1 Introduction

Early in 1987, the AIDS epidemic was still in its relatively early stages in the United Kingdom, although well underway in the United States. In preparation for a one-day Royal Statistical Society meeting on the statistical aspects of AIDS (Royal Statistical Society, 1988) later that year, an informal workshop was organised by David Cox to discuss the various mathematical modelling approaches then being employed. At David’s instigation, I prepared a review paper (Isham, 1988) on models of the transmission dynamics of HIV and AIDS for the one-day meeting, based in part on the workshop, and thus began my involvement with epidemic modelling. It has proved to be an area full of theoretical challenges and complex applied problems. Understanding the dynamics of the spread of an infectious disease brings possibilities for its control. Human infections such as influenza, malaria and HIV are still major worldwide causes of morbidity and mortality. In 2001, the UK experienced the severe economic and social effects of an epidemic of foot and mouth disease, while early in 2003 the world became aware of a new infection, Severe Acute Respiratory Syndrome (SARS), spreading rapidly across the globe. In both cases, the ease and speed of modern communications meant that considerable transmission had already taken place before the threat of an epidemic was recognised. In such situations, strict control strategies must be imposed rapidly if they are to be effective in curtailing the spread of infection and preventing a large-scale epidemic. A clear understanding of the basic theory of epidemic models is a prerequisite for this and, equally, for informing the development of the successful vaccination policies against a wide range of infections (Anderson and May, 1991) routinely used around the world.

In this chapter, I will concentrate on a review of some of the topics that have preoccupied researchers working on models for the transmission of infection over the last fifteen years or so, looking at progress that has been made and identifying continuing challenges. The topics discussed will be a personal selection and the references given will be indicative rather than complete. I will concentrate on model structure and general modelling issues rather than on models for specific infections, although some of these will be used for illustration, and will

---

*This is a preliminary version of an article to appear as Chapter 1 of *Celebrating Statistics: Papers in Honour of Sir David Cox on his 80th Birthday*, edited by A. C. Davison, Y. Dodge and N. Wermuth, to be published by Oxford University Press in 2005.
focus almost entirely on models of infections for animal/human hosts and mention processes on regular spatial lattices, more relevant for plant diseases, only in passing. I will also touch only briefly on statistical aspects. It is interesting to look back a decade to the issues identified in the discussion paper of Mollison et al. (1994): heterogeneity, thresholds and persistence, nonstationarity, control. In spite of considerable advances, many of these are equally topical today and are destined to be the focus of research effort for some while to come. As is appropriate for this volume, my focus is on stochastic models. Those contributing to that Royal Statistical Society workshop in 1987 included Norman Bailey and Maurice Bartlett, both of whom had already played pivotal roles in developing the theory of stochastic epidemic models (Bailey, 1975; Bartlett, 1960). Nevertheless, at that stage, most applied work on specific infections relied on deterministic models, as described in the key text of Anderson and May (1991) and the later review of Hethcote (2000). One of the most noticeable changes of the last fifteen years has been the increased acceptance by very many biologists, not only of the role that mathematical modelling has to play in the solution of many of their most pressing problems, but also that such models will often need to incorporate intrinsic stochasticity. Thus, not only is there the long-acknowledged need to incorporate uncertainty in parameter values — generally achieved via a numerical sensitivity analysis — but, more importantly, the fact that the size of a population of infected hosts or of infecting parasites within a single host is random and integer-valued has many important consequences. These include effects on such issues as the possible fade-out and re-emergence of an infection, the prediction of the course of an individual realisation of an epidemic, and the determination of the appropriate period of application of a control treatment.

The purposes for which models are developed are many and varied. The suitability of a model and the appropriateness of the assumptions on which it is based depend entirely on its purpose. Models may be used for a careful exposition of issues and general understanding of transmission dynamics, to reach general qualitative conclusions, or for real-time use in a particular epidemic. They can also help in formulating what data should be collected in order to answer particular questions of interest. Simple models can lead to qualitative understanding and provide robust approximations. More complex models inevitably require more detailed assumptions about the nature of the underlying process and tend to have correspondingly more unknown parameters. Such models often require numerical solution or simulation, but can still allow those factors and sources of heterogeneity largely responsible for driving the dynamics to be distinguished from those having little influence. A particular strength of models is their ability to answer ‘what if’ questions. Thus they have an important role to play in determining control strategies, formulating ways to detect a future outbreak and developing contingency plans to deal with such an outbreak — see, for example Keeling et al. (2003) in the context of foot and mouth disease. In recent years, drug resistance has become an important issue, and the evolution of such resistance, and means of its control, are major concerns (Austin et al., 1999). Epidemic models have an important role to play in underpinning work in this area.

In some applications, however, recognition of the complex structure of the contacts that may result in transmission of infection has resulted in numerical simulation of individual-based models, in which each host is represented separately. Such models are often wholly, or largely, deterministic and seek explicitly to represent the detailed complexity of real-world contacts, rather than subsuming less-important details in random variation. In a few cases, individual-based models are founded on very detailed observation of a particular population, although even then their construction may only depend on marginal rather than joint properties of the population. Nevertheless, in the context of a specific infection, such models may give useful information about the potential spread and control of the infection in that population, although that information is unlikely to be applicable in any other situation. More often, though, individual-based mod-
els require vast numbers of poorly substantiated assumptions, especially concerning parameter values, to which the robustness of the results obtained is uncertain. In addition, such models generally convey little insight into which mechanisms are the most important in driving the transmission of infection, and those to which the dynamics are relatively insensitive. The claim is that such models are needed because of their ‘realistic’ assumptions, in contrast to the simple population structure and within-subgroup homogeneous mixing that underlie most population models. It is therefore vitally important to explore how model properties are affected by population heterogeneity and structure, and in recent years much effort has been directed at addressing such questions.

2 Historical background

Mathematical modelling of infectious diseases has a long history; see, in particular, Bailey (1975). The starting point is generally taken to be a paper by Daniel Bernoulli (Bernoulli, 1760) on the prevention of smallpox by inoculation; an account of Bernoulli’s model-based analysis of data can be found in Daley and Gani (1999). However, as Bailey (1975) points out, it was another hundred years or so before the physical basis for the cause of infectious diseases became well-established. Thus, the pace of progress only really picked up early in the 20th century, with the work of people such as Hamer (1906), Ross (Ross, 1911, 1916; Ross and Hudson, 1917a,b), and Kermack and McKendrick (1927), which established the principle of mass action or homogeneous mixing — by which the rate of new infections is proportional to the current numbers of susceptibles and infectives in the population — and the now-familiar deterministic equations for the general epidemic model. Homogeneously mixing models are also referred to as mean field models. Very many of the epidemic models in use today have this general epidemic (SIR) model, or its stochastic counterpart, as their basis and thus it can be used to illustrate many of the main issues. In the SIR model, if \( x(t), y(t), \) and \( z(t) \) represent, respectively, the numbers of susceptible, infective and removed individuals (i.e. those recovered but immune, isolated, or playing no further part in the transmission of infection for whatever reason) in the population at time \( t \), then

\[
\begin{align*}
\frac{dx(t)}{dt} &= -\alpha x(t)y(t)/n, \quad (2.1) \\
\frac{dy(t)}{dt} &= \{\alpha x(t)/n - \gamma\} y(t), \quad (2.2) \\
\frac{dz(t)}{dt} &= \gamma y(t). \quad (2.3)
\end{align*}
\]

The variables \( x, y \) and \( z \) are not restricted to integer values and thus it is only appropriate to use this deterministic model for large populations. In some versions of the model, the parameter \( \alpha \) above is replaced by \( n\beta \), but the formulation given here is more appropriate when an infection is spread by direct contact and an individual makes potentially infectious contacts at a fixed rate \( \alpha \) regardless of the population size. When \( n \) is fixed, the distinction is not important but confusion may arise if \( \beta \) is regarded as fixed if the limiting situation when \( n \to \infty \) is considered. Note however that the latter parameterisation is appropriate if the variables \( x, y \) and \( z \) represent population proportions. The parameter \( \gamma \) can be interpreted as the reciprocal of the mean infectious period. In the natural stochastic formulation of this model, infectious periods are independently and exponentially distributed with parameter \( \gamma \).

In the deterministic SIR model, the number of infectives grows as long as the proportion of susceptibles in the population exceeds \( \gamma/\alpha = 1/R_0 \), where the reproduction ratio, \( R_0 = \alpha/\gamma \), represents the mean number of effective contacts made by an infective during an infectious
period. If $R_0 < 1$, then this condition cannot be met and only a minor outbreak can result. Thus the target, in controlling a particular outbreak of infection, is to take steps to bring $R_0$ below one by reducing the rate of infectious contacts or the duration of the infectious period. With an infection for which $R_0 > 1$, a population will be protected from epidemic outbreaks as long as the proportion of susceptibles is kept below the threshold by vaccination; this effect is known as herd immunity. For an endemic infection, a population of fixed size, but subject to demographic changes and without an effective vaccination strategy, will be subject to recurrent epidemics.

Herd immunity. For an endemic infection, a population of fixed size, but subject to demographic changes and without an effective vaccination strategy, will be subject to recurrent epidemics when the proportion of susceptibles in the population builds up sufficiently. Estimates of $R_0$ are sometimes based on the final size of an epidemic, that is, the total number infected, $\lim_{t \to \infty} z(t)$. It is easy to show that this limit is the solution of the equation $n - z - x(0)e^{-zR_0/n} = 0$. Various approximations are then possible depending on assumptions about $R_0$, $x(0)$, and $n$. For example, if $R_0 > 1$ and $x(0) \approx n$ exceeds the threshold value $(n/R_0)$ by an amount $\nu$ then, approximately, as $t \to \infty$, the limiting value of $x(t)$ will be $\nu$ below the threshold giving a final size of $2\nu$. However, this approximation requires $\nu$ to be small, that is, $R_0$ close to 1; for large $R_0$, the final size behaves as $n(1 - e^{-R_0})$.

Stochastic epidemic models were also being developed early in the 20th century, alongside the deterministic ones, and McKendrick (1926) discussed a stochastic version of the general epidemic model. However, at that time, there was rather more interest in discrete-time models, and chain binomial models in particular, and further discussion of continuous time stochastic SIR models will be postponed to Section 3. The best-known chain binomial model was proposed by Reed and Frost in lectures given in 1928 (Wilson and Burke, 1942, 1943). However, their work had been preceded some 40 years earlier by En’ko in 1889 (Dietz, 1988, 1989). In the standard Reed–Frost model, given the numbers $S_t, I_t$ of susceptibles and infectives at time $t$, $S_{t+1}$ has a binomial distribution with index $S_t$ and mean $S_t(1 - p)^{K}$, and $I_{t+1} = S_t - S_{t+1}$. In effect, individuals are assumed to be infective for a single time unit and in that time can infect (independently, with probability $p$) any member of the population who is susceptible. This means that the number of potentially infectious contacts scales with the population size, although the Reed–Frost model is usually only applied to small populations, e.g., households, where this is not a problem. In contrast, in the En’ko formulation it is assumed that each individual makes a fixed number, $K$, of contacts at random with replacement from among the $n-1$ remaining individuals in the population, infection being transmitted if the person contacted is infectious, whereby $E(S_{t+1} | S_t, I_t) = S_t \{1 - I_t/(n - 1)\}^K$. Dietz and Schenzle (1985) discuss generalisations where $K$ is a random variable. For example, if $K$ has a Poisson distribution with mean $-(n - 1)\ln(1 - p)$, then the expected value of $S_{t+1}$ given $(S_t, I_t)$ is the same as for the Reed–Frost model, although the corresponding distributional relationship does not hold. More generally, if appropriate assumptions are made about the distribution of $K$ — essentially, that the number of contacts per unit time is suitably small — then

$$E(I_{t+1} | S_t, I_t) = S_t E \left[ 1 - \{1 - I_t/(n - 1)\}^K \right] \approx S_t I_t E(K)/(n - 1),$$

which corresponds to the homogeneous mixing assumption of the continuous-time SIR models described above.

An alternative modification of the Reed–Frost model is to assume that the probability that a susceptible escapes infection by a single, specified infective is $(1 - p)^{1/n}$, rather than $(1 - p)$. In this case, given $S_t, I_t$, the mean number of infectives at time $t+1$ is given by

$$E(I_{t+1} | S_t, I_t) = S_t \{1 - (1 - p)^{1/n}\} \approx S_t I_t p/n$$

for small $p$. Thus this modification is in a sense equivalent to the use of the factor $1/n$ in equation (2.2) for the continuous time deterministic SIR model, and allows comparison of epidemics in
host populations of different sizes. In particular, if two populations have identical proportions of infectives and susceptibles at time \( t \), then the expected proportions at time \( t + 1 \) will also be the same. If \( n \) is large and a single infective is introduced into a population of \( n - 1 \) susceptibles then, to a first approximation, \( p \) can be interpreted as the probability that there will be some spread of infection.

In the last 80 years, deterministic and stochastic epidemic models, in both discrete and continuous time, have been much studied. Together with the texts already cited (Bartlett, 1960; Bailey, 1975; Anderson and May, 1991; Daley and Gani, 1999), the following books between them contain a wealth of information on epidemic models, as well as providing an invaluable source of further references, covering many more topics than there will be scope to touch upon in this chapter: Becker (1989), Renshaw (1991), Mollison (1995), Grenfell and Dobson (1995), Isham and Medley (1996), Andersson and Britton (2000), Diekmann and Heesterbeek (2000).

3 Stochastic models for homogeneous populations

Any stochastic epidemic model has a deterministic counterpart, obtained by setting the deterministic population increments equal to the expected values of the conditional increments in the stochastic model. Thus, many stochastic models will have the same deterministic analogue. Useful insight into the stochastic model can often be gained from its corresponding deterministic model, as long as the population size is large. For example, in the stochastic SIR model, given the current state \((X(t), Y(t), Z(t))\) specifying the numbers of susceptible, infective and removed individuals respectively, we have

\[
E\{Y(t+dt) - Y(t) \mid X(t) = x(t), Y(t) = y(t), Z(t) = z(t)\} = \{\alpha x(t)/n - \gamma\}y(t)dt + o(dt) \tag{3.1}
\]

corresponding to the deterministic equation (2.2) for \( y(t) \). However, it follows from (3.1) that

\[
\frac{dE\{Y(t)\}}{dt} = \left\{\frac{\alpha E\{X(t)\}}{n} - \gamma\right\}E\{Y(t)\} + \frac{\alpha}{n}\text{cov}\{X(t), Y(t)\}, \tag{3.2}
\]

from which it is clear that the solution of the deterministic equations is not simply the mean of the stochastic process. This is an inevitable consequence of the nonlinearity of the transition rates of the stochastic model. Nevertheless, for a broad class of processes, the deterministic solution is a good approximation to the stochastic mean of a major outbreak when \( n \) is large.

A thorough discussion of the properties of deterministic and stochastic versions of the SIR model is given by Bailey (1975); valuable accounts and updates for many model variants can be found in Lefèvre (1990), Andersson and Britton (2000) and Diekmann and Heesterbeek (2000). Even for these simple models, explicit results are often difficult to obtain, and much effort and ingenuity has been expended in deriving methods of solution and approximations. In particular, the threshold behaviour of the stochastic SIR model for large populations (Whittle, 1955; Bailey, 1975; Andersson and Britton, 2000) is broadly analogous to that of the deterministic model described in Section 2. If the initial number of infectives is small, essentially all contacts of infectives are with susceptibles and a branching process approximation is appropriate. Thus, if \( R_0 \leq 1 \), there will be only a small outbreak of infection and the distribution of the final size of the epidemic will be J-shaped, i.e. with a mode at or close to zero and decreasing monotonically thereafter. The time to extinction of the infection will be \( O(1) \) as \( n \to \infty \). If \( R_0 > 1 \) and \( I_0 \) denotes the initial number of infectives, then with probability \( 1 - (1/R_0)^{I_0} \) the branching process explodes, corresponding to a major outbreak of infection. As the number of infectives builds
up, the approximation ceases to be appropriate and a central limit effect takes over. In this case, the final size distribution will be \( U \)-shaped, i.e. bimodal, and the time to extinction will be \( O(\log n) \) as \( n \to \infty \) (Barbour, 1975). The central limit effect can be examined formally by looking at a sequence of models where the initial proportions of susceptibles and infectives are kept fixed, so that the process is kept well away from the boundary of the state space, as the population size \( n \) increases to infinity. In this case, the SIR process tends to a Gaussian diffusion about the deterministic solution. Both the means and the variances of the numbers in the three states scale with the population size, and thus the proportions tend to the deterministic solution with probability one. These results were first obtained by Whittle (1957) and later extended (Daley and Kendall, 1965; Kurtz, 1970, 1971, 1981; Barbour, 1972, 1974; Daniels, 1991) to a very general class of density-dependent Markov jump processes. This threshold behaviour at \( R_0 = 1 \) is, in fact, an asymptotic result for large populations, with a major outbreak of infection being essentially one in which the size of the epidemic can be arbitrarily large. Many authors (e.g., Bailey (1975) using model simulations) have observed that threshold values strictly larger than one are appropriate for finite populations. In a discussion of this issue, Näsell (1995) suggests defining the threshold as the value of \( R_0 \) for which the distribution of the total size of the epidemic switches from J-shape to U-shape. Based on numerical work, he conjectures that this threshold has the form \( R_0 = 1 + O(n^{-1/3}) \).

In the past, the view seems to have been widely held that a deterministic model gives the mean behaviour of the corresponding stochastic system, at least asymptotically, and that for large populations there is therefore little to be gained from using a stochastic model which will generally be more difficult to analyse. However, it is now widely accepted that both deterministic and stochastic models have their strengths and can provide insight and understanding. Even with large populations, chance fluctuations do not always average out to have rather little overall effect. But even when they do, it may be important to take the variability of individual realisations into account, for example in predicting the course of an individual outbreak (Isham, 1991, 1993a). The latter paper discusses variants of the SIR model used in connection with the AIDS epidemic: as both means and variances scale with the population size \( n \), the variation about the deterministic limit is of order \( \sqrt{n} \), so that there is considerable uncertainty about predicted values, even apart from the effects of parameter uncertainty. Similarly, stochastic effects can be important in determining a control strategy where some measure is to be implemented ‘until the epidemic is over’. Once numbers of cases are small, there is considerable uncertainty in the residual time to extinction that cannot be conveyed by a deterministic treatment. Stochastic effects also play an essential role in questions of recurrence and extinction (or fade-out) of infections. As already noted, with a deterministic epidemic model governed by a set of differential equations, an open population with recruitment of susceptibles will be subject to recurrent epidemics if initially exposed to a source of infection for which the reproduction ratio \( R_0 \) exceeds unity. The infection never wholly dies out and can regenerate from arbitrarily small amounts of residual infection. Thus, for example, tiny fractions of an infected fox — the ‘attofox’ of Mollison (1991) — can generate repeated waves of rabies. In contrast, a stochastic epidemic that reaches a state in which only a single infective remains may well soon die out completely, and is in any case certain to do so eventually, unless there is a continuing external source of infection.

The attention on fade-out has focussed mainly on childhood infections, especially measles. Measles is a seasonal infection driven largely by the school year, and time series of pre-vaccination data typically show an annual cycle together with major outbreaks recurring at rather regular intervals of somewhere between two to five years. In small populations the infection often goes extinct. Much of the early work seeking to explain these recurrent epidemics is due to the semi-
inal work of Bartlett (Bartlett, 1949, 1956, 1957, 1960), who introduced the idea of a critical community size below which an infection will rapidly go extinct unless it is reintroduced from outside; in a data-analytic investigation he describes the critical size as that for which there is a 50:50 chance of fade-out following a major outbreak. A more recent discussion of the role of community size in measles fade-out is given by Keeling and Grenfell (1997). In theoretical studies, terms such as ‘fade-out’ and ‘major outbreak’ need formal definition. For example, Nåsell (1999) considers an open SIR model (i.e. with demography) and looks at the population size for which the expected time to extinction from the quasi-stationary distribution (see below) takes a fixed value. In comparing theoretical results with empirical data, however, allowance must be made for complicating features of real infections, such as population structure, vaccination coverage and the durations of disease stages --- the assumed exponential distributions of simple models are seldom appropriate.

In a basic measles model, an extra latent stage is usually included in the SIR model to give an SEIR model. Those in the latent (i.e. exposed) class are infected but not yet infectious. Demography is also included, and a steady-state population is generally assumed, achieved via equal per capita birth and death rates applying to all members of the population. In a deterministic version of the model, using reasonably realistic parameter values, the number infected oscillates about a stable point but, in contrast to the data, these oscillations are damped. In addition, Schenzle (1984) commented that it is difficult to find a set of parameters that gives an acceptable fit simultaneously to all aspects of an epidemic. Damped oscillations were first noted by Soper (1929) in a model variant with no latent class and an exponentially growing population since no deaths were included in the model. Although this model is in one sense rather unrealistic, it gives useful qualitative results, and the corresponding stochastic formulation has been used by Bartlett and many subsequent authors. In particular, Bartlett (1956) showed by simulation that the stochastic formulation of the model does lead to undamped oscillations, at a time scale that depends on population size, and to the possibility of fade-out. An alternative resolution within the deterministic framework is to assume that the contact rate $\alpha$ varies seasonally, see for example Dietz (1976), which can also generate the required undamped oscillations. Nevertheless, it is generally accepted that grouping children into age bands has an important effect over and above the timing effects of the school year, and thus Schenzle (1984) used an age-structured model, explicitly incorporating such effects, successfully to reproduce the stable biennial cycle observed in data for England and Wales.

Measles data for large populations show an apparently chaotic behaviour (Schaffer, 1985), due to interaction between the natural periodicity of the nonlinear dynamics and the seasonal periodicity of the contact rate (Bartlett, 1990). In this respect, it is important to note that the level of spatial and/or temporal aggregation of the data will have a considerable effect on the observed properties. For a while, the chaotic properties of deterministic measles models were the focus of considerable activity; for a review of some of this work see Isham (1993b). However, as already noted, demographic stochasticity plays an essential role in the recurrence and extinction properties of epidemic models. Recently, Nåsell (2002b) has carefully investigated deterministic and stochastic versions of an SEIR model, with a periodic contact rate and allowing reintroduction of infection from an external source, using a mixture of analytic approximation and simulations. He concludes that stochasticity is necessary for acceptable representation of recurrent epidemics. In particular, as is already well-known (Grenfell, 1992) the high level of seasonal forcing needed if the SEIR model is to be chaotic gives an unrealistically low incidence of infections in the troughs between the epidemics. In this work Nåsell also shows how the amount of damping in the oscillations of the nonseasonal deterministic model reflects the variability in the times between outbreaks in the corresponding stochastic model. In an alternative
approach (Finkenstädt et al., 2002) measles data for England and Wales are explored by use of a model that is a combination a mechanistic stochastic SIR epidemic model and a statistical time-series. The model runs in discrete time and uses an algorithm to reconstruct the size of the susceptible class from information on births and cases. The number of infectives at time \( t + 1 \) then has a negative binomial distribution whose expectation is a function of the numbers of susceptibles and infectives at time \( t \), including an additional Poisson-distributed number of immigrant infectives, and is proportional to \( X^\theta Y^\psi \). The constant of proportionality is allowed to vary seasonally and the parameters \( \theta \) and \( \psi \) allow possible departures from homogeneous mixing, that is \( \theta = \psi = 1 \). It is found that the model, when fitted to data, can successfully reproduce the dynamics of measles outbreaks over a range of population sizes from small towns to large cities, with a seasonal transmission rate that closely corresponds to the effects of the school year.

Many of the issues touched upon above in the context of stochastic models for measles present substantial theoretical challenges more generally. In particular, current research continues to be concerned with important questions on thresholds for invasion — under what conditions does a small amount of initial infection invade an almost entirely susceptible population? — and persistence — if it does so, what is the necessary condition for the infection to persist and become endemic? These questions have been addressed with reference to a number of different models — see Näsell (2002a), for example — and various properties have been used to examine threshold behaviour. In a population with no immigration of infectives after the initial infection, an endemic situation can result for long time periods, even when eventual extinction is certain, as long as there is replacement of susceptibles. This can be achieved through births of susceptibles, as in the open SIR model discussed in the context of measles above, or in a closed population if individuals can be infected more than once, as for example in an SIS model, where infectives return to the susceptible state on recovery instead of becoming immune. The latter model is mathematically more tractable although generally less obviously realistic. In models such as these where eventual extinction is certain, the distributions of the time to extinction from a variety of initial conditions have been studied. From any particular initial state, essentially there are two qualitatively different possibilities, either the infection will die away rather quickly, or it will reach a state of quasi-stationary equilibrium — that is, an equilibrium distribution conditional on extinction not having occurred (Darroch and Seneta, 1967) — about the endemic equilibrium of the deterministic model, in which it will remain for some random time, before eventually a sufficiently large random fluctuation will lead to extinction.

The certainty of eventual extinction for the stochastic SIS model contrasts with the deterministic version of the model, where if \( R_0 > 1 \), any initial infection, however small, will invade a susceptible population and reach an endemic equilibrium. Nevertheless, various other properties of this model do exhibit threshold behaviours. For example, starting from a single infective in an otherwise wholly susceptible population, it is easy to see that there will an initial growth in the expected number of infectives if \( R_0 > 1 + 1/(n - 1) \), which could be thus be defined as a threshold value (Jacquez and Simon, 1993). Alternatively, the time to that extinction also exhibits a threshold behaviour. Andersson and Djeulic (1998) consider an initial condition in which \( I_0 \) goes to infinity with \( n \), and show that the time to extinction is asymptotically exponentially distributed with a mean that grows exponentially with \( n \) if \( R_0 > 1 \). While this same behaviour may result if \( R_0 > 1 \) but \( I_0 = O(1) \), it is no longer certain. When \( R_0 \leq 1 \), the time to extinction stays small or behaves as \( \log n \) as \( n \to \infty \) depending on the initial number of infectives. Näsell (1996, 1999) discusses approximations for the mean time to extinction when the initial distribution is the quasi-stationary distribution and also shows that the quasi-stationary distribution itself exhibits a threshold behaviour, being approximately geometric for \( R_0 < 1 \) and
approximately normal when $R_0 > 1$; see also Clancy and Pollett (2003).

The properties of the SIR model with demography have similarly been much studied — see Bailey (1975) and Andersson and Britton (2000), for example — although results are rather more difficult to obtain than for the SIS model discussed above; this is especially true of the time to extinction, although approximations are available (Nåsell, 1999). As expected, in the limiting $n \to \infty$ case and assuming that the initial proportion of infectives scales with $n$, there is a threshold at $R_0 = 1$ below which an infection dies out rather rapidly and above which it increases to an endemic level with Gaussian fluctuations about that level.

In the last few paragraphs, some recent results to do with invasions and persistence, times to extinction, thresholds and critical community sizes have been outlined. There are many other questions that could be asked involving different models or different initial conditions. For example, one might be interested in the probability that an infection becomes extinct, or the further time in which it does so, conditionally upon a major outbreak having already resulted from the introduction of a single infective. A particular challenge for the future is to extend results such as those described above, to allow seasonally varying parameters or structured populations where assumptions of homogeneous mixing do not apply (see Section 4). If individuals are of several types depending, for example, on behavioural or spatial characteristics, with different rates of contact within and between groups, how does this affect the time to extinction? As mentioned earlier, this is particularly relevant for the case of the recurrent epidemics such as measles.

The models described in this section have all been variations on an SIR theme, with a simple and relatively tractable mathematical structure. In particular, they assume that the hosts are identical and homogeneously mixing; models for heterogeneous populations will be discussed in the next section. They also assume that the times that an individual spends in distinct states — latent, infectious, and so forth — are exponentially distributed. In order to make models more nearly reflect the realities of disease progression it is important to allow alternative distributions with a non-zero mode; for example, Keeling and Grenfell (1998), Lloyd (2001) and Keeling and Grenfell (2002) discuss the effect on the persistence of measles. In some cases the variation of the distribution may be very small, as for the incubation period for childhood diseases such as measles or chicken pox, while in others, such as the period between HIV infection and AIDS diagnosis, a very wide range of values is possible. The usual ‘forward equation’ approach for Markov models can be used to cover variable rates of disease progression if the states of the system are extended to include information about the elapsed time from infection; the gamma distribution — via Erlang’s ‘method of stages’, see Cox (1955), for example — and the Weibull distribution are especially tractable possibilities algebraically. In addition, explicitly incorporating elapsed time from infection into the models allows the rate of transmission to depend on this time, as is appropriate for HIV for example, where infectives are thought to be at their most infectious soon after infection and then again later when developing the symptoms of AIDS. However, an alternative and very elegant approach is due to Sellke (1983); see also Andersson and Britton (2000). In this, an individual becomes infected when the total infection pressure $(\alpha/n) \int_0^t Y(t)dt$ — in which infectives are weighted by their infectious periods, which can have an arbitrary distribution — reaches an exponentially distributed level. More generally, Clancy (1999a,b) uses martingale methods to extend results on the SIR model, allowing a broad class of infection and removal rate functions, and discusses optimal strategies for intervention via isolation/or and immunisation.
4 Heterogeneity

4.1 Sources of heterogeneity

So far, we have concentrated almost entirely on models for populations of identical, homogeneously mixing hosts. However, for applied purposes, it is often vital to incorporate sources of heterogeneity, even though the aim is a simple, parsimonious model. It is helpful, then, to distinguish intrinsic heterogeneity of the hosts, for example a variation in susceptibility due to genetics or immune status, from heterogeneity of mixing whereby the rate of contact between a pair of hosts depends on their spatial separation or relative positions in a social network. In the next subsection, we concentrate on intrinsic host heterogeneity, postponing discussion of heterogeneous mixing to Section 4.3 which concentrates on structured populations in which mixing within local groups takes place at a higher rate than with the rest of the population, and Section 4.4 where mixing determined by contact networks is discussed. Models are almost always temporally homogeneous, although for endemic diseases and epidemics over very long time periods it may be necessary to allow for changing behaviour and contact patterns in the host population, the introduction or alteration of control policies, and so forth. Although it is usually assumed that changes to the host population occur independently of the infection process, there is also the possibility that changes happen in response to the infection process itself. An additional source of heterogeneity, that will not be discussed in any detail here, comes from the infecting parasites themselves. There are many different co-existing parasite strains causing infections such as influenza, malaria and dengue, and it is necessary to take account of the interactions between these and the host's immunological system, as infection with one strain may confer subsequent immunity to other closely related strains. It has been shown, for example, that treating the strains of a malaria parasite species as indistinguishable can have misleading consequences for the transmissibility of the infection (Gupta et al., 1991, 1994). From a modelling perspective, parasite heterogeneity means that the state space of the process has to be substantially increased to include the past infection history of the hosts, and in any particular case the aspects of this history that are most important in determining the dynamics of the infection will need to be established. The increased complexity of these models means that most work in this area so far has been deterministic; an exception is the recent simulation study of influenza evolution by Ferguson et al. (2003).

4.2 Heterogeneity of hosts

Age is one obvious source of variation between hosts that may be relevant to transmission of infection and/or disease progression. Thus, for example, the period from infection with HIV to diagnosis of full AIDS is known to vary with host age (Billard et al., 1990). In such cases, it may be important to model the age structure of the population. Of course, it may also be that individuals mix differentially according to their age, as in the case of measles discussed earlier. For sexually transmitted diseases, hosts will need to be divided into male and female, and in some cases it will be necessary to distinguish couples in monogamous relationships and to model the changes in these partnerships over time. In general, such population structure is easily incorporated, by dividing the population first on the basis of relevant explanatory variables such as age and gender, and then subdividing these categories by infection status: susceptible, infective, and so forth. The infection states themselves may be further divided. For example, the degree of susceptibility may vary with the immunological status of the host, and a further
important area for future research is to link susceptibility to infection in epidemic models with
detailed models of the workings of the host’s immune system. It is simplest to assume that
the host heterogeneity can be described by discrete variables, although continuously varying
characteristics can also be considered if necessary. The overall framework is often essentially
that of the SIR model, but with a much larger state space. The large number of variables
and parameters may mean that model properties have to be determined numerically or via
simulation.

Most of the models allowing host heterogeneity focus on the complexities of a particular infection.
However, Ball and Clancy (1995) discuss a multitype model in which infectives are randomly
allocated a type on infection, and obtain asymptotic results for the final size and severity, that is,
the area under the trajectory of infectives, of the epidemic, as the initial population size increases.
A special case of this model, with just two types, allows the infection to be carrier-borne — that
is, a proportion of infectives develop symptoms immediately and are withdrawn from circulation,
while the others remain symptom-free and able to transmit infection. Host heterogeneity is also a
feature of a class of models for contagion proposed by Dodds and Watts (2004). In these models,
which have a homogeneously mixing SIR basis, contacts take place at regular time points and
each infectious contact between an infective and susceptible results in a random ‘dose’ of infection
being transmitted, with the susceptible only being infected when the accumulated dose over the
last $T$ time points reaches some randomly assigned threshold. A more natural alternative might
be to include all past doses with an exponentially decaying weight. Individuals are able to
recover once this accumulated dose falls below the threshold. In an essentially deterministic
analysis, Dodds and Watts (2004) show that for a range of model specifications the resulting
dynamics of the system falls into one of three universal classes depending on the values of two
basic parameters: the probabilities that an individual will be infected as a result of one, and
two, infectious contacts.

A vast number of models have been constructed to describe the dynamics of specific infections,
both for human and animal infections, although until recently their analysis has mainly focussed
on deterministic approximations. For the most part, the simple homogeneous SIR model has to
be extended to allow for host heterogeneity and for extra infection states. An enormous amount
of work has gone into modelling sexually-transmitted diseases, such as gonorrhoea, hepatitis B,
and especially HIV and AIDS in various homosexual, heterosexual, and intravenous drug-using
communities. As well as measles, there are many other childhood diseases, such as mumps,
chicken pox (varicella), rubella, and pertussis, and infections such as influenza and tuberculosis
that cause considerable morbidity and mortality, particularly in developing countries. Models
for smallpox have received recent attention because of bioterrorism concerns. In the case of
tropical diseases such as malaria, schistosomiasis and onchocerciasis, there is no direct host-
host transmission of the parasite, and models must include complicated interactions between
parasites, human hosts, secondary hosts and intermediate vectors. For malaria, it is the process
of biting by the mosquito vector that must be modelled; for onchocerciasis, the vector is a
blackfly; for schistosomiasis, the schistosome fluke spends part of its life-cycle in a second host
(a snail) with infection transmitted between the humans and snails through water contact.

Both schistosomiasis and onchocerciasis are macroparasitic diseases, which adds an extra comp-
lication to the models. All the infections that we have discussed so far have been of micropa-
rasites (viruses and bacteria are familiar examples), which replicate within the host and rapidly
reach an equilibrium level, so that it is appropriate to describe the state of the host population
by the numbers of susceptibles, infectives, and so forth. This equilibrium level may depend
on properties of the hosts such as the competence of their immune systems, in which case it
will be appropriate to include such sources of host heterogeneity in the model. Macroparasites such as nematodes, on the other hand, have part of their life cycle outside the host, so that the parasite load only builds up through reinfection. Both this reinfection process, the individual parasite burdens of the hosts, and the external parasite population must be incorporated in a full stochastic model, which will involve many complex dependencies and nonlinearities. Thus most models for macroparasite infections concentrate on a particular part of the process, while making drastically simplifying assumptions about other aspects. One approach is to use a hybrid model (Nåsell, 1985), in which some part of the process, for example the dynamics of the host population (Kretzschmar, 1989b,a), is replaced by a deterministic approximation while allowing for random variability of the hosts’ parasite loads. An alternative approach is not to model the parasite population external to the hosts, but to assume that infections of hosts occur in a non-homogeneous Poisson process that is independent of the parasite loads of the hosts. This may be appropriate, together with an additional simplification — to ignore the demography of the host population — for a managed animal population early in the season. In this case (Herbert and Isham, 2000; Cornell et al., 2003) many other features of the physical process such as clumped infections, immune responses affecting parasite establishment, fertility and mortality, and non-exponentially distributed durations of parasite stages, can all be incorporated in an analytically tractable model. Other approximations have also been discussed; for example, infection may be assumed to be transmitted directly between hosts, using a branching process to approximate the initial stages of the spread of infection between identical (and immortal) hosts (Barbour and Kafetzaki, 1993; Barbour et al., 1996). In models of specific diseases, interest generally focusses on control issues, the effects of prospective vaccination strategies or treatment interventions, and sometimes on short-term prediction. This presupposes that the mechanisms controlling the dynamics of the particular infection are already well understood, so that the model can be used as a firm basis for such investigations, but this may not be the case; for example, the age-intensity profiles of parasite loads from cross-sectional data on schistosomiasis (Chan and Isham, 1998) show that the load builds up during the childhood to reach a maximum before decreasing with age. To what extent is this due to a protective acquired immunity rather than to behaviour changes? For example, children may spend more time in/near the snail-infested water than adults. In onchocerciasis, the parasite loads increase with age and there is a debate over the opposing roles of protective acquired immunity and parasite-induced immunosuppression, with experimental investigations giving contradictory findings. Duerr et al. (2002) use a model-based analysis to support the importance of immunosuppressive mechanisms. More generally, Duerr et al. (2003) compare the effects of a range of different mechanisms — including age-dependent exposure, parasite-induced host mortality, clumped infections and density-dependent effects on the parasites — on age-intensity profiles and dispersion patterns. The latter are of interest because macroparasite loads are generally overdispersed with most hosts having few parasites, and a few hosts having large numbers of parasites. An additional concern and challenge for the future is that almost all epidemic models are concerned with a single infection, and while this may be appropriate for many short-lived microparasitic infections, in other cases between-species immunological interactions may be important. Thus, the susceptibility of a particular host to infection or reinfection by a specific parasite may depend not only on the host’s immune response to their infection history with the same parasite species, but also to that with other species. Cross-species immune responses may be immuno-suppressive or protective: Behnke et al. (2001) discuss both types of response for multi-species helminth infections, while Bruce et al. (2000) provide evidence for cross-protective interactions between different malaria species. As yet there is little or no stochastic modelling of such mechanisms.

While analysis of the deterministic version corresponding to any stochastic model is a useful
starting point, less stringent approximations can be helpful in providing information about variability and are often straightforward even for relatively complicated processes. For simplicity, consider the closed SIR model, which is a bivariate process. The nonlinearity of the model means that the forward equations for the expected numbers of infectives and susceptibles (see equation (3.2) for the former) involve second order terms; setting the covariance terms in these equations to zero recovers the equations for the deterministic approximation. Similarly the forward equations for second order moments involve third order terms, and so on. In the method of moment closure this set of equations is closed by imposing a suitable relationship between these moments. For example, for the SIR model one might choose to assume that the joint distribution of susceptibles and infectives is bivariate normal, thus obtaining a set of five moment equations to be solved (Isham, 1991, 1993a); these equations are precisely those that apply in the limiting Gaussian diffusion discussed in Section 3. Moment closure has been found to give acceptable approximations in other other cases too; for example Nåsell (2003) considers the quasi-stationary distribution of the stochastic logistic model, a finite state birth-death process. The success of the moment closure method can sometimes be attributed to central limit effects, as in the case of the density-dependent Markov processes. However, the method can often work surprisingly well for small populations, and close to the boundary of the state space (Isham, 1995); Stark et al. (2001) explore the connection with the existence of inertial manifolds as a possible explanation for this. Distributions other than the multivariate normal may also be used to suggest suitable relations between the moments: for example, in macroparasite infections, the parasite load is often well approximated by a negative binomial distribution. A problem is that once one moves away from the multivariate normal distribution, it may be difficult to find a joint distribution with the right marginals, but where the covariances are not inappropriately constrained; see Herbert and Isham (2000) for a discussion of a negative binomial approximation. Alternatively, Keeling (2000) suggests an approximation based on standardised so-called multiplicative moments such as $V^*, T^*$ where $E(X^2) = [E(X)]^2 V^*$, $E(X^3) = [E(X)]^3 V^{*3} T^*$ and so forth, which he finds gives good results in cases where the multivariate normal approximation is unacceptable — for example, it can give negative mean population sizes. For a univariate model, this approach is equivalent to assuming a lognormal distribution.

Space constraints mean that it has not been possible to refer to even a tiny proportion of the models on specific single strain, multistrain and multispecies infections here; Anderson and May (1991) and collections of workshop papers such as Gabriel et al. (1990) and Isham and Medley (1996) provide a convenient general starting point for models of many infections, but a quick scan through recent journals or an internet search will generate a wealth of fascinating examples. Historically, the complexities of specific infections has led to a focus on the use of deterministic models, but a greater understanding of the importance of intrinsic stochasticity together with considerable advances in computing power means that the use of stochastic models is becoming increasingly feasible. However, this does not necessarily mean resorting to detailed stochastic simulations. There is also a role for toy models to illuminate particular aspects of a complex process as well as for a combination of analytic approximations and numerical evaluation to investigate properties of interest. A single example will have to suffice. A simple model (Herbert and Isham, 2000) for the early-season dynamics of a macroparasite infection in a managed age-cohort of hosts was mentioned earlier. The stochastic feedback of the current parasite population via the reinfection process of the hosts can be simulated (Cornell et al., 2004), and shows complicated interactions between stochastic effects, including the need for the parasites to mate within hosts, and spatial effects governed by the size of the host population. A toy model — a bisexual Galton-Watson metapopulation process (Cornell and Isham, 2004), whose properties can be explored using a mixture of algebraic approximations and numerical
4.3 Heterogeneity of mixing

In all the models described so far, apart from the age structured measles models mentioned in Section 3, homogeneous mixing has been assumed, with all members of the host population equally likely to be in potentially infectious contact. For the pre-vaccination measles data discussed earlier, the inclusion of age structure in the models plays an important role. However, most measles cases are in young children, with transmission taking place at pre-school, school or in the family, and at school children mix predominantly with in their own age group. Thus age may be a surrogate for social structure in the population and an alternative strategy is to model this structure directly. A general approach is then to divide the host population into groups (or types), where it is assumed that hosts mix homogeneously within each group. Contacts between groups are modelled by use of a mixing matrix whose \((i, j)\)th element \(p_{ij}\) specifies the probability that a host in group \(i\) will have an potentially infectious contact, \(i.e.\) one in which transmission will take place if one host is infected and the other is susceptible, with a host in group \(j\). This group \(j\) host is chosen at random from members of the group. Then, for example, if there are \(X_i(t)\) susceptibles and \(Y_j(t)\) infectives in group \(i\) at time \(t\), and all individuals make such contacts at rate \(\alpha\), the total rate at which infections take place in group \(i\) is \(\alpha X_i \sum_j p_{ij} Y_j / n_j\), where \(n_j\) is the size of group \(j\). As usual in such models, the chance of a within-group infection is taken to be \(p_i Y_i / n_i\) rather than using the more realistic factor \(n_i - 1\) in the denominator — a negligible approximation in most cases. Clearly the numbers of \(i-j\) contacts must balance, so the elements of the mixing matrix must satisfy the constraints \(n_i p_{ij} = n_j p_{ji}\). Various structures for the mixing matrix have been proposed, including proportional mixing where \(p_{ij} = n_j / \sum_l n_l\), which is just homogeneous mixing of the whole population; restricted mixing where \(p_{ij} = \delta_{ij}\) \((\delta_{ij} = 1\) if \(i = j\), \(\delta_{ij} = 0\) otherwise\) so that individuals only mix within their subgroup; preferred mixing where \(p_{ij} = \epsilon \delta_{ij} + (1 - \epsilon) n_j / \sum_l n_l\), where individuals have a reserved proportion of contacts with their own subgroup and the rest are spread proportionately through the whole population; and various more complicated possibilities such as mixing schemes that model the social or geographical locations in which contacts take place (Jacquez et al., 1989; Koopman et al., 1989).

In some cases the mixing matrix is estimated from empirical data. It is then relevant to note that if demography is included in the model so that the group sizes, \(n_j\), are not fixed, the constraints on the mixing matrix mentioned above mean that its elements will vary with time; one possibility is then to model \(n_i p_{ij}\), using a log-linear model for example (Morris, 1991, 1996).

It is straightforward to generalise the above mixing structure to allow the distinct subgroups to have different contact rates \(\alpha_i\) \((\alpha_i \neq \alpha_j)\). The constraints on the elements of the mixing matrix are then \(n_i \alpha_i p_{ij} = n_j \alpha_j p_{ji}\), with corresponding changes to the various special cases, \(e.g.,\) proportional mixing corresponds to \(p_{ij} = \alpha_j n_j / \sum_i \alpha_i n_i\), so that it is the ‘contact’ that is chosen at random rather than the person contacted. The variability of contact rates is particularly important for sexually transmitted diseases where the host population may consist of large numbers of individuals with low contact rates and a rather smaller number with very high rates. In this case, the spread of infection is determined not just by the average contact rate over the population but by its variability (Anderson et al., 1986). For example, Jacquez and Simon (1990) show that, for a deterministic model with proportional mixing, the rate of growth of infectives at the start of the epidemic is determined by \(\mu_\alpha + \sigma_\alpha^2 / \mu_\alpha\) where \(\mu_\alpha\) and \(\sigma_\alpha^2\) denote respectively the mean and variance of the contact rate over the population.

An important concern of both theoretical and practical interest is how to generalise the concept of
the reproduction ratio, $R_0$, for general heterogeneous and structured populations. The definition of $R_0$ was first put on a firm footing by Heesterbeek (1992), see also Diekmann and Heesterbeek (2000). Essentially, $R_0$ is the spectral radius (dominant eigenvalue) of the first generation matrix, which specifies the expected numbers of type $j$ infections directly caused by a type $i$ infective; the formulation allows hosts to move between types during their lifetimes and applies equally with heterogeneity of hosts due to disease status or immune level, and with heterogeneity of host mixing. As for a homogeneous population (see Section 3), threshold results for a stochastic model are less straightforward than for its deterministic counterpart. The condition for invasion in a deterministic model is that $R_0 > 1$, but for stochastic models the threshold results are rather more complicated. Keeling and Grenfell (2000) discuss the effect of population structure, and in particular the number of neighbours of each individual, in the context of a simple stochastic SIR-type model in a closed population. As expected, the smaller the size of the neighbourhood, the higher the threshold for invasion.

Much of the theoretical work on epidemic models for metapopulations is due to F. Ball and his collaborators. The emphasis is on models for small groups (households) within a large population, where the assumed mixing structure is relatively simple; properties such as the final size and severity of an epidemic, and asymptotic results when the number of households becomes large, are of particular interest. One important distinction is whether infection spreads between groups through the movement of infectives, infection then usually being restricted to susceptibles within the same group, or whether the group membership is fixed but infectious contacts occur between groups. For example, in a series of papers (Ball and Clancy, 1993; Clancy, 1994, 1996), the model includes movement of infectives between groups, with the rate of contact of the infective with a susceptible depending on the infective’s original and current groups and the group of susceptible. A simplifying special case is then to assume that infection is only transmitted within groups. An important example of the alternative approach is provided by Ball et al. (1997) who consider epidemics with two levels of mixing: local and global. Individuals mix mainly within their own (local) group, but have occasional contacts with the rest of the population. The key result is that the reproduction ratio for the epidemic is then simply a product of two factors: the reproduction ratio for the homogeneously mixing epidemic when the local contact rate is set to zero, and an ‘amplification factor’ $A$, say, which is the mean size of an epidemic started by a single randomly chosen infective when the global contact rate is set to zero. Thus if $A$ — and hence necessarily the group sizes — are sufficiently large, a substantial outbreak can result even if the level of global contacts is extremely low. This result is not restricted to when the groups form a partition of the population, and applies equally when individuals have overlapping sets of local neighbours, a scenario explored in more detail in Ball and Neal (2002). It can be seen that $A$ needs to be reduced to control the spread of infection, and a vaccination strategy that aims to reduce the number of each individual’s unvaccinated neighbours to be as near as possible equal is intuitively optimal. In more recent work, Ball and Lyne (2002) look at optimal vaccination strategies for a household model, where the costs of the vaccination programme are included. In a further level of generality, Ball and Lyne (2001) consider a multitype household model where the infection rate depends on the types of the infected and susceptible individuals as well as on whether the contact is a local or global one. Most of the work on household models considers SIR-type epidemics, but Ball (1999) investigates the spread of an SIS model in a population with a large number of small households, determining an appropriate threshold parameter for invasion in this case, while Arrigoni and Pugliese (2002) obtain limiting results when both the household size and their number go to infinity.
4.4 Spatial models

A particular challenge in epidemic modelling is the appropriate way to allow for spatial population structure, whereby the rate of contact between hosts depends on their spatial separation. The hierarchical framework of the metapopulation models described above can be used as a surrogate for spatial structure. In the case of measles, outbreaks seldom take place in isolated communities, and it is often important to look jointly at spatially coupled epidemics (Grenfell, 1992). In general, the chance of extinction of infection from the whole population decreases as the mixing between the subpopulations is reduced, due to the so-called ‘rescue’ effect. However, this effect will be lessened if the epidemics in the sub-populations are nearly in phase due to a strong degree of seasonal forcing. One of the earliest explicitly spatial models for the spread of infection was a metapopulation model for the spread of influenza, first in the USSR and later globally (Rvachev and Longini, 1985), where data on the transportation networks were used to provide information about migration routes. The more general setting of Ball et al. (1997) in which there is no partition of the population into fixed groups, but in which each individual has their own set of neighbours, is particularly appropriate in a spatial context. In small worlds models (Watts and Strogatz, 1998; Newman et al., 2002a), individual hosts have some simple spatial structure, with most contacts reserved for spatially neighbouring hosts and just a few long distance contacts. Such models are highly relevant to epidemics characterised by local spread plus a few large jumps, as occurred with the recent outbreaks of foot and mouth disease and SARS. The great circle model, in which hosts are equally spaced round a circle, is used as a simple illustrative example by Ball et al. (1997), and further explored by Ball and Neal (2003) who derive asymptotic results including a threshold theorem for invasion to occur and a central limit theorem for the final epidemic conditionally upon its occurrence.

In the metapopulation models, it is normally the case that any individual host may be directly infected by any other, even though the chance of this may reflect spatial separation and could be extremely small. Contact networks provide a way of representing a more complex population structure, where only hosts connected by edges of the network have any direct contact. These networks can reflect structure in geographical or social space and, traditionally, their study has been the province of social scientists who have generally been concerned with the collection of data and characterisation of empirical networks in relatively small populations. Social networks are particularly relevant to the spread of infections transmitted by direct contact — for example, sexually transmitted diseases (Johnson, 1996) — and study has focussed especially on HIV. However, in recent years mathematicians have become increasingly involved in network models. Two sorts of networks are discussed: directed graphs showing ‘who is infected by whom’ and undirected graphs representing the underlying contact structure of the population. Much theoretical work is devoted to the latter, and particularly to investigating the properties of simple random graphs (Bollobás, 1985) and of Markov random graphs (Frank and Strauss, 1986). In a simple random graph, an edge exists independently between each pair of vertices with probability $p$. In order that the degree of a vertex (i.e. the number of edges emanating from it) is well-behaved as the size of the network $(n)$ increases, it is usual to assume that $p = \beta/n$, so that for large $n$ the degree distribution is approximately Poisson. Alternatively, the degree distribution may be specified and/or estimated from empirical data; a graph can then be constructed by a random pairing of the arms from each vertex (Newman et al., 2002b). The problem when the sum of all the degrees is odd can be removed by resampling the degree of one node. If the degree distribution has a power-law tail then the graph is said to be scale-free (Barabasi and Albert, 1999; Barabasi et al., 1999). Scale-free networks will be characterised by rather few individuals having very large numbers of contacts and acting as ‘hubs’. The degree
distributions of empirical networks are frequently far from Poisson and power-law tails with exponents of between $-2$ and $-3$ are typical (Newman, 2003), in which case the degree has an infinite variance. If such a graph were used to model an epidemic, this would indicate an infinite $R_0$ (see e.g., Jones and Handcock (2003)). However, this is not a practical concern since homogeneity assumptions are almost certainly inappropriate in an arbitrarily large population. In any case, simply specifying the local properties of a network by an empirically reasonable degree distribution is not sufficient to guarantee that the most probable global networks resemble observed data. For example, empirical networks typically have many more short loops than is the case for a random graph. Since it is much easier to collect data, estimate parameters and apply control policies at a local level, this raises the general question of how local properties determine global structure. Even for a homogeneously mixing population, where all pairs of individuals are in potential contact, a ‘who is infected by whom’ random graph framework can be useful in investigating epidemic properties that are not time dependent (Ball et al., 1997), such as the probability of a large outbreak, or the final size of an epidemic. In particular, a simple random graph structure is appropriate for an SIR epidemic with a constant infectious period and Barbour and Mollison (1990) use this representation to obtain asymptotic results for the Reed–Frost model.

A Markov random field, defined on a set of sites equipped with a neighbourhood structure, has the property that the values of the field on disjoint sets of sites are conditionally independent, given the values on a neighbourhood region (Dobrushin, 1968). In its most familiar specification, the sites are the vertices of a graph (often a regular lattice) whose edges define the neighbourhood structure. Thus the graph is prespecified and the field — which might be binary in a special case — has a Markov property. In contrast, for a Markov random graph the sites on which a neighbour relation is defined are the edges (or ties) between vertices. For example, a common assumption is that two edges are neighbours if they share a common vertex. A binary Markov random field is then defined on these edges, the resulting random graph being determined by those edges for which the field takes the value 1. Thus, in contrast with a simple random graph, a Markov random graph has dependent edges. The Hammersley–Clifford theorem (Preston, 1973) shows that for Markov random graphs the global properties are completely determined by the local conditional independence properties, and that the joint distribution of the graph has a simple exponential form, where the exponent is a sum of contributions from cliques of neighbouring edges. This exponential form for the joint distribution makes Markov random graph models particularly tractable from a statistical perspective and likelihood-based methods of inference using Markov chain Monte-Carlo algorithms are being developed (Snijders, 2002). It also means that many simple special cases can be considered as possible models for networks with varying amounts of local dependence, and these can be easily simulated and explored. However, the homogeneity assumptions involved are usually unrealistically strong and lead to realisations that are unrepresentative of empirical networks. In particular, most of the probability may be concentrated on small parts of the state space with either very few or very large numbers of edges (Handcock, 2003). Such models are therefore inappropriate except for very small populations. For larger networks, empirically plausible ways to construct neighbourhoods that might lead to processes more closely resembling observed networks are proposed by Pattison and Robins (2002). One possibility is that network ties take place in a social setting, while another is that the neighbourhood structure depends on the realisation of the process — so that, for example, two edges might be neighbours if they are separated by a single edge. The use and interpretation of fitted random graph models needs care, even supposing that completely observed network data are available for model fitting. In particular, there is the question of how appropriate the fitted model is to other scenarios, and whether it provides more than a description of a specific social
structure. Even if it does have some general applicability, it will be important to ascertain the robustness of the network properties to uncertainty in the values of the fitted parameters.

Most attention has concentrated on the network structure of the population *per se*, for example the size of the largest component, or the degree of assortive mixing (in which a high-degree node is likely to be connected to another such node) exhibited, and relatively little to considering realistic disease dynamics. The most intuitively natural approach is to view a contact network as the underlying structure on which an infection evolves. In this case, it may be assumed simply that if an arbitrary node in the graph is infected, the size of the epidemic will be that of the connected component containing the index case. Alternatively, the edges of the graph can be given associated probabilities specifying the chances that node $B$ will become infected if $A$ is, and *vice versa*. In either case, the timing of contacts, duration of infectivity and concurrency of partnerships are all ignored, which is certainly not always appropriate (Morris and Kretzschmar, 1989). Whether an epidemic takes off will depend not on the simple mean number of neighbours but, rather, on a size-biased mean because those having more contacts will themselves be more likely to become infected; for example, Andersson (1998) obtains asymptotic results for a simple discrete time random graph model. An alternative approach is to model the directed ‘who infects whom’ graph as a single entity (Andersson and Britton, 2000). Care is needed in this approach because of the dependencies due to the fact that an infective with a long infectious period will tend to infect many neighbours. When the contact network is regarded as the framework on which an infection spreads, the stationarity of that network may be an issue. Except over short time periods, at the very least it will be necessary to include births of susceptibles and deaths of all individuals together with their associated network edges. Very little attention has been paid to the question of the evolution of the underlying contact networks — but see Read and Keeling (2003) who use simulation to look at the effects of different network structures on disease transmission — and even less to the idea that the network may evolve in response to the disease dynamics. One of the obvious difficulties here is that it is hard enough to collect complete data on a network of contacts at (more or less) a single time instant, and collecting longitudinal data to support a model of network evolution is generally infeasible, except in the smallest population. However, in recent work, Snijders (2001) has been developing continuous-time Markov chain models of network dynamics, albeit in relatively small closed populations of perhaps a few hundred individuals. The idea is to start from a fixed initial configuration, and then to model the evolution of the network from this starting point. The creation of links between individuals in the network is allowed to depend on explanatory variables, which themselves evolve, that can represent individual characteristics such as behaviour or infection status. Each individual seeks to create or dissolve links so as to optimise an objective function.

None of the models described so far is a fully spatial model for the spread of infection. Much of the work most closely meeting this description is that on regular spatial lattices (Durrett and Levin, 1994; Durrett, 1995), which arises from work in statistical physics, but unfortunately there is no room to discuss this here. These lattice processes are most obviously applicable for models of plant diseases and, although they are sometimes applied in other contexts, the underlying homogeneous space makes them generally inappropriate for human/animal diseases. The challenge of producing satisfactory stochastic spatial epidemic models has been around a long time, and Bartlett (1960) notes their intractability. Thus, work allowing for general spatial transmission has traditionally been deterministic. There is a need for simple spatial stochastic models that allow a non-uniform population density in continuous space and, in recent years, plausible stochastic models have often been explored by simulation. For example, Bolker (2003) uses a mixture of moment approximation and simulation in a model that allows both underlying spatial heterogeneity and population density to affect transmission. For spatial processes, the
velocity of the spread of infection and the diameter of the infected region are added to the usual concerns relating to thresholds for invasion and persistence, duration and final size of the epidemic.

In building models of spatial-temporal processes, there is the whole question of which spatial, temporal and population scales to use. A good example of this issue is provided by measles data which are available, for England and Wales, in the form of weekly reports of numbers of cases on a town-by-town basis, the sizes of which are highly variable. The data can be analysed as they stand or aggregated into larger spatial units. In a very small town, a homogeneous mixing assumption may not be too unrealistic, but the data for a large town or city will inevitably already consist of the superposition of coupled epidemics in many subgroups. Properties such as the persistence of infection (or the chance of fade-out) will depend on the level of aggregation of the data (Keeling and Grenfell, 1997). In general, the appropriate scale or scales to use will depend on the purposes of the modelling, but it may be sensible to think in terms of a hierarchy of models, so that local spread over short time scales is represented by an individual-based model, and then these fine scale models are used as the building blocks in a larger scale model, and so on. This general approach is particularly suitable for metapopulation models as, for example, in Ball et al. (1997).

Spatial models are complicated to analyse, and techniques such as pair approximation (Keeling, 1999a,b) that attempt to encapsulate complicated spatial information within a deterministic framework are sometimes used. The basic idea is as follows. In a homogeneously mixing model, the rate of new infections is proportional to the product $X(t)Y(t)$ of current numbers of susceptibles and infectives. In a spatial model, the chance that a particular susceptible-infective pair will generate a new infection needs to be weighted appropriately by a function of their spatial separation. With an underlying contact network, these weights are binary and $X(t)Y(t)$ needs to be replaced by the number of susceptible-infective pairs of neighbours $XY(t)$, say. Thus, a forward equation for $E(X(t))$ will involve $E(\{XY\}(t))$, that for the expected number, $E(\{XX\}(t))$, of susceptible-susceptible pairs will involve the expected number of triples $E(\{XXY\}(t))$, while that for $E(\{XY\}(t))$ will involve $E(\{XXY\}(t))$ and $E(\{YY\}(t))$, and so on. By considering increasingly long strings of susceptible and infective nodes from the network, where adjacent elements of the string are neighbours, an arbitrary spatial structure can be represented. However, the pair approximation method simplifies this process by replacing the expected number of triples of the form $ABC$ by a simple formula only involving pairs, thus closing the set of simultaneous equations. The simplest version is to use $E(ABC) \approx E(AB)E(BC)/E(B)$, where we drop the dependence on $t$ for notational simplicity. However, this approximation will give an overestimate, essentially because it is based on a conditional independence assumption and does not allow for the fact that if node $j$ is a neighbour of both $i$ and $k$, then there is a good chance that $i$ and $k$ are neighbours. A better approximation attempts to allow for this by using a correction factor of the form $(1 - 1/m)(1 - \phi(n/m))E(AC)/E(AC)$ where $\phi$ is the probability that $i$ and $k$ are neighbours given that they are both neighbours of $j$ (for example, $\phi = 0.4$ for a hexagonal lattice), $m$ is the average number of neighbours per node, and $n$ is the size of the network. There is a similarity between the methods of pair approximation and moment closure. For moment closure, there is no spatial structure and the transition rates are functions of the total numbers in various subpopulations. The set of moment equations characterising the joint distribution of these numbers can be closed by an approximation whereby higher order moments such as $E(\{XXY\}(t))$ are specified in terms of lower-order moments; these relationships may be suggested by distributional properties. Thus moment closure gives a set of approximations to the moments of the joint distribution of the state of the system. With pair approximation, spatial information is included, but the allowance for spatial structure is at the
expense of stochasticity. Only mean properties are considered, so that the effect of the spatial structure of, say, the infectives on the mean number of susceptibles is approximated but there is no attempt to look at other moments of the distribution, although this would be possible, at least in principle. Thus the approximation is two-fold, involving both the use of a deterministic model and the approximation of expected numbers of spatially neighbouring nodes with particular infection states. In models for populations in continuous spaces, a method similar to pair approximation — rather confusingly sometimes called the method of moments (Law and Dieckmann, 2000; Dieckmann and Law, 2000) — that looks at the correlation structure of the spatial field is used. There, an equation for the evolution of the expected number of pairs of particular types at a specific separation, $\zeta$ say, will involve the number of triples $[ABC; \zeta, \xi]$ at separations $\zeta$ and $\xi$. This is then approximated by $[AB; \zeta][BC; \xi]/[B]$.

Pair approximation methods were used (Ferguson et al., 2001a,b) to model the recent UK foot and mouth epidemic, using information on farm positions and the network of contacts between farms to give numbers of pairs of neighbouring farms in the various infection states, as a means of exploring possible intervention strategies such as ‘ring’ culling around infected farms, and vaccination. This approach contrasts with a more explicitly spatial and stochastic approach (Keeling et al., 2001) involving the full spatial structure of the farm network. In this approach, a matrix of probabilities that each farm will infect each other farm, if infected itself, is constructed that evolves as an infection spreads. Realisations of an epidemic can therefore be simulated to predict its future course, using appropriate parameter values estimated externally. Again the model was used as a means of investigating possible control strategies; see Kao (2002) for comparative discussion of these issues.

5 Statistical issues and further discussion

I have concentrated here on the development of stochastic models for the transmission of infection. If these models are to be used to draw conclusions about specific infections, rather than to gain mathematical understanding of general classes of infection processes, they will usually have to be fitted and validated against data. This section will be confined to a few general remarks; for further discussion and a more detailed introduction to the statistical issues touched on briefly below see Becker (1989), Becker and Britton (1999) and Andersson and Britton (2000). Some of the simplest models, such as the SIR model, are broadly applicable in many situations. In that case, one or both easily interpretable parameters — the reproduction ratio, $R_0$, and the mean infectious period, $1/\gamma$ — can often be assumed known or can be determined experimentally, although if necessary they can be easily estimated from data on the course of an epidemic; see e.g., Becker (1989). In general, if parameter values are to be estimated from experiments then these need careful statistical design. If such estimates are available externally — i.e. independently of the model — then model properties can be compared with those of empirical data to validate the model. With relatively complicated models of specific infections, there will generally be a need not only to estimate at least some of the parameters of the model from epidemic data, and to test the sensitivity of the inferences made from the model to uncertainties in their values, but also to validate the assumed structure of the model. It is particularly helpful if an inferential framework can be constructed within which a hierarchy of increasingly complex models can be compared. It is important to ensure that any conclusions drawn are robust to model uncertainty. In this regard, and especially when strong simplifying assumptions have been made, it can be reassuring if similar conclusions result from a range of different modelling approaches; for example, this was the case in the recent UK foot and mouth outbreak.
In an ideal world, one would have complete observation of a single epidemic, or better still, multiple replications thereof, or a series of linked outbreaks of an endemic infection. More often, the information is incomplete and data are aggregated over time periods — even the whole course of the epidemic as with final size data — or over large heterogeneous populations. Alternatively perhaps only cross sectional data are available. In such cases, a complete epidemic model specification may be unnecessary. In this regard, it is interesting to ask to what extent population structure can be inferred from incomplete data, especially from a single realisation of an epidemic process. For example, if there is very little mixing between the two homogenously mixing subgroups in a simple metapopulation model, and a single infective is seeded in one subgroup, there may be a clear time-lag between the epidemics in the two groups. As the between-subgroup mixing increases the time-lag will tend to disappear. To what extent can such population heterogeneity be inferred from data for a single epidemic realisation, especially when only partial information — perhaps the total numbers of new cases in the population in discrete time intervals — is recorded?

In predicting the future course of a partially observed epidemic, a range of predictions can be given incorporating such uncertainties and these should, additionally, include the effects of random variation between realisations. If, as is generally the case, parameters are estimated by model fitting, then the more parameters are fitted, the more the capability to validate the model by comparison of observed and fitted properties is reduced, as is the predictive power of the fitted model with regard to future epidemics. As with any spatial and/or temporal stochastic process, the properties to be used for fitting and for validation need to be carefully considered and will depend on the purposes for which the modelling is being carried out. When there is complete observation of the times of infection and removal (recovery) for an epidemic outbreak, likelihood methods can be used to obtain parameter estimates. Maximum likelihood estimation is also sometimes possible with incomplete observation, for example with final size data. More generally though, likelihood methods tend to be unwieldy with anything other than the smallest populations. Martingale estimating equations or the use of the EM algorithm to estimate the missing data then provide useful approaches. However, in recent years, the very rapid growth of computer-intensive methods in general and of Markov chain Monte Carlo methods in particular has led increasingly to their application in fitting epidemic models (Gibson and Renshaw, 2001; O’Neill, 2002), one of the earliest uses being parameter estimation for a partially observed SIR model in a Bayesian framework (O’Neill and Roberts, 1999). Nevertheless, in particular cases, the need for rapid answers during the course of an outbreak has lead to much simpler approaches, such as choosing parameter values to maximise the likelihood over a fixed grid (as in Riley et al. (2003) for SARS).

One of the most important purposes in fitting models to epidemic data is the investigation of potential control strategies. Control may be achieved by reducing the rate of contacts between susceptibles and infectives, the probability that a contact results in transmission, or the duration of infectivity. As, for example, in the case of foot and mouth disease and SARS, these reductions might be achieved by travel restrictions, biosecurity measures, or quarantine arrangements. Another possibility is to reduce the pool of susceptibles by vaccination. In this case, models can be used to determine the impact of a range of vaccination strategies, by making various assumptions about the effects of the vaccine — does it confer indefinite total protection, total protection for a proportion of the population and none for the rest, partial protection for each individual, does the immunity wane over time, or perhaps the effect is not to protect the vaccinated individual from being infected but to make an infective less infectious to others, and so on? Allowing for vaccination status in the models adds an extra layer to the heterogeneity of the population. Various measures of vaccine efficacy have been proposed, and the design of
studies and methods of their estimation are important statistical issues; see e.g., Datta et al. (1999) and Becker et al. (2003). Heterogeneity of mixing is an added complication in assessing possible vaccination strategies. Even for a perfect vaccine, if there is a limited supply, or limited resources with which to administer it, how should the vaccine be allocated? In the case of the simple metapopulation models described in Section 4.3, aiming to reduce the numbers of susceptibles in each subgroup to the same level makes intuitive sense, but assumes that numbers of susceptibles are known and that it is socially and operationally feasible to carry out such a strategy.

For almost all of the research described above, it has been assumed that a small number of infectives — perhaps just one — is introduced into a wholly susceptible population. Then the interest is in the probability that the infection will spread and lead to a major outbreak, the properties of that outbreak conditionally on it doing so, and how best to control transmission. But models also have a role to play in answering another important problem: how best to detect an emerging new infection, such as SARS; or a new outbreak of a pre-existing one, such as foot and mouth disease. This topic that has become especially topical with the current emphasis on detection of bioterrorism attacks, and has spurred renewed interest in models for smallpox, which was successfully eradicated worldwide by 1980. Returning closer to home however, and with the reason for this volume in mind, those who have followed David Cox’s interests over the last few years will know of his preoccupation with badgers and the spread of tuberculosis in cattle (Donnelly et al., 2003). Potential models for this exemplify many of the complications discussed in this chapter including the presence of more than one host species, uncertain transmission routes, complex spatial considerations and incomplete observations. This work has recently led David to think about the question of how to detect a new outbreak of infection, and to know when, and how, to instigate control measures. Perhaps he will soon provide the answer to this question in yet another seminal contribution to applied probability!

Acknowledgements

My thanks are due to all those, far too numerous to list, who have shared with me their considerable insights on epidemic models. In particular, I am grateful to the participants of many epidemic modelling workshops, and their organisers and sponsors, including Sydney (1990), Skokloster (1990), Oberwolfach (1991, 1994, 1999), INI (1993), Stockholm (1995), Woudschoten (1996), Skye (1997), Canberra (1998), PIMS (1999), Trieste (1999), DIMACS( 2002, 2003), Copenhagen (2003), Mariefred (2003) and IMA (2003), from whom I have learned much. But above all, my thanks go to David Cox, without whose gentle spur I might never have discovered the fascination of epidemic models and without whose friendship and collaboration my life would have been immeasurably the poorer.

References


