(\nu)$ random variables (similar methods have been used to model epidemics for closed populations, e.g. Anderson & Watson, 1980). Thus, let $X_\ell(t)$ denote the number of susceptible individuals in 'age-stage' ℓ at time t, $\ell = 1, \dots, i$. For latent/infectious individuals one should in principle keep track of both the age-stage and latent/infectious stage thus giving rise to two indices. However, when the length of the latent and infectious periods are very short in comparison to the life-length, i.e. $(1+r)\epsilon \ll 1$ where $\epsilon := \gamma^{-1}/\mu^{-1}$,

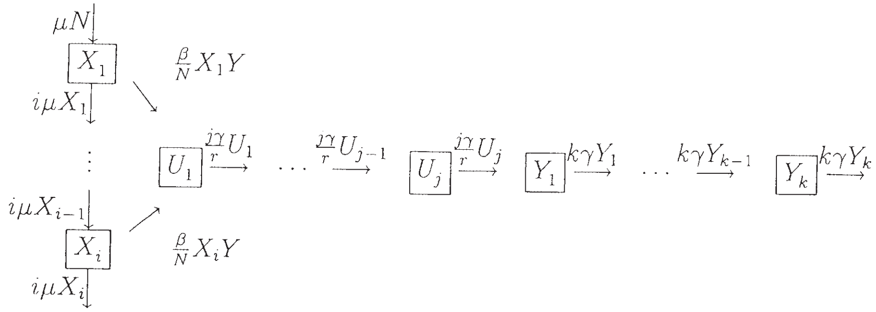

Fig. 1. Schematic picture of the model.

as is the case for childhood diseases, then the chance of dying while latent or infective is negligible (remember that we consider non-fatal diseases). From now on we therefore modify the model, by not allowing for deaths while latent/infectious, in order to obtain explicit results. Mathematically this means we only consider leading terms in expansions in ϵ. It hence suffices to keep track of the latent/infectious stage of infected individuals since they will (most likely) not die while latent/infectious. Thus, let $U_\ell(t)$ and $Y_m(t)$ respectively denote the total number of individuals latent in 'latent stage' ℓ and infectious stage m at time t, $\ell = 1, \dots, j$, $m = 1, \dots, k$. Further we let $X(t) = \sum_\ell X_\ell(t)$, $U(t) = \sum_\ell U_\ell(t)$, $Y(t) = \sum_\ell Y_\ell(t)$, respectively denote the total numbers of susceptible, latent and infectious individuals. The epidemic process is a Markovian vector jump process which is specified by its jump intensities. In Figure 1 we illustrate the approximate model, neglecting ageing while latent/infectious. The numbers by the arrows are the transition rates (which coincide with the full model up to leading term in ϵ). For example, susceptible individuals in stage ℓ are infected at rate $X_\ell(t)Y(t)\beta/N$ since between each infective–susceptible pair a contact occurs at rate β/N.

The number of recovered (and immune) individuals does not enter into the transition rates, so we need not keep track of this quantity. Thus, the state at time t of the epidemic process is specified by the $(i + j + k)$-dimensional vector $(\mathbf{X}(t), \mathbf{U}(t), \mathbf{Y}(t)) = (X_1(t), \dots, X_i(t), U_1(t), \dots, U_j(t), Y_1(t), \dots, Y_k(t))$. It is worth noting that the set of disease-free states $\{(m_1, \dots, m_i, 0, \dots, 0), m_\ell \geq 0\}$ is an absorbing class. All other states are hence transient.

The parameters to be used in the sequel, and their interpretation, are as follows: N= population size, μ^{-1}= average life-length, i^{-1}= the squared coefficient of variation of the life-length, $\epsilon = \gamma^{-1}/\mu^{-1}$= average length of the infectious period relative to the average life-length (in applications a very small quantity), r= the average length of the latency period relative to the infectious period, j^{-1} and k^{-1} which are the squared coefficients of variation of the latency period and infectious period respectively, and finally $R = \beta/\gamma$ which is the average number of contacts an individual has with other individuals during his/her infectious period, a fundamental parameter for epidemics often called the basic reproduction number. In the present paper it is assumed that $R > 1$, otherwise only few individuals will ever be infected before the epidemic goes extinct. The value of R varies for different communities and diseases (see Anderson & May, 1991 p 70 for some numerical values). Also ϵ and r (as well as j and k) depend on the disease. For measles a typical value of R is 15 and the latency period and infectious period are approximately one week each, so if the life expectancy is $\mu^{-1} = 70$ years this results in $\epsilon \approx 0.0003$ and $r = 1$.

3. Large population results

3.1. Deterministic approximation

To find the exact distribution of the epidemic process is not manageable. Instead we use the theory of diffusion approximation of population processes as for example described in Ethier & Kurtz (1986), approximations relying on weak convergence of stochastic processes. To this end we first study the corresponding deterministic system, and in the second subsection the scaled, diffusion-like, process.

In order to find the endemic level of the model, we assume that N is large and approximate the Markov process

$$\left(\frac{\mathbf{X}(t)}{N}, \frac{\mathbf{U}(t)}{N}, \frac{\mathbf{Y}(t)}{N}\right), \qquad t \geq 0,$$

with the solution $(\mathbf{x}(t), \mathbf{u}(t), \mathbf{y}(t))$, $t \geq 0$, of the following deterministic system of differential equations:

$$\begin{aligned}
\frac{dx_1}{dt} &= \mu - \beta x_1 y - i\mu x_1, \\
\frac{dx_2}{dt} &= i\mu x_1 - \beta x_2 y - i\mu x_2, \\
&\vdots \\
\frac{dx_i}{dt} &= i\mu x_{i-1} - \beta x_i y - i\mu x_i, \\
\frac{du_1}{dt} &= \beta x y - j\frac{\gamma}{r}u_1, \\
\frac{du_2}{dt} &= j\frac{\gamma}{r}u_1 - j\frac{\gamma}{r}u_2, \\
&\vdots \\
\frac{du_j}{dt} &= j\frac{\gamma}{r}u_{j-1} - j\frac{\gamma}{r}u_j, \\
\frac{dy_1}{dt} &= j\frac{\gamma}{r}u_j - k\gamma y_1, \\
\frac{dy_2}{dt} &= k\gamma y_1 - k\gamma y_2, \\
&\vdots \\
\frac{dy_k}{dt} &= k\gamma y_{k-1} - k\gamma y_k,
\end{aligned}$$

where x_ℓ, u_ℓ and y_ℓ are the asymptotic proportions of susceptible, latent and infectious individuals, respectively, in stage ℓ, and where $x = \sum x_\ell$, $u = \sum u_\ell$ and $y = \sum y_\ell$. The differential equations correspond to the jump intensities of the model (see Figure 1) with everything divided by N since we now consider population proportions.

The endemic level is obtained by equating these differential equations to zero. To simplify notation we introduce $p(i) := i/(i + R - 1)$ which should be interpreted as the probability of moving on to the next 'age-stage' rather than being infected, i.e. following a down-arrow rather than a right-arrow from a susceptible box in Figure 1. We obtain

$$\begin{aligned}
\hat{x}_\ell &= p(i)^{\ell-1}\frac{1-p(i)}{1-p(i)^i}\frac{1}{R},\\
\hat{u}_\ell &= \frac{r\mu}{j\gamma}\frac{R-1}{R} = \epsilon\frac{r}{j}\frac{R-1}{R},\\
\hat{y}_\ell &= \frac{\mu}{k\gamma}\frac{R-1}{R} = \epsilon\frac{1}{k}\frac{R-1}{R},
\end{aligned}$$

which implies

$$\begin{aligned}
\hat{x} &= \frac{1}{R},\\
\hat{u} &= \epsilon r\frac{R-1}{R},\\
\hat{y} &= \epsilon\frac{R-1}{R}.
\end{aligned}$$

The last three equations are of main importance since the different stages in terms of age, latency period and infectious period are only for modelling purposes to obtain a Markov process. It is seen that the susceptible proportion $\hat{x}$ is much larger than the latent and infectious proportions – remember that for childhood diseases ϵ is very small. Further, the endemic level is independent of the variances in life-length, latency period and infectious period, i.e. it does not depend on i, j or k. The solution $(\mathbf{x}, \mathbf{u}, \mathbf{y}) = (\hat{\mathbf{x}}, \hat{\mathbf{u}}, \hat{\mathbf{y}})$ is a stable solution to the set of differential equations, whereas the other solution $(\mathbf{x}, \mathbf{u}, \mathbf{y}) = (i^{-1}\mathbf{1}, \mathbf{0}, \mathbf{0})$, corresponding to the disease-free equilibrium, is unstable.

3.2. Diffusion approximation

We proceed by studying the centered and scaled process

$$\begin{aligned}
(\tilde{\mathbf{X}}_N, \tilde{\mathbf{U}}_N, \tilde{\mathbf{Y}}_N) &= (\tilde{X}_1, \ldots, \tilde{X}_i, \tilde{U}_1, \ldots, \tilde{U}_j, \tilde{Y}_1, \ldots, \tilde{Y}_k)\\
&= \sqrt{N}\left(\frac{X_1}{N} - \hat{x}_1, \ldots, \frac{X_i}{N} - \hat{x}_i, \frac{U_1}{N} - \hat{u}_1, \ldots, \frac{U_j}{N} - \hat{u}_j,\right.\\
&\qquad\left.\frac{Y_1}{N} - \hat{y}_1, \ldots, \frac{Y_k}{N} - \hat{y}_k\right).
\end{aligned}$$

Denote by $\mathscr{F}_t$ the σ-algebra (i.e. the information) generated by the process up to time t. Working out the first and second order infinitesimal moments, such as for example

$$\frac{1}{h}E\left[\tilde{X}_\ell(t+h)-\tilde{X}_\ell(t)\middle|\mathscr{F}_t\right] \qquad \text{and} \qquad \frac{1}{h}E\left[\left(\tilde{X}_\ell(t+h)-\tilde{X}_\ell(t)\right)^2|\mathscr{F}_t\right] \tag{1}$$

for small h, using the transition rates given in Figure 1, and applying theory for diffusion approximation (Ethier & Kurtz, 1986) suggests that the process $(\tilde{\mathbf{X}}_N(t), \tilde{\mathbf{U}}_N(t), \tilde{\mathbf{Y}}_N(t))$, $t \geq 0$, may be approximated by a $(i+j+k)$-dimensional Ornstein-Uhlenbeck process which is the limiting process as the population size tends to infinity. The local drift and covariance matrices B and S of the limiting process, obtained using (1) and likewise for the other components, are given in the Appendix. Applying results from diffusion theory (e.g. Karatzas & Shreve, 1991, p 357) shows that the stationary distribution of this process is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\Sigma = (\sigma_{ij})$, where Σ solves the matrix equation

$$B\Sigma + \Sigma B^T = -S.$$

We make the result precise in the following proposition.

Proposition 3.1. *Consider the model defined above and suppose that the initial point of the process* $(\tilde{\mathbf{X}}_N(0), \tilde{\mathbf{U}}_N(0), \tilde{\mathbf{Y}}_N(0))$ *converges in probability to some deterministic point* $(\mathbf{x}_0, \mathbf{u}_0, \mathbf{y}_0)$. *Then*

$$(\tilde{\mathbf{X}}_N, \tilde{\mathbf{U}}_N, \tilde{\mathbf{Y}}_N) \Longrightarrow (\tilde{\mathbf{X}}, \tilde{\mathbf{U}}, \tilde{\mathbf{Y}}) \qquad \text{as} \quad N \to \infty,$$

on any finite time interval, where $(\tilde{\mathbf{X}}, \tilde{\mathbf{U}}, \tilde{\mathbf{Y}})$ *is an Ornstein-Uhlenbeck process with local drift matrix B and local covariance matrix S (given in the Appendix) and with starting point* $(\mathbf{x}_0, \mathbf{u}_0, \mathbf{y}_0)$. *The stationary distribution of this process is multivariate normal with mean* $\mathbf{0}$ *and covariance matrix Σ defined above.*

It is not tractable to try to find exact solutions of Σ for general i, j, k. An expression for Σ, valid for small ϵ (i.e. $\gamma^{-1}/\mu^{-1} \ll 1$), is derived in the Appendix. In the next section we study the distribution of the time to extinction of the disease in a finite population. To obtain explicit results we use the limiting normal distribution given above. Of particular interest there is the mean and variance of the total number of latent and infectious individuals, i.e. the expectation and variance of $U+Y := \sum_\ell U_\ell + \sum_\ell Y_\ell$, denoted μ_{U+Y} and σ^2_{U+Y} respectively. To get an idea of the variation relative to the expected level of the diffusion we also give the coefficient of variation, $CV_{U+Y} = \sigma_{U+Y}/\mu_{U+Y}$. These moments are, up to first order, given by

$$\mu_{U+Y} = N(\hat{u}+\hat{y}) = N\epsilon(1+r)\frac{R-1}{R}, \tag{2}$$

$$\begin{aligned}\sigma^2_{U+Y} &= N\sum_{\ell,m=i+1}^{i+j+k}\sigma_{\ell m}\\ &\approx \frac{N(1+r)^2}{R\left(\frac{k+2}{3k}+r+\left(1+\frac{1}{j}\right)\frac{kr^2}{k+1}+\frac{(i/(i+R-1))^i}{1-(i/(i+R-1))^i}\left(1+\frac{2kr}{k+1}\right)^2\frac{k+1}{2k}\right)}\end{aligned} \tag{3}$$

$$CV_{U+Y} = \frac{\sqrt{R}}{\sqrt{N}\epsilon(R-1)\left(\frac{k+2}{3k} + r + \left(1+\frac{1}{j}\right)\frac{kr^2}{k+1} + \frac{(i/(i+R-1))^i}{1-(i/(i+R-1))^i}\left(1+\frac{2kr}{k+1}\right)^2\frac{k+1}{2k}\right)^{1/2}}. \tag{4}$$

The coefficient of variation indicates how ‘far away’ from absorbtion the epidemic process is at equilibrium. A closer look at (4) shows that CV_{U+Y} is decreasing in: $N,\ \epsilon,\ R,\ r$ and increasing in i and j; nothing general can be said about k except when there is no latency period ($r = 0$), when it is increasing in k. Some special cases of CV_{U+Y} are given in Section 5.

4. Quasi-stationarity and the time to extinction

4.1. The quasi-stationary distribution

In the previous section it was shown that when the population size N is fairly large then the epidemic process may be approximated by an Ornstein-Uhlenbeck process with a specified multivariate normal distribution as its stationary distribution. This approximation can only be valid before the epidemic goes extinct, that is, when there are no latent or infectious individuals present (i.e. $U = Y = 0$). After this time there will be no more infections so the spread of disease completely stops. In the present section we show that for any N the time to extinction starting in quasi-stationarity (i.e. conditional on not having gone extinct) is exponentially distributed. An exact expression for the parameter of the exponential distribution is not available. In order to obtain an approximate expression we proceed by approximating the quasi-stationary distribution by the limiting Ornstein-Uhlenbeck process (c.f. Nåsell, 1999).

We are interested in the distribution of the time to extinction, defined by

$$T = \inf\{t \geq 0 \,:\, \mathbf{U}(t) = \mathbf{0}, \mathbf{Y}(t) = \mathbf{0}\}.$$

This distribution depends on the initial state $(\mathbf{X}(0), \mathbf{U}(0), \mathbf{Y}(0))$. Two situations are of special interest: $(\mathbf{X}(0),\ \mathbf{U}(0), \mathbf{Y}(0)) = (i^{-1}N\mathbf{1}, \mathbf{0}, \mathbf{e}_1)$, i.e. a virgin population to which one infectious individual is introduced, or that the initial distribution is at ‘equilibrium’. In the present paper we restrict our attention to the latter case. A natural choice for equilibrium distribution is the so-called *quasi-stationary distribution* which we now define. Let $p_{\mathbf{x},\mathbf{u},\mathbf{y}}(t)$ be the probability that the process is in state $(\mathbf{x}, \mathbf{u}, \mathbf{y})$ at time t for some given initial distribution and let $p_{\bullet,\mathbf{u},\mathbf{y}}(t) = \sum_{\mathbf{x}} p_{\mathbf{x},\mathbf{u},\mathbf{y}}(t)$ denote the marginal distribution of $(\mathbf{U}(t), \mathbf{Y}(t))$. Then define $q_{\mathbf{x},\mathbf{u},\mathbf{y}}(t) := p_{\mathbf{x},\mathbf{u},\mathbf{y}}(t)/(1 - p_{\bullet,\mathbf{0},\mathbf{0}}(t))$ to be the probability that the process is in state $(\mathbf{x}, \mathbf{u}, \mathbf{y})$ at time t given that it has not yet become absorbed into the disease-free class of states. The quasi-stationary distribution $Q = \{q_{\mathbf{x},\mathbf{u},\mathbf{y}}\}$, which lives on the set of transient states, is then defined by

$$q_{\mathbf{x},\mathbf{u},\mathbf{y}} = \lim_{t\to\infty} q_{\mathbf{x},\mathbf{u},\mathbf{y}}(t).$$

This distribution is sometimes referred to as the quasi-limiting distribution. For more about quasi-stationary or quasi-limiting distributions we refer the reader to

Pollett & Roberts (1990). In words, the quasi-stationary distribution is the distribution after a long time, conditioned on not having gone extinct. In the next subsections we study the time to extinction having $Q = \{q_{\mathbf{x},\mathbf{u},\mathbf{y}}\}$ as initial distribution. This time duration is denoted T_Q.

4.2. The exact distribution of T_Q

Let T_Q denote the time to extinction starting with the quasi-stationary distribution. We have the following result which is an application of a general result for Markov processes.

Proposition 4.1. *T_Q is exponentially distributed with rate parameter $k\gamma q_{\bullet,\mathbf{0},\mathbf{e}_k} = k\gamma \sum_{\mathbf{x}} q_{\mathbf{x},\mathbf{0},\mathbf{e}_k}$.*

Proof. Indeed, T_Q is memoryless since

$$\begin{aligned} &P(T_Q > t + s \mid T_Q > t, (\mathbf{X}(0), \mathbf{U}(0), \mathbf{Y}(0)) \sim Q) \\ &\quad = P(T_Q > t + s \mid T_Q > t, (\mathbf{X}(t), \mathbf{U}(t), \mathbf{Y}(t)) \sim Q) \\ &\quad = P(T_Q > s \mid (\mathbf{X}(0), \mathbf{U}(0), \mathbf{Y}(0)) \sim Q), \end{aligned}$$

by the Markov property, and the exponential character of T_Q follows immediately. The rate parameter of the exponential distribution is simply the sum of rates into the disease-free class of states, weighted by the quasi-stationary distribution. It is only possible to enter the disease-free class when there are no latent individuals and exactly one infectious individual which is in the last infectious stage. The jump rate from this state is $k\gamma$, independent of how many susceptibles there are. Therefore the total jump rate is $k\gamma q_{\bullet,\mathbf{0},\mathbf{e}_k}$.

It remains to find an approximation of the expected time to extinction,

$$\tau = E(T_Q) = \frac{1}{\gamma k q_{\bullet,\mathbf{0},\mathbf{e}_k}}. \tag{5}$$

4.3. An approximate formula for $E(T_Q)$

Let us return to the formula given in (5). To estimate $kq_{\bullet,\mathbf{0},\mathbf{e}_k}$ using the stationary multivariate normal distribution of the limiting diffusion presents severe difficulties. Instead we estimate $q_{\bullet,1} := \sum_\ell q_{\bullet,\mathbf{e}_\ell,\mathbf{0}} + \sum_\ell q_{\bullet,\mathbf{0},\mathbf{e}_\ell}$, the quasi-stationary probability that there is exactly 1 individual who is either latent or infectious. Then we derive an approximate relation between $kq_{\bullet,\mathbf{0},\mathbf{e}_k}$ and $q_{\bullet,1}$. An available estimate for $q_{\bullet,1}$ is to use the marginal normal distribution for $U + Y$ of the limiting stationary distribution derived in Section 3.2, an idea adopted from Nåsell (1999). Because we have conditioned on not having gone extinct we truncate the distribution at the point 0.5 (conditional on non-extinction with continuity correction of the integer-valued variable $U + Y$) to obtain the approximation

$$q_{\bullet,1} \approx \frac{\varphi\left((1 - \mu_{U+Y})/\sigma_{U+Y}\right)}{\sigma_{U+Y}\left(1 - \Phi\left((0.5 - \mu_{U+Y})/\sigma_{U+Y}\right)\right)}, \tag{6}$$

where $\Phi(\cdot)$ and $\varphi(\cdot)$ denote the standard normal distribution function and density function, respectively.

It remains to relate $kq_{\bullet,\mathbf{0},\mathbf{e}_k}$ to $q_{\bullet,1}$ in order to evaluate (5). Thus we define $\rho := q_{\bullet,\mathbf{0},\mathbf{e}_k}/q_{\bullet,1}$ and heuristically derive an approximation for ρ, the relative (quasi stationary) probability that, given that there is exactly one latent or infectious individual, this individual is in the infectious stage k.

Start the epidemic process according to the quasi-stationary distribution. Label the visits at $U+Y=1$ before extinction by $1,\dots,M_1$ (M_1 is random). Start a new epidemic and label its visits at $U+Y=1$ by $M_1+1,\dots,M_1+M_2$, and so forth. For each such visit i at $U+Y=1$ let V_i denote the time spent in $(\mathbf{U},\mathbf{Y})=(\mathbf{0},\mathbf{e}_k)$ before leaving the state $U+Y=1$, and W_i the total time spent in the state $U+Y=1$. Then the time spent in state $(\mathbf{0},\mathbf{e}_k)$ relative to the total time spent in $U+Y=1$ among the first n visits is $\rho_n=(V_1+\ldots+V_n)/(W_1+\ldots+W_n)$. Since each such visit is obtained from an epidemic in quasi-stationarity it follows that $\lim_{n\to\infty}\rho_n=q_{\bullet,\mathbf{0},\mathbf{e}_k}/q_{\bullet,1}=\rho$ almost surely. On the other hand, by the strong law of large numbers, ρ_n converges almost surely to $E(V)/E(W)$, so $\rho=E(V)/E(W)$ and it remains to calculate $E(V)$ and $E(W)$.

We calculate these means conditioned on where the state $U+Y=1$ was entered from, and these conditional means are derived assuming that the proportion susceptible is at the endemic level $\hat{x}=1/R$ and consequently $X=N/R$. Then $E(W|(\mathbf{0},\mathbf{e}_\ell))=\lambda^{-1}(1+\nu+\ldots+\nu^{k-\ell})=\lambda^{-1}(1-\nu^{k-\ell+1})/(1-\nu)$, where $\lambda=k\gamma+\beta/R=(k+1)\gamma$ is the rate with which the individual leaves a given infectious state and $\nu=k\gamma/(k+1)\gamma=k/(k+1)$ is the probability that this happens by moving on to the next infectious state rather than a new infection which would take the process out of $U+Y=1$. One way to interpret the expression for $E(W|(\mathbf{0},\mathbf{e}_\ell))$ is that it remains in any state for a time with mean λ^{-1}, and the probability to reach m infectious states ahead is ν^m. Similarly the other conditional moments are shown to satisfy: $E(W|(\mathbf{e}_\ell,\mathbf{0}))=\tilde{\lambda}^{-1}(1-\tilde{\nu}^{j-\ell+1})/(1-\nu)+\lambda^{-1}\tilde{\nu}^{j-\ell+1}(1-\nu^k)/(1-\nu)$, $E(V|(\mathbf{0},\mathbf{e}_\ell))=\lambda^{-1}\nu^{k-\ell}$ and $E(V|(\mathbf{e}_\ell,\mathbf{0}))=\lambda^{-1}\tilde{\nu}^{j-\ell+1}\nu^{k-1}$, where $\tilde{\lambda}=(j+r)\gamma/r$ and $\tilde{\nu}=j/(j+r)$. A natural estimate for the probabilities determining which state $U+Y=1$ is entered to is to assume that the probability that the individual is latent is $r/(r+1)$ and $1/(r+1)$ for being infectious, since this relation holds for the expected lengths of the two periods, and to assume uniform distribution within the latent and infectious stages respectively. This implies that $P(\mathbf{e}_\ell,\mathbf{0})=r/j(r+1)$ and $P(\mathbf{0},\mathbf{e}_\ell)=1/k(r+1)$. We then use that $E(V)=\sum_\ell E(V|\mathbf{e}_\ell,\mathbf{0})P(\mathbf{e}_\ell,\mathbf{0})+E(V|\mathbf{0},\mathbf{e}_\ell)P(\mathbf{0},\mathbf{e}_\ell)$, and similarly for $E(W)$. After simplifying the formulas we conclude that $k\rho=k\rho(r,j,k)=kE(V)/E(W)$ can be approximated by

$$k\rho(r,j,k)\approx\frac{1-\left(\frac{k}{k+1}\right)^k\left(\frac{j}{j+r}\right)^j}{r+\left(\frac{k}{k+1}\right)^k\left(\frac{j}{j+r}\right)^j} \tag{7}$$

Simple algebra shows that $k\rho(r,j,k)$ is increasing in j and k but decreasing in r. Inserting (6) and (7) into (5) gives the expression

$$\tau\approx\frac{\sigma_{U+Y}\left(1-\Phi\left((0.5-\mu_{U+Y})/\sigma_{U+Y}\right)\right)}{\gamma k\rho(r,j,k)\varphi\left((1-\mu_{U+Y})/\sigma_{U+Y}\right)}\approx\frac{\sigma_{U+Y}\Phi\left(\mu_{U+Y}/\sigma_{U+Y}\right)}{\gamma k\rho(r,j,k)\varphi\left(\mu_{U+Y}/\sigma_{U+Y}\right)}$$

where μ_{U+Y} and σ^2_{U+Y} are given by (2) and (3) respectively. The last approximation is valid because the endemic level μ_{U+Y} is much larger than 1 when considering persistence. Introduce

$$\begin{aligned} f(R, r, i, j, k) = {} & \frac{k+2}{3k} + r + \left(1 + \frac{1}{j}\right)\frac{kr^2}{k+1} \\ & + \frac{(i/(i+R-1))^i}{1-(i/(i+R-1))^i}\left(1 + \frac{2kr}{k+1}\right)^2 \frac{k+1}{2k}, \end{aligned}$$

appearing in σ^2_{U+Y}. Later we will use the observations that f is globally increasing in r and decreasing in i, j, R, and also in k if $r = 0$. Then, after some further rearrangement and replacing μ_{U+Y} and σ^2_{U+Y} by the corresponding expression given in equations (2) and (3), one obtains

$$\tau \approx \mu^{-1} \frac{(1+r)\sqrt{N\epsilon^2}\,\Phi\left(\left(N\epsilon^2 \frac{(R-1)^2}{R} f(R, r, i, j, k)\right)^{1/2}\right)}{k\rho(r, j, k)\sqrt{R\, f(R, r, i, j, k)}\,\varphi\left(\left(N\epsilon^2 \frac{(R-1)^2}{R} f(R, r, i, j, k)\right)^{1/2}\right)}. \tag{8}$$

This expression is the final approximation of $\tau = \tau(\mu^{-1}, N, \epsilon, R, r, i, j, k)$, the expected time to extinction starting in quasi-stationarity. Because T_Q follows the exponential distribution, $\tau = E(T_Q)$ in (8) determines the distribution. Expressions for, and behaviour of, τ are of interest to epidemiologists (e.g. Keeling & Grenfell, 1997). In particular it is of interest to study the influence of different parameters on τ.

The first conclusion is that τ grows linearly in the expected life-length μ^{-1}. This is trivial since all other time parameters are expressed relative to the expected life-length which hence is just the time unit. Another conclusion is that τ is increasing in the population size N and ϵ, the duration of the infectious period relative to the life-length. These are the only global monotonicities. However, if the common argument of φ and Φ in (8) exceeds 1, or equivalently the coefficient of variation CV_{U+Y} defined in (4) is less than 1, then we can say more (this holds in large enough populations and it is the situation where endemicity and quasi-stationarity are most relevant). Using stated properties of ρ and f it then follows that τ is decreasing in i and j and increasing in r. Further, for the same situation, τ is increasing in R if $R \gg 1$ (for most childhood diseases R exceeds 10, cf Anderson & May, 1991 p 70). The only parameter where no general conclusion holds is k, the squared inverse of the coefficient of variation of the infectious period, except if $r = 0$ when τ is decreasing in k.

We have concluded the following qualitative results. In a large enough population and for relevant parameter values the expected time to extinction is increasing in: the basic reproduction number, the average life-length and its variance, the expected length of the latency period and its variance and in the average length of the infectious period. It is also increasing in the variance of the infectious period when there is no or a short latency period. The magnitude of influence of the parameters is studied briefly in the next section where τ is evaluated as a given parameter varies over an interval while the remaining parameters are kept at fixed typical values.

Remember that i, j and k are the squared inverse of the coefficients of variation of the life-length, latency period and infectious period respectively, so the opposite relations are valid for the coefficients of variation, or equivalently the variances.

4.4. Vaccination and the critical community size N_c

Suppose that a vaccine is available which prevents individuals from becoming infected. For simplicity we only consider vaccines that give complete and life-long immunity (see for example Halloran *et al.*, 1992, for more realistic effects of vaccination) . Suppose further that a vaccination program is initiated which continuously vaccinates a proportion v of the newly born individuals. How does this affect T_Q and τ? It is not hard to show that the same model can be used for this situation, only with new parameter values. The only jump-intensity that is affected in Figure 1 is the top arrow on the left: now new susceptible individuals enter the population at rate $\mu N(1-v)$ since a proportion v is vaccinated. But then we have to write $N(1-v)$ elsewhere in the figure to have balance, so then β becomes $\beta(1-v)$. Thus, if we let $N' = N(1-v)$ and $\beta' = \beta(1-v)$ we have the model defined in Section 2, only with N' and β' replacing N and β. Because $R = \beta/\gamma$ this also reduces R to $R' = R(1-v)$, all other parameters are unaffected. This implies that the same approximation of the epidemic process is valid. For example the stationary distribution of the approximating diffusion is Gaussian with mean, variance and coefficient of variation given by (2) – (4) where N is replaced by $N(1-v)$ and R by $R(1-v)$. It also follows that T_Q, the time to extinction starting in quasi-stationarity, is exponentially distributed and $\tau = E(T_Q)$ is obtained from (8) with the same replacements:

$$\tau \approx \mu^{-1} \frac{(1+r)\sqrt{N\epsilon^2}\Phi\left(\left(N\epsilon^2 \frac{(R(1-v)-1)^2}{R} f(R(1-v), r, i, j, k)\right)^{1/2}\right)}{k\rho(r, j, k)\sqrt{Rf(R(1-v), r, i, j, k)}\varphi\left(\left(N\epsilon^2 \frac{(R(1-v)-1)^2}{R} f(R(1-v), r, i, j, k)\right)^{1/2}\right)}. \tag{9}$$

It follows that τ is decreasing in v and its influence is quite large, which is illustrated in Section 5.

Next we define the critical community size N_c. The critical community size $N_c = N_c(t, p)$ with time horizon t and extinction probability p is the solution N that satisfies

$$P(T_Q > t) = 1 - p. \tag{10}$$

In words, N_c is an upper bound on the community size for the epidemic to have a fair chance of going extinct before a given time point. Since T_Q is exponentially distributed, its mean τ specifies the distribution and hence also N_c. In fact, it follows immediately that $N_c(t, p)$ is the solution N to the equation $\tau(\mu^{-1}, N, \epsilon, R, r, i, j, k) = t/(-\ln(1-p))$. How N_c depends on the different parameters is not very transparent. The only simple relationship is that N_c is inversely proportional to ϵ^2. This is immediate since N and ϵ^2 always appear together. This relation is demonstrated by simulations in the next section. The dependence on other parameters is more complicated. However, if in the definition of $N_c(t, p)$, t is large enough and p is not too large, say $t = \mu^{-1}$ (one life-length) and $p = 1/2$, then the conclusions for τ relying on large N may be interpreted in terms of N_c: if τ is increasing this means a smaller community would result in the same τ, so the monotonic relations for N_c are opposite of the ones stated for τ in the end of the previous subsection. For example, we then have that N_c is decreasing in r. That is, with a longer latency period endemicity will occur in smaller communities.

For any fixed parameter values and choice of time horizon t and extinction probability p the critical community size N_c may of course be easily evaluated using a mathematical software program. Similarly, the critical immunity coverage, i.e. the proportion necessary to vaccinate in order to have extinction within some time horizon likely is simple to calculate numerically.

5. Examples and simulations

First we study the stationary distribution of the limiting diffusion process specified in Section 3. The single most important quantity when interested in excursions away from the stationary endemic level is the coefficient of variation CV_{U+Y} defined in (4). The larger CV_{U+Y} the larger excursions should be likely. Below we give this quantity for some parameter choices of special interest, where we have extended the formulas to cover the case where a proportion v of all new born individuals are vaccinated.

No latent period $r = 0$. Our calculations are only valid for $r > 0$, but CV_{U+Y} has a well defined limit as $r \to 0$:

$$CV_{U+Y} \approx \frac{\sqrt{R}}{\sqrt{N}\epsilon(R(1-v)-1)\left(\frac{k+2}{3k} + \frac{(i/(i+R(1-v)-1))^i}{1-(i/(i+R(1-v)-1))^i}\,\frac{k+1}{2k}\right)^{1/2}}.$$

As mentioned in Section 3.2 CV_{U+Y} is then increasing in k, i.e. decreasing in the variances of the infectious period.

The 'exponential' case $(i, j, k) = (1, 1, 1)$. This means that the life-length, latent and infectious periods are modelled by exponential distributions.

$$CV_{U+Y} \approx \frac{\sqrt{R}}{\sqrt{N}\epsilon(R(1-v)-1)\left(1 + r + r^2 + \frac{1}{R(1-v)-1}(1+r)^2\right)^{1/2}}.$$

The 'constant' case $(i, j, k) = (\infty, \infty, \infty)$. This means that the life-length, latent and infectious periods are constant, the other extreme compared to the example above.

$$CV_{U+Y} \approx \frac{\sqrt{R}}{\sqrt{N}\epsilon(R(1-v)-1)\left(1/3 + r + r^2 + \frac{e^{-R(1-v)+1}}{1-e^{-R(1-v)+1}}(1+2r)^2/2\right)^{1/2}}.$$

If latter two examples are compared, for $R = 15$ and $r = 0.5$, it is seen that CV_{U+Y} is 75% larger in the exponential case implying that the variation of the life-length, latent and infectious periods do have influence on the behaviour of the epidemic.

Now we compute τ, using the approximation given in (9), to study the magnitude of influence of each parameter. Each parameter is varied separately over a relevant interval, keeping the other parameters fixed at some typical value. This has been done for a community of size 10^5 and a population with 10^6 individuals. The results are reported in Table 1 where the time unit is years and the expected life-length was set to 75 years. The typical values were chosen to be $R = 15$ (a common estimate for measles, e.g. Anderson & May 1991, p 70), $\epsilon = 0.0003$ (corresponding to 1 week infectious period), $r = 0.5$ (half as long latency period), $CV_{\text{life}} = 0.3$, $CV_{\text{lat}} = 0.5$ and $CV_{\text{inf}} = 0.5$ (corresponding to standard deviations of 20 years, 3 and 1.5 days respectively). The type value for vaccinations was chosen to be no vaccinations ($v = 0$) but when varied it has been evaluated up to a 90 % vaccination coverage ($v = 0.9$). The reason for not evaluating τ for even larger coverage is that as soon as $R(1 - v) \leq 1$ the approximations break down. When this happens a major outbreak is no longer possible and the epidemic will die out immediately so the notion of quasi-stationarity becomes less important.

In Table 1 we observe that the expected time to extinction of the smaller population is only a few years ($\tau = 3.6$ years for type values) why it would die out reasonably quick. It is seen that the most influential parameters are ϵ and r and to some extent R and v. In the larger community we see that τ is larger ($\tau = 32$ years for the type values) i.e. more like an endemic situation. The parameters are in general more influential; most influential are the same parameters as in the smaller community together with CV_{inf}. Another observation is that τ is now increasing in R and CV_{life} contrary to the smaller community. An explanation for the somewhat surprising observation that τ is *decreasing* in R for moderate N goes as follows: even though μ_{U+Y} increases with R, seeming to make extinction less likely and hence τ larger, the effect of increasing R also makes the standard deviation σ_{U+Y} increase which has a greater influence when the endemic level μ_{U+Y} is not too far from 0. If τ is computed for even larger communities it rapidly grows as does the influence of each parameter. For example $\tau = 98000$ years for the type values in a community with $N = 10^7$ individuals, so there the disease would definitely become endemic. Of course, the assumption of homogeneous mixing is less realistic in large communities. However, the presence of heterogeneities are believed

Table 1. Computation of τ, using (9), for different parameter choices with $\mu^{-1} = 75$ years

Parameter	Type value	Parameter range	Monotone $N = 10^5$	Monotone $N = 10^6$	τ range $N = 10^5$	τ range $N = 10^6$	Rel. change $N = 10^5$	Rel. change $N = 10^6$
R	15	5–20	$\searrow$	$\nearrow$	3.2–5.4	22–41	0.6	1.8
ϵ	0.0003	0.00015–0.0006	$\nearrow$	$\nearrow$	1.6–10.3	15–525	6.4	35
r	0.5	0–2	$\nearrow$	$\nearrow$	2.4–11.6	11–2000	4.9	170
CV_{life}	0.3	0–1	$\searrow$	$\nearrow$	3.4–3.6	31–36	0.96	1.1
CV_{lat}	0.5	0–1	$\nearrow$	$\nearrow$	3.5–3.6	31–36	1.01	1.1
CV_{inf}	0.5	0–1	$\nearrow$	$\nearrow$	3.5–3.7	26–45	1.08	1.5
v	0	0–0.9	$\searrow$	$\searrow$	1.7–3.6	5–32	0.46	0.17

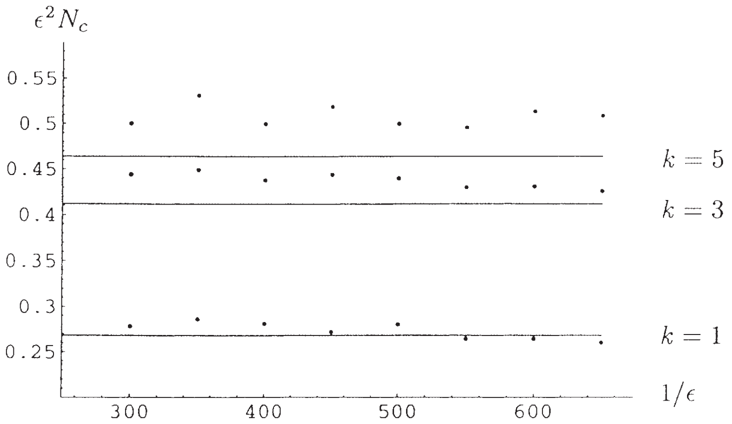

Fig. 2. $\epsilon^2 N_c$ plotted against $1/\epsilon$ for different values of k. Simulated values $(\cdots)$ and theoretical approximation (—).

to increase the expected time to extinction τ thus making endemicity even more likely (cf. Section 6).

Simulations have been performed in order to check the validity of the approximate formula for τ. Instead of simulating τ we have simulated N_c also making it possible to verify the observation that N_c is proportional to $1/\epsilon^2$. For simplicity no latency period was assumed and the life-length was modelled by the exponential distribution (i.e. $r = 0$ and $i = 1$). Further, $R = 12$ and $v = 0$ (no vaccinations) were used in the simulations which were performed for different values of ϵ and k. The time horizon was chosen as $t = \mu^{-1}$ (one average life-length) and the extinction probability was set to $p = 1/2$. For a given population size N, 3000 simulations were performed and N was accepted as the critical community size if the observed proportion of fade-outs lied between 47 % and 53 %. The result is shown in Figure 2 where, for a given k, $\epsilon^2 N_c$ is plotted for different lengths of the infectious period (in order to avoid too long simulation times rather large values of ϵ were used). The result shows that N_c seems inversely proportional to ϵ^2 as our approximation suggests, even though the horizontal level obtained using (8) is somewhat off for $k = 3$ and 5. It is also seen that N_c is increasing in k, i.e. decreasing in the variance of the infectious period. If on the other hand much more likely and frequent fade-outs were studied the last conclusion is no longer true. The opposite may even be the case as Keeling and Grenfell (1997) observed in simulations modelling the spread of measles in England and Wales.

6. Discussion

The present paper tries to extend an epidemic model studied recently (van Herwaarden & Grassman, 1995, Nåsell, 1999) to a more realistic setup allowing for a latency period but also letting the variance of the infectious period, latency period and life-length be close to arbitrary. The time to extinction starting in quasi-stationarity is shown be exponentially distributed. An approximate expression for the mean parameter is derived from the diffusion approximation of the epidemic process, equation (8), or equation (9) when vaccinations are considered. It should be stressed that this is only an approximation: when the process is close to the absorbing barrier, i.e. when there are few infectives, it moves at a slower rate than the approximating diffusion. Thus the approximation should only serve as a qualitative guidance and not be relied on in detail.

The technique of the paper, to define the life and the latent and infectious periods using a series of successive exponentially distributed stages, can be extended further. For example, the rates to jump between successive stages could vary, thus allowing for a wider class of distributions than the Γ-distributions in the present setup. Secondly, one could allow for an infection rate that varies between different stages of infectivity. Presently this is only done in two steps: the latent stages, where there is no infectivity, and the infectious stages, over which the infectivity stays constant. The technical level will of course be higher under this extended model and to obtain an explicit approximate expression for τ under that scenario appears to be very complicated.

The approximate formula for τ works adequately, but certainly not perfectly, as simulations in Section 5 show. However, perhaps more important for approaching real-life epidemics than improving the approximation is to generalize the model. For example, spatial, social and individual heterogeneities play an important role in the spread of infectious diseases, and so do seasonal effects (see Anderson & May, 1991, p 83). Sound knowledge and simulation results for models taking such effects into account indicate that the expected time to extinction is longer in a heterogeneous community implying that the critical community size decreases (e.g. Keeling & Grenfell, 1997, and references therein). For seasonal effects the converse relation holds. Further, vaccination can be modelled more realistically by assuming partial and waning immunity over time. It is our belief however that the qualitative statements of the present paper, for the parameters studied, remain valid even in such more realistic settings. A thorough study of this is yet to be carried out.

Acknowledgements. The present work was initiated at the RSS Epidemic Workshop held in Skye in April 1997. We thank the organizers for inviting us to an interesting workshop and Odo Diekmann, Klaus Dietz, Ingemar Nåsell and Anders Martin-Löf for stimulating discussions on persistence.

A. Appendix

To start, we give the local drift matrix B and the local covariance matrix S simply obtained as the conditional moments of the scaled epidemic process $(\tilde{\mathbf{X}}(t), \tilde{\mathbf{U}}(t), \tilde{\mathbf{Y}}(t))$ over small time increments, as in (1). Thereafter we derive the leading terms in the covariance matrix, Σ, of the stationary distribution of the approximating Ohrnstein-Uhlenbeck process with which $(\tilde{\mathbf{X}}(t), \tilde{\mathbf{U}}(t), \tilde{\mathbf{Y}}(t))$ is approximated by, a matrix defined as the unique solution to the matrix equation $B\Sigma + \Sigma B^T = -S$.

For a general square matrix A of order $i + j + k$, write

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix},$$

where rows and columns have been divided into blocks of size i, j and k, respectively. As mentioned above the local drift matrix is obtained by computing the conditional infinitesimal moments of the components of the scaled epidemic process (e.g. equation 1) which should be expressed in terms of the various $d\tilde{X}_\ell(t)$, $d\tilde{U}_\ell(t)$ and $d\tilde{Y}_\ell(t)$. This turns out to be equivalent to differentiating the set of differential equations, given in Section 3.1, for the corresponding deterministic system and evaluate this matrix of partial derivatives at the equilibrium point. The local drift matrix B is thus given by

$$B_{11} = \begin{pmatrix} -\beta\hat{y} - i\mu & 0 & 0 & \cdots \\ i\mu & -\beta\hat{y} - i\mu & 0 & \cdots \\ 0 & i\mu & -\beta\hat{y} - i\mu & \cdots \\ \vdots & & & \end{pmatrix}, \; B_{12} = \mathbf{0},$$

$$B_{13} = -\beta \begin{pmatrix} \hat{x}_1 & \hat{x}_1 & \hat{x}_1 & \cdots \\ \hat{x}_2 & \hat{x}_2 & \hat{x}_2 & \cdots \\ \hat{x}_3 & \hat{x}_3 & \hat{x}_3 & \cdots \\ \vdots & & & \end{pmatrix},$$

$$B_{21} = \beta\hat{y} \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & & & \end{pmatrix}, \; B_{22} = j\frac{\gamma}{r} \begin{pmatrix} -1 & 0 & 0\cdots \\ 1 & -1 & 0\cdots \\ 0 & 1 & -1\cdots \\ \vdots & & \end{pmatrix},$$

$$B_{23} = \beta\hat{x} \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & & & \end{pmatrix},$$

$$B_{31} = \mathbf{0}, B_{32} = j\frac{\gamma}{r} \begin{pmatrix} 0 & \cdots & 0 & 0 & 1 \\ 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 \\ \vdots & & & & \end{pmatrix}, B_{33} = k\gamma \begin{pmatrix} -1 & 0 & 0 & \cdots \\ 1 & -1 & 0 & \cdots \\ 0 & 1 & -1 & \cdots \\ \vdots & & & \end{pmatrix}.$$

The (symmetric) local covariance matrix S, also defined through its submatrices, is obtained by computing the conditional infinitesimal second moments of the scaled epidemic process. For the diagonal elements (the variance terms) this is equivalent to differentiating the system of differential equations of Section 3.1, each with respect to the same coordinate once more, and making all terms of a derivative become positive – an increase or decrease has the same (positive) effect on the variance. Concerning the the covariance terms they will be 0 unless a jump between the two states in question is possible. In the latter case the covariance will be transition rate between the two states in the deterministic system, but with negative sign because one of the components increases while the other decreases. For example, the transtion rate between x_1 and x_2 is $i\mu x_1$ (an individual in age state 1 jumps into the next age state). It follows that S has the following form:

$$S_{11}=\begin{pmatrix} \mu+\beta\hat{x}_1\hat{y}+i\mu\hat{x}_1 & -i\mu\hat{x}_1 & 0 & 0 & \cdots \\ -i\mu\hat{x}_1 & \beta\hat{x}_2\hat{y}+i\mu(\hat{x}_1+\hat{x}_2) & -i\mu\hat{x}_2 & 0 & \cdots \\ 0 & -i\mu\hat{x}_2 & \beta\hat{x}_3\hat{y}+i\mu(\hat{x}_2+\hat{x}_3) & -i\mu\hat{x}_3 & \cdots \\ \vdots & & & & \end{pmatrix},$$

$$S_{12}=\beta\hat{y}\begin{pmatrix} -\hat{x}_1 & 0 & 0 & \cdots \\ -\hat{x}_2 & 0 & 0 & \cdots \\ -\hat{x}_3 & 0 & 0 & \cdots \\ \vdots & & & \end{pmatrix}, \quad S_{13}=\mathbf{0},$$

$$S_{21}=S_{12}^T, \quad S_{22}=j\frac{\gamma}{r}\begin{pmatrix} \frac{r\beta}{j\gamma}\hat{x}\hat{y}+\hat{u}_1 & -\hat{u}_1 & 0 & 0 & \cdots \\ -\hat{u}_1 & \hat{u}_1+\hat{u}_2 & -\hat{u}_2 & 0 & \cdots \\ 0 & -\hat{u}_2 & \hat{u}_2+\hat{u}_3 & -\hat{u}_3 & \cdots \\ \vdots & & & & \end{pmatrix},$$

$$S_{23}=j\frac{\gamma}{r}\begin{pmatrix} \vdots & & & \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ -\hat{u}_j & 0 & 0 & \cdots \end{pmatrix},$$

$$S_{31}=S_{13}^T=\mathbf{0}, \quad S_{32}=S_{23}^T \quad S_{33}=k\gamma\begin{pmatrix} \frac{j}{kr}\hat{u}_j+\hat{y}_1 & -\hat{y}_1 & 0 & 0 & \cdots \\ -\hat{y}_1 & \hat{y}_1+\hat{y}_2 & -\hat{y}_2 & 0 & \cdots \\ 0 & -\hat{y}_2 & \hat{y}_2+\hat{y}_3 & -\hat{y}_3 & \cdots \\ \vdots & & & & \end{pmatrix}.$$

Remembering that the parameter ϵ is in general a small number, we expand the equation $B\Sigma+\Sigma B^T=-S$ in ϵ. To simplify notation we divide by the time scale parameter μ. Write

$$\begin{aligned} \mu^{-1}B &= B^+/\epsilon + B^0, \\ \mu^{-1}S &= S^0, \\ \Sigma &= \Sigma^+/\epsilon + \Sigma^0 + \epsilon\Sigma^- + O(\epsilon^2). \end{aligned}$$

Introducing the notation $[A, B]=AB+B^TA^T$, we get the following equations:

$$[B^+, \Sigma^+]=\mathbf{0}, \tag{11}$$

$$[B^+, \Sigma^0]=-[B^0, \Sigma^+] \tag{12}$$

$$[B^+, \Sigma^-]=-\left(S+[B^0, \Sigma^0]\right). \tag{13}$$

The matrix B^+ is defined as follows:

$$B_{11}^+=\mathbf{0}, \quad B_{12}^+=\mathbf{0}, \quad B_{13}^+=\frac{1-p(i)}{1-p(i)^i}\begin{pmatrix} -1 & -1 & -1 & \cdots \\ -p & -p & -p & \cdots \\ -p^2 & -p^2 & -p^2 & \cdots \\ \vdots & & & \end{pmatrix},$$

$$B^+_{21} = \mathbf{0}, \quad B^+_{22} = \frac{j}{r}\begin{pmatrix} -1 & 0 & 0 & \cdots \\ 1 & -1 & 0 & \cdots \\ 0 & 1 & -1 & \cdots \\ \vdots & & & \end{pmatrix}, \quad B^+_{23} = \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & & & \end{pmatrix},$$

$$B^+_{31} = \mathbf{0}, \quad B^+_{32} = \frac{j}{r}\begin{pmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & & & \end{pmatrix}, \quad B^+_{33} = k\begin{pmatrix} -1 & 0 & 0 & \cdots \\ 1 & -1 & 0 & \cdots \\ 0 & 1 & -1 & \cdots \\ \vdots & & & \end{pmatrix}.$$

Also, for B^0 we have

$$B^0_{11} = \begin{pmatrix} -(R-1+i) & 0 & 0 & \cdots \\ i & -(R-1+i) & 0 & \cdots \\ 0 & i & -(R-1+i) & \cdots \\ \vdots & & & \end{pmatrix}, B^0_{12} = \mathbf{0}, \quad B^0_{13} = \mathbf{0},$$

$$B^0_{21} = (R-1)\begin{pmatrix} 1 & 1 & 1 & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & & & \end{pmatrix}, B^0_{22} = \mathbf{0}, \quad B^0_{23} = \mathbf{0},$$

$$B^0_{31} = \mathbf{0}, \quad B^0_{32} = \mathbf{0}, \quad B^0_{33} = \mathbf{0}.$$

The (symmetric) local covariance matrix is given by $S = \mu S^0$, where

$$S^0_{11} = \frac{i(1-p)}{R(1-p^i)}\begin{pmatrix} \frac{R(1-p^i)}{i(1-p)} + \frac{1}{p} & -1 & 0 & 0 & \cdots \\ -1 & 2 & -p & 0 & \cdots \\ 0 & -p & 2p & -p^2 & \cdots \\ \vdots & & & & \end{pmatrix},$$

$$S^0_{12} = \frac{(R-1)(1-p)}{R(1-p^i)}\begin{pmatrix} -1 & 0 & 0 & \cdots \\ -p & 0 & 0 & \cdots \\ -p^2 & 0 & 0 & \cdots \\ \vdots & & & \end{pmatrix}, \qquad S^0_{13} = \mathbf{0},$$

$$S^0_{21} = (S^0_{12})^T, S^0_{22} = \frac{R-1}{R}\begin{pmatrix} 2 & -1 & 0 & 0 & \cdots \\ -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \\ \vdots & & & & \end{pmatrix},$$

$$S^0_{23} = \frac{R-1}{R}\begin{pmatrix} \vdots & & & \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ -1 & 0 & 0 & \cdots \end{pmatrix},$$

$$S^0_{31} = (S^0_{13})^T, \quad S^0_{32} = (S^0_{23})^T, \quad S^0_{33} = \frac{R-1}{R}\begin{pmatrix} 2 & -1 & 0 & 0 & \cdots \\ -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \\ \vdots & & & & \end{pmatrix}.$$

Define Σ^+ as follows:

$$\Sigma^+_{11} = \sigma^+_{11}\begin{pmatrix} 1 & p & p^2 & \cdots \\ p & p^2 & p^3 & \cdots \\ p^2 & p^3 & p^4 & \cdots \\ \vdots & & & \end{pmatrix},$$

and $\Sigma^+ = \mathbf{0}$ otherwise. Then Σ^+ solves (11). We next solve (12). Define

$$\xi = \frac{1-p^i}{(1+r)\left(\binom{k+1}{2} + k^2 r\right)(1-p)}\sigma^+_{11}.$$

Then the solution matrix Σ^0 is as follows:

$$\Sigma^0_{12}(1,m) = -r\xi\left[\frac{i}{j}\left(\binom{k+1}{2} + k^2 r\right) + \frac{k^2 rm(R-1)}{j^2}\right];$$

$$1 \le m \le j,$$

$$\Sigma^0_{12}(\ell,m) = p^{\ell-1}(R-1)r\xi\left[\frac{1}{j}\left(\binom{k+1}{2} + k^2 r\right) - \frac{k^2 rm}{j^2}\right];$$

$$2 \le \ell \le i, 1 \le m \le j,$$

$$\Sigma^0_{13}(1,m) = -\xi\left[\frac{i}{k}\left(\binom{k+1}{2} + k^2 r\right) + (R-1)(m+kr)\right];$$

$$1 \le m \le k,$$

$$\Sigma^0_{13}(\ell,m) = p^{\ell-1}(R-1)\xi\left[\frac{k+1}{2} - m\right];$$

$$2 \le \ell \le i, 1 \le m \le k,$$

$$\Sigma^0_{22} = \frac{(R-1)(1-p^i)}{1-p}\left(\frac{kr}{j}\right)^2\xi\begin{pmatrix} 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ \vdots & & & \end{pmatrix},$$

$$\Sigma^0_{23} = \frac{(R-1)(1-p^i)}{1-p}\frac{kr}{j}\xi\begin{pmatrix} 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ \vdots & & & \end{pmatrix},$$

$$\Sigma^0_{33} = \frac{(R-1)(1-p^i)}{1-p}\xi\begin{pmatrix} 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ \vdots & & & \end{pmatrix}.$$

(We are only interested in leading terms so Σ_{11}^0 is irrelevant because $\Sigma_{11}^+ \neq \mathbf{0}$.) To obtain these results was by far the most tedious part of the present work. They were derived by first using symbol manipulating software to get solutions for some special cases, then guessing the general formulas, and finally checking the formulas by inserting them in the equations. Now note that

$$\mathrm{Var}(U+Y) \approx N \sum_{\ell,m=i+1}^{i+j+k} \sigma_{ij} = N \frac{(R-1)(1-p^i)^2 k^2\,(1+r)\,\sigma_{11}^+}{\left(\binom{k+1}{2} + k^2 r\right)(1-p)^2} + O(\epsilon N),$$

thus it remains to find σ_{11}^+. We do this, not by solving (13), but by finding a criterion for (13) to be solvable. The matrix V given by

$$V_{22} = k^2 \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ 1 & 1 & 1 & \cdots \\ \vdots & & & \end{pmatrix}, \qquad V_{23} = V_{32}^T = \begin{pmatrix} & & & \vdots \\ \cdots & 3k & 2k & k \\ \cdots & 3k & 2k & k \\ \cdots & 3k & 2k & k \end{pmatrix},$$

$$V_{33} = \begin{pmatrix} & & & \vdots \\ \cdots & 9 & 6 & 3 \\ \cdots & 6 & 4 & 2 \\ \cdots & 3 & 2 & 1 \end{pmatrix},$$

and zeros otherwise solves the matrix equation

$$[(B^+)^T, V] = 0. \tag{14}$$

For arbitrary systems of linear equations we have that $A\mathbf{x} = \mathbf{b}$ and $A^T\mathbf{v} = \mathbf{0}$ implies $\mathbf{b} \cdot \mathbf{v} = 0$. If we regard (13) and (14) as such linear systems and apply this observation, the relation

$$-\sum_{\ell,m=1}^{i+j+k} \left(S^0 + [B^0, \Sigma^0]\right)_{\ell m} v_{\ell m} = 0$$

is obtained. This criterion boils down to the relation

$$\sigma_{11}^+ = \frac{\left(\binom{k+1}{2} + k^2 r\right)}{R(R-1)k^2 \left(\frac{1-p^i}{1-p}\right)^2 \left(\frac{k+2}{3k} + r + \left(1 + \frac{1}{j}\right)\frac{kr^2}{k+1} + \frac{p^i}{1-p^i}\left(1 + \frac{2kr}{k+1}\right)^2 \frac{k+1}{2k}\right)},$$

and the desired equation follows readily. Also this result is the product of clever guess-work, followed by verification of the guessed formula.

References

Anderson, D.A., Watson, R.K.: On the spread of disease with gamma distributed latent and infectious periods. Biometrika, **67**, 191–198 (1980)

Anderson, R.M., May, R.M.: Infectious diseases of humans; dynamics and control. Oxford University Press, Oxford (1991)

Bartlett, M.S.: Deterministic and stochastic models for recurrent epidemics. Proc. Third Berkeley Symp. Math. Statist. & Prob. Univ. California Press, Berkely and Los Angeles **4**, 81–109 (1956)

Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence. Wiley, New York (1986).

Halloran, M.E., Haber, M., Longini, I.M.: Interpretation and estimation of vaccine efficacy under heterogeneity Amer. J. Epidemiol. **136**, 328–343 (1992)

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

Karatzas, I., Shreve, S.E.: Brownian motion and stochastic calculus. 2nd edn., Springer-Verlag, New York (1991)

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

Nåsell, I.: On the time to extinction in recurrent epidemics. J. Roy. Statist. Soc. B **61**, 309–330 (1999)

Pollett, P.K., Roberts, A.J.: A description of the long-term behaviour of absorbing continuous-time Markov chains using a centre manifold. Adv. Appl. Prob. **22**, 111–128 (1990)