Chapter 0

Introduction

In this chapter we first introduce our terminology regarding structure in both infectivity and susceptibility of individuals. We trace the genesis of an epidemic after an infectious disease is introduced into a population where it was not present before, and in passing briefly discuss some of the questions that are studied by mathematical methods. We introduce the concept of R_0, discuss some of its history, and indicate some of the most important applications of R_0 in mathematical epidemiology. Finally, we describe an active research area concerning the incorporation of acquired immunity in models as an illustration of the most ideal form of mathematical modelling in epidemiology.

0.1. Structure in epidemic models

The basic idea of (physiologically) structured population dynamic models is to distinguish individuals from one another according to characteristics that determine the birth-, death- and resource consumption rates (or, more generally, the interaction with the environment), as well as the rates at which these characteristics change themselves. In this thesis we are solely interested in the spread of a contagious disease, therefore we limit ourselves to those individual characteristics that influence the *force of infection* of the disease in question (i.e., the rate at which susceptible uninfected individuals become infected through contacts with infected individuals). In this section we discuss these individual characteristics in some more detail.

The first step in modelling a structured population is to choose the characteristics that are relevant to the problem one is concerned with. In mathematical jargon this is called the choice of the i-state representation, where i denotes 'individual'. In the context of epidemic models this is a double task

(we have to deal with population dynamics and disease transmission) and we shall accordingly refer to the components of the i-state which describe the development of the disease within individuals as the d-state, where d denotes 'disease'. From the point of view of the disease the rest of the i-state may lead to *heterogeneity* in the population. Accordingly, we shall call this part of the i-state the h-state.

Building a model starts at the individual level: specify and describe (mathematically) the processes which change the i-state, either in a continuous manner or by jumps (like the susceptible $\rightarrow$ infected transition), and the processes which change the number of individuals (birth and death). We shall always restrict ourselves to deterministic models for continuous i-state change. Jump-, birth- and death-processes are in principle stochastic, but we use a law-of-large-numbers argument to describe chance processes by rates (like in chemical kinetics, radio-active decay etc.)

The population state, or p-state, is by definition the function (or measure) which describes the number of individuals and how they are distributed over the various i-states. Given a description of the dynamics at the individual level it is a matter of bookkeeping to calculate the changes in the p-state. This is easy to do on an infinitesimal basis (so we end up with differential equations) since then the contributions of the various processes become independent of each other and can simply be added. For the present purpose there is no need to become more specific about the theory behind structured population models. For an extensive overview of the theory and its applications see Metz and Diekmann (1986). For a more recent discussion of an alternative approach to general physiologically structured population models see Diekmann, Gyllenberg, Metz and Thieme (1992).

Let us elaborate on the i-state characteristics that matter for disease transmission. As far as the d-state is concerned there are basically two possibilities: 'age of infection', and 'degree of infection'.

1. 'Age of infection'.

Suppose that after infection the disease develops as an autonomous process within the infected individual (so here we have in mind that the invading organism reproduces within the host at such a high rate that further infections are irrelevant; this concerns viral, bacterial, and most protozoan diseases such as measles, influenza, rabies and HIV, and fall under the general heading of 'micro-parasites', see e.g. Anderson and May (1991)). This assumption is due to Kermack and McKendrick (1927). In the notation we will use in this thesis, their key idea was to describe the *expected infectivity of an individual τ units of time after it became infected* by a (non-negative) function $A(\tau)$. The assumption is that infection runs its course within the host without any further influence of the environment and that, consequently, we can use an 'age' representation to describe the expected output, i.e. the infectivity of the infected individual towards other individuals in the population. All relevant aspects of

the detailed stochastic time evolution of the internal population of viral particles or bacteria and the concommitant reaction occuring in the immune system are incorporated in A. The precise nature of $A(\tau)$ is of course determined by the particularities of the disease one studies. It should ideally be derived on the basis of a submodel that reflects the relevant characteristics of the disease. Two examples of possible shapes of $A(\tau)$ are given in figure 0.1 below (note the difference in time-scale).

The function $A(\tau)$ can either be interpreted as a deterministic property or as an expectation (where we imagine very many 'stochastic' individuals, distributed in a manner which does not change as time proceeds). In this way the usual 'compartment' models of the S-E-I-R type (also called 'prevalence' models) are included. Here the class S contains the *susceptible* individuals, E those that are *exposed* (i.e. infected but not yet infectious or, in other words, where the infection is in a latency period), I contains the *infectious* individuals and R the ones that are *removed* (containing the infected individuals that are no longer infectious). The interpretation of 'removed' varies from (temporarily or permanently) immune to dead.

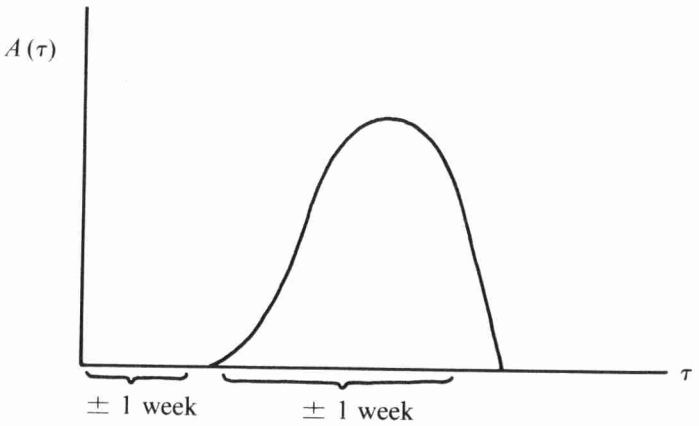

Figure 0.1a Measles infectivity as a function of time

Example 0.1. We regard a compartmental model as described above. Assume that the time periods spent in each compartment are exponentially distributed with some appropriate rate constant. Suppose after infection individuals enter class E and then make the transition to class I at a rate σ whereafter they are removed at a rate α. While being in I they have transmission rate constant β (see furtheron). We claim that

$$A(\tau) = \beta \frac{\sigma}{\alpha - \sigma}(e^{-\sigma\tau} - e^{-\alpha\tau}) \tag{0.1.1}$$

The argument goes as follows. The probability to be in class E at time τ_0 after infection is $e^{-\sigma\tau_0}$ and the density function for entering class I is therefore $\sigma e^{-\sigma\tau_0}$. In order to be in class I at d-time τ one should:

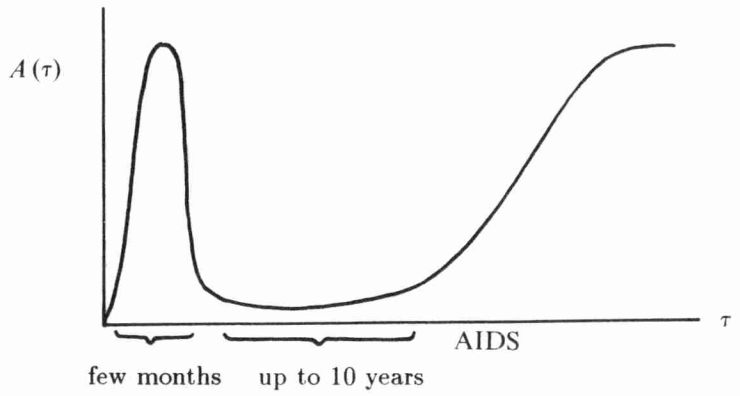

Figure 0.1b HIV-infectivity as a function of time

i) have entered I at some d-time $\tau_0 \in (0, \tau)$
ii) have remained in I in the interval (τ_0, τ)

The probability that ii) holds is $e^{-\alpha(\tau-\tau_0)}$. Hence the probability to be in I at d-time τ is

$$\int_0^\tau \sigma e^{-\tau_0\sigma} e^{-\alpha(\tau-\tau_0)} d\tau_0 = \frac{\sigma}{\alpha-\sigma}(e^{-\sigma\tau} - e^{-\alpha\tau})$$

Upon multiplying by β we find A. Note that $\int_0^\infty A(\tau)d\tau = \frac{\beta}{\alpha}$, independent of σ.

So, we can either imagine a discrete d-state variable with a stochastic jump process to describe the progression of the disease, or a continuous d-state variable τ (with steady progression) and an infectivity function A, see figure 0.2.

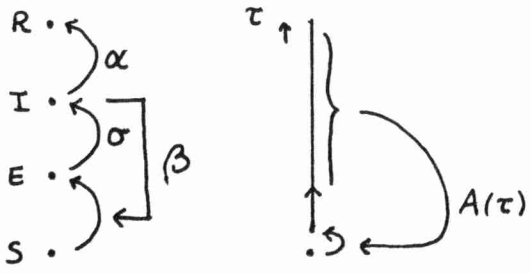

Figure 0.2

The two representations are equivalent when (0.1.1) holds. The τ-representation is suited for a more general class of models since A can be any

non-negative function.

$\diamondsuit$

Example 0.2. Let us look at a slightly more complicated example where we try to incorporate the shape of the function A in the case of HIV, see figure 0.1. In this example A is also an expectation. We recognise three levels (classes) of infectivity, which we express as three different values of the infection probability per contact, h_1, h_2, h_3. Note that this example is not very different from example 0.1, we could take $h_1 = 0$ and call the first class 'exposed', we could call the second class 'infected', and the third class, with $h_3 = 0$, could be thought of as the 'removed' class.

We assume that the time spent in the infectivity classes 1 and 2 respectively, is exponentially distributed, with parameters θ_1 and θ_2, and that we have an HIV induced death-rate ρ_i in class i ($i \in \{1, 2, 3\}$). Let $P_i(\tau)$ be the probability that an infected individual is still alive a time τ after it became infected, and that it currently has infection-level i, then the following system describes the dynamics of the P_i's

$$\begin{aligned}
\frac{dP_1}{dt} &= -(\theta_1 + \rho_1)P_1(t) \\
\frac{dP_2}{dt} &= \theta_1 P_1(t) - (\theta_2 + \rho_2)P_2(t) \\
\frac{dP_3}{dt} &= \theta_2 P_2(t) - \rho_3 P_3(t).
\end{aligned}$$

We assume that each newly infected individual starts in class 1, so the initial condition is taken to be

$$P(0) = (P_1(0), P_2(0), P_3(0))^T = (1, 0, 0)^T.$$

Let c_i denote the average number of contacts an infected individual in class i makes per unit of time. A given infected individual is now expected to have infectivity

$$A(\tau) = \sum_{i=1}^{3} h_i c_i P_i(\tau),$$

at time τ after it became infected.

$\diamondsuit$

2. 'Degree of infection'.

In this case infection is not a unique event but rather a repeated process. Meanwhile reproduction within the host can or cannot take place. The d-state is defined as the number of parasites a host harbours and so the d-state is usually taken to be discrete (non-negative integers). Examples include schistosomiasis and other worm diseases (gathered under the name 'macro-parasites', see Anderson and May (1991)). One colloquially speaks of 'wormload'-models.

For analysis of this type of models see e.g. Hadeler and Dietz (1983), Anderson and May (1991), Kretzschmar (1989).

In this thesis we will only be concerned with category 1 infections.

Remarks.

(1) The corresponding class of models is sometimes described by the term 'density models'. The difference between 'prevalence' and 'density' is, however, much more a matter of data collection: does one only count how many individuals suffer from a disease or does one also take into account how severely an individual is affected by the disease (e.g. by estimating the wormload). The essential mechanistic difference between models of type 1 and type 2 derives from the environmental impact after the first infection.

(2) While 1 and 2 seem to be exclusive categories, there is at least one important disease, malaria, that belongs to both. Superficially speaking, malaria would belong to category 1 because the protozoan species that cause the disease multiply very rapidly within the host. However, connected to malaria is the phenomenon of acquired immunity (see papers by McGregor and Wilson, and Dietz in Wernsdorfer and McGregor (1988), and Aron (1983)). The more additional infections with the parasite (possibly different strains of the same species) that an individual acquires, the higher its level of immunity will rise (leaving aside intricacies that concern the required length of the time-period between successive infections). The immunity does not protect against re-infection, but individuals with a high level of immunity do not experience the severe disease symptoms, they do, however, remain infectious. The immunity-phenomenon places malaria in category 2. We will return to this in section 0.4 below.

$\diamondsuit$

We now turn to the possible h-states. Such traits can be static (like male-female) or dynamic (suffering from another disease, stage of development), they can take discrete values (like homo-/hetero-/bisexual) or continuous values (like spatial position of a plant, or age). In particular cases h-states can be very complicated, and contain both continuous and discrete, static and dynamic components. We will denote the state space of heterogeneity characteristics by Ω. If the h component of the i-state has more than one possible value, then we have to take into account that the expected infectivity function A may depend on both the h-state of the susceptible and the h-state of the infected individual taking part in a contact. We therefore postulate, for category 1 diseases, a function $A(\tau, \xi, \eta)$ which gives the expected infectivity of an individual that was infected τ time-units ago while having h-state $\eta \in \Omega$, towards susceptibles with h-state $\xi \in \Omega$.

Remarks.

(3) Admittedly, our splitting of the i-state in a d-state and a rest is somewhat ambiguous since the rest may contain information which is only relevant for calculating transmission rates (like propensity to make sexual contacts) and

not for the population dynamics per se. To clarify the terminology we attempt to formulate a definition: the d-state is that part of the i-state which describes the difference between susceptible and infected individuals, see figure 0.3.

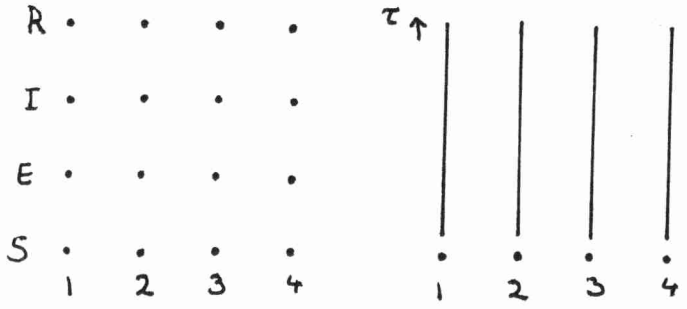

Figure 0.3

Here the d-state variable is taken in the vertical direction and the rest of the i-state is discrete and static and taken in the horizontal direction. The numbers may, for example, stand for

1 male, making many sexual contacts
2 female, making many sexual contacts
3 male, making few sexual contacts
4 female, making few sexual contacts

(4) The word 'individual' allows a broad interpretation and may, apart from the obvious reference to single humans, animals or plants, refer to e.g., a family, village, focus in a field of plants or 'married' couple.

(5) Note that in $A(\tau, \xi, \eta)$ we parametrise the infected individuals by the h-state at the moment of their own infection (τ time-units ago). This saves us, for the moment, from specifying the precise dynamics, if any, of the individual's h-state. An individual that has just become infected while having h-state η will be colloquially said to have been 'born with h-state η' ('born' with regard to the disease).

(6) For age we may have an additional reason to include it in the i-state: one can often use data about the age at which the disease is contracted to estimate the force of infection (Anderson and May, 1991).

$\Diamond$

Finally, we take a closer look at the *force of infection*. On the individual level, it describes the probability per unit of time for a susceptible individual to become infected. On the population level this is, due to a law of large numbers argument, equivalent to the fraction of the susceptible population that the infected individuals are able to meet, and subsequently infect, per unit of

time. So, this function consists of a part that describes the frequency of meeting susceptibles, and a part that gives the probability that transmission of the infection during a contact (at meeting) is successful. During the time period of a meeting the two individuals can have a number of contacts. What actually constitutes a 'meeting' and a 'contact' is determined by the disease in question. For example, travelling in the same compartment of a train constitutes a meeting for influenza, and coughing your fellow passengers in the face constitutes a contact where the virus could be transmitted. For HIV infection however, this does not count as a meeting or contact. We will call the product of the frequency of meeting (a per couple probability per unit of time of meeting), and the number of (relevant) contacts during a meeting, the *contact frequency*, and the probability that a contact during a meeting indeed leads to transmission, the *success ratio*. The success ratio need not be symmetric with respect to the population structure (e.g., for many diseases which spread through heterosexual contact the success ratio for the case where the man is infectious and the woman susceptible is different from the case where it is the other way around; likewise the malaria transmission between mosquito and man is asymmetric).

Remark (Law of mass-action). The principle of mass-action kinetics was first introduced for application to the spread of infectious diseases by Hamer (1906) in a paper in the Lancet. In modern terminology it states that the rate of appearance of new infecteds is proportional to the number of 'meetings' per unit of time between susceptibles and infectives, and that the density of 'meetings' (i.e., number per unit area) is proportional to both the density of susceptibles and the density of infectives (i.e., one has a bilinear term to describe the interactions between the individuals). This is based on the assumptions that the infectives are (spatially) homogeneously distributed in the population and that locally the probability of 'meeting' per unit of time is the same for any pair of individuals. ♢

If there is no heterogeneity in susceptibility in the population, if we assume mass-action kinetics for the frequency of meeting and no dependence on the total population density N, then we can write βSI for the rate with which the susceptible population becomes infected (where S and I are the (spatial) densities of susceptible and infected individuals respectively), β is called the transmission rate constant. The force of infection, commonly denoted by $\lambda(t)$, is then given by $\lambda(t) = \beta I(t)$. The assumption concerning the total population density leads for sexually transmitted diseases to the phenomenon that an increase in this density automatically implies that individuals will have more sexual contacts per unit of time. To remedy this, one usually allows $\beta = \beta(N)$ to depend on the total population density. The question then is: what does $\beta(N)$ look like? Typically, it will be a saturating function of N. With increasing density N the number of meetings an individual can take part in will rise slower and slower, and ultimately saturate, because the individual simply runs out of time for any additional meetings. The function $\beta(N)$ is an important ingredient

in calculating the basic reproduction ratio R_0 (see section 0.3) and one should therefore try to obtain a closed expression of it. In the past many different expressions have been suggested for $\beta(N)$, but all these where 'drawn out of the hat' to meet a number of presupposed intuitive requirements. It was an open problem to mechanistically derive an expression for $\beta(N)$. One should of course ideally first derive an expression and then show that it behaves as one would intuitively expect it to behave. In chapter 4 we give a solution to the open problem and show that

$$\beta(N) = \beta \frac{1 + 2\theta N - \sqrt{1 + 4\theta N}}{2\theta N^2},$$

where β is the average probability of transmission of the infection between two individuals taking part in a meeting, and θ is a measure for the relation between the average duration of meetings and the time periods between meetings (see chapter 4).

As far as a type 1 disease is concerned, modelling consists of first specifying the relevant h-state, and subsequently devising a submodel to describe how the infectivity function $A(\tau, \xi, \eta)$ (which encompasses the contact frequency, the success ratio, details about recovery, etc.) depends on this h-state. Making a submodel that reflects the way in which the disease is transmitted, while in addition paying attention to relevant aspects on which data are available, is usually a very complicated matter. However, aside from this complication there is already a difficulty in the interpretation of 'relevant h-states'. How would one be able to decide what, of the many possibilities, are the differences in the population that are the most important for the transmission process? In other words, what is the influence, on for example R_0, of various mechanisms or phenomena acting in the population that one could take into account in describing the transmission processes (and which call for the introduction of a certain heterogeneity structure to make their adequate modelling possible)? This is still very much an open research area.

Let us give one example. In modelling sexually transmitted diseases one could argue that one should take into account the fact that individuals have the habit of establishing longer lasting relationships. In calculating R_0 for these diseases we adopt this approach (see chapter 3). From the point of view of the disease it is uneconomic that the individuals form pairs. We assume that during the existence of a pair the individuals have no sexual contacts with individuals 'outside'. If one of the members is infected, and infects its partner, then all sexual contacts these two members have after the transmission, and before the pair breaks up, are 'wasted' from the disease's point of view. Therefore, one expects that pair formation will influence the spread of the disease. One could then argue that larger units than pairs, like triplets, should be taken into account, to accomodate for 'extra-marital' relationships, that can be important in establishing links between different groups (e.g bisexuals link the heterosexual population to the homosexual population, and can therefore be important

for the spread of the infection). Going still further, it has proven to be important to take the entire social network structure of sexual relationships into account if one wants to predict the future behaviour of an established sexually transmitted disease like HIV. This is the very nice graph-theoretical approach of Blanchard, Bolz and Krüger (1990, and the references to their earlier work given there). It is still an open problem however to determine what the influence is on the *invasion* (described by R_0) of a sexually transmitted disease of the incorporation of increasingly more complex (and realistic) relationship networks.

0.2. Basic questions of mathematical epidemiology

It is tempting to think that the main use of mathematical models in epidemiology is to predict future trends in the spread of infectious diseases. However, broadly speaking this expectation cannot be fulfilled. Firstly, the most complex models for specific diseases are still highly oversimplified to base realistic predictions on. But secondly, making models still more complicated to highten their realism, leads to a rapid proliferation of parameters, hardly any of which can be 'guesstimated' with any accuracy, let alone be measured. Short term predictions can often be made (for example, estimating the expected number of AIDS cases arising in the next five years, from past trends, to determine the number of hospital beds that have to be reserved for these patients), but this is very much the realm of statistics. Long term predictions are very difficult.

The main use of mathematics in epidemiology is to obtain insight, in particular concerning the relative importance of factors that influence the spread of an infection. More generally and formally, one is looking for insight into the relation between mechanisms that operate on the individual level and the phenomena that result on the level of the population. The advantage of mathematics is that it is made of paper, and one can, for example, rather cheaply and without ethical connotations perform 'experiments' to evaluate the efficacy of control measures. Hethcote (1990) extensively discusses the philosophy of applying mathematics to epidemiological questions.

There are some five main areas where mathematical models are used to answer epidemiologically relevant questions. We discuss them briefly (along the lines of Diekmann, (1991)) in the order in which they occur 'naturally' after an infectious disease has entered a population where it was not present before. We assume that this disease confers permanent immunity to individuals that have recovered from the infection (think of childhood diseases such as measles and rubella). We also assume that all individuals in the population are equally susceptible.

We start with the case where the total number of individuals in our population is constant on the time-scale on which the epidemic processes of infection and recovery (or death caused by the disease) occur, i.e. we disregard demography. The first question that arises is: if the infectious disease enters our population, will it cause a spreading epidemic, or will it die out right away? This is referred to as the *invasion* question. The existence of a threshold quantity (called the *basic reproduction ratio*) that can be used to answer this question, is a major insight that mathematical thinking has brought to epidemiology. The main part of this thesis is concerned with this quantity, commonly denoted by the symbol R_0.

Suppose that an epidemic does occur. If we still ignore births of new susceptible individuals, then we can picture a typical epidemic as in figure 0.4. The I symbolises the infected part of the population and S the susceptible part. At first, the infecteds will increase slowly in number, then more rapidly, and at some point in time their number will decrease again because of (i) lack of sufficient susceptibles, and (ii) by the fact that the infecteds are only infectious for a certain amount of time, whereafter they become immune (alternatively they die along the way). The relevant questions are: when does I reach its maximum value?; how large will this maximum be?; what fraction (if any) of the susceptibles ultimately escapes from getting the disease?

It was long believed that an epidemic would only stop at the moment that all susceptible individuals had become infected. The second major insight furnished by mathematical modelling, is that this is generally not the case; there will be a positive (albeit possibly very small) fraction of the population that never contracts the infection (Kermack and McKendrick, 1927).

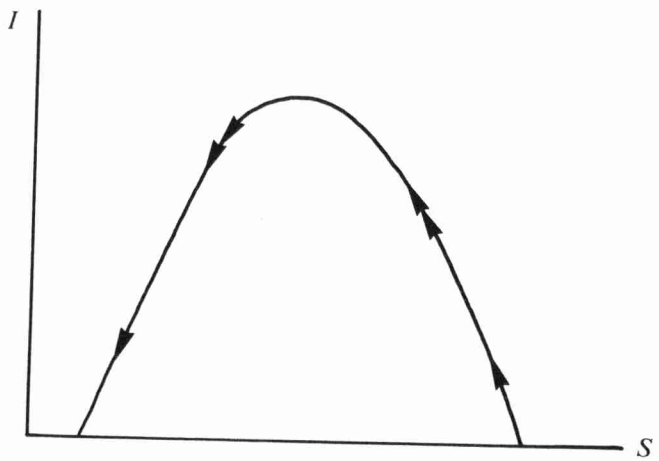

Figure 0.4 Epidemic outbreak

Now let us move one step further and take the demography into account, we allow for an inflow of new susceptibles. However, we assume that the demography and the disease transmission are decoupled by a time-scale argument.

We then get generically a situation as pictured in figure 0.5., called 'recurrent behaviour'.

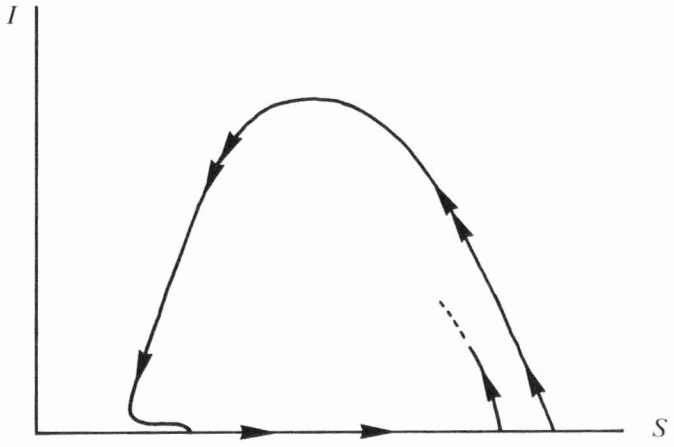

Figure 0.5 Recurrent behaviour

After a rapid epidemic outbreak as described in the previous paragraph, the disease will disappear (locally) from the population. To describe this accurately one would have to take stochasticity into account. (While not going into details, we mention the very useful distinction that is made in this respect between *endemic fade-out* and *epidemic fade-out* by Anderson and May (1991, page 20).) After the disease has gone extinct locally, the population will gradually be replenished by the birth of new susceptibles on the demographic time-scale. When the susceptible population is 'large enough' again, i.e. when the threshold quantity mentioned above is above threshold, a re-introduction of the disease from outside the population will lead to a new epidemic. The standard example of this behaviour is measles in Iceland, see the beautiful book of Cliff and Haggett (1988); there are not enough inhabitants in Iceland to 'create' new susceptibles fast enough to 'keep up' with the epidemic. It is an as yet unsolved theoretical problem to find a nice characterisation of the 'borderline' between such behaviour and the behaviour we describe next.

When the disease is continually being transmitted at the demographic time-scale (no decoupling of demography and transmission) it is called *endemic*, in the SI-plane an internal equilibrium can exist, see figure 0.6 for the generic picture. An obvious question is: will there be a stable steady state or are there oscillations? If the latter is the case, then how does the period of oscillations depend on the relevant parameters?

Finally, in the setting of the previous paragraph, there is the *regulation* question. How, if at all, does the disease affect the growth rate of the population? There is currently much interest in this problem, for example with respect to the AIDS-epidemic in certain African countries.

The questions discussed above can all be answered, more or less easily,

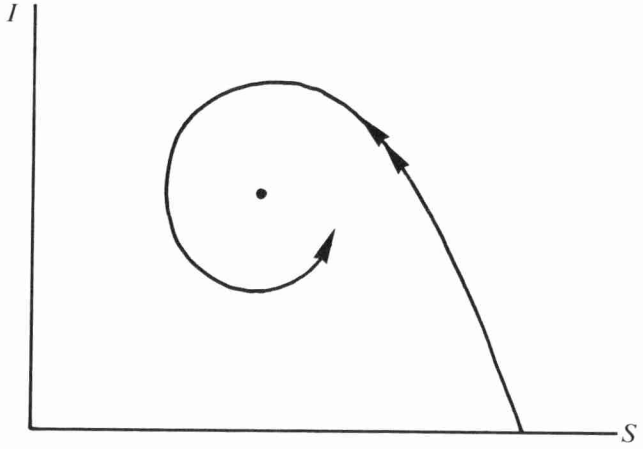

Figure 0.6 Endemic situation

when we make no distinctions in susceptibility between the individuals in the population. When we take relevant heterogeneity into account, the questions become much more difficult, and in fact only the invasion question can, at the present state of the art, be answered in great generality. With regard to the other questions progress has been made for specific h-states, most notably for discrete finite h-state variables (endemic equilibrium: e.g. Beretta and Capasso (1986), Lin and So (1990); demographic interaction: e.g. Busenberg, Iannelli and Thieme (1991b)), and in the case of age (endemic equilibrium: e.g. Busenberg, Iannelli and Thieme (1991a), Inaba (1990), Greenhalgh (1988); regulation: e.g. Andreasen (1989), May, Anderson and Mclean (1988ab)). More information on treating questions raised in this section, with and without heterogeneity, can be found in Diekmann, Heesterbeek, Kretzschmar and Metz (1989) (to be elaborated into a book).

0.3. The basic reproduction ratio

The main concern of this thesis is to answer the invasion question posed in section 0.2 and to explain the methodology of calculating the basic reproduction ratio R_0 in heterogeneous populations. The concept of R_0 in epidemiology is due to Sir Ronald Ross (1909) in connection to his work on malaria (Ross had received the Nobel prize in 1902 for his discovery that malaria was caused by a protozoan species that was transmitted from individual to individual by mosquitoes, and that the disease was therefore not caused by 'bad air' as its name reflects). The introduction of this concept is probably the earliest example of insightful application of mathematics to epidemiology. Ross observed that eradication of malaria is possible by decreasing the number of mosquitoes

present in a certain area. Prior to that it was generally believed that malaria would always survive as long as some mosquitoes were still present and that total eradication of mosquitoes was impossible. Ross showed, using a simple model, that there was a quantity that, when suppressed below unity, would guarantee the disappearance of malaria from the area, and that this quantity depended on the ratio of mosquito density to human density. Thus he found a critical mosquito density. Empirical corroboration was later obtained in India with the discovery of neighbouring areas with and without malaria and mosquito densities respectively above and below the critical level.

In 1923, Lotka used, in a demographic context, the term 'reproduction rate per generation' for a quantity that described the expected number of offspring that a generation of individuals will produce (Lotka, 1923), see also Fisher (1930). In demography this concept appears to go back to Böckh in 1886 (Smith and Keyfitz (1977)). It does not take long to see that this 'reproduction rate' and the quantity Ross has in mind are intimately related.

The next step on the epidemiological side was made in a paper by W.O. Kermack and A.G. McKendrick which appeared in 1927 in the Proc. Roy. Soc. Edinb. which has had a major influence on mathematical modelling of epidemics. It is probably both the most cited and least well read paper in the whole of mathematical epidemiology. Sadly, it has become common practice for people to refer to a certain simple compartmental model formulated in terms of ordinary differential equations as *the* Kermack-McKendrick model whereas in fact (see the remark following example 0.3 below) Kermack and McKendrick treated a much more sophisticated model.

Assuming the law of mass action to describe the interaction between susceptibles and infectives Kermack and McKendrick were led to introduce, in the notation of this thesis, $R_0 = S\int_0^\infty A(\tau)d\tau$ as the expected number of secondary cases (new infected individuals) produced by one infectious individual during its entire infective life in a susceptible population of density S. Clearly the R_0 has threshold value one, i.e. when $R_0 < 1$ (each infected individual does on average not even replace itself in the population) no epidemic develops upon introduction of the infection, whereas when $R_0 > 1$ (each infected individual on average replaces more than itself) an epidemic gets started. Kermack and McKendrick did not name their threshold quantity, nor did they attach a symbol to it. MacDonald (1952), in his work on malaria, called the threshold quantity the 'basic reproductive value' and denoted it by z_0. Dietz (1975), used the symbol R. It is unclear who was the first, in an epidemiological context, to use the symbol R_0. The earliest reference in a demographic context we could find is a paper by A.J. Coale (1957) as reprinted in Smith & Keyfitz (1977).

The subscript 'zero' could be taken to reflect the fact that we presuppose a 'virgin' population (with respect to the disease). If not all individuals are susceptible (for example because they have recovered and became immune, or because they are vaccinated) then one should discount for this. One then obtains the *net* reproduction ratio, usually denoted by R today (see Hethcote (1990) for a discussion). However, in demography an alternative reason for

the subscript 'zero' as the zero'th moment of the probability distribution that describes the expected future offspring, is given (Smith and Keyfitz (1977)).

Remark. Often R_0 is called the 'basic reproductive *rate*', but clearly, as has been pointed out by many others, it is not a rate, it is a number. To be even more precise, it is a ratio: number of individuals per individual. It is not only a dimensionless but even a quasi-dimensionless quantity. Moreover, and we thank Mats Gyllenberg for pointing this out to us, it is grammatically more correct to speak of 'reproduction ratio' instead of 'reproductive ratio', because the 'ratio' does not reproduce. (This remark also applies to terms like 'latent period' where it should be 'latency period'.) See Hethcote (1990) for a host of other names denoting the same quantity. The word 'rate' is probably the most misused word in the whole of theoretical biology.

$\diamondsuit$

We now give the following biological 'definition' of the basic reproduction ratio R_0:

> *R_0 is the expected number of secondary cases produced by a typical infected individual during its entire infectious period, in a suceptible population.*

In simple cases without heterogeneity, it is easy to write down a formula that expresses R_0 in the parameters that describe the transmission processes, by directly interpreting the biological definition.

Example 0.3. We return to the example 0.1 from section 0.1. Suppose we have an infected individual. While it is in class I, it has, by assumption, a probability per unit of time β to meet and infect a susceptible. The number of secondary cases it is expected to produce *per unit of time* is therefore βS, where S is the density of susceptible individuals. By assumption, the infected individual remains in class I for an average period of length $1/\alpha$, therefore it is expected to remain producing βS new cases per unit of time for $1/\alpha$ units of time. So clearly,

$$R_0 = \frac{\beta}{\alpha} S. \tag{0.3.1}$$

In section 0.1 we already noted that $\int_0^\infty A(\tau)d\tau = \beta/\alpha$, and so the formula of Kermack and McKendrick brings us once more to (0.3.1).

$\diamondsuit$

Remark. In the formulation with a more general function $A(\tau)$, the Kermack-

McKendrick model can be written as

$$\begin{aligned} \frac{dS}{dt} &= -\lambda(t)S(t) \\ \frac{\partial j}{\partial t} + \frac{\partial j}{\partial \tau} &= 0 \\ j(t,0) &= -\frac{dS}{dt}(t) \\ \lambda(t) &= \int_0^\infty A(\tau) j(t,\tau) d\tau \end{aligned} \tag{0.3.2}$$

(Kermack-McKendrick (1927), Metz (1978); see also Lauwerier, (1984)). Here, λ is the force of infection acting on the susceptible population, and $j(t,\tau)$ denotes the number of infected individuals at time t that have disease-age τ. The differential equations model that is usually called 'the' Kermack-McKendrick model is obtained as the special case

$$A(\tau) = \beta e^{-\alpha\tau},$$

by defining

$$I(t) := \int_{-\infty}^{t} e^{-\alpha(t-s)} j(s,0) ds.$$

The resulting ODE system is then given by

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \alpha I. \end{aligned}$$

$\diamondsuit$

Example 0.4. We return to example 0.2 in section 0.1. If we define a matrix G of transition probabilities per unit of time between the three infectivity levels, then the ODE-system for the probabilities P_i can be written as $P(\tau) = e^{G\tau}P(0)$. For R_0 we find, using Kermack-McKendrick (see also Metz (1978)),

$$R_0 = \int_0^\infty \sum_{i=1}^{3} h_i c_i P_i(\tau) d\tau = -\langle \tilde{h}, G^{-1}P(0) \rangle$$

where $\tilde{h} = (h_1c_1, h_2c_2, h_3c_3)^T$ and where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^3$. In this example we can also derive R_0 directly by a heuristic argument. The total amount of time that an individual is expected to be in class 1 is $1/(\theta_1 + \rho_1)$, and during that period it is expected to make $h_1c_1/(\theta_1 + \rho_1)$ new cases. A fraction $\theta_1/(\theta_1 + \rho_1)$ of the individuals passes into class 2. If an individual indeed enters class 2 then the expected time it remains there

is $1/(\theta_2 + \rho_2)$ and $h_2 c_2 \theta_1/[(\theta_1 + \rho_1)(\theta_2 + \rho_2)]$ new victims are expected to be made. Finally, a fraction $\theta_2/(\theta_2 + \rho_2)$ passes into class 3. There the individual is expected to remain for a time period $1/\rho_3$, and the number of expected new cases is $h_3 c_3 \theta_2 \theta_1/[(\theta_1 + \rho_1)(\theta_2 + \rho_2)\rho_3]$. If we add up the three results for the different classes we arrive at R_0 and we have derived $\int_0^\infty A(\tau)d\tau$.

$\diamondsuit$

In general one can state that if one disregards heterogeneity in susceptibility, and only takes variation in infectivity into account, then R_0 can be obtained by straightforward averaging over the infected population (see examples 0.3 and 0.4). If we include differences in susceptibility as well as in infectivity then one can still take averages if one assumes that susceptibility and infectivity do not influence each other (e.g. with respect to the probability of meeting). We call this 'separable mixing' (a generalisation of what is usually called 'proportionate mixing' in the literature) and discuss it in chapter 1. Problems arise with the approach to R_0 discussed above when we apply it to populations where we recognise a non-trivial h-state space, as well as variation in infectivity, and where the two influence each other. One could calculate the expected number of cases produced among susceptibles of h-state ξ, say, by infecteds that were 'born' with h-state η. One would obtain numbers for all combinations $(\xi, \eta) \in \Omega \times \Omega$. The problem then is how to average all these numbers to arrive at a quantity that has the same interpretation as R_0. More precisely, we want to average all the numbers in such a way that for the average the threshold condition holds. This problem is solved in chapter 1 (and for sexually transmitted diseases, if we take the length of relationships into account, in chapter 3). That finding the right average is indeed a problem is evident in many places in the literature (even recently, see e.g. Jacquez, Simon and Koopman (1991); Anderson and May (1991), section 9.3, 10.3.1).

Mathematical 'treatments' have previously been published only for special h-states, a systematic approach was lacking. In retrospect, there were both correct definitions for special cases (R_0 as dominant eigenvalue of a matrix, if only a finite number of different types of individuals are considered, see e.g. Hethcote and Yorke (1984), and R_0 as the spectral radius of an operator with age as h-state, see e.g Greenhalgh (1990)), and incorrect definitions that sometimes gave the correct answer only because the assumptions happened to be just right (e.g. R_0 as a weighted average, see e.g. Anderson and May (1991), section 10.3.1).

The crucial fact that makes it possible to answer the invasion question from section 0.2 in a general setting, is that in the very beginning, when the disease first arrives in the population, there are many more susceptibles than infecteds. Therefore we can, to a first approximation, assume that the susceptible population is of constant density in the initial phase of a (possible) epidemic (it does not decrease appreciably, even if some susceptibles contract the infection). Moreover, we can assume that the infected part of the population is initially so diluted that any new contact that an infected individual makes is necessarily

with a susceptible (the spatial density of infecteds can be assumed as being too low, in that phase, for two infecteds to meet with non-negligible probability). These are the key features of the invasion question that make the calculation of R_0 possible.

It is difficult to compute the precise value of R_0 for a given disease in a given population (e.g because one often lacks accurate estimates of the parameters involved). Fortunately, for the most important applications we do not need to know the accurate value of R_0; what is needed is the ability to know the dependence of R_0 on the parameters that one deems important for the disease transmission dynamics. The main, and very fruitful, use of R_0 is in testing the efficacy of control measures and vaccination programs that are aimed at eradicating a disease that is already present in the population (see Anderson and May (1991), for many examples). Due to the threshold condition, in order to locally eradicate a disease it is needed to suppress R_0 below unity for a sufficiently long period of time (see the Ross example above). R_0 depends on the density of susceptibles, the rate of recovery, the success ratio, to name but a few ingredients, and it is quantities like these whose values are affected by control measures. For example, using condoms during sexual contacts decreases the success ratio for sexually transmitted diseases; if the right medication can be given to an infected individual, then the individual will recover faster and consequently will have a shorter infectious period in which to transmit the disease to others (e.g. gonorrhoea); if a vaccine is available for a certain disease, then the probability that an infected individual meets a susceptible will decrease if (part of) the population is vaccinated. If we allow for heterogeneity, then the questions arise: which 'distribution' of control measures over all the possible types of individuals would be most efficacious for lowering transmission?; if a vaccine is available, which types of individuals should we administer it to, and in what frequency (vaccination strategy)?

Example. We take a brief look at the evaluation of different vaccination strategies for common childhood diseases. The information is taken from Van Druten et al. (1986). For additional information see Hethcote (1989), Greenhalgh (1990), Anderson and May (1983). For rubella (german measles), which is a mild illness in most individuals but is a serious threat to pregnant women and their unborn offspring, there are three strategies in use. One is to vaccinate all young children, a second is to vaccinate only prepubertal girls, and a third is a combination: vaccination of all children at a young age (about 1 year old) and again at about 9 years old. Originally in the Netherlands the second strategy was followed, but as of 1984 one follows the third (with a cocktail of measles, mumps and rubella vaccines). Mathematical analysis can show that the adherence to certain strategies can in fact *increase* the fraction of serious cases. The reason is that complications of rubella infection depend on the age at infection. If the chosen vaccination strategy does not eliminate the disease but only brings it to a certain low level, then the force of infection in the population will decrease and consequently the average age at which infection is contracted

can rise, and more women run the risk of receiving the infection during pregnancy. Important questions that can be answered by mathematical analysis, for example by studying the effect of vaccination on the threshold quantity, are: which immunisation coverage is required to eliminate rubella? (around 95% of the young children in the Netherlands); if elimination cannot be attained what consequences does this have for the fraction of unvaccinated individuals? It was thought that with the current combined vaccination strategy against measles, mumps and rubella in the Netherlands it should be possible to eliminate these three diseases in the first half of the nineties. However, recent discoveries show that the measles virus can persist in (and be transmitted through) saliva even if the individual involved is vaccinated and does not show any symptoms of the disease (A.D.M.E. Osterhaus, RIVM, personal communication). This suggests that the virus could persist in the population, even with a high vaccination coverage (100% coverage is always unattainable because, e.g., some groups of people refuse for religious reasons).

We can summarise the strategy for the application of R_0 to 'real-world' problems as follows. For a concrete disease one determines the infectivity function $A(\tau, \xi, \eta)$ and proceeds to calculate R_0 (if necessary under some special assumptions) along the lines of chapter 1. Next, one investigates, for example graphically, the qualitative way in which R_0 depends on key parameters (such as the recovery rate, the fraction of the population that is vaccinated, the success ratio, depending of course on the control strategy under scrutiny). Which (combination of) parameter(s) leads to the fastest decrease in R_0?

In this thesis we are mainly concerned with methodology. Although we give examples that show how the function A can be composed from more basic modelling ingredients, we do not, for the larger part, discuss any real epidemiological applications that could be directly useful to, for example, base public health decisions on (see also Diekmann (1991)). The applications to 'real world' problems we do discuss, along the lines suggested above, both concern sexually transmitted diseases. The first is given in section 1 of chapter 2, the second takes up sections 3.5 and 3.6 of chapter 3. Apart from these applications, our treatment of the problem remains quite abstract from an epidemiological point of view. The results presented here can provide a basis for subsequent more quantitative work. The application of the abstract theory to reach meaningful biological conclusions about actual epidemiological questions, is far from trivial. A particularly nice example of the kind of follow-up that could also be envisaged for the abstract theory in this thesis, is given by Van den Bosch in his papers on the spatial spread of plant diseases (see Van den Bosch, Zadoks and Metz (1988a,b), Van den Bosch, Frinking, Metz and Zadoks (1988)). He gives a successful biological operationalisation of the abstract calculation (from, among others, Diekmann (1978)) of the asymptotic propagation speed of solutions to certain integral equations. This operationalisation is carried through right to the level of experimental data, where the, formerly abstract, mathematics yields biological insight about the spread of plant diseases in a field.

0.4. A superinfection model

Before we start with the thesis proper in the next chapter, we briefly return to the problems with acquired immunity to malaria as mentioned in section 0.1. We only indicate here how the spread of an infectious disease with superinfection could be modelled; this section should be read as a report on work in progress (with M. Kretzschmar).

We assume that there is no heterogeneity in susceptibility. The idea, originally due to J.A.J. Metz, is to extend the approach for diseases with 'age of infection' as d-state, going back on Kermack-McKendrick, and introduce the more complicated d-state $(\tau_1, \tau_2, ..., \tau_n)$. Here, the individual is thought to have received a total of n doses of infection (of the same disease causing organism, but possibly of different strains), where the first infection is taken to be contracted τ_1 time-units ago, the second one τ_2 time-units ago, etcetera, and where the last infection arrived τ_n time-units ago. Instead of a function $A(\tau)$ we now introduce an infectivity function $A_n(\tau_1, ..., \tau_n)$ which we assume to fully describe the conditional infectivity of individuals who have experienced n infections in the past, and whose last infection was received τ_n time-units ago, given that the next-to-last was received τ_{n-1} time-units ago, etcetera, given that the first infection was received τ_1 time-units ago. Here 'infection' is defined as a new dose of infectious material that enters the body of the host-individual, irrespective of the fact that perhaps the immune system could be able to eradicate this material immediately. For every dose that enters we start a new clock. Once infected these τ-clocks keep on ticking and the function A_n determines the individual's 'infectivity status'. We keep track of all re-infections, whether or not they have already been cleared from the body of the host, the contribution of a particular re-infection to the infectivity at some later point in time being described entirely by A_n. We always keep track of the total number of infections received, irrespective of the possibility that an individual can be disease-free for some time-intervals (A_n does not need to be a increasing function of n).

Let $i_n(t, \tau_n, ..., \tau_1)$ be the number of individuals that at time t have experienced n infections, where the first infection was contracted τ_1 time-units ago, the second one τ_2 time-units ago, etcetera, and where the last infection arrived τ_n time-units ago. The word 'susceptible' no longer has the same meaning as before; we propose to call individuals who have never been infected, *virgins*, and indicate them by S. For convenience we introduce the following interval-notation: let $s_i = \tau_{i-1} - \tau_i$, where τ_i is the time elapsed since the ith infection (τ_0 is taken as the age of the individual). Then s_1 is the length of the time-interval that the individual is a virgin. The time elapsed since the last incidence of infection is indicated by s ($s = \tau_n$). The age of an individual is given by $s + s_1 + \cdots + s_n$. Define the n-tuple $\sigma_n := (s_1, ..., s_n)$, then σ_n does not change with time, and we retain s and t as the only time variables.

We denote the individuals with n infections by $j_n(t, s, \sigma_n)$, $n \geq 0$, where j_n is the function i_n transformed to interval-notation. We will write $\mathcal{A}_n$ for the transformed infectivity function A_n. As state space we take

$$\Omega := \cup_{n=0}^{\infty} \Omega_n,$$

with $\Omega_n := \{(s, \sigma_n) \in \mathbb{R}^{n+1} | s \geq 0, s_i > 0, i = 1, ..., n\}$.

If we let $\lambda(t)$ describe the force of infection at time t, then, for $n > 0$,

$$j_n(t, 0, \sigma_n) = \lambda(t) j_{n-1}(t, s, \sigma_{n-1}) \tag{0.4.1}$$

where $\sigma_n := (s, \sigma_{n-1})$ (i.e. s_n is set equal to s at the moment of a renewed infection, the clock s is subsequently put to zero again). If we allow a natural per-capita death rate μ, we can describe the infection process by

$$\frac{\partial j_n}{\partial t} + \frac{\partial j_n}{\partial s} = -(\lambda(t) + \mu) j_n(t, s, \sigma_n), \tag{0.4.2}$$

with boundary condition given above for $n \geq 1$, and

$$i_0(t, 0) = \Lambda(t), \tag{0.4.3}$$

for a given function $\Lambda(t)$ that describes births of new virgins. For the force of infection we take the most general formulation in that we let all infectivity in the population count

$$\lambda(t) = \sum_{n=1}^{\infty} \int_{\Omega_n} \mathcal{A}_n(s, \sigma_n) j_n(t, s, \sigma_n) d(s, \sigma_n). \tag{0.4.4}$$

One feels that both 'single infection models with immunity' (where the d-state is the time elapsed since the first infection) and certain 'wormload models' (where the d-state is the *number* of worms in the body of the individual), should be special cases of our model. The idea is that in the first case one has to assume that all infections after the first one have no influence whatsoever, and that in the second case, we interpret our model in such a way that we start a 'new clock' for every new worm that enters the body of the host. With some redefining of model-ingredients the special cases can, not surprisingly, indeed be obtained.

In the 'single infection with immunity'-case we obtain the Kermack-Mckendrick model discussed in section 0.3. (system (0.3.2)) (in this case a version with vital dynamics). For this, define $\tau := s + s_n + \cdots + s_2$ and let $\mathcal{A}_n(s, \sigma_n) = \mathcal{A}(\tau)$ depend only on time elapsed since the first infection. Let $\sigma_n' := (s_2, s_3, ..., s_n)$. If we then take

$$j(t, \tau) := \sum_{n=1}^{\infty} \int_{s+s_n+\cdots+s_2=\tau} \int_0^{\infty} j_n(t, s, \sigma_n) ds_1 d(s, \sigma_n')$$

then (0.4.1)-(0.4.4) 'collapses' into system (0.3.2).

In the 'wormload'-case we first have to slightly generalise the boundary condition (0.4.1)

$$j_n(t,0,\sigma_n) = f_{n-1}(\lambda(t))j_{n-1}(t,s,\sigma_{n-1}).$$

Furthermore we allow the infectivity function $\mathcal{A}_n$ to depend on time, and we will write the infectivity as $B_n(t,s,\sigma_n)$. The function f, which can of course be taken as the identity for the case described in the previous paragraph, is necessary because in the life-cycle of many worm-diseases an intermediate host-species operates. For example, eggs of adult schistosome parasites in humans are shed from the body of the host; in water, free living larvae (miracidia) develop that infect snails (the intermediate host); the snail releases a second aquatic stage of the parasite (cercaria) that can infect humans that come into contact with infected water (see e.g., Anderson and May (1991)). The function f is needed to be able to describe the effect on infectivity of 'passing through' the intermediate host (how is the amount of eggs shed by the human host related to the amount of cercaria). For assumptions on f relevant in the context of worm-infections, see e.g. Kretzschmar (1989). For example, one could take $f_n(\lambda(t)) = \kappa_n g(\lambda(t))$ where κ_n describes the contact rate of the host with infectious stages coming from the intermediate host (e.g. in case of schistosomes some measure for watercontact), and g describes the connection between average parasite load of a human host individual and the resulting production of stages infective to humans by the intermediate host).

In the function B_n one would, among other things, take the death of individual worms in the host into account, and the production by worms inside the host of stages of the parasite that are infectious to the intermediate host. We look at the simplest case where $B_n(t,s,\sigma_n) = \bar{B}_n(t)$ (here we assume that the lifetime of the adult parasite is of the same order of magnitude as the lifetime of the host, and that adult worms are not shed from the host body, though eggs are). Define

$$j_n(t,a) := \int_{s+s_n+\cdots+s_1=a} j_n(t,s,\sigma_n)d(s,\sigma_n)$$

as the number of individuals of age a that carry n worms at time t. Then we can write the force of infection $\lambda(t)$ as

$$\lambda(t) = \int_0^\infty \sum_{n=1}^\infty \bar{B}_n(t)j_n(t,a)da.$$

If we choose

$$\bar{B}_n(t) = \frac{n}{\int_0^\infty \sum_{n=1}^\infty j_n(t,a)da}$$

then $\lambda(t)$ corresponds to the average parasite load of an infected host individual. For more realism, one should retain the full unsimplified equation for $\lambda(t)$.

Remark. The basic reproduction ratio for the superinfection model (0.4.1)-(0.4.4) is easy to calculate with the following heuristic considerations. Remember that for R_0 we regard the beginning of an epidemic when the number of infected individuals is very small. In that phase therefore, practically all individuals are virgins and we can assume that the probability that an individual is infected twice (in the malaria case: a mosquito takes two infected bloodmeals/a human is fed upon twice by infected mosquitoes) is negligible. Thus we can, for the invasion problem, restrict ourselves to two disease classes, the uninfected virgins and the individuals who have received a single infection. We are then in the same situation as Kermack-McKendrick (1927), albeit that we have a special form for the function A, and $R_0 = S\int_0^\infty A_1(\tau_1)d\tau_1$.

$\diamondsuit$

Future research on the superinfection model (0.4.1)-(0.4.4) takes two directions: a mathematical route, and a modelling route. By putting the t-derivatives to zero one can easily derive an equation for the existence of an endemic equilibrium

$$1 = \sum_{n=1}^{\infty} \int_{\Omega_n} \mathcal{A}_n(s,\sigma_n)\Lambda\lambda^{n-1}e^{-(\lambda+\mu)(s+s_1+\cdots+s_n)}d(s,\sigma_n)$$

and subsequently investigate under which conditions on $\mathcal{A}_n$ an endemic equilibrium actually exists (here the theory of Laplace transforms appears to be useful); ultimatly we want to obtain insight into the dynamic behaviour of the model. It would, for example, be interesting to determine the influence of vaccination in the context of the model.

As far as the modelling route is concerned, the underlying aim is to apply the model to malaria. This has to be made concrete in the determination of a submodel to obtain insight in the characteristics of A_n specific for malaria transmission. The mechanics of acquired immunity seem to be based on repeated infection by different strains of the same parasite species (Wernsdorfer and McGregor (1988)). In this way the antibody repertoire of the infected individual grows and grows, until most of the (locally) present strains have been encountered. After that, most additional infections are 'familiar' to the immune system and can be dealt with more effectively.

Instead of going into details about the further analysis, we use the superinfection model to linger for a moment on the philosophy of doing mathematical epidemiology. The above 'progress report' is included because it illustrates nicely what, in our opinion, research, that one would classify as *mathematical epidemiology*, should ideally be like.

On the one hand the superinfection model is very general, and may also be of use to other diseases than malaria (the phenomenon of acquired immunity also plays a role in the epidemiology of worm diseases, see Anderson and May, (1991), chapter 18), and therefore clearly calls for the purely mathematical

aspects of proving theorems of the kind: 'Under these and these conditions on A_n an endemic equilibrium is possible', and 'Under conditions so and so on A_n this equilibrium is globally/locally stable'. The general model can be used to investigate in what way superinfection affects the dynamics of the disease, and evaluate how acquired immunity interacts with given control measures.

On the other hand the model needs, to apply it to malaria (or any other disease where gradual rise in immunity status occurs), pure modelling endeavour to describe, on an immunological basis, the waxing and waning of the acquired immunity level in dependence of the time-chain of infections arriving in the body of the host, in order to suggest a not unreasonable form of the function $A_n(\tau_1, ..., \tau_n)$. In this part, one of the difficulties (read: challenges) is to try to base the immunological sub-model on parameters that can be measured or at least allow for an educated guess, otherwise the whole undertaking is of academic value only (this, incidentally, need not be a negative thing in general).

The interplay between the mathematical and the modelling part can lead to an interesting paradox. It is better to determine a reasonable A_n first, because if that fails the model will be of no use to investigate the dynamics of malaria. However, it is also better to prove theorems first about the behaviour of solutions of the general model, because if this fails, or if the model does not allow any 'interesting' dynamics, then the whole exercise to determine a reasonable A_n will be obsolete. The mathematics and the modelling should somehow be balanced.

If the balance tips to the mathematical side one obtains results like existence and uniqueness, by well-known fixed point arguments, of yet another variation from the set of complicated ODE or PDE systems which allow an epidemiological interpretation. Reading papers of this kind, which invariably start with the flimsiest of explanations about the profound biological background of the equations only to completely ignore these lines for the rest of the paper once the equations have been stated (omitting any form of feedback to the biology), quickly leads to a feeling of: 'you showed that these equations have a solution, now what?' One should not pass off purely mathematical papers as mathematical epidemiology (or indeed biomathematics). Really interesting are these papers only if they introduce new techniques, provide new concepts and offer insights, or if they describe a methodology that can be useful to others.

If the balance tips toward the modelling side, one gets extremes such as research that spends a huge amount of time on determining a model for the spread of vivax malaria between the even-numbered houses of Wurno, Nigeria. The problem here is that the methods involved, and the insights gained, have almost always a wider applicability to other problems. Researchers at the other side of the globe could be wrestling with similar problems albeit for a different situation (e.g., a completely different disease and its spread between odd-numbered houses). Trying to place ones details in a more general framework is what science is all about. Furthermore, as is fittingly illustrated in chapter 1, the most insightful way to regard a problem is often through

generalisation and abstraction.

It is in the delicate balance and continous cross-fertilisation between abstract methodolgy that has large generality, and specific applications to given 'real-world' biological problems, that the heart of mathematical biology lies. Whether or not there exist any researchers (including yours truly) who carry their heart in the right place is another matter entirely.

0.5. References

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

Andreasen, V. (1989): Disease regulation of age-structured host populations. *Theor. Pop. Biol.* **36**: 214-239.

Aron, J.L. (1983): The dynamics of immunity boosted by exposure to infection. *Math. Biosc.* **64**: 249-259.

Beretta, E. & V. Capasso (1986): On the general structure of epidemic systems. Global asymptotic stability. *Comp. & Maths. with Appls.* **12A**: 677-694.

Blanchard Ph., Bolz, G.F. and T. Krüger (1990): Modelling AIDS-Epidemics or any venereal disease on random graphs. In: *Stochastic Processes in Epidemic Theory*, J.-P. Gabriel, C. Lefèvre & P. Picard (eds.), Lecture Notes in Biomathematics No. 86, pp. 104-117.

Busenberg, S.N., Iannelli, M. & H.R. Thieme (1991a): Global behavior of an age-structured epidemic model. *SIAM J. Math. Anal.* **22**: 1065-1080.

Busenberg, S.N., Iannelli, M. & H.R. Thieme (1991b): Demographic change and persistence of HIV/AIDS in a heterogeneous population. *SIAM J. Appl. Math.* **51**: 1030-1052.

Cliff, A.D. & P. Haggett (1988): *Atlas of Disease Distributions.* Basil Blackwell, Oxford.

Diekmann, O. (1978): Thresholds and travelling waves for the geographical spread of infection. *J. Math. Biol.* **6**: 109-130.

Diekmann, O. (1991): Modelling infectious diseases in structured populations. In: B.D. Sleeman & R.J. Jarvis (eds) *Ordinary and Partial Differential Equations, vol. III.* Pitman RNiMS 254: 67-79. Longman, Harlow.

Diekmann, O., Gyllenberg, M., Metz, J.A.J. & H.R. Thieme (1992): The 'cumulative' formulation of (physiologically) structured population models. preprint, CWI-report AM9205.

Diekmann, O., Heesterbeek, J.A.P., Kretzschmar, M. & J.A.J. Metz (1989): Building Blocks and Prototypes for Epidemic Models. Preprint.

Dietz, K. (1975): Transmission and control of arbovirus diseases. In: *Epidemiology.* D. Ludwig & K.L. Cooke (eds), pp. 104-121, SIAM, Philadelphia

Fisher, R.A. (1930): *The Genetical Theory of Natural Selection.* Clarendon, Oxford.

Greenhalgh, D. (1988): Threshold and stability results for an epidemic model with an age-structured meeting rate. *IMA J. Math. Appl. Med. Biol.* **5**: 81-100.

Greenhalgh, D. (1990): Vaccination campaigns for common childhood diseases. *Math. Biosc.* **100**: 201-240.

Hadeler, K.P. & K. Dietz (1983): Nonlinear hyperbolic partial differential equations for the dynamics of parasite populations. *Comp. Math. Appl.* **9**: 415-430.

Hamer, W.H. (1906): Epidemic disease in England. *Lancet*, March 17: 733-739.

Hethcote, H.W. & J.A. Yorke (1984): *Gonorrhea Transmission Dynamics and Control.* Lect. Notes Biomath. Vol. 56, Springer-Verlag, Berlin.

Hethcote, H.W. (1989): Rubella. In: S.A. Levin, T.G. Hallam, L.J. Gross (eds.) *Applied Mathematical Ecology*, Biomathematics Texts 18: 212-234, Spinger-Verlag, Berlin.

Hethcote, H.W. (1990): Current issues in epidemiological modeling. Preprint.

Inaba, H. (1990): Threshold and stability results for an age-structured epidemic model. *J. Math. Biol.* **28**: 411-434.

Jacquez, J.A., Simon, C.P. & J.S. Koopman (1991): The reproduction number in deterministic models of contagious disease. *Comments on Theor. Biol.* **2**: 159-209.

Lauwerier, H.A. (1984): *Mathematical Models of Epidemics*, Mathematical Centre Tracts 138, Amsterdam.

Lin, X. & J.W.-H. So (1990): Global stability of the endemic equilibrium in epidemic models with subpopulations. Preprint.

Lotka, A.J. (1923): Contributions to the analysis of malaria epidemiology *Am. J. Hyg.* **3** (suppl. 1): 1-121.

Kermack, W.O. & A.G. McKendrick (1927): Contributions to the mathematical theory of epidemics, part I. *Proc. Roy. Soc. Edinb. A* **115**: 700-721.

Kermack, W.O. & A.G. McKendrick (1932): Contributions to the mathematical theory of epidemics, part II (the problem of endemicity). *Proc. Roy. Soc. Edinb. A* **138**: 55-83.

Kermack, W.O. & A.G. McKendrick (1933): Contributions to the mathematical theory of epidemics, part III (further studies of the problem of endemicity). *Proc. Roy. Soc. Edinb. A* **141**: 94-122.

M. Kretzschmar (1989): A renewal equation with a birth-death process as a model for parasitic infections. *J. Math. Biol.* **27**: 191-221.

MacDonald, G. (1957): *The Epidemiology and Control of Malaria.* Oxford University Press.

May, R.M., Anderson, R.M. & A.R. McLean (1988a): Possible demographic consequences of HIV/AIDS epidemics: I. Assuming HIV infection always leads to AIDS. *Math. Biosc.* **90**: 475-505.

May, R.M., Anderson, R.M. & A.R. McLean (1988b): Possible demographic consequences of HIV/AIDS epidemics: II. Assuming HIV infection does not necessarily lead to AIDS. In: C. Castillo-Chavez, S.A. Levin, C.A. Shoemaker (eds.) *Mathematical Approaches to Problems in Resource Management and Epidemiology*, Lect. Notes Biomath. vol. 81: 220-248.

Metz, J.A.J. & O. Diekmann (eds) (1986): *The Dynamics of Physiologically Structured Populations.* Lecture Notes in Biomath. vol. 68, Springer, Heidelberg.

Ross, R. (1909): *The Prevention of Malaria.* Murray, London.

Smith, D. & N. Keyfitz (1977): *Mathematical Demography, selected papers.* Biomathematics vol. 6, Springer Verlag, Berlin.

Van den Bosch, F., Zadoks, J.C. & J.A.J. Metz (1988a): Focus formation in plant disease, I: The constant rate of focus expansion. *Phytopathology* **78**: 54-58.

Van den Bosch, F., Zadoks, J.C. & J.A.J. Metz (1988b): Focus formation in plant disease, II: Realistic parameter-sparse models. *Phytopathology* **78**: 59-64.

Van den Bosch, F., Frinking, H.D., Metz, J.A.J. & J.C. Zadoks (1988): Focus formation in plant disease, III: Two experimental examples. *Phytopathology* **78**: 919-925.

Van Druten, J.A.M., Th. de Boo, A.D. Plantinga (1986): Measles, mumps and rubella: control by vaccination. In: *Develop. Biol. Standard* vol. 65: 53-63, Karger, Basel.

Wernsdorfer, W. & I.A. McGregor (eds.) (1988): *Principles and Practice of Malariology,* Livingstone, Edinburgh.

Chapter 1

Definition and calculation of R_0

The biological 'definition' of R_0 is the expected number of secondary cases produced by a typical infected individual during its entire period of infectiousness in a demographically steady susceptible population. In this chapter we explain that the mathematical counterpart of this 'definition' is to define R_0 as the spectral radius of a certain positive linear operator on $L_1(\Omega)$, where Ω is the heterogeneity state space. In special cases one can easily calculate or estimate R_0 using the mathematical definition. Several examples involving various different types of heterogeneity states are briefly discussed.

1.1. Introduction

Suppose we want to know whether or not a contagious disease can 'invade' into a population of humans, animals or plants, which is in a demographic steady state (at the time-scale of disease transmission) with all individuals susceptible*. To settle this question we first of all linearise, i.e. we ignore the fact that the density of susceptibles decreases due to the infection process. It has become common practice in the analysis of the simplest models to consider next the associated *generation process* and to define the basic reproduction ratio R_0 as the expected number of secondary cases produced, in a demographically steady susceptible population, by a *typical* infected individual during its entire period of infectiousness. The famous *threshold criterion* then states:

* It could be that not all individuals are susceptible to the disease. For example, a fraction of the population might be successfully vaccinated, or recovered and immune. This does not complicate matters as long as we assume that the population is in a steady state with respect to the susceptible/not susceptible subdivision, and as long as we have many more susceptibles than initially infecteds.

the disease can invade if $R_0 > 1$, whereas it cannot if $R_0 < 1$.

(Of course, it could well happen that even when $R_0 < 1$ the initial infected population will still increase a little before the disease dies out, but this increase is then less than the exponential increase that is characteristic of beginning epidemics.) It is the aim of this chapter to demonstrate how these ideas extend to less simple (though probably still highly oversimplified) models involving *heterogeneity* in the population, and to explain the meaning of 'typical' in the 'definition' of R_0. Subsequently, we shall deal with the actual computation of R_0 in certain special cases, in particular the so-called 'separable' or 'weighted homogeneous mixing' case.

1.2. The definition

Let the individuals be characterised by a variable ξ, which we shall call the h-state variable, taking values in some state space Ω. Let $S = S(\xi)$ denote the density function of susceptibles, describing the steady demographic state in the absence of the disease (in order to avoid confusion we emphasise that S is not a probability density function; that is, its integral equals the total population size in the demographic steady state, and not one). We will start with a general formulation (in terms of integral operators) of a model with heterogeneity. Define $A(\tau, \xi, \eta)$ as the expected infectivity of an individual which was infected τ units of time ago, while having h-state η, towards a susceptible with h-state ξ. We define further $i(t, \xi)$ to be the number of newly infected individuals with h-state ξ arising per unit of time, evaluated at time t.

We want to find an expression relating $i(t, \xi)$ to its past history. If we assume pure mass-action we can take $i(t, \xi)$ to be a factor $M(t, \xi)$ times $S(\xi)$: $i(t, \xi) = S(\xi)M(t, \xi)$, where we have already linearised by taking for S the steady demographic state. The factor $M(t, \xi)$ is a measure for the 'total infective pressure' per unit of time acting on the susceptibles with h-state ξ (force of infection) and will in general depend on ξ. To express M in previously defined quantities we note that individuals with d-state τ were themselves infected at time $t - \tau$, and that these infecteds have at time t infectivity $A(\tau, \xi, \cdot)$ towards susceptibles with h-state ξ. We then get a contribution $A(\tau, \xi, \eta)i(t - \tau, \eta)$ to the infective pressure per unit of time on the susceptibles of h-state ξ. Integrating over all possible h-state values and all τ gives the total infective pressure. This leads to

$$i(t, \xi) = S(\xi) \int_0^\infty \int_\Omega A(\tau, \xi, \eta) i(t - \tau, \eta) d\eta d\tau \tag{1.2.1}$$

for the linearised real time equation. In the Appendix we shall explain the link with models formulated in terms of differential equations.

Remark. In order to have a unified notation for various cases we write integrals to denote sums whenever Ω is discrete (either completely, or just with respect to some component of ξ).

$\diamondsuit$

Instead of the real-time process (1.2.1), we will consider the associated generation process. We will regard *generations* of infected individuals. We call the initially infected individuals (the ones that bring the disease into the 'virgin' population) the first generation, all individuals whose infection is caused by members of this first generation, will collectively be called the second generation, etcetera. The expected number of infections produced during its entire infective life by an individual which was 'born' with h-state η is given by

$$\int_\Omega S(\xi) \int_0^\infty \int_\Omega A(\tau, \xi, \eta) d\tau d\xi.$$

We may call this quantity the next-generation factor of η.

Since all new cases arise, in general, with h-states different from η, these numbers do not tell us exactly what happens under iteration, i.e. in subsequent generations (although it is clear that the supremum with respect to η yields an upper estimate for R_0). So we abandon the idea of introducing an infected individual with a particular well-defined h-state and start instead with a 'distributed' individual, described by a density $\phi \in L_1(\Omega)$ (which is of course nonnegative). This density ϕ denotes the present generation of infected individuals, distributed over the h-state space Ω. The next generation is then given by

$$(K(S)\phi)(\xi) = S(\xi) \int_\Omega \int_0^\infty A(\tau, \xi, \eta) d\tau \phi(\eta) d\eta \tag{1.2.2}$$

which tells us both how many secondary cases arise from the generation ϕ and how these new cases are distributed over Ω. We will call $K(S)$ the *next-generation operator* and we assume that we have sufficient conditions on A and S to guarantee that $K(S)$ is a positive operator on $L_1(\Omega)$.

Remark. The following conditions are sufficient. Let $S(\cdot) \in L_\infty(\Omega)$ and nonnegative; $A(\cdot, \cdot, \cdot)$ is defined and nonnegative on $[0, \infty) \times \Omega \times \Omega$; $A(\cdot, \cdot, \eta) \in L_1([0, \infty) \times \Omega)$ and uniformly bounded in norm with respect to η. Then it is easy to show that $K(S)$ is a positive (continuous) operator on $L_1(\Omega)$ (nonnegative functions are mapped onto nonnegative functions).

$\diamondsuit$

We note that the next-generation factor of ϕ is simply the $L_1(\Omega)$-norm of $K(S)\phi$, i.e.,

$$\int_\Omega S(\xi) \int_\Omega \int_0^\infty A(\tau, \xi, \eta) d\tau \phi(\eta) d\eta d\xi$$

(we do not have to write absolute value signs because of the remark above). If we take the supremum of the next-generation factor over all ϕ with $\|\phi\| = 1$,

then we obtain, by definition, the operator norm of $K(S)$. This yields an upper estimate of R_0 for the same reason as above: the distribution with respect to ξ is changed in the next generation and consequently the factor of ϕ does not predict accurately what happens under iteration.

Example. As a concrete example consider a host-vector model. (For the purpose of exposition, we adopt here the strict version of the law of mass action, even though this does not necessarily yield a good model in this case, see, e.g., chapter 7 of Bailey (1982).) Taking $\Omega = \{1, 2\}$ and write $\int_0^\infty A(\tau, i, j)d\tau = a_{ij}$ with $a_{ij} > 0$ if and only if $i \neq j$, we find that $K(S)$ is represented by the matrix

$$\begin{pmatrix} 0 & a_{12}S_1 \\ a_{21}S_2 & 0 \end{pmatrix}$$

and the operator norm is $\max\{a_{12}S_1, a_{21}S_2\}$ (the two numbers correspond to vector $\rightarrow$ host, and host $\rightarrow$ vector transmission, respectively). No matter which of the two is the larger one, in the next generation it is necessarily the other of the two numbers which is the relevant *factor*. Therefore, the operator norm of $K(S)$ is not a good definition of R_0. Since $a_{12}S_1a_{21}S_2$ is the two-generation factor, the *average* next-generation factor* is

$$\sqrt{a_{12}S_1a_{21}S_2},$$

which is $\leq \|K(S)\|$. How can we define such an average quantity in general?

After m generations the magnitude of the infected population is (in the linear approximation) $K(S)^m\phi$ and consequently, the per-generation growth factor is $\|K(S)^m\|^{1/m}$. We want to know what happens to the population in the long run, so we let $m \rightarrow \infty$. The spectral radius of $K(S)$, $r(K(S))$, is defined as $\sup\{|\lambda| | \lambda \in \sigma(K(S))\}$, where $\sigma(K(S))$ denotes the spectrum of $K(S)$. Then the well-known relations

$$r(K(S)) = \inf_{m \geq 1} \|K(S)^m\|^{1/m} = \lim_{m \rightarrow \infty} \|K(S)^m\|^{1/m}$$

hold (e.g. Rudin (1973)). Starting from the first generation ϕ, the mth generation $K(S)^m\phi$ converges to zero for $m \rightarrow \infty$ if $r(K(S)) < 1$, whereas it can be made arbitrarily large by a suitable choice of ϕ and m, when $r(K(S)) > 1$.

Moreover, the *positivity* guarantees that in the latter case there is not really a restriction on ϕ. The positivity of $K(S)$ implies that $r(K(S))$ is an eigenvalue of $K(S)$ (Schaefer (1960, 1974)), and in fact it is then a dominant

* One could argue, as MacDonald did (see Bailey (1982), p. 100, and the references given there), that one should consider the average number of cases in the host population arising from one case in the host population *via* vector cases. From our point of view this amounts to looking *two* generations ahead. Indeed, one obtains exactly MacDonald's result if one writes out $a_{12}S_1a_{21}S_2$ in terms of biting rates, etc.

eigenvalue (in the sense that $|\lambda| \leq r(K(S))$ for all λ in the spectrum of $K(S)$) and we denote it by ρ_d. As a rule (see remark 1 below) one has in addition convergence to a steady distribution

$$K(S)^m\phi \sim c(\phi)\rho_d^m\phi_d \quad \text{for} \quad m \to \infty,$$

where ϕ_d is the eigenvector (which is positive) corresponding to ρ_d, and $c(\phi)$ is a scalar which is positive whenever ϕ is nonnegative and not identically zero. So, after a certain period of transient behaviour, each generation is (in an approximation which improves as time proceeds) ρ_d times as big as the preceding one, and distributed over the h-state space as described by ϕ_d. We call the h-state of an individual at the moment it becomes infected the 'state at birth', and we normalise $\|\phi_d\| = 1$. Then ϕ_d allows the following interpretation. Given that an individual becomes infected, ϕ_d is the probability distribution for the state at birth of that individual.

If we rephrase this as: 'the *typical* number of secondary cases is ρ_d', then we are ready for the

Definition $R_0 = r(K(S)) = \rho_d =$ the dominant eigenvalue of $K(S)$.

Note that here we run with the hare and hunt with the hounds (Diekmann (1991)). We have linearised because we are only interested in the beginning of the epidemic, but at the same time we iterate very often because we do not want our result to depend on the precise details of the introduction of the infection. If it takes many iterations before one observes the stable distribution ϕ_d and the multiplication factor ρ_d for the linearised problem, it may be that the nonlinearity in the 'full' problem is already noticeable. This happens, for example, for the spatial spread of epidemics on large domains (see Van den Bosch, Diekmann, and Metz, (1990); see also example 1.4.4). In general, the speed of convergence to the stable distribution is initially determined by the 'degree of (ir)reducibility' of $K(S)$. For example, if the infection enters the population in a subgroup that has only very weak links to other subgroups ($K(S)$ 'almost' reducible), it can take many generations before the infection passes from the initial subgroup into other subgroups. After this initial phase the speed of convergence is determined by the difference, in absolute value, between the dominant eigenvalue and the remainder of the spectrum of $K(S)$.

Where in the homogeneous case the threshold criterion is a direct consequence of the 'definition' of R_0, we now actually have to prove a threshold theorem that relates the R_0 defined on a generation basis to the development of the epidemic in real time (both in the linearised version). Before we prove this result, we make a number of remarks about the definition of R_0.

Remarks

(1) In order to guarantee that *any* introduction of infection into the population leads to an epidemic when $R_0 > 1$, we need to assume *irreducibility*. Otherwise it could, for example, happen that the infection enters the population in a

subgroup that has no contacts with any other sub-groups. Irreducibility of $K(S)$ is defined as follows (Schaefer, 1974). Let $X \subset \Omega$ be any subset of Ω of positive (Lebesgue) measure, and such that also Ω/X has positive measure. Then, for each X satisfying this assumption, the expected infectivity A must satisfy

$$\int_{\Omega/X} \int_X \tilde{A}(\xi, x) dx d\xi > 0,$$

(where we have written $\tilde{A}(\cdot, \cdot) = \int_0^\infty A(\tau, \cdot, \cdot) d\tau$). One can rephrase this by saying that for a given generation ϕ of infected individuals that 'lives' on a subset X of Ω, one must have that the next generation $K(S)\phi$ is strictly positive on some set of positive measure disjoint from X. As an example, think of spatial coordinates in $\mathbb{R}^2$ as h-state, and consider a fungal plant disease in a field. Suppose we have a focus of positive area anywhere in that field, then irreducibility means that we expect that plants in a positive area outside the focus will become infected through spores coming from the plants within the focus. 'Age' as h-state also provides an illustrative example with analogous reasoning. In practical cases one will normally have irreducibility.

In case of a discrete h-state space, i.e. when $K(S)$ can be represented by a positive matrix, irreducibility is equivalent to the property that for any element of $K(S)$ there must exist a $p > 0$ such that the corresponding element in $K(S)^p$ is strictly positive. We can interpret this as saying that there must be some chain of consecutive infections leading from 'each h-state to any other h-state'.

To assure convergence to a stable distribution we need that $\rho_d > 0$, that ρ_d is strictly dominant and simple, and that ϕ_d is strictly positive. A sufficient condition that guarantees this is that a power of $K(S)$ is compact, and $K(S)$ is irreducible (Jentzsch's Theorem, see Schaefer (1974)). In the case of a discrete h-state it is sufficient that the matrix $K(S)$ is primitive. This is a stronger condition than irreducibility and demands that a number $p > 0$ can be found such that $K(S)^p$ is strictly positive.

(2) Parameters similar to R_0 determine the asymptotic behaviour in branching processes with general state space. See Jagers and Nerman (1984); Mode (1971).

(3) Let $\hat{S}$ denote the susceptible population in a steady *endemic* state. Then necessarily $r(K(\hat{S})) = 1$. See section 2.2.2.

(4) We have restricted our attention to the bilinear case. However, replacing $S(\xi)$ in the definition (1.2.2) of $K(S)$ by $h(S(\xi))$ or some saturating functional response $S(\xi)/(1+\int_\Omega S(\eta)d\eta)$ or something similar, does not make any essential difference. For sexually transmitted diseases one would for example take

$$h(S)(\xi) = \frac{S(\xi)}{\int_\Omega c(\eta)S(\eta)d\eta},$$

where c is some appropriate function, because we somehow have to account for the observation that, if the total population size increases, this does not

necessarily lead to more *sexual* contacts for a given individual. Note that for the invasion problem one will always have an expression involving the (known) function S only. Of course, things are different if one wants to characterise endemic states, like in remark 3 above.

(5) To obtain a complete model one has to specify the demographic processes, and in particular, how per capita birth- and death rates are affected by the disease. If one makes the obvious assumption that the disease leads to a lower (or equal) birth rate and to a higher (or equal) death rate, one can use the linearised problem to obtain upper estimates for the nonlinear problem. If we have monotonicity in S in our operator $K(S)$ then the fact that $R_0 < 1$ for the linear problem (this implies exponential decrease of the infected population in the linear approximation, see Theorem 1.4 below) leads to a negative exponential upper-estimate for the non-linear problem. Thus one can prove, in general, *global* rather than local stability of the disease-free equilibrium for $R_0 < 1$. Or, in other words, broadly speaking, endemic states are impossible when $R_0 < 1$. In the case where there is no monotonic dependence on S in the linear operator $K(S)$ (and therefore also not in the dominant eigenvalue) global stability of the trivial equilibrium can break down. For example, if we consider sexually transmitted diseases with the function $h(S)(\cdot)$ mentioned in remark 4 then monotonicity is lost. In this case it is possible that a subcritical bifurcation occurs at the endemic state, and that therefore a non-trivial equilibrium is possible for $R_0 < 1$ (a mathematical treatment of this phenomenon is given in Huang, Cooke and Castillo-Chavez (1992) and Castillo-Chavez, Cooke, Huang and Levin (1989)).

(6) If the environment is not constant, then we can no longer define R_0 as the dominant eigenvalue of an operator based on a generation process. It can happen that the various rate 'constants' are subject to stochastic changes, but that we retain the linearisation (S constant). In that case, instead of a single generation operator, we have a stochastic sequence of operators. The invasion problem is decided by the dominant Lyapunov exponent of the invading infected individuals (see e.g. Tuljapurkar, (1990); Metz and De Roos, (1991); Metz, Nisbet and Geritz (1992)). Invasion is only successful if this exponent is larger than zero. The dominant Lyapunov exponent corresponds, in the constant-environment case, to the logarithmic growth-rate of the infected population (as opposed to the geometric growth-rate described on a generation basis by the dominant eigenvalue R_0).

(7) Many useful results on estimating the spectral radius of a positive linear operator, including integral operators on L_p spaces, can be found in Krasnosel'skij, Lifshits and Sobolev (1989). ◇

At the end of this section we consider the threshold theorem. In the proof of this theorem we basically follow Inaba (1990) and Heijmans (1986), where similar results are treated for different operators. We return to the linearised real time equation (1.2.1). It is easy to see that this equation has a solution of

the form $i(t,\xi) = e^{\lambda t}\psi(\xi)$ if and only if ψ is an eigenvector of the operator K_λ defined by

$$(K_\lambda \phi)(\xi) := S(\xi) \int_\Omega \int_0^\infty A(\tau, \xi, \eta) e^{-\lambda \tau} d\tau \phi(\eta) d\eta \tag{1.2.3}$$

with eigenvalue one, where $\phi \in C_+$, and the cone C_+ is defined as the set of nonnegative functions in $L_1(\Omega)$. Let C_+^* denote the dual cone, this is the set of all positive linear functionals on $L_1(\Omega)$ (i.e., the positive bounded measurable functions on Ω). Let $\langle F, \psi \rangle$ denote the duality pairing between an arbitrary element $F \in C_+^*$ and $\psi \in C_+$. Then C_+^* consists of all $F \in L_\infty(\Omega)$ such that $\langle F, \psi \rangle \geq 0$ for all $\psi \in C_+$. An element $\psi \in C_+$ is called a *quasi-interior point* if $\langle F, \psi \rangle > 0$ for all $F \in C_+^* \backslash \{0\}$ (for those in the know, we mention that this is equivalent to the usual definition which demands that the ideal generated by ψ is norm dense, Aliprantis and Burkinshaw (1985); this usual definition states, roughly speaking, that the infection, starting from a distribution that is a quasi-interior point, can 'reach almost all of Ω'). A functional $F \in C_+^*$ is strictly positive if $\langle F, \psi \rangle > 0$ for all $\psi \in C_+ \backslash \{0\}$.

Let the family of positive operators $\{K_\lambda\}$, $\lambda \in \mathbb{C}$, on $L_1(\Omega)$ be defined by (1.2.3). Let $r(K_\lambda)$ denote the spectral radius of K_λ.

We assume that for all $\lambda \in \mathbb{R}$ we have sufficient conditions on the corresponding operators K_λ for the following to hold:

(i). $r(K_\lambda)$ is a simple eigenvalue of K_λ, and $r(K_\lambda) > 0$, with eigenvector $\psi_\lambda \in C_+ \backslash \{0\}$ which we fix by normalising as $\|\psi_\lambda\| = 1$.

(ii). ψ_λ is a quasi-interior point of C_+.

(iii). The dual eigenvector $F_\lambda \in C_+^*$ corresponding to $r(K_\lambda)$ is a strictly positive functional.

Remark. These assumptions are satisfied for example when $r(K_\lambda)$ is a pole of the resolvent $(K_\lambda - \mu)^{-1}$ and the operator K_λ is so-called *semi-nonsupporting* (Sawashima (1964)). The latter is like the irreducibility condition for matrices, it demands that for every pair $\psi \in C_+ \backslash \{0\}$, $F \in C_+^* \backslash \{0\}$, there exists a positive integer $p = p(\psi, F)$ such that $\langle F, K^p \psi \rangle > 0$. This means, roughly speaking, that 'the h-states are well-mixed' in the infection process.

$\diamondsuit$

Furthermore we will use two additional assumptions in the proof of Lemma 1.1.

A_1. The function $A(\cdot, \xi, \eta)$ has compact support, uniformly in ξ and η.

A_2. Define the number p_λ by

$$p_\lambda := \inf_{\tau \geq 0} \int_\Omega \int_\Omega S(\xi) A(\tau, \xi, \eta) \psi_\lambda(\eta) d\eta d\xi$$

then there should exist an $\varepsilon > 0$ such that $p_\lambda > \varepsilon$ for all $\lambda < 0$.

Remark. These assumptions are, together with (i-iii), sufficient conditions for Lemma 1.1 to hold, but they are by no means necessary. Biologically speaking,

assumption A_1 poses no restriction on generality because no individual has an infinitely long infectious period, if only for the fact that it has a finite lifetime. However, from a modelling point of view it is unsatisfying, because it is, for example, not fulfilled in the frequently used Markov-transition models. The assumption guarantees that the integrals in (1.2.3) converge for all $\lambda \in \mathbb{R}$ and that therefore the limit $\lim_{\lambda \to -\infty} r(K_\lambda)$ is defined. Assumption A_2 is a technical condition. Of course one has that $p_\lambda > 0$ for every $\lambda < 0$ because ψ_λ is a nonnegative, nonzero, eigenvector of K_λ.

$\diamondsuit$

Define

$$\Sigma := \{\lambda \in \mathbb{C} | 1 \in P_\sigma(K_\lambda)\},$$

where $P_\sigma(K_\lambda)$ denotes the point spectrum of K_λ. Observe that if $\lambda \in \mathbb{R}$ and $r(K_\lambda) = 1$, this implies $1 \in P_\sigma(K_\lambda)$ and so: $\lambda \in \Sigma$. We first show that $\Sigma \cap \mathbb{R} \neq \emptyset$.

Lemma 1.1 $\exists\, !\ \lambda \in \mathbb{R}$ *with* $r(K_\lambda) = 1$.

Proof: Suppose $\lambda, \mu \in \mathbb{R}$, and $\lambda > \mu$. Let $\psi \in C_+ \backslash \{0\}$ be arbitrary, then $(K_\mu \psi)(\xi) > (K_\lambda \psi)(\xi)$ for $\xi \in \mathit{support}(K_\lambda \psi)$ (note that $K_\lambda \psi(\xi) > 0$ on a subset of Ω of positive measure, because if this was not the case then taking the duality pairing of $K_\lambda \psi$ with F_λ leads to a contradiction with the strict positivity of F_λ). Take $\psi = \psi_\lambda$ then we find that on $\mathit{support}(K_\lambda \psi)$ we have $K_\mu \psi_\lambda > r(K_\lambda)\psi_\lambda$. For this relation we take the duality pairing with the strictly positive dual eigenvector associated with $r(K_\mu)$: $\langle F_\mu, K_\mu \psi_\lambda \rangle > \langle F_\mu, r(K_\lambda)\psi_\lambda \rangle \Rightarrow r(K_\mu)\langle F_\mu, \psi_\lambda \rangle > r(K_\lambda)\langle F_\mu, \psi_\lambda \rangle$, so $r(K_\mu) > r(K_\lambda)$, where we have used $\langle F_\mu, \psi_\lambda \rangle > 0$. We find that the map $\lambda \longmapsto r(K_\lambda)$ is strictly decreasing on $\mathbb{R}$.

The continuity of $\lambda \longmapsto r(K_\lambda)$ follows from the uniform continuity of $\lambda \longmapsto K_\lambda$, where we use assumption A_1 and the norm-boundedness of $A(\cdot, \cdot, \eta)$, and the following argument. Let $\{\lambda_n\}$, $\lambda_n \to \mu$ be a sequence of real numbers converging to some number μ, and let the spectral radii r_n of the corresponding operators K_{λ_n} converge to some real number r_∞, i.e. $r_n \to r_\infty$. Suppose that $r_\infty \neq r_\mu$, then either there is a jump to the right in the sequence $\{r_n\}$ and $r_\mu > r_\infty$, or there is a jump to the left and $r_\mu < r_\infty$. We show that both cases lead to a contradiction. In the first case we can choose a λ near μ such that $r_\mu \in \rho(K_\lambda)$, the resolvent set of K_λ, and furthermore $\|K_\lambda - K_\mu\|$ is small. A theorem of Kato (Kato, 1976, page 208) on the continuous dependence of the resolvent set $\rho(K)$ on the operator K, then implies that $r_\mu \in \rho(K_\mu)$, which contradicts the fact that r_μ is an eigenvalue of K_μ. In the second case the argument is slightly different. If $r_\infty > r_\mu$ then $r_\infty \in \rho(K_\mu)$. Then also $r_\infty \in \rho(K_\lambda)$ for some λ near μ, because of the same result by Kato mentioned above. The resolvent set is open so there is a neighbourhood U of r_∞ with $U \subset \rho(K_\lambda)$. Now we can choose λ sufficiently close to μ to get $r_\lambda \in U$ and this contradicts the fact that r_λ is in the spectrum of K_λ.

It is clear that $\lim_{\lambda \to \infty} r(K_\lambda) = 0$. We now show that also $\lim_{\lambda \to -\infty} r(K_\lambda) = \infty$. By using the eigenvector ψ_λ of $r(K_\lambda)$ we find $r(K_\lambda) = \|K_\lambda \psi_\lambda\|$, which after

applying Fubini's Theorem twice can be written as

$$r(K_\lambda) = \int_0^\infty e^{-\lambda\tau} \int_\Omega \int_\Omega S(\xi)A(\tau,\xi,\eta)\psi_\lambda(\eta)d\eta d\xi d\tau.$$

If we use assumption A_2 we obtain that for every $N > 0$ there exists a $\lambda < 0$ such that $r(K_\lambda) > \varepsilon N$, and therefore $\lim_{\lambda\to-\infty} r(K_\lambda) = \infty$.

We conclude that there is a unique real root λ of $r(K_\lambda) = 1$.

$\diamondsuit$

We denote the unique real root from Lemma 1.1 by λ_d and the eigenvector corresponding to $r(K_{\lambda_d}) = 1$ by ψ_d: $(K_{\lambda_d}\psi_d)(\xi) = \psi_d(\xi)$ for $\xi \in \Omega$ a.e..

Remark. The question what assumptions, like A_1 and A_2 above, guarantee that there exist $\lambda \in (-\infty, 0]$ for which $r(K_\lambda) > 1$, is, from a biological point of view, almost academic. Ultimately what one wants to show is that $R_0 > 1$ leads to the existence of a $\lambda_d > 0$, and that $R_0 < 1$ leads to exponential decrease of the infected population. But this follows even without special assumptions from the first part of the proof of Lemma 1.1. If $R_0 > 1$ then $r(K_0) > 1$ and the existence of a unique $\lambda_d > 0$ follows from the fact that $\lambda \longmapsto r(K_\lambda)$ is strictly decreasing and continuous. On the other hand, if $R_0 < 1$ then there will not exist a $\lambda_d > 0$, in the meaning of Lemma 1.1, and therefore one has exponential decrease. The only thing one does not obtain from this argument is the precise rate of exponential decrease, but this is, biologically speaking, of minor importance.

$\diamondsuit$

In the proof of Lemma 1.2 below, we use a result from (Rudin (1974), Thm. 1.39) which we now state explicitly for our case:

Lemma 1.2 (Rudin) *Let $g \in L_1(\Omega)$ and $|\int_\Omega g(x)dx| = \int_\Omega |g(x)|dx$. Then there exists a $\beta \in \mathbb{C}$ such that $\beta g(x) = |g(x)|$ a.e. on Ω.*

We will use the notation $\psi > \phi$ if $\psi(\xi) > \phi(\xi)$ a.e. on Ω.

Lemma 1.3 $\lambda \in \Sigma,\ \lambda \neq \lambda_d \quad \Rightarrow \quad \mathrm{Re}\lambda < \lambda_d.$

Proof: Let $\lambda \in \Sigma$, $\lambda \neq \lambda_d$. We first show that $\mathrm{Re}\lambda \leq \lambda_d$. Let ψ be the eigenvector associated with $1 \in P_\sigma(K_\lambda)$. Write $\lambda = \lambda_R + i\lambda_I$. Define an absolute value $|\cdot|$ for $\phi \in L_1(\Omega)$ by $|\phi|(\xi) := |\phi(\xi)|$. Then $|\psi| = |K_\lambda\psi|$ and $K_{\lambda_R}|\psi| \geq |\psi|$ a.e. on Ω. For this last inequality we take the duality pairing with the dual eigenvector F_{λ_R} associated with $r(K_{\lambda_R})$: $\langle F_{\lambda_R}, K_{\lambda_R}|\psi|\rangle \geq \langle F_{\lambda_R}, |\psi|\rangle \Rightarrow \langle {K^*}_{\lambda_R}F_{\lambda_R}, |\psi|\rangle \geq \langle F_{\lambda_R}, |\psi|\rangle \Rightarrow r(K_{\lambda_R})\langle F_{\lambda_R}, |\psi|\rangle \geq \langle F_{\lambda_R}, |\psi|\rangle \Rightarrow r(K_{\lambda_R}) \geq 1$. From the proof of Lemma 1.1 we find that $\lambda_R(= \mathrm{Re}\lambda) \leq \lambda_d$.

Now suppose that $\lambda_R = \lambda_d$, then it is easy to see that $K_{\lambda_d}|\psi| = |\psi|$ (for suppose $K_{\lambda_d}|\psi| > |\psi|$ then take the duality pairing with F_d, corresponding to $r(K_{\lambda_d}) = 1$, and find $\langle F_d, |\psi|\rangle > \langle F_d, |\psi|\rangle$, which contradicts the strict positivity of F_d), so we have $|\psi| = c\psi_d$ for some constant c which we can

assume to be 1 without loss of generality. Then $\psi(\xi) = \psi_d(\xi)e^{if(\xi)}$ for some function $f : \Omega \to \mathbb{R}$. If we substitute this into the relation $K_{\lambda_d}\psi_d = |K_\lambda \psi|$ we obtain

$$S(\xi)\int_\Omega\int_0^\infty A(\tau,\xi,\eta)e^{-\lambda_d\tau}d\tau\psi_d(\eta)d\eta =$$
$$|S(\xi)\int_\Omega\int_0^\infty A(\tau,\xi,\eta)e^{-(\lambda_d+i\lambda_I)\tau}d\tau\psi_d(\eta)e^{if(\eta)}d\eta|.$$

From this it follows, using Lemma 1.2, that $-\lambda_I\tau + f(\eta) = \alpha$ for some constant $\alpha \in \mathbb{C}$. If we use $K_\lambda\psi = \psi$ then

$$e^{i\alpha}K_{\lambda_d}\psi_d(\xi) = \psi_d(\xi)e^{if(\xi)} \Rightarrow f(\xi) = \alpha \text{ a.e. in } \Omega.$$

It follows that $\lambda_I = 0$ and so $\lambda = \lambda_R = \lambda_d$.

♢

From Lemmas 1.1 and 1.3 we conclude that there is a unique real λ_d among the $\lambda \in \Sigma$, that moreover has the largest real part. Finally we find

Theorem 1.4 $\lambda_d \geq 0 \quad \Leftrightarrow \quad r(K_0) \geq 1.$

Proof: This immediate from the proof of Lemma 1.1: the map $\lambda \mapsto r(K_\lambda)$ is decreasing on $\mathbb{R}$. So $0 < \lambda_d \Rightarrow r(K_0) > r(K_{\lambda_d}) = 1$ and $\lambda_d < 0 \Rightarrow 1 > r(K_0)$.

♢

Remarks

(8) λ_d is a rate, whereas R_0 is a number (or better: a ratio).

(9) Note that λ_d and ψ_d describe the growth and the h-state distribution in the exponential phase of an epidemic, when the influence of the precise manner in which the epidemic started has died off, and the influence of the nonlinearity is not yet perceptible. ♢

1.3. Computational aspects: easy special cases

1.3.1 Separable mixing

To compute the dominant eigenvalue of a positive operator is, in general, not an easy task. However, there is one special case in which the task is a triviality: when the operator has one-dimensional range. Biologically this corresponds to the situation where the distribution (over the h-state space Ω) of the 'offspring' (i.e. the ones who become infected) is *independent* of the h-state of the 'parent' (i.e. the one who transmits the infection). We will call

this *separable mixing* or separable infectivity and susceptibility, or (separably) weighted homogeneous mixing.

Assume that

$$\int_o^\infty A(\tau,\xi,\eta)d\tau = a(\xi)b(\eta) \tag{1.3.1}$$

then

$$(K(S)\phi)(\xi) = S(\xi)a(\xi)\int_\Omega b(\eta)\phi(\eta)d\eta = c(\phi)S(\xi)a(\xi),$$

where c is a constant depending only on ϕ. So, $K(S)$ has a one-dimensional range and there can only be one eigenvector (up to normalisation): $S(\cdot)a(\cdot)$. Since

$$K(S)Sa = \Big(\int_\Omega b(\eta)S(\eta)a(\eta)d\eta\Big)Sa$$

we conclude that

$$R_0 = \rho_d = \int_\Omega b(\eta)S(\eta)a(\eta)d\eta. \tag{1.3.2}$$

Remarks

(1) Note that assumption (1.3.1) is satisfied when

$$A(\tau,\xi,\eta) = a(\xi)B(\tau,\eta)$$

but that this is a slightly more restrictive requirement.

(2) A convenient normalisation is

$$\int_\Omega S(\xi)a(\xi)d\xi = 1. \tag{1.3.3}$$

Then Sa is the probability density function for the h-state at infection, while $b(\eta)$ is the total expected number of 'offspring' of an individual which was infected while having h-state η. This interpretation yields once more that

$$R_0 = \int_\Omega b(\eta)S(\eta)a(\eta)d\eta.$$

(3) The special case where the functions a and b are the same up to a multiplicative constant, is usually referred to in the literature as proportionate mixing (Barbour, 1978). ♢

1.3.2 Separable mixing with enhanced within group infection

A second case in which it is easy to derive an explicit threshold criterion, even if we cannot calculate R_0 explicitly, occurs when individuals preferentially mix with their own kind and otherwise practise separable mixing. If

we moreover assume that the h-state of an individual remains constant over epidemiological time then $K(S)$ is of the form

$$(K(S)\phi)(\xi) = S(\xi)\Big\{a(\xi)\int_\Omega b(\eta)\phi(\eta)d\eta + c(\xi)\phi(\xi)\Big\}, \tag{1.3.4}$$

where $c(\xi)S(\xi)$ is the number of first generation 'offspring' produced 'directly' in one's own group.

The eigenvalue problem $K(S)\phi = \rho\phi$ can be rewritten as

$$\frac{1}{\rho - c(\xi)S(\xi)}S(\xi)a(\xi)\int_\Omega b(\eta)\phi(\eta)d\eta = \phi(\xi). \tag{1.3.5}$$

If we multiply both sides of (1.3.5) by $b(\xi)$, and then intergrate over Ω, we obtain the characteristic equation

$$\int_\Omega \frac{b(\xi)S(\xi)a(\xi)}{\rho - c(\xi)S(\xi)}d\xi = 1.$$

The left hand side defines a decreasing function of ρ which tends to zero for $\rho \to \infty$. The unique real root R_0 is larger than one if and only if either:

$$c(\xi)S(\xi) > 1 \quad \text{for some } \xi \in \Omega, \tag{1.3.6i}$$

or, otherwise,

$$\int_\Omega \frac{b(\xi)S(\xi)a(\xi)}{1 - c(\xi)S(\xi)}d\xi > 1. \tag{1.3.6ii}$$

(Of course a more precise formulation of (i) is ess sup $c(\xi)S(\xi) > 1$.) If (i) holds, a single infected individual with h-state ξ will already start a full blown epidemic among its likes. If, on the other hand, $c(\xi)S(\xi) < 1$ for all $\xi \in \Omega$, any epidemic has to be kept going by the additional cross infections among different types. To understand (ii) we distinguish between cross infections and direct infections within the own group, and argue as follows. As before Sa is, with the normalisation (1.3.3), the probability density function for the h-state at *cross* infection. The expected total number of cases, including its own, produced by an individual of h-state ξ, through chains of infectives which stay wholly among its likes, is $(1 - c(\xi)S(\xi))^{-1}$. Each of these produces an expected number of cross infections equal to $b(\xi)$. So, by treating the 'clans' as a kind of individuals, we are in the situation of our original separable mixing problem and we find from (1.3.2)

$$\int_\Omega b(\xi)\frac{1}{1 - c(\xi)S(\xi)}S(\xi)a(\xi)d\xi$$

as the expected number of offspring at the clan level. An epidemic occurs if and only if this number exceeds one.

J.A.J. Metz had already derived the criterion (1.3.6-i,ii) in the context of the geographical spread of plant diseases (think of foci within fields)(Hans Metz, personal communication). Our attention was drawn to this special case by Andreasen and Christiansen (1989) (in which they derive the same result in the context of a finite h-state space) and Blythe and Castillo-Chavez (1989). Combinations like (1.3.4) can also be found in Nold (1980), Hethcote and Yorke (1984), Hyman and Stanley (1988) (where it is called biased mixing), and Jacquez et al. (1988) (where it is called preferred mixing).

1.3.3 Multigroup separable mixing

An obvious mathematical generalisation of separable mixing is to assume that $K(S)$ has a finite dimensional range. The general form of this statement need not have a biological interpretation. Therefore we restrict our elaboration to a special example in this category which does allow a biological interpretation.

Let ξ be of the form (i, ξ_i), where i can take the values $1, 2, ..., n$ and ξ_i takes values in Ω_i. So, $\Omega = \cup_{i=1}^{n} \{i\} \times \Omega_i$. Assume that

$$\int_0^\infty A(\tau, (i, \xi_i), (j, \xi_j)) d\tau = a_i(\xi_i) b_{ij}(\xi_j)$$

which one could call a *local* form of weighted homogeneous mixing, since, with the normalisation

$$\int_\Omega a_i(\xi_i) S((i, \xi_i)) d\xi_i = 1$$

the *conditional* (on the first component being i) probability density function for the h-state at infection is independent of the h-state of the one who infects and is given by $a_i(\cdot) S((i, \cdot))$. The mixing condition allows us to average over one of the components of the h-state. The next-generation operator under this mixing condition is

$$(K(S)\phi)(i, \xi_i) = S(i, \xi_i) a_i(\xi_i) \sum_{j=1}^{n} \int_{\Omega_j} b_{ij}(\xi_j) \phi(j, \xi_j) d\xi_j, \tag{1.3.7}$$

and we conclude that, in order to be an eigenvector, necessarily

$$\phi(i, \xi_i) = \sigma_i S(i, \xi_i) a_i(\xi_i). \tag{1.3.8}$$

If we substitute (1.3.8) into (1.3.7) we deduce that in addition the vector σ should be an eigenvector of the matrix M with entries

$$m_{ij} = \int_{\Omega_j} b_{ij}(\xi_j) S(j, \xi_j) a_j(\xi_j) d\xi_j. \tag{1.3.9}$$

In particular R_0 is the dominant eigenvalue of the matrix M, and we have decreased the dimension of the eigenvalue problem from '∞' to n. Note that in principle we could allow the first component of the h-state to be continous, but that this of course leads to an infinite dimensional operator M instead of a matrix (although possibly the new infinite dimensional eigenvalue problem is 'easier' than the original one).

1.4. Some examples

1.4.1 Discrete and static h-state

In this case the operator $K(S)$ is represented by a matrix (of course, this also holds in the case that the h-state is discrete and *dynamic*). We shall first show how this matrix can be derived in the special case of the conventional $SEIR$ compartmental model.

Let M be the diagonal matrix of the per capita standard death rates of the various types. After infection individuals enter the exposed class E. From there they make the transition to the infective class I at a rate described by a diagonal matrix Σ whereupon they are removed at a rate described by a diagonal matrix D. Finally, let $T(S)$ be the *transition matrix*, i.e. the matrix such that

$$\frac{dE}{dt} = T(S)I - ME - \Sigma E$$

(beware: I denotes the vector of infectives, not the identity matrix). We claim that

$$K(S) = T(S)\Sigma(\Sigma + M)^{-1}(D + M)^{-1}. \tag{1.4.1}$$

The (easy) argument is as follows. The fraction of the exposed individuals which enter I (before dying) is the diagonal of $\Sigma(\Sigma + M)^{-1}$. The average time an individual remains in I is given by the diagonal of $(D + M)^{-1}$. Within class I the transmission is described by $T(S)$.

If $M = 0$ the expression (1.4.1) simplifies to

$$K(S) = T(S)D^{-1}.$$

(In Sect. 9 of Jacquez et al. (1988), a special case of this matrix is introduced with $T(S)$ written out in some more detail.) Note that, as is to be expected, Σ is irrelevant for the computation of R_0 in case $M = 0$, even though it may of course have substantial influence on the magnitude of λ_d.

Now regard an SIR model, with $M = 0$. We present an alternative proof of the version of Theorem 1.4 for the discrete case. The result connects the 'generation growth' described by the matrix $K(S) = T(S)D^{-1}$ with the continuous time growth described by the matrix $H := T(S) - D$.

Let us first illustrate this in the 1-group case. We write the linearised (i.e. S constant) equation for I as $\dot{I} = (\beta S - \alpha)I$. Integrating this we find $I(t) = I_0 e^{(\beta S - \alpha)t}$ from which we see that $\beta S - \alpha < 0$ will let the disease die out. This is the same as saying $\frac{\beta S}{\alpha} (= R_0) < 1$. This generalises very naturally.

Definition: The quantity $s(H) := \sup\{Re\lambda | \lambda \in \sigma(H)\}$ is called the spectral bound of a matrix H.

For the proof of Theorem 1.6 we use the following lemma.

Lemma 1.5 *Let H be a real $n \times n$ matrix with $h_{ij} \geq 0$, $i \neq j$. Then $s(H) < 0 \Leftrightarrow \det H \neq 0$ and $H^{-1} \leq 0$.*

Proof:

$\Rightarrow$

First of all $s(H) < 0 \Rightarrow \det H \neq 0$ because $\det H = 0$ implies $0 \in \sigma(H)$ which contradicts $s(H) < 0$. Then for H we have by assumption $h_{ij} \geq 0, i \neq j$, and the following holds (see e.g. Berman & Plemmons (1979) Theorem 3.12): $e^{Ht} \geq 0$ for all $t \geq 0$. So, as $\det H \neq 0$ and all eigenvalues have negative real parts we have $\int_0^\infty e^{Ht} dt \geq 0 \Rightarrow -H^{-1} \geq 0 \Rightarrow H^{-1} \leq 0$.

$\Leftarrow$

By assumption there exists $\theta > 0$ such that $H + \theta I \geq 0$ where I is the identity matrix. We can now deduce from the Perron-Frobenius Theorem (see e.g. Minc (1988)) that $s(H)$ is an eigenvalue with nonnegative eigenvector. So $Hx = s(H)x$ for some vector $x \geq 0$. Applying H^{-1} to both sides we find $x = s(H)H^{-1}x$. Since $H^{-1}x \leq 0$, $x \geq 0$ and $x \neq 0$ (because x is an eigenvector) this requires $s(H) < 0$.

$\diamondsuit$

Theorem 1.6 *Let $H = M - N$ be a matrix with nonnegative off-diagonal elements, $K = MN^{-1}$ with $M \geq 0$, and N a diagonal matrix with positive diagonal elements. Then: $s(H) < 0 \Leftrightarrow r(K) < 1$.*

Proof: By Lemma 1.5, $s(H) < 0 \Leftrightarrow \det H \neq 0$ and $H^{-1} \leq 0$. We have the following chain of inferences

$H^{-1} \leq 0 \Leftrightarrow (M - N)^{-1} = ((MN^{-1} - I)N)^{-1} = N^{-1}(MN^{-1} - I)^{-1} \leq 0$

$\Leftrightarrow N^{-1}(K - I)^{-1} \leq 0$

$\Leftrightarrow (K - I)^{-1} \leq 0.$

As $\det(K-I) \neq 0 \Leftrightarrow \det H \neq 0$ (because $0 \in \sigma(H)$ if and only if $1 \in \sigma(K)$), we can apply Lemma 1.5 again, this time to $K - I$. We find: $s(H) < 0 \Leftrightarrow s(K - I) < 0 \Leftrightarrow s(K) < 1 \Leftrightarrow r(K) < 1$, where we use that for $K \geq 0$, $r(K)$ is an eigenvalue of K by the Perron-Frobenius Theorem.

$\diamondsuit$

In the separable mixing case, the entries of $T(S)$ are of the form

$$a_i S_i b_j,$$

and according to (1.3.2) R_0 equals the trace of the matrix $K(S)$. See Hethcote and Yorke (1984) for another derivation of this fact.

1.4.2 Sexually transmitted diseases

Heterosexual transmission only. Let the index 1 refer to females and the index 2 to males. For each sex we distinguish individuals according to some variable ξ_i which is static (the interpretation of ξ_1 may or may not be the same as the

interpretation of ξ_2). If we adopt the multigroup separable mixing assumption and neglect homosexual transmission, then we arrive at the matrix

$$M = \begin{pmatrix} 0 & m_{12} \\ m_{21} & 0 \end{pmatrix},$$

where

$$m_{12} = \int_{\Omega_2} b_{12}(\xi_2) S(2, \xi_2) a_2(\xi_2) d\xi_2,$$
$$m_{21} = \int_{\Omega_1} b_{21}(\xi_1) S(1, \xi_1) a_1(\xi_1) d\xi_1.$$

We conclude that

$$R_0 = \sqrt{m_{12} m_{21}}.$$

(See Hethcote and Yorke (1984) for a 'discrete' version of this result.)

If we distinguish not only males and females, but on top of that hetero-, bi-, and homosexuality, we easily arrive at a 6×6-matrix whose dominant eigenvalue one has to compute to obtain R_0.

Sexual activity. Frequently, the variables ξ_i are used to describe sexual activity (in the sense of: propensity to make sexual contacts), and a_i and b_{ji} are taken to be proportional to ξ_i. In the context of the heterosexual transmission model above we would, more precisely, take

$$a_2(\xi_2) = \frac{\xi_2}{\int_{\Omega_1} \xi_1 S(1, \xi_1) d\xi_1}$$
$$b_{12}(\xi_2) = \beta_{12} \xi_2$$

with formulas for a_1 and b_{21} obtained by exchanging 1's and 2's throughout (one may argue that $\int_{\Omega_1} \xi_1 S(1, \xi_1) d\xi_1 = \int_{\Omega_2} \xi_2 S(2, \xi_2) d\xi_2$ is required if ξ is interpreted as the actual number of sexual contacts per unit of time). Thus one arrives at

$$m_{12} = \frac{\beta_{12} \int_{\Omega_2} \xi_2^2 S(2, \xi_2) d\xi_2}{\int_{\Omega_1} \xi_1 S(1, \xi_1) d\xi_1}$$

and mutatis mutandis at an expression for m_{21}, and therefore

$$R_0 = \left(\beta_{12} \beta_{21} \frac{\int_{\Omega_1} \xi_1^2 S(1, \xi_1) d\xi_1}{\int_{\Omega_1} \xi_1 S(1, \xi_1) d\xi_1} \frac{\int_{\Omega_2} \xi_2^2 S(2, \xi_2) d\xi_2}{\int_{\Omega_2} \xi_2 S(2, \xi_2) d\xi_2} \right)^{1/2}.$$

Note that

$$\frac{\int_{\Omega_i} \xi_i^2 S(i, \xi_i) d\xi_i}{\int_{\Omega_i} \xi_i S(i, \xi_i) d\xi_i} = \text{mean} + \frac{\text{variance}}{\text{mean}}$$

and that this result is analogous to a result of Dietz (1980) and identical to formula (5.7) in May and Anderson (1988).

1.4.3 Dynamic h-states

Now we look explicitly at the case where the h-state is dynamic. In $A(\tau, \xi, \eta)$, η refers to the h-state of an infected individual, say individual 'x', at the moment it contracted its own infection. The infective 'force' is then evaluated a time τ later. In the time interval $[0, \tau)$ that has passed since infection, the h-state of x has changed to, say, state θ in Ω. The current infectivity of x may depend on θ. Assume that the infectivity towards an individual in state ξ indeed depends on θ, but not on the history of h-state transitions by which it reached θ nor on the time elapsed since infection τ (i.e. assume transitions in h-state are like in a Markov process where the future depends only on the present state and not on the past). Let $a(\xi, \theta)$ denote this infectivity and let $P(\tau, \theta, \eta)$ be the conditional probability that x is still alive at time τ and that the h-state of x is now θ, given that it was η at time 0. Then the expected infectivity towards ξ's of x at time τ since infection is

$$A(\tau, \xi, \eta) = \int_\Omega a(\xi, \theta) P(\tau, \theta, \eta) d\theta$$

and for $K(S)$ we find

$$(K(S)\phi)(\xi) = S(\xi) \int_\Omega \int_0^\infty \Big(\int_\Omega a(\xi, \theta) P(\tau, \theta, \eta) d\theta \Big) \phi(\eta) d\tau d\eta.$$

If the h-state is static then $P(\tau, \theta, \eta) = \delta(\theta - \eta)$ (where δ denotes the Dirac-delta 'function') and we see that 'a' is equal to 'A'.

If we assume $a(\xi, \theta) = a_1(\xi) b(\theta)$ then $K(S)$ has one-dimensional range and R_0 can be calculated explicitly. For $P(\tau, \theta, \eta)$ we have to calculate the probability that our individual x is still alive at time τ after he was infected ánd that his h-state is now θ. Let μ be a death rate that is independent of the h-state x is in. Then

$$P(\tau, \theta, \eta) = e^{-\mu\tau} P_1(\tau, \theta, \eta).$$

Remark. Note that if we take *age* as our h-state, $\Omega = [0, \infty)$, we can easily write down an expression for $\mathcal{P}$. Let $\mathcal{F}(\eta)$ be the probability that an individual survives up to age η, then

$$\mathcal{P}(\tau, \theta, \eta) = \frac{\mathcal{F}(\eta + \tau)}{\mathcal{F}(\eta)} \, \delta(\theta - (\eta + \tau)).$$

So

$$(K(S)\phi)(\xi) = S(\xi) \int_0^\infty \int_0^\infty a(\xi, \eta + \tau) \frac{\mathcal{F}(\eta + \tau)}{\mathcal{F}(\eta)} \phi(\eta) d\tau d\eta. \qquad \diamondsuit$$

Let us restrict ourselves to the finite discrete case $\Omega = \{1, ..., n\}$. Then $K(S)$, $\mathcal{A} := (a_{ij})_{1 \le i,j \le n}$, and $\mathcal{P}_1(\tau) := (P_1(\tau, i, j))_{1 \le i,j \le n}$ are $n \times n$ matrices. Let G describe the transition probabilities per unit of time between all pairs of h-states (i, j) in Ω. Then

$$\mathcal{P}_1(\tau) = e^{G\tau}$$

and so

$$\mathcal{P}(\tau) = e^{(G-\mu)\tau}.$$

For the matrix $K(S)$ we finally find

$$K(S) = diag(S_1, ..., S_n) \int_0^\infty \mathcal{A} e^{(G-\mu)\tau} d\tau = diag(S_1, ..., S_n)\mathcal{A}(\mu - G)^{-1}.$$

In applications one has to find an expression for the matrices $\mathcal{A}$ and G and calculate the steady demographic state distribution of the susceptible population $S_1, ..., S_n$.

Remark. In the above expression for $K(S)$ we have assumed a constant production of 'infectious material' h (incorporated in '$\mathcal{A}$'), by the infectives. We can allow this quantity h to be a function of the d-age τ of an infective. If we assume that this function is the same for all possible h-states, then $h(\tau)$ can be factored out of '$\mathcal{A}$' and we can write

$$K(S) = diag(S_1, ..., S_n) \int_0^\infty h(\tau)\mathcal{A} e^{(G-\mu)\tau} d\tau.$$

$\diamondsuit$

In chapter 2 we will discuss at length an example which makes use of the ideas in this section.

1.4.4 Heterogeneity in spatial position

Finally we look at the case where $\Omega = \mathbb{R}^2$. This would for example correspond to a model for the spatial spread of an infectious disease. In this case $A(\tau, x, \eta)$ describes the infectivity at static position x due to one infective with d-state τ at static position η. This situation occurs, for example, when one studies the spread of an infectious plant disease in a field (see Van den Bosch, Metz and Diekmann (1990), and the references given there, for more details).

Assume that $S(x) = S_0$ a constant, and that $\tilde{A}(x, \eta) = \int_0^\infty A(\tau, x, \eta) d\tau$ is rotation symmetric i.e. $\tilde{A}(x, \eta) := V(|x - \eta|)$, for some appropriate V. We write $\int_{\mathbb{R}^2} V(|x - \eta|)dx = \int_{\mathbb{R}^2} V(|y|)dy =: \gamma$.

From (1.2.2) we get for the next-generation operator

$$(K(S)\phi)(x) = S_0 \int_{\mathbb{R}^2} V(|x - \eta|)\phi(\eta)d\eta$$

on some appropriate function space (bounded continuous functions on $\mathbb{R}^2$, see Diekmann (1978) for details). We observe that the constant function $\phi(x) \equiv 1$ is a positive eigenvector of $K(S)$ with eigenvalue γS_0 and conclude: $R_0 = \gamma S_0$. Note that, if we start with a small localised initial 'patch' of infection, it can take infinitely many iterations before the stable distribution is reached. For the full problem with non-linear S, the non-linearity will then already have become noticeable. Therefore, the definition of R_0 as the dominant eigenvalue of the next-generation operator is problematic in the spatial case. However, in the case of spatial heterogeneity one is more interested in the asymptotic speed of propagation.

1.5. Appendix: relation to differential equation models

We shall indicate how the integral equation (1.2.1) is related to the differential equation model for the progress of a disease in a stratified population, which one often finds in the literature. To this end we first observe that A can be decomposed as

$$A(\tau, \xi, \eta) = \int_\Omega \beta(\tau, \xi, \rho) k(\tau, \rho, \eta) d\rho. \tag{1.5.1}$$

Here $k(\tau, \rho, \eta)$ is the probability that an individual has not yet lost all (future) infectivity and has now h-state ρ, while $\beta(\tau, \xi, \rho)$ is the infectivity of such an individual towards a susceptible with h-state ξ. We limit ourselves to two special cases. The first concerns h-states that are fixed over time, and the second concerns the dynamic h-state *age*. It is an as yet unsolved problem to give sufficient conditions on A to obtain similar 'reductions' to differential equation models when one considers an arbitrary dynamic h-state.

General static h-state variables

If the h-state variable is static then (1.5.1) can be simplified to

$$A(\tau, \xi, \eta) = \beta(\tau, \xi, \eta) k(\tau, \eta). \tag{1.5.2}$$

We assume that β is independent of τ: $\beta(\tau, \xi, \eta) = \beta(\xi, \eta)$, and that k depends on τ in an exponential manner $k(\tau, \eta) = e^{-\alpha(\eta)\tau}$ for some nonnegative function $\alpha(\eta)$.
Now define

$$I(t, \xi) := \int_{-\infty}^{t} e^{-\alpha(\xi)(t-\tau)} i(\tau, \xi) d\tau \tag{1.5.3}$$

which is the total number of infectives with h-state ξ at time t.

Differentiation of (1.5.3) with respect to time t gives

$$\frac{dI}{dt} = i(t,\xi) - \alpha(\xi)I(t,\xi). \tag{1.5.4}$$

On the other hand we still have expression (1.2.1) for $i(t,\xi)$ (in the nonlinearised version), into which we substitute (1.5.2) together with the assumptions on β and k above. We obtain

$$i(t,\xi) = S(t,\xi)\int_0^\infty \int_\Omega \beta(\xi,\eta)e^{-\alpha(\eta)\tau} i(t-\tau,\eta)d\eta d\tau.$$

Exchange of the order of integration using Fubini's Theorem and substitution of $\tau \leftrightarrow t-\tau$ leads to

$$i(t,\xi) = S(t,\xi)\int_\Omega \beta(\xi,\eta)\int_{-\infty}^t e^{-\alpha(\eta)(t-\tau)} i(\tau,\eta)d\tau d\eta.$$

So, together with the definition of $I(t,\eta)$ we have

$$i(t,\xi) = S(t,\xi)\int_\Omega \beta(\xi,\eta)I(t,\eta)d\eta. \tag{1.5.5}$$

Finally, substitution of (1.5.5) into (1.5.4) gives the differential equation for $I(t,\xi)$ we were looking for

$$\frac{dI}{dt}(t,\xi) = S(t,\xi)\int_\Omega \beta(\xi,\eta)I(t,\eta)d\eta - \alpha(\xi)I(t,\xi).$$

The operator $K(S)$ is given by

$$(K(S)\phi)(\xi) = S(\xi)\int_\Omega \frac{\beta(\xi,\eta)}{\alpha(\eta)}\phi(\eta)d\eta.$$

Age dependent models

We now let $\Omega = \mathbb{R}_{\geq 0}$ and we take the age of an individual as the h-state. We shall write a for ξ, as this is conventional notation.

In that case (1.5.2) becomes $A(\tau,a,a') = \beta(\tau,a,a'+\tau)k(\tau,a')$. Assume that β is independent of τ and that the survival $k(\tau,\eta)$ has the following form: $k(\tau,\eta) = exp(-\int_\eta^{\eta+\tau}\alpha(s)ds)$.

Define

$$I(t,a) := \int_0^a i(t-\tau,a-\tau)e^{-\int_{a-\tau}^a \alpha(s)ds}d\tau \tag{1.5.6}$$

then

$$i(t,a) = S(t,a)\int_0^\infty \int_0^\infty \beta(a,\eta+\tau)e^{-\int_\eta^{\eta+\tau}\alpha(s)ds} i(t-\tau,\eta)d\eta d\tau.$$

Substitution of $a' \leftrightarrow \eta + \tau$ gives

$$i(t,a) = S(t,a) \int_0^\infty \int_\tau^\infty \beta(a,a') e^{-\int_{a'-\tau}^{a'} \alpha(s)ds} i(t-\tau, a'-\tau) da' d\tau$$

$$\Rightarrow i(t,a) = S(t,a) \int_0^\infty \beta(a,a') \int_0^{a'} e^{-\int_{a'-\tau}^{a'} \alpha(s)ds} i(t-\tau, a'-\tau) d\tau da',$$

where we have used Fubini's Theorem.We can rewrite this as

$$i(t,a) = S(t,a) \int_0^\infty \beta(a,a') I(t,a') da'.$$

If we differentiate (1.5.6) we obtain the differential equation for the number of infectives

$$\frac{\partial I}{\partial t} + \frac{\partial I}{\partial a} = i(t,a) - \alpha(a) I(t,a).$$

The operator $K(S)$ is given by

$$(K(S)\phi)(a) = S(a) \int_0^\infty \int_0^\infty \beta(a, a'+\tau) k(\tau, a') \phi(a') d\tau da'.$$

1.6. References

Aliprantis, C.D. & O. Burkinshaw (1985): *Positive Operators.* Academic Press, Orlando.

Andreasen, V. & F.B. Christiansen (1989): Persistence of an infectious disease in a subdivided population. *Math. Biosc.* **96**: 239-253.

Bailey, N.T.J. (1982): *The Biomathematics of Malaria.* London, Griffin.

Barbour, A.D. (1978): MacDonald's model and the transmission of bilharzia. *Trans. Roy. Soc. Trop. Med. Hyg.* **72**: 6-15.

Berman, A. & R.J. Plemmons (1979): *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York.

Blythe, S.P. & C. Castillo-Chavez (1989): Like-with-like preference and sexual mixing models. *Math. Biosc.* **69**: 221-238.

Castillo-Chavez, C., Cooke, K., Huang, W. & S.A. Levin (1989): Results on the dynamics for models for the sexual transmission of the human immunodeficiency virus. *Appl. Math. Lett.* **2**: 327-331.

Diekmann, O. (1978): Thresholds and travelling waves for the geographical spread of infection. *J. Math. Biol.* **6**: 109-130.

Diekmann, O. (1991): Modelling infectious diseases in structured populations. In: B.D. Sleeman & R.J. Jarvis (eds.) *Ordinary and Partial Differential Equations, vol. III.* Pitman RNiMS 254: 67-79. Longman, Harlow.

Dietz, K. (1980): Models for vector-borne parasitic diseases. In: Barigozzi, C. (ed.) *Vito Volterra Symposium on Mathematical Models in Biology.* (Lect. Notes Biomath., vol. 39, pp. 264-277), Berlin: Springer.

Heijmans, H.J.A.M. (1986): The dynamical behaviour of the age-size distribution of a cell population. In: Metz, J.A.J., Diekmann, O. (eds.) *The Dynamics of Physiologically Structured Populations.* (Lect. Notes Biomath., vol. 68, pp. 185-202) Berlin: Springer.

Hethcote, H.W. & J.A. Yorke (1984): *Gonorrhea, Transmission Dynamics and Control.* (Lect. Notes Biomath., vol. 56) Berlin: Springer.

Hethcote, H.W. & J.W. van Ark (1987): Epidemilogical models for hetrogeneous populations: proportionate mixing, parameter esimation, and immunisation programs. *Math. Biosc.* **84**: 85-118.

Huang, W., Cooke, K.L. & C. Castillo-Chavez (1992): Stability and bifurcation for a multiple group model for the dynamics of HIV/AIDS transmission. *SIAM J. Appl. Math.* **52:** 835-854.

Hyman, J.M. & E.A. Stanley (1988): Using mathematical models to understand the AIDS epidemic. *Math. Biosc.* **90**: 415-473.

Inaba, H. (1990): Threshold and stability results for an age-structured epidemic model. *J. Math. Biol.* **28**: 411-434.

Jacquez, J.A., Simon, C.P., Koopman, J., Sattenspiel, L. & T. Perry (1988): Modeling and analyzing HIV transmission: the effect of contact patterns. *Math. Biosc.* **92**: 119-199.

Jagers, P. & O. Nerman (1984): The growth and composition of branching populations. *Adv. Appl. Probab.* **16**: 221-259.

Kato, T. (1976): *Perturbation Theory for Linear Operators*, Springer Verlag, Berlin.

Krasnosel'skij, M.A., Lifshits, Je. A. & A.V. Sobolev (1989): *Positive Linear Systems (the Method of Positive Operators).* Heldermann Verlag, Berlin.

R.M. May & R.M. Anderson (1988): The transmission dynamics of human immunodeficiency virus (HIV). *Phil. Trans. R. Soc. Lond.* **B 321**: 565-607.

Metz, J.A.J. & A.M. De Roos (1991): The role of physiologically structured population models within a general individual-based modelling perspective. In: *Proceedings Workshop Individual based modelling, Knoxville, Tenessee.*

Metz, J.A.J., Nisbet, R.M. & S.A.H. Geritz (1992): How should we define

'fitness' for general ecological scenarios? *TREE* **7**: 198-202.

Minc, H. (1988): *Nonnegative Matrices.* Wiley, New York.

Mode, C.J. (1971): *Multitype Branching Processes, Theory and Applications.* New York: Elsevier.

Nold, A. (1980): Heterogeneity in disease-transmission modeling. *Math. Biosc.* **52**: 227-240.

Rudin, W. (1973): *Functional Analysis*, McGraw-Hill, New York.

Rudin, W. (1974): *Real and Complex Analysis*, McGraw-Hill, New York.

Sawashima, I. (1964): On spectral properties of some positive operators. *Nat. Sci. Dep. Ochanomizu Univ.* **15**: 53-64.

Schaefer, H.H. (1960): Some spectral properties of positive linear operators. *Pacific J. Math.* **10**: 1009-1019.

Schaefer, H.H. (1974): *Banach Lattices and Positive Operators.* Berlin: Springer.

Tuljapurkar, S. (1990): *Population Dynamics in Variable Environments.* Lect. Notes Biomath., vol. 85, Springer-Verlag, Berlin.

Van den Bosch, F., Metz, J.A.J. & O. Diekmann (1990): The velocity of spatial population expansion. *J. Math. Biol.* **28**: 529-565.

Chapter 2

Examples of R_0 calculations

We discuss two extended examples of situations where it would prove to be difficult to calculate R_0 without the use of the theory as developed in chapter 1. The first example, section 2.1, investigates the possible influence of other sexually transmitted diseases in the population on the spread of the AIDS-virus HIV. The second example, section 2.3, shows the flexibility of the multigroup separable mixing assumption from chapter 1; we discuss the calculation of R_0 for an infectious disease spreading through a realistic type of pig breeding farm. Section 2.2 focusses on age as the h-state variable, and known expressions from the literature are derived in our way.

2.1. Two is worse than one

2.1.1 Introduction

It has been observed in, for example, Africa that certain genital ulcer diseases, particularly chancroid and syphilis, can increase the risk of HIV-infection (Piot et al., 1988). The damage these ulcerative sexually transmitted diseases (we write USTD's for short) cause to the genital skin and membranes may facilitate both HIV transmission and acquisition. AIDS has clearly been able to establish itself in the heterosexual population in Africa, as opposed to the situation in Europe and the USA. There transmission is highest in other subpopulations. Because of the higher prevalence of other STD's in Africa, one can pose a few obvious questions. To what degree can USTD's that are endemic in a population facilitate the spread of HIV into that population (for example the heterosexual population in Europe or the USA)? What is the efficacy of control measures aimed at these USTD's in halting the spread of HIV in that

population? At which aspects of the USTD should control measures then be aimed to be most effective?

For simple models the characterisation of R_0 is also simple. One can usually find an expression for R_0 by heuristic arguments. When more realism is introduced in the description of the transmission-rates by incorporating some of the heterogeneity among individuals, it is in general unclear how R_0 should be defined. Still one would like to use R_0 in analysing more complicated situations. In chapter 1 it was shown what in the general case the precise mathematical counterpart of the original biological interpretation of R_0 is and how it can be calculated from the model assumptions. In section 2.1.2 of the present chapter, we apply the formalism to calculate R_0 for a simple HIV-transmission model when a USTD is already endemic in the population. This was motivated by a paper of May and Anderson (1988) which gives for the same situation an implicit equation for the exponential growth rate of the HIV-infectives. This growth rate is the quantity λ_d from chapter 1, it is related to R_0 via Theorem 1.4. We end section 2.1.2 with a brief discussion of some qualitative conclusions. In section 2.1.3 we extend the calculations of section 2.1.2 in three directions: first to the case where HIV in turn influences the USTD; secondly to the case where we distinguish male/female heterogeneity; and finally to the case where we additionally take the sexual activity of the individuals into account.

2.1.2 HIV and cofactors

Let us first look at a very simple HIV-model. Write I for the population of infectives and regard,

$$\frac{dI}{dt}(t) = \sigma p I(t) - (\mu + \rho) I(t). \tag{2.1.1}$$

So we assume that contacts are totally random, with σ new contacts per unit of time for each individual. Let p be the probability that HIV-transmission is successful upon contact, μ be the natural death rate and let ρ be an additional disease induced death rate. We assume for R_0 calculations that the density of susceptibles does not change over time (so every new contact an infective makes is with a susceptible).

For equation (2.1.1) we can derive R_0 by looking at a single infected individual and directly interpret the 'biological' definition. This leads to

$$R_0^{hiv} = \frac{\sigma p}{\mu + \rho}, \tag{2.1.2}$$

because the number of successful new contacts an infective makes per unit of time is σp and the infectious period lasts on average $1/(\mu + \rho)$ time units. As there is no heterogeneity all infected individuals are the same and consequently any individual is 'typical'. This changes if we include heterogeneity among the individuals of the population. We have seen that it is still possible to develop a meaningful R_0-concept in that situation.

We divide the individuals in the population into classes (h-states) according to a set Ω of heterogeneity characteristics. Let $S(\xi)$ denote the susceptible individuals with h-state $\xi \in \Omega$ in the disease free steady demographic state. In the general case one needs the following ingredients to calculate R_0: (i) the structure Ω in the population relevant to the transmission of the disease that one wants to investigate (i.e. which kind of states/classes can an individual be in and how does the state of an individual change over time if it is not static); (ii) the way these classes determine the contact structure; and (iii) the way the state at the moment of infection and the time elapsed since infection determine the infectivity. The last two ingredients can be combined into the function $A(\tau, \xi, \eta)$ describing the infectivity, towards susceptibles of h-state ξ, of an infective which was infected τ time-units ago and had h-state η at the time it contracted the infection.

In the simple HIV-model mentioned above we recognised only one h-state and we have $A(\tau) = \sigma p\ e^{-(\mu+\rho)\tau}$. In the general case 'typical' signifies that we have a particular probability distribution for the heterogeneity class of the individual. This distribution is (asymptotically) generated by the process of epidemic progress itself.

If we assume separable mixing, i.e.

$$\int_0^\infty A(\tau, \xi, \eta) d\tau = a(\xi) b(\eta)$$

we have seen that we obtain (under mass-action kinetics)

$$R_0 = \int_\Omega a(\eta) S(\eta) b(\eta) d\eta.$$

For STD's the appropriate expression depends to some extent on the way population *size* is incorporated and, moreover, on the interpretation of the variable ξ. As a rule the expression is

$$R_0 = \frac{\int_\Omega a(\eta) S(\eta) b(\eta) d\eta}{\int_\Omega S(\eta) c(\eta) d\eta},$$

where $c(\eta)$ is some appropriate function (see remark 4 in section 1.2). Biologically, separable mixing means that the heterogeneity classes of the individuals that *become* infected are independent of the particular class of the one that *causes* these new infections. The infectivity, mentioned above, is then a product of a term that only depends on the class of the susceptible in a contact and a term that only depends on the class of the infective.

If there are a finite number of states ($\Omega = \{1, ..., n\}$) we read the integrals as sums. Note that for the simple HIV-model we have but one state, $ab = \sigma p/(\mu + \rho)$ and $c = 1$, leading once more to (2.1.2).

Assume there are two diseases in a population: d and D. Assume that disease d is in an endemic steady state. We want to calculate R_0 for the disease

D, assuming that the susceptibility to D is, for individuals having d, v times as large as for individuals without d. What we have in mind is that meetings between individuals are totally random, but that the success ratio for disease transmission, given that contact takes place, is enlarged by a factor v. By $w > 1$ we denote the factor by which the success ratio is enlarged when a D infectious individual is also suffering from d, and let p be the success ratio when both individuals involved in the contact are free from d. Consistency demands that $pvw \leq 1$. We assume that meetings occur independently of the h-state of the individuals involved and that the rate is given by σ/S. For our h-state space we take $\Omega = \{0, +\}$, where '0' means free of d, and '+' means having d. Then the matrix $\mathcal{A}$ is given by

$$\mathcal{A} = \frac{\sigma p}{S} \begin{pmatrix} 1 & w \\ v & vw \end{pmatrix}.$$

Describe by S_0 and S_+ the steady (with respect to d) state population sizes of '0' and '+' individuals. Let ζ denote the force of d-infection in the steady state and let γ be the probability per unit of time that d is cured (whereupon susceptibility to d returns). We do not concern ourselves with the question of how these parameters arise, we assume that they completely describe the dynamics of d phenomenologically. Let furthermore μ be the natural death rate. Then any individual undergoes, as long as it does not die, transitions between '0' and '+' according to the rate matrix

$$G = \begin{pmatrix} -\zeta & \gamma \\ \zeta & -\gamma \end{pmatrix}.$$

If the disease D is associated with an extra, τ-independent, mortality rate ρ we find

$$\mathcal{P} := \int_0^\infty \mathcal{P}(\tau) d\tau = \int_0^\infty e^{(G-\mu-\rho)\tau} d\tau = (\mu + \rho - G)^{-1}.$$

So

$$K(S) = diag(S_0, S_+)\, \mathcal{A}\, \mathcal{P} = \frac{\sigma p}{S} \begin{pmatrix} S_0 & wS_0 \\ vS_+ & vwS_+ \end{pmatrix} \mathcal{P}$$

with

$$\mathcal{P} = \frac{1}{(\mu + \rho)(\mu + \rho + \gamma + \zeta)} \begin{pmatrix} \mu + \rho + \gamma & \gamma \\ \zeta & \mu + \rho + \zeta \end{pmatrix}.$$

Note that $K(S)$ has one-dimensional range, because matrix $\mathcal{A}$ has a one-dimensional range spanned by $\binom{1}{v}$. Note that this can only happen because we have assumed a product vw in the success ratio for the case of a contact between two '+' individuals. Any other function of v and w does not correspond to separable mixing and so does not lead to a one-dimensional range.

The range of $K(S)$ is spanned by $\binom{S_0}{vS_+}$. The one eigenvector of $K(S)$ is given by

$$\phi^* = \frac{\sigma p}{S}\begin{pmatrix} S_0 \\ vS_+ \end{pmatrix}.$$

Then by rewriting matrix $\mathcal{A}$,

$$\begin{aligned} K(S)\phi^* &= \frac{\sigma p}{S}\begin{pmatrix} S_0 \\ vS_+ \end{pmatrix}\begin{pmatrix} 1 \\ w \end{pmatrix}^T \mathcal{P}\phi^* \\ &= \phi^*\begin{pmatrix} 1 \\ w \end{pmatrix}^T \mathcal{P}\phi^* \end{aligned}$$

(where x^T denotes the transpose of a vector x). So the only eigenvalue of $K(S)$ is

$$R_0 = \frac{\sigma p}{S}\begin{pmatrix} 1 \\ w \end{pmatrix}^T \mathcal{P}\begin{pmatrix} S_0 \\ vS_+ \end{pmatrix}.$$

To obtain an explicit formula for R_0 we still have to find expressions for the steady states (with respect to d), S_0 and S_+. Since

$$\frac{dS_+}{dt} = \zeta S_0 - \gamma S_+ - \mu S_+$$

we deduce that in steady state $\frac{S_+}{S_0} = \frac{\zeta}{\gamma+\mu}$ or

$$S_0 = \frac{\gamma+\mu}{\gamma+\mu+\zeta}S, \quad S_+ = \frac{\zeta}{\gamma+\mu+\zeta}S,$$

where $S = S_0 + S_+$. So finally we find

$$\begin{aligned} R_0 &= \sigma p \frac{(\gamma+\mu)(\mu+\rho+\gamma+\zeta w) + \zeta v w\left(\frac{\gamma}{w}+\mu+\rho+\zeta\right)}{(\gamma+\mu+\zeta)(\mu+\rho)(\mu+\rho+\gamma+\zeta)} \\ &=: R_0^{hiv} F. \end{aligned} \tag{2.1.3}$$

Note that the special case $w = v = 1$ yields $R_0 = R_0^{hiv}$ as to be expected (since in that case the '0', '+' distinction is totally irrelevant).

In the following we study only the multiplication factor F in (2.1.3). If all parameters are positive then $F \geq 1$. One can easily see, by taking appropriate limits, that if either γ is very large or ζ is very small, then $R_0 \approx R_0^{hiv}$. This is obvious from the biological interpretation. One can show that F is a strictly decreasing function of γ, and a strictly increasing function of ζ. If γ is small (i.e. slow recovery from d) then

$$R_0 \approx R_0^{hiv}\frac{vw(\zeta^2+\zeta(\mu+\rho)) + w\zeta\mu + \mu(\mu+\rho)}{(\mu+\zeta)(\mu+\rho+\zeta)},$$

and if ζ is large (many new victims per unit of time) then $R_0 \approx R_0^{hiv} vw$. Note that the product of v and w determines how large F can become.

We now look at the case where $v = w$. Under that assumption we can rewrite F as follows

$$\begin{aligned} F &= (1 - \frac{S_+}{S})\Big(\frac{\mu + \rho + \gamma + \zeta v}{\mu + \rho + \gamma + \zeta}\Big) + \frac{S_+}{S} v^2 \Big(\frac{\frac{\gamma}{v} + \mu + \rho + \zeta}{\mu + \rho + \gamma + \zeta}\Big) \\ &=: (1 - \frac{S_+}{S}) a_1 + \frac{S_+}{S} v^2 a_2, \end{aligned} \tag{2.1.4}$$

where $a_1 \geq 1$ and $a_2 \leq 1$. If we make the following approximations

$$\frac{\gamma}{\gamma + \mu} \approx 1, \qquad \frac{\rho}{\gamma + \mu} \approx 0,$$

then we find, after some rearranging,

$$F \approx 1 + \frac{S_+}{S}(v - 1)\big(\frac{S_+}{S}(v - 1) + 2\big).$$

We see that, for $v = w$, F approximately increases quadratically with both v and the prevalence of d in the population.

Let us take, for $v = w$, $\mu = 0.02$ per year and $\rho = 0.1$ per year. Figure 2.1 shows the graph of F as a function of γ for $\zeta = 0.5$ per year, and $\zeta = 4$ per year, with $v = 1, .., 4$. Figure 2.2 shows the graph of F as a function of ζ for $\gamma = 0.5$ per year and $\gamma = 3$ per year, with $v = 1, .., 4$.

One can see from figure 2.1 that if recovery from d is slow, say in the order of a year or more, and ζ is small, then F can be decreased significantly by only slightly raising γ, for example by minor improvement of medical care. However if ζ is larger than say 4 per year, then γ must be increased much more for the same effect, making it in the order of 3 months or less to cure d. From figure 2.2 one can see that if γ is small (slow recovery), then trying to decrease F by decreasing ζ, for example by lowering the success-probability of USTD-transmission by campaigning against unprotected sex, is unsuccessful over a wide range of ζ-values. The effect of such campaigns is noticeable much sooner if recovery from the USTD is in the order of a few months or less.

With the lack of reasonable estimates for many of the major parameters occuring in models continuing up to the present day, the value of the concept of R_0 does not always lie in associating an actual number with it, but often more in the qualitative behaviour of R_0 seen as a function of the model-parameters. For this purpose it is convenient to have an explicit expression for R_0.

The question whether the values of the parameters for Africa occuring in the HIV-USTD model are enough to bring R_0 above unity, while for Europe they are not, is very difficult to settle (the value of R_0^{hiv} is already uncertain). The question is most interesting of course for the heterosexual population. It is possible to give an expression similar to equation (2.1.3) for the case that

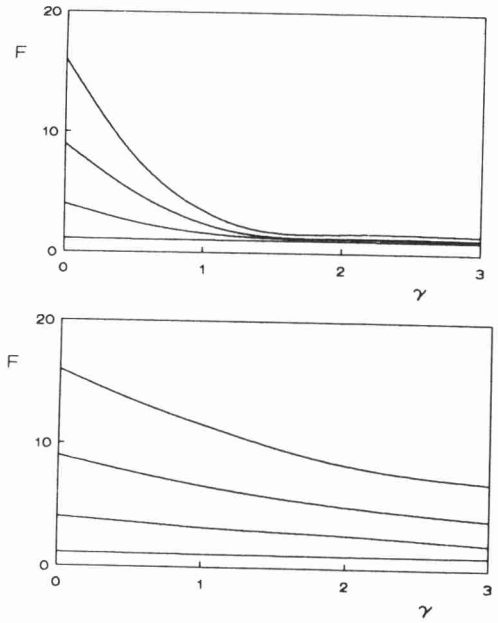

Figure 2.1 Graph of F as a function of γ for $\zeta = 0.5$ per year (top) and $\zeta = 4$ per year (bottom), both for $v = 1$ (lowest line) up to $v = 4$ (highest line)

males and females are distinguished, where every parameter is allowed to be different for the two classes (this collapses to equation (2.1.3) if the values are equal for both types, the derivation is given in section 2.1.3). However, as long as we have no idea of the parameter ranges this generality does not yield any information that the simpler model could not give.

The conclusion is that the presence of USTD's in Africa could theoretically increase R_0^{hiv} significantly in light of the fact that USTD's have a higher prevalence in African countries as opposed to Europe and the USA, and the recovery-rate for these USTD's will probably be lower in Africa due to a lower level of medical care. The value of the parameters v and w is crucial. Because the transmission mechanisms of HIV and the USTD's are intimately related it could prove very difficult to estimate these enhancement-factors.

2.1.3 Extensions

In this section we treat three extensions of the model from section 2.1.2.

In Aral and Holmes (1991) it is stated that, 'in persons who have been exposed to HIV, chancroid often fails to respond to some therapies that are otherwise highly successful. Thus, HIV infection may help the spread of a bacterial STD that in turn helps to spread HIV'. Let us incorporate into our

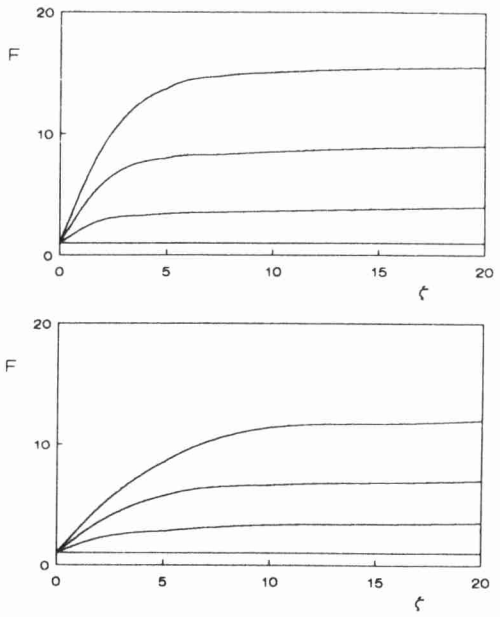

Figure 2.2 Graph of F as a function of ζ for $\gamma = 0.5$ per year (top) and $\gamma = 3$ per year (bottom), both for $v = 1$ (lowest line) up to $v = 4$ (highest line)

model in section 2.1.2 the influence of disease D on the cure rate of disease d by introducing a factor z ($0 \leq z \leq 1$) that describes to what extent the probability per unit of time to recover from d is decreased by the presence of D. With this change, the steady state population sizes S_0 and S_+ with respect to d are the same as in section 2.1.2, because these are calculated in the situation where the invading disease D is not yet present. The only change relative to section 2.1.2, is that the transition rate matrix G for the Markov process on $\{0, +\}$ now reads

$$G = \begin{pmatrix} -\zeta & z\gamma \\ \zeta & -z\gamma \end{pmatrix}.$$

The matrix $\mathcal{P}$ then becomes

$$\mathcal{P} = \frac{1}{(\mu+\rho)(\mu+\rho+z\gamma+\zeta)} \begin{pmatrix} \mu+\rho+z\gamma & z\gamma \\ \zeta & \mu+\rho+\zeta \end{pmatrix}$$

and R_0 is given by

$$\begin{aligned} R_0 &= \sigma p \frac{(\gamma+\mu)(\mu+\rho+z\gamma+\zeta w)+\zeta v w\left(\frac{z\gamma}{w}+\mu+\rho+\zeta\right)}{(\gamma+\mu+\zeta)(\mu+\rho)(\mu+\rho+z\gamma+\zeta)} \\ &=: R_0^{hiv} f. \end{aligned} \tag{2.1.5}$$

(Note that for $v = w = 1$ we find that $f = 1$ independent of z, which is to be expected.) Regard the multiplication factor f, with $v = w$, as a function of z. Then it is trivial to verify that $df/dz \leq 0$ with equality only if $v = 1$. It is obvious that $f(1) = F$, and $f(z) \geq F$ for all $z \in [0, 1)$. An easy way to see this is to rewrite f, for $v = w$, in a similar form as (2.1.4). We obtain

$$f = (1 - \frac{S_+}{S})b_1 + \frac{S_+}{S}v^2 b_2$$

where $b_1 \geq a_1$ and $b_2 \geq a_2$. We see that R_0 depends monotonically on z. A feeling for the rapidness of increase in the value of R_0 as one decreases the value of z can be obtained by making graphs similar to those in figures 2.1 and 2.2.

$\diamondsuit$

As a second extension we illustrate the use of the assumption of multigroup separable mixing by calculating R_0 for the case where we include the gender of an individual in the model of section 2.1.2. In our discussion in section 1.3.3 we take $i \in \{1, 2\}$ to denote gender, '1' denoting females and '2' denoting males (such in defiance of epidemiological tradition), and let $\Omega_1 = \Omega_2 = \{0, +\}$. We will use variables q, r, s for elements in Ω_1, Ω_2, and variables i, j for elements of $\{1, 2\}$. The h-state of an individual is then given by pairs of the form (i, q). Let $P_j(\tau, r, s)$ be the probability that an infected individual who became infected with D while its h-state was (j, s), has h-state (j, r) at time τ later. Note that the second component of the h-state is dynamic, where in chapter 1 it was static, and that we therefore have to modify our multigroup seperable mixing assumption. As before $A(\tau, (i, q), (j, s))$ denotes the expected infectivity of an infected individual that was itself infected with h-state (j, s), towards a susceptible with h-state (i, q), a time τ after becoming infected. Then we assume

$$A(\tau, (i, q), (j, s)) = \sum_{r=0,+} a_i(q) b_{ij}(s) P_j(\tau, r, s).$$

We allow all parameters from section 2.1.2 to be dependent on gender, and one can easily derive expressions for the functions $a_i(q)$ and $b_{ij}(s)$ for all combinations of i, j and q, s, by comparison with section 2.1.2. For example, for $i = 1, j = 2$ we find: $a_1(0) = 1, b_{12}(0) = \sigma_2 p_2 / S^1$; $a_1(+) = v_2, b_{12}(+) = \sigma_2 p_2 w_2 / S^1$, where $S^i = S_0(i) + S_+(i)$ and

$$S_0(i) = \frac{\gamma_i + \mu_i}{\gamma_i + \mu_i + \zeta_i} S^i.$$

The expected number of new cases with h-state (i, q) is given by

$$(K(S)\phi)(i, q) = S_q(i) a_i(q) \sum_{j=1}^{2} \sum_{r=0,+} \sum_{s=0,+} \int_0^\infty b_{ij}(r) P_j(\tau, r, s) d\tau \phi(j, s).$$

In biological terms this expression states, from the right end: there are $\phi(j,s)$ infected individuals that had h-state (j,s) at the moment of their own infection; during the time-period $[0,\tau)$ this h-state has changed to (j,r); the expected infectivity towards susceptible individuals with h-state (i,q) is then given by $a_i(q)b_{ij}(r)$; and finally, the new infected individuals with h-state (i,q) are a fraction of the susceptible population $S_q(i)$.

Because the 4×4-matrix $K(S)$ has a 2-dimensional range, we obtain from the theory in chapter 1 that R_0 is given by the dominant eigenvalue of the 2×2-matrix M with entries

$$m_{ij} = \sum_{r=0,+} \sum_{s=0,+} \int_0^\infty b_{ij}(r) P_j(\tau, r, s) S_s(j) a_j(s) d\tau$$

$i,j \in \{1,2\}$. What remains is to determine the P_j's. The matrix G, that describes the Markov transition rates between the four different h-states (i,q), is given by

$$G = \begin{pmatrix} -\zeta_1 & \gamma_1 & 0 & 0 \\ \zeta_1 & -\gamma_1 & 0 & 0 \\ 0 & 0 & -\zeta_2 & \gamma_2 \\ 0 & 0 & \zeta_2 & -\gamma_2 \end{pmatrix}.$$

If we define $\mathcal{P} = diag(P_1, P_2)$ with $P_i = (\int_0^\infty P_i(\tau, r, s) d\tau)_{r,s \in \{0,+\}}$, and $\mu = (\mu_1, \mu_2)^T$, $\rho = (\rho_1, \rho_2)^T$, then

$$\mathcal{P} = (\mu + \rho - G)^{-1}.$$

The dominant eigenvalue of M can be calculated explicitly. For example, in the case of heterosexual contacts only we obtain

$$R_0 = \sqrt{m_{12} m_{21}}.$$

$\diamondsuit$

As a third extension, let us consider the variant of the model in section 2.1.2 in which the (fixed) sexual activity level of an individual figures as another component of the h-state. We take the sexual activity level to be a discrete variable (but this is not a necessity). The possible h-states are indicated by $(i,0)$ and $(i,+)$ with $i = 0, 1, 2, \ldots$. Assuming

$$\frac{dS_{(i,+)}}{dt} = i\zeta S_{(i,0)} - \gamma S_{(i,+)} - \mu S_{(i,+)}$$

we find

$$S_{(i,0)} = \frac{\gamma + \mu}{\gamma + \mu + i\zeta} S_i$$

$$S_{(i,+)} = \frac{i\zeta}{\gamma + \mu + i\zeta} S_i$$

where S_i denotes the size of class i. Let

$$\phi_i := \begin{pmatrix} \phi_{(i,0)} \\ \phi_{(i,+)} \end{pmatrix}, \quad \phi := (\phi_0, \phi_1, ...)^T.$$

The operator $K(S)$ is now represented by an infinite matrix acting on ϕ, an infinite sequence of two-vectors. As we have assumed a fixed sexual activity level for an individual, the h-state dynamics can be considered for each ϕ_i of this sequence separately. We assume that for the ith two-vector these changes are governed by the matrix

$$G_i = \begin{pmatrix} -i\zeta & \gamma \\ i\xi & -\gamma \end{pmatrix}.$$

The meeting rate of an $(i, \cdot)$ individual with a $(j, \cdot)$ individual is assumed to be

$$\frac{\sigma ij}{\sum_k kS_k}.$$

The success ratios for D transmission are equal to those in the case with no diversity in sexual activity level treated before. If we write $\mathcal{P}_i = (\mu + \rho - G_i)^{-1}$, $\mathcal{P} := \text{diag}(\mathcal{P}_0, \mathcal{P}_1, ...)$, and $\mathcal{A} = (a_{ij})_{0 \le i,j \le \infty}$ with

$$a_{ij} = \frac{\sigma pij}{\sum_k kS_k} \begin{pmatrix} 1 & w \\ v & wv \end{pmatrix}$$

then $K(S)$ is the infinite matrix

$$K(S) = \text{diag}(S_{(0,0)}, S_{(0,+)}, ...)\mathcal{A}\mathcal{P},$$

with

$$K(S)_{ij} = \frac{\sigma pij}{\sum_k kS_k} \begin{pmatrix} S_{(i,0)} & wS_{(i,0)} \\ vS_{(i,+)} & wvS_{(i,+)} \end{pmatrix} \mathcal{P}_j.$$

So, the ith element of the image $K(S)\phi$ of ϕ is given by the two-vector

$$(K(S)\phi)_i = \frac{\sigma pi}{\sum_k kS_k} \begin{pmatrix} S_{(i,0)} & wS_{(i,0)} \\ vS_{(i,+)} & wvS_{(i,+)} \end{pmatrix} \sum_{j=0}^{\infty} j\mathcal{P}_j\phi_j.$$

The range of $(K(S)\phi)_i$ is spanned by the two-vector

$$\frac{\sigma pi}{\sum_k kS_k} \begin{pmatrix} S_{(i,0)} \\ vS_{(i,+)} \end{pmatrix} \sim \frac{ip_i}{\gamma + \mu + i\zeta} \begin{pmatrix} \gamma + \mu \\ iv\zeta \end{pmatrix}$$

where $p_i := S_i/(\sum_k S_k)$, the fraction of susceptible individuals with sexual activity level i. As before, this leads directly to an expression for R_0,

$$R_0 = \frac{\sigma p}{\sum_k kS_k} \begin{pmatrix} 1 \\ w \end{pmatrix}^T \sum_j j^2 \mathcal{P}_j \begin{pmatrix} S_{(j,0)} \\ vS_{(j,+)} \end{pmatrix}$$

or, in more detail,

$$R_0 = \frac{\sigma p}{\sum_k k p_k} \sum_j j^2 p_j \frac{(\gamma+\mu)(\mu+\rho+\gamma+j\zeta w)+j\zeta v w\left(\frac{\gamma}{w}+\mu+\rho+j\zeta\right)}{(\gamma+\mu+j\zeta)(\mu+\rho)(\mu+\rho+\gamma+j\zeta)}.$$

Formula (4.32) in May and Anderson (1988) is the analogue of this expression when one starts from (1.2.1) in chapter 1 and sets out to find the initial exponential growth rate λ_d.

2.2. Age structure

In this example we focus our attention on a continuous dynamic h-state variable.

Let $\mathcal{F}(a)$ denote the survival probability as function of age a, in the absence of the disease. Then, at population dynamical equilibrium

$$S(a) = S(0)\mathcal{F}(a).$$

Let $\gamma(\tau, a, \alpha)$ be the expected infectivity of an infected individual of age α and d-age τ, towards a susceptible individual of age a. Then

$$A(\tau, a, \alpha) = \gamma(\tau, a, \alpha+\tau)\frac{\mathcal{F}(\alpha+\tau)}{\mathcal{F}(\alpha)}$$

and

$$(K(S)\phi)(a) = S(0)\mathcal{F}(a)\int_0^\infty \int_0^\infty \gamma(\tau, a, \alpha+\tau)\frac{\mathcal{F}(\alpha+\tau)}{\mathcal{F}(\alpha)}\phi(\alpha)d\alpha d\tau.$$

2.2.1 Separable mixing

Under the separable mixing assumption

$$\gamma(\tau, a, \alpha) = f(a)g(\tau, \alpha) \tag{2.2.1}$$

we find

$$R_0 = S(0)\int_0^\infty \int_0^\infty g(\tau, \alpha+\tau)\mathcal{F}(\alpha+\tau)f(\alpha)d\alpha d\tau.$$

2.2.2 Endemic steady states

We shall now consider an endemic steady state. Let

$$\lambda(a) = \text{age specific force of infection,}$$

i.e. the age specific probability per unit of time of becoming infected. The survival function

$$\mathcal{F}_i(a) = e^{-\int_0^a \lambda(\alpha)d\alpha} \tag{2.2.2}$$

describes tha probability of being suceptible for those who did not die. Hence

$$\hat{S}(a) = \hat{S}(0)\mathcal{F}(a)\mathcal{F}_i(a)$$

where $\hat{S}$ describes the susceptible population in a steady *endemic* state. The age specific incidence rate is $\lambda(a)\hat{S}(a)$ and consistency now requires that

$$\begin{aligned}\lambda(a) &= \int_0^\infty \int_0^\infty A(\tau, a, \alpha)\lambda(\alpha)\hat{S}(\alpha)d\alpha d\tau \\ &= \hat{S}(0)\int_0^\infty \int_0^\infty \gamma(\tau, a, \alpha+\tau)\mathcal{F}(\alpha+\tau)\mathcal{F}_i(\alpha)\lambda(\alpha)d\alpha d\tau\end{aligned}$$

which can be considered as a nonlinear (recall (2.2.2)) integral equation for the (unknown) function λ. Nota that linearisation at the trivial solution $\lambda \equiv 0$ and the transformation $\phi \to \mathcal{F}\lambda$ lead us back to the eigenvalue problem for $K(\hat{S})$, as to expected. If we assume separable mixing (2.2.1), we find necessarily

$$\lambda(a) = Qf(a),$$

where the scalar Q has to satisfy the characteristic equation

$$1 = \hat{S}(0)\int_0^\infty \int_0^\infty g(\tau, \alpha+\tau)\mathcal{F}(\alpha+\tau)e^{-Q\int_0^\alpha f(\sigma)d\sigma} f(\alpha)d\alpha d\tau. \tag{2.2.3}$$

2.2.3 Vaccination

Dietz and Schenzle (1985) consider the effect of vaccination and take

$$\hat{S}(a) = \hat{S}(0)\mathcal{F}(a)\mathcal{F}_v(a)\mathcal{F}_i(a),$$

where $\mathcal{F}_v(a)$ denotes the probability that an individual which did not die is immune due to vaccination. The analogue of (2.2.3) then is

$$1 = \hat{S}(0)\int_0^\infty \int_0^\infty g(\tau, \alpha+\tau)\mathcal{F}(\alpha+\tau)\mathcal{F}_v(\alpha)e^{-Q\int_0^\alpha f(\sigma)d\sigma} f(\alpha)d\alpha d\tau$$

which can alternatively be written as

$$1 = \hat{S}(0) \int_0^\infty \int_0^\infty g(\tau,\theta)\mathcal{F}(\theta)\mathcal{F}_v(\theta-\tau)e^{-Q\int_0^{\theta-\tau} f(\sigma)d\sigma} f(\theta-\tau)d\tau d\theta.$$

If we adopt the further assumption that

$$g(\tau,\alpha) = h(\alpha)k(\tau)\mathcal{F}_r(\tau),$$

where k describes the infectivity as a function of d-age and $\mathcal{F}_r$ describes the 'removal' from the infected class, we finally arrive at

$$1 = \hat{S}(0) \int_0^\infty h(\theta)\mathcal{F}(\theta) \int_0^\infty k(\tau)\mathcal{F}_r(\tau)\mathcal{F}_v(\theta-\tau)e^{-Q\int_0^{\theta-\tau} f(\sigma)d\sigma} f(\theta-\tau)d\tau d\theta$$

which is, apart from the notation, identical to formula (3) in Dietz and Schenzle (1985). These authors introduce yet two other simplifications:

(i) $h = f$, i.e. susceptibles and infectives have the same age dependence in activity level;
(ii) the duration of the disease is short on the time-scale of ageing.

Then the last equation above can be approximated by

$$1 = C\hat{S}(0) \int_0^\infty f^2(\theta)\mathcal{F}(\theta)\mathcal{F}_v(\theta)e^{-Q\int_0^{\theta} f(\sigma)d\sigma} d\theta$$

where C is a constant (describing the 'magnitude' of the total infectivity). One can now use data about the endemic state to estimate f, Q and $\mathcal{F}$, and subsequently calculate whether or not a given $\mathcal{F}_v$ suffices to eradicate the disease. We refer once more to Dietz and Schenzle (1985) for additional information.

2.2.4 Separable mixing with enhanced within age group infection

To conclude this subsection we show how to compute the analogue of the threshold condition (1.3.6 i,ii) in chapter 1 in the case of age dependence (recall that in deriving the condition in section 1.3.2 we assumed that the h-state is constant which it obviously is not if we consider age). Assume that

$$\gamma(\tau,a,\alpha) = f(a)g(\tau,\alpha) + h(\tau,\alpha)\delta(a-\alpha),$$

where δ denotes Dirac's delta 'function'. Then

$$\begin{aligned}(K(S)\phi)(a) =& S(a)\Big\{f(a)\int_0^\infty \int_0^\infty g(\tau,\alpha+\tau)\frac{\mathcal{F}(\alpha+\tau)}{\mathcal{F}(\alpha)}\phi(\alpha)d\alpha d\tau \\ &= \int_0^a h(\tau,a)\frac{\mathcal{F}(a)}{\mathcal{F}(a-\tau)}\phi(a-\tau)d\tau\Big\}.\end{aligned}$$

define an operator L by

$$(L\psi)(a) = S(a)\int_0^a h(a-\alpha,a)\frac{\mathcal{F}(a)}{\mathcal{F}(\alpha)}\psi(\alpha)d\alpha$$

and rewrite the eigenvalue problem $K(S)\phi = \rho\phi$ as

$$\theta(\psi)Sf + L\phi = \rho\phi,$$

where θ is the $\mathbb{C}$-valued mapping defined by

$$\theta(\psi) = \int_0^\infty \int_0^\infty g(\tau,\alpha+\tau)\frac{\mathcal{F}(\alpha+\tau)}{\mathcal{F}(\alpha)}\psi(\alpha)d\alpha d\tau.$$

For ρ real and sufficiently large we can invert $\rho I - L$. In fact, one has the series expansion

$$(\rho I - L)^{-1} = \sum_{n=0}^{\infty}\frac{L^{(n)}}{\rho^{n+1}}. \tag{2.2.4}$$

By substituting $\phi = (\rho I - L)^{-1}\theta(\phi)Sf$ in the definition of θ we find the characteristic equation

$$1 = \int_0^\infty \int_0^\infty g(\tau,\alpha+\tau)\frac{\mathcal{F}(\alpha+\tau)}{\mathcal{F}(\alpha)}((\rho I - L)^{-1}Sf)(\alpha)d\alpha d\tau.$$

If we assume that (2.2.4) keeps converging up to $\rho = 1$ (this is the analogue of assumption $c(\xi)S(\xi) < 1$ for all $\xi \in \Omega$ in chapter 1) we find that $R_0 > 1$ if and only if

$$\int_0^\infty \int_0^\infty g(\tau,\alpha+\tau)\frac{\mathcal{F}(\alpha+\tau)}{\mathcal{F}(\alpha)}\sum_{n=0}^{\infty}(L^{(n)}Sf)(\alpha)d\alpha d\tau > 1.$$

This condition allows an interpretation similar to that of condition (1.3.6 ii) given in section 1.3.2.

2.3. Multigroup separable mixing and pigs

Suppose one regards a viral disease in a heterogeneous population of animals (of the same species) that are reared in different stables on farms, where the animals are moved from one stable to another (and possibly also between farms) at regular intervals. Suppose furthermore that a vaccine is available for this particular disease and that one would like to determine, using R_0, which vaccination strategy, if any, is 'optimal' in the sense that it leads to eradication of the disease under given economic constraints. Basically, one would like

to know which subgroups of animals should be vaccinated how often and in which order, and with which vaccine (if more than one kind is available). The situation we have in mind concerns the pseudorabies virus (causing Aujeszky's disease in pigs) on a special type of pig breeding farm, a so-called farrow-to-finish operation (see, e.g., Grenfell and Smith, 1990). Against the pseudorabies virus vaccines are available since 1975, but only recently 'deleted' vaccines have been developed. Antibodies against these deleted vaccines can be distinguished from antibodies induced by a wild-type infection. Only these vaccines are allowed against the psuedorabies virus in the Netherlands. This makes it possible to accurately trace the natural spread of infection in vaccinated herds. A future aim is to investigate whether an efficient vaccination campaign can be devised that leads to eradication of pseudorabies in the Netherlands.

Of course, a prerequisite for ultimately carrying out a program as described above, is that one can actually compute R_0 as a function of relevant and 'measurable' parameters. In the present section we describe, for the farrow-to-finish operation, the construction of the next-generation operator, but we do not go into details about the nature of the infectivity function $A(\tau, \xi, \eta)$. For the case where we let time proceed in discrete steps, we give in De Jong, Diekmann, Heesterbeek (in preparation) an explicit algorithm to calculate R_0 for the more general situation (apart from the discreteness of time) where we allow an arbitrary finite number of different units (e.g., stables) that the individuals can visit in any sequence desired.

Imagine that we have three stables: a farrowing house, a nursery unit and a finishing unit, see figure 2.4.

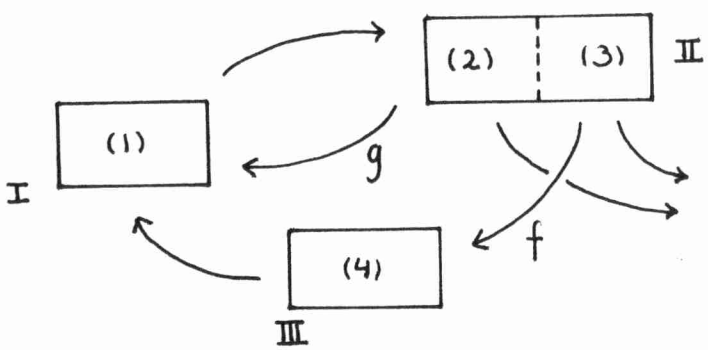

Figure 2.3 Flow diagram of individuals between the various compartments, see text for details

The pregnant sows are kept in the farrowing house (I) for a period of time T_1, until the moment that they are moved to the nursery unit (II) to give birth. The sows are kept in unit II together with their gilts for a certain period of

time T_2. After that a fraction g of the sows is impregnated again and returned to unit I for a new cycle, the remaining sows are eaten. Of the gilts, upon leaving unit II, a fraction f (of course only females) is reared in the finishing unit (III) to let them mature into sows that enter the pregnancy cycle I $\rightarrow$ II $\rightarrow$ I ... etc. after a certain amount of time T_3. The remainder of the female gilts, and the male gilts are sold to other type of breeding farms or are eaten (the overall effect is that all pigs get eaten, but some pigs get eaten sooner than others).

How can we calculate R_0 given that sows and gilts react differently to the disease, and when the age of an individual is an important determinant for the expected course of the infection? Let us assume: there is no vertical transmission; we have homogeneous mixing within each of the three different units; there is no extra death rate due to the disease; the animals in the fractions g and f are selected irrespective of being infected or not. As h-state we recognise:

(i) four types of individuals: sows in unit I (1), sows in unit II (2), gilts in unit II (3), and gilts in unit III (4).
(ii) the 'unit-age' of an individual, i.e. the time that has passed since the individual entered a certain unit. We indicate the unit-age by α_1 for unit I ($0 \leq \alpha_1 \leq T_1$), by α_2, α_3 for unit II ($0 \leq \alpha_2, \alpha_3 \leq T_2$), and by α_4 for unit III ($0 \leq \alpha_4 \leq T_3$).

An individual is characterised by the pair (i, α_i), $i \in \{1, 2, 3, 4\}$.

First we assume, for the moment, that an individual that leaves a certain unit, cannot return to it later. We ask ourselves who can infect whom. This is easy, all types of individuals can infect all other types except for the fact that sows (1 and 2) cannot infect gilts of type 4, because there are no sows in unit III. We want to find the operator $K(S)$. Let $\phi = \phi(i, \alpha_i)$ be the distribution of infected individuals over the h-state space $\Omega = \cup_{i=1}^{4} \{i\} \times \Omega_i$, with $\Omega_1 = [0, T_1]$, $\Omega_2 = \Omega_3 = [0, T_2]$, and $\Omega_4 = [0, T_3]$. With $[j \rightarrow (i, \alpha_i); \phi]$ we indicate the expected number of new cases with h-state (i, α_i) that are caused by individuals that were themselves infected as type j with their distribution with respect to unit-age given by ϕ. The next-generation operator $K(S)$, acting on $L_1(\Omega)$, is then described by

$$(K(S)\phi)(i, \alpha_i) = \sum_{j=1}^{4} [j \rightarrow (i, \alpha_i); \phi], \qquad i \in \{1, 2, 3, 4\}.$$

To determine the entries $[j \rightarrow (i, \alpha_i); \phi]$ we start with one fixed infected individual in a certain unit, and follow by the right book-keeping the infections that this individual causes on its travels through the various units. As always in R_0 calculations, we start from a susceptible population in a stable demographic state in absence of the disease. Let $S = S(i, \alpha_i)$ denote that stable state with respect to type and unit-age. Let us regard an individual with h-state $(4, \alpha_4')$ that has just become infected , i.e. a gilt in unit III who becomes

infected after spending an amount of time α_4' in unit III. The book-keeping now proceeds as follows: for $[4 \to (4,\alpha_4);\phi]$ we keep track of the infections that our individual causes in its own unit III, in the remaining time $T_3 - \alpha_4'$ that is left before it has to leave the unit; it then proceeds to unit I where it can infect pregnant sows $[4 \to (1,\alpha_1);\phi]$; finally, it arrives in unit II where it can infect newborns $[4 \to (3,\alpha_3);\phi]$, and nursing sows $[4 \to (2,\alpha_2);\phi]$. If we define $A(\tau,(i,\alpha_i),(4,\alpha_4'))$ as the expected infectivity of our individual towards susceptibles with h-state (i,α_i) at a time τ after it became infected, then, for $i = 4$,

$$[4 \to (4,\alpha_4);\phi] = S(4,\alpha_4)\int_0^{T_3}\int_0^{T_3-\alpha_4'} A(\tau,(4,\alpha_4),(4,\alpha_4'))\phi(4,\alpha_4')d\tau d\alpha_4',$$

(remember that we have, temporarily, assumed that an individual can visit each unit only once). The inner integral accounts for the fact that the individual we are following can be infectious in unit III for a time-period of length anywhere in the interval $[0, T_3 - \alpha_4']$. Our infected individual subsequently enters unit I and there

$$[4 \to (1,\alpha_1);\phi] = S(1,\alpha_1)\int_0^{T_3}\int_{T_3-\alpha_4'}^{T_1+T_3-\alpha_4'} A(\tau,(1,\alpha_1),(4,\alpha_4'))\phi(4,\alpha_4')d\tau d\alpha_4'.$$

The next unit our infected individual enters is unit II, and there it could infect both sows and newborns, respectively described by $[4 \to (2,\alpha_2);\phi]$:

$$S(2,\alpha_2)\int_0^{T_3}\int_{T_1+T_3-\alpha_4'}^{T_2+T_1+T_3-\alpha_4'} A(\tau,(2,\alpha_2),(4,\alpha_4'))\phi(4,\alpha_4')d\tau d\alpha_4'$$

and $[4 \to (3,\alpha_3);\phi]$:

$$S(3,\alpha_3)\int_0^{T_3}\int_{T_1+T_3-\alpha_4'}^{T_2+T_1+T_3-\alpha_4'} A(\tau,(3,\alpha_3),(4,\alpha_4'))\phi(4,\alpha_4')d\tau d\alpha_4'.$$

The expected number of infections caused by individuals 'born' with types $(1,\alpha_1')$, $(2,\alpha_2')$ and $(3,\alpha_3')$ are, straightforwardly, given by analogous expressions. In the cases $[3 \to (4,\alpha_4);\phi]$ and $[3 \to (1,\alpha_1);\phi]$ we have to multiply by f, because only a fraction f of the newborns will be reared in the finishing unit.

Of course, the precise nature of the function A will have to be determined by a sub-model that takes the particularities of the disease one studies into account. Let us assume, for here, that we have local separable mixing, i.e.

$$A(\tau,(i,\alpha_i),(j\alpha_j)) = a_i(\alpha_i)b_{ij}(\tau,\alpha_j),$$

where we normalise by taking

$$\int_{\Omega_i} a_i(\alpha_i)S(i,\alpha_i)d\alpha_i = 1, \qquad i = 1,2,3,4.$$

The next-generation operator $K(S)$ has a 4-dimensional range under this condition, with eigenvectors

$$\phi(i,\alpha_i) = \sigma_i S(i,\alpha_i)a_i(\alpha_i), \qquad i = 1,2,3,4,$$

where σ is an eigenvector of a 4×4-matrix $M = (m_{ij})_{1\le i,j\le 4}$, and the m_{ij} are determined by the $[j \to (i,\alpha_i);\phi]$'s. For example, for m_{14} we find from $[4 \to (1,\alpha_1);\phi]$

$$m_{14} = \int_0^{T_3} a_4(\alpha_4')S(4,\alpha_4') \int_{T_3-\alpha_4'}^{T_1+T_3-\alpha_4'} b_{14}(\tau,\alpha_4')d\tau d\alpha_4',$$

and in general

$$m_{ij} = \int_{\Omega_j} a_j(\alpha_j')S(j,\alpha_j') \int_{upper_{ij}}^{lower_{ij}} b_{ij}(\tau,\alpha_j')d\tau d\alpha_j',$$

(where, in the cases m_{43}, m_{13} we have to multiply by f). The basic reproduction ratio is the dominant eigenvalue of matrix M.

Remember that we assumed before embarking upon the above calculation that our infected individual could only tour the three units once. In reality of course, sows can repeatedly go around in the circle $II \to I \to II \to \cdots$. The expressions for the m_{ij}'s then become more complicated, but can still be written down by bookkeeping. We only treat m_{12} as an example: the infection of sows in unit I by a sow that was 'born' in unit II. The infected sow can re-enter unit I, after leaving unit II, with probability θ, from I it will then go to II once more, etc.. We obtain

$$m_{12} = \int_{\Omega_2} a_2(\alpha_2')S(2,\alpha_2') \sum_{n=1}^{\infty} \theta^n \int_{nT_2+(n-1)T_1-\alpha_2'}^{n(T_2+T_1)-\alpha_2'} b_{12}(\tau,\alpha_2')d\tau d\alpha_2'.$$

The other m_{ij}'s are extended in a similar vein.

Remark. Because we do not take death of individuals into account, and we assume a constant inflow of new individuals, the $S(i,\alpha_i)$'s will be constants S_i, that are independent of the unit-age. In every unit there are an equal number of individuals with any given unit-age. The relations that then have to hold between the S_i's are:

$$\begin{aligned} fS_3 &= S_4 \\ S_1 &= S_4 + S_2 \\ \frac{S_3}{S_2} &= \text{average litter-size} \\ S_1 &= S_2. \end{aligned}$$

$\diamondsuit$

2.4. References

S.O. Aral & K.K. Holmes (1991): Sexually transmitted diseases in the AIDS era. *Sc. American* **264**: 18-25.

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

M.C.M. De Jong, O. Diekmann, J.A.P. Heesterbeek (1992): Computation of R_0 for discrete-time epidemic models with contact structure. (manuscript)

K. Dietz & D. Schenzle (1985): Proportionate mixing models for age-dependent infection transmission. *J. Math. Biol.* **22**: 117-120.

B.T. Grenfell & G. Smith (1990): Mathematical model for the impact of a pseudorabies epizootic on the productivity of a farrow-to-finish operation. *Am. J. Vet. Res.* **51**: 156-164.

R.M. May & R.M. Anderson (1988): The transmission dynamics of human immunodeficiency virus (HIV). *Phil. Trans. R. Soc. Lond.* **B 321**: 565-607.

P. Piot, F.A. Plummer, F.S. Mhalu, J.L. Lamboray, J. Chin & J.M. Mann (1988): AIDS: An International Perspective. *Science* **239**: 573-579.

Chapter 3

R_0 for sexually transmitted diseases

In this chapter it is shown how one can calculate the basic reproduction ratio R_0 for infectious disease models where an arbitrary but finite number of disease-states are recognised and where the phenomena of pair formation and separation are taken into account. Several examples are discussed. We apply the theory to investigate the effects of variable infectivity on the spread of HIV in a heterosexual population. We calculate R_0 as a function of the number of new partners during the infectious period, keeping the total number of contacts fixed. Numerical evidence suggests that R_0 decreases for variable infectivity, if the average infectivity is kept constant. Finally, we make a preliminary attempt to study the effects of behaviour change on the spread of HIV.

3.1. Introduction

Biologically speaking the basic reproduction ratio R_0 is the expected number of secondary cases caused by one typical infected individual during its entire period of infectiousness, in a population consisting of susceptibles only. Mathematically speaking one investigates whether or not, starting from a few infected individuals, the disease can invade into a susceptible population that is in its demographic steady state (chapter 1). Because the initial number of infectious individuals is low, one can assume that every contact an infectious individual will make is necessarily with a susceptible, and that, during the initial phase, the infectious process will not cause an appreciable decrease in the density of susceptibles. This makes the calculation of R_0 into a linear problem, and it is the reason that its determination can be carried out allowing for arbitrary complexity in the description of the transmission dynamics. The main idea is to regard generations of infected individuals and to construct a certain operator

that describes the transmission dynamics of the disease as a discrete process relating subsequent generations. In chapter 1, it was shown what this operator looks like and that R_0, as 'defined' above, equals the operator's spectral radius (R_0 can equivalently be characterised as the dominant eigenvalue of this 'next-generation operator').

For sexually transmitted diseases it has been advocated by Dietz and Hadeler (1988) that a 'realistic' model should take into account the fact that individuals form partnerships for non-negligible periods of time. During that time period the two partners only have contacts with each other and in this way they are, momentarily, not in danger of receiving the infection from individuals outside the pair. Perhaps even more important, from the point of view of the disease, is that part of the 'infective potential' of an infected individual in a pair is 'wasted' if the partner is also infected. In chapter 1, the possibility of pair formation and separation was not incorporated. Two assumptions underlying the 'construction' of the next-generation operator fail if we allow individuals to form pairs and this entails that we cannot incorporate pair formation by a *direct* generalisation of the next-generation operator. Implicitly it was assumed that every contact an infected individual has, is with a 'new' susceptible, which is by definition no longer true in the pair formation case. Explicitly it was assumed that the only relevant 'output' of an individual (i.e. that what one has to know of an infected individual to determine its influence on the spread of the infection) was the expected *infectivity* A (where the average is taken over all possible sample paths of disease progress). As a consequence, an age representation for the expected infectivity status of an individual could be used. In other words, in a sufficiently large population, A can be considered as a deterministic function of τ, where τ measures the time that has passed since the individual became infected (for homogeneous populations this is the approach of Kermack-Mckendrick (1927)). In the case of pair formation a second output quantity, survival, comes into play. Of course, the survival of an infected individual is always important, since it influences its infectivity, and as such needs to be incorporated in A. Within the context of pair formation models however, the survival of the partner has a second influence on the spread of infection: if your partner dies, you yourself become available for new contacts. For models incorporating pair formation an equivalent age representation of disease-state is not always possible.

However, we can make use of the ideas underlying the R_0 calculation in chapter 1. We can still construct an operator that describes the transmission dynamics as a discrete process on generations of infected individuals. In this chapter we show what this operator looks like in the case of pair formation if we recognise an arbitrary but finite number of possible disease-states $\{1, ..., n\}$ which are passed through sequentially, always starting in state 1. The disease-state of an individual determines its current infectivity level and probability of dying. For the pair formation processes we practically follow the simplest model described in Dietz (1989), Dietz and Hadeler (1988). The difference is that in those papers a pair starts, by definition, with a sexual contact. In this

chapter we take a 'period of courtship', in which the pair is not yet sexually active, into account.

In section 3.2 we describe the model assumptions, explain the ingredients that are necessary for determining R_0 and give the 'algorithm' for its calculation. The next-generation operator turns out to be an $m \times m$-matrix, where m is the number of disease-states with positive infectivity, and R_0 will be the dominant eigenvalue of this matrix. These results are generalised in section 3.3 to include arbitrary heterogeneity (in susceptibility) among the individuals in the population. As an example we treat the characteristics male/female. In section 3.4 we briefly consider the results of various limit procedures applied to the present pair formation model. Among other things it is shown how the appropriate models that neglect pair formation can be obtained as limiting cases. In sections 3.5 and 3.6 we apply the theory to investigate the effect of variable infectivity on the spread of HIV, and we show some preliminary results concerning the incorporation of behaviour change.

This chapter treats a pair formation analogue of a multi-stage variable infectiousness model developed by Blythe and Anderson (1988) and the calculations are therefore a generalisation of the results in Dietz (1989), where the cases of one and two disease-states were considered.

3.2. Description of the model and calculation of R_0

In our model we distinguish two classes of pairs: those that are in a courtship period, and those that are in the sexually active phase. The courtship period is characterised by the absence of sexual contacts. So individuals in a courtship are, just as single individuals, not at risk of either receiving or transmitting the infection. Individuals only have sexual contacts in the sexually active phase of a partnership.

In addition to the courtship label, we recognise the following characteristics of an *infected* individual:

disease-state: $i \in \{1, ..., n\}$
partnership-state: $j \in \{-1, 0, 1, ..., n\}$.

Here '-1' means that the individual is single (no partner at the moment of observation); '0' means that the individual is paired with a susceptible; $j \in \{1, ..., n\}$ means that the individual is paired with an infected individual that has disease-state j. Together the two characteristics determine the *type* (i, j) of an infected individual. Types of individuals in a courtship period are indicated by $(i, j)_c$. For the moment we assume that all individuals are equally susceptible (so we disregard any heterogeneity other than the single-pair dichotomy).

As we are only interested in R_0, we assume that every new partner of a single infected individual is necessarily a susceptible. The consequence of

this is that the only courtship-types that we have to consider are $(i,0)_c$, for $i \in \{1,...,n\}$, where the infected individual has a susceptible partner. We assume that all pairs start with a courtship phase.

Let $\Lambda := \{(i,j)|1 \leq i \leq n,\ -1 \leq j \leq n\} \cup \{(i,0)_c : 1 \leq i \leq n\}$ be the set of all possible types. Then $|\Lambda| = n(n+3)$ and consequently our type-space is $\mathbb{R}^{n(n+3)}$. Let $\Sigma := \{1,2,...,n(n+3)\}$. We will call Σ the state space of infectives and the elements of Σ are called *states*. Let $L : \Lambda \to \Sigma$ describe the lexicographic ordering on Λ, with the side condition that a courtship-type precedes the corresponding sexually active type, i.e.

$$L(i,j) < L(i',j') \iff \{i < i'\} \vee \{i = i',\ j < j'\},$$
$$L(i,-1) < L(i,0)_c < L(i,0),\quad 1 \leq i \leq n.$$

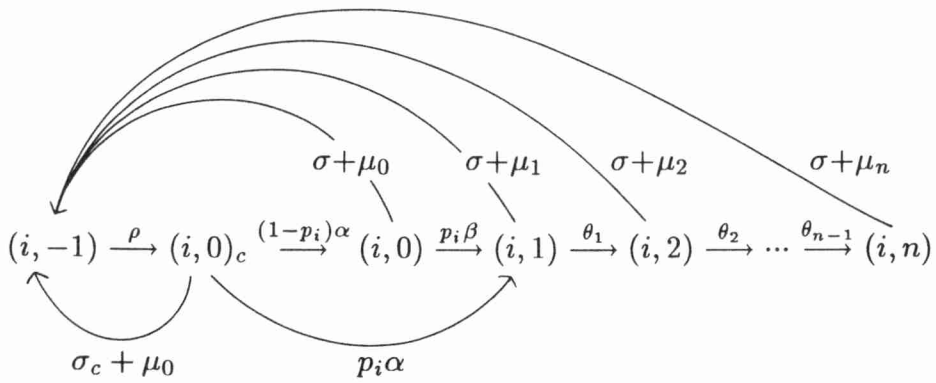

Figure 3.1 Possible type-changes of an infected individual of fixed disease-state i, provided the individual stays alive

We make the following assumptions:

1. The disease-states are passed through in natural order. In particular, a freshly infected individual starts its life (in disease sense) with type $(1,j)$ for some $j \in \{1,...,n\}$.

2. Given that the infected individual does not die, the time that its disease-state is i is exponentially distributed with parameter θ_i (where $\theta_n = 0$, i.e. disease-state n is retained until death).

3. The infectivity in disease-state i is described by the probability $p_i \geq 0$ that a sexual contact with a susceptible leads to transmission.

4. μ_0 is the death-rate of susceptibles, μ_i is the death-rate of an infected individual with disease-state i.

5. Every single individual has a constant probability ρ per unit of time to become a member of a courtship-pair. The divorce-rate is σ_c in the courtship phase, and σ in the sexually active phase.

6. A pair always starts with a courtship phase. The length of this phase is, conditional on the survival and no divorce of the two partners, exponentially distributed with parameter α. By definition the sexually active phase starts with one sexual contact.

7. During the sexually active phase, the partners have β sexual contacts per unit of time.

The graph in figure 3.1 traces the possible changes in the type of an infected individual of fixed disease-state i, as long as it does not die. Note that the only two types of individual that can *cause* an infection are $(i,0)_c$ (first contacts), and $(i,0)$, $i \in \{1,...,n\}$.

Define a matrix $M : \mathbb{R}^n \to \mathbb{R}^n$ with elements m_{ij}, $(1 \leq i,j \leq n)$, as follows:

> m_{ij} is the sum of the expected number of type-transitions $(i,0)_c \to (i,1)$, and $(i,0) \to (i,1)$, during the entire remaining life, of an individual that just became type $(1,j)$.

The matrix M is the next-generation operator, mapping a generation of infectives, distributed with respect to the disease-state of the partner at the moment infection took place, onto the next such generation. Or, in other words, M yields the next generation, given the present one, while keeping account of the state at 'birth'. As was shown in chapter 1 we have to carry out the right averaging over the m_{ij} to arrive at R_0: R_0 is the dominant eigenvalue of the matrix M. In the following we determine the precise nature of M on the basis of the assumptions listed above.

It is convenient to work both with types in Λ (in cases where we use the interpretation to make inferences) and with states in Σ (if we just do straightforward linear algebra), and we will accordingly choose the representation that is the easiest in a given situation.

We regard the changes in disease-state and partnership-state of an individual as a Markov process on the state space Σ. Let a matrix $G : \mathbb{R}^{n(n+3)} \to \mathbb{R}^{n(n+3)}$ describe the transition probabilities per unit of time between the states, i.e. g_{kl} gives the rate of leaving state $l \in \Sigma$ to go to state $k \in \Sigma$ (note that in the probabilistic literature on Markov processes it is usually the other way around), and $g_{ll} = -\sum_{k \neq l} g_{kl}$ – rate of dying. If we let $P(\tau) : \mathbb{R}^{n(n+3)} \to \mathbb{R}^{n(n+3)}$ be the matrix containing the probabilities $P_{kl}(\tau)$ of being in state k and alive at time τ after starting in state l at time zero, then we have

$$P(\tau) = e^{G\tau}.$$

The interpretation of the m_{ij} tells us that for their calculation we need to know the probability that a freshly infected individual, 'born' in state $L(1,j)$, is still

alive at time τ after the infection occured and that its state is $L(i,0)_c$ or $L(i,0)$ at that time. Then,

$$m_{ij} = p_i\beta \int_0^\infty P_{L(i,0)L(1,j)}(\tau)d\tau \; + \; p_i\alpha \int_0^\infty P_{L(i,0)_cL(1,j)}(\tau)d\tau, \qquad (3.2.1)$$

or, in other words, the expected number of times that an infected individual becomes of type $(i,1)$, given that it is 'born' with type $(1,j)$, is

$$m_{ij} = -p_i\beta(G^{-1})_{L(i,0)L(1,j)} - \; p_i\alpha(G^{-1})_{L(i,0)_cL(1,j)} \qquad (3.2.2)$$

for $1 \leq i,j \leq n$.

The next task is to specify G and to calculate the right elements of G^{-1}. The structure of G is determined by the assumption that all infected individuals start their 'infected life' with disease-state 1 and that their disease-state from there on rises from time to time by one, up to n, as long as the individual does not die 'along the way'. Exploiting the structure we can explicitly write down the inverse of G in a very simple way.

Example 3.1. We work out the case with three disease-states, $n = 3$, because this is a prototype for all $n \geq 1$. We have $\Sigma = \{1, ..., 18\}$ and

$$G = \begin{pmatrix} A_1 & 0 & 0 \\ D_1 & A_2 & 0 \\ 0 & D_2 & A_3 \end{pmatrix}$$

where '0' is the 6×6 zero-matrix, $D_j = diag(\theta_j) = \theta_j Id$, and A_i, $i \in \{1,2,3\}$ is given by

$$\begin{pmatrix} a_1(i) & \mu_0 + \sigma_c & \mu_0 + \sigma & \mu_1 + \sigma & \mu_2 + \sigma & \mu_3 + \sigma \\ \rho & a_2(i) & 0 & 0 & 0 & 0 \\ 0 & (1-p_i)\alpha & a_3(i) & 0 & 0 & 0 \\ 0 & p_i\alpha & p_i\beta & a_4(i) & 0 & 0 \\ 0 & 0 & 0 & \theta_1 & a_5(i) & 0 \\ 0 & 0 & 0 & 0 & \theta_2 & a_6(i) \end{pmatrix}$$

where $a_1(i) = -\mu_i - \theta_i - \rho$; $a_2(i) = -\mu_i - \mu_0 - \theta_i - \sigma_c - \alpha$; $a_3(i) = -\mu_i - \mu_0 - \theta_i - \sigma - p_i\beta$; $a_4(i) = -\mu_i - \mu_1 - \theta_i - \theta_1 - \sigma$; $a_5(i) = -\mu_i - \mu_2 - \theta_i - \theta_2 - \sigma$; $a_6(i) = -\mu_i - \mu_3 - \theta_i - \sigma$. It is easily verified that G^{-1} can be expressed in the 6×6 matrices that constitute G as follows

$$G^{-1} = \begin{pmatrix} {A_1}^{-1} & 0 & 0 \\ -D_1{A_2}^{-1}{A_1}^{-1} & {A_2}^{-1} & 0 \\ D_1D_2{A_3}^{-1}{A_2}^{-1}{A_1}^{-1} & -D_2{A_3}^{-1}{A_2}^{-1} & {A_3}^{-1} \end{pmatrix}.$$

The matrix M can now be determined. Let us consider the special case $p_2 = 0$. The assumption that individuals with disease-state 2 are not infectious implies

both that no individual in state $L(2,0)$, or $L(2,0)_c$, can infect its partner *and* that the individual itself cannot have been 'born' in the state $L(1,2)$. We find that

$$M = \begin{pmatrix} m_{11} & 0 & m_{13} \\ 0 & 0 & 0 \\ m_{31} & 0 & m_{33} \end{pmatrix}.$$

The dominant eigenvalue of this matrix equals the dominant eigenvalue of

$$M' = \begin{pmatrix} m_{11} & m_{13} \\ m_{31} & m_{33} \end{pmatrix}.$$

In terms of the elements of G^{-1} we can write

$$\begin{aligned} m_{11} &= -p_1\beta(G^{-1})_{L(1,0)L(1,1)} - p_1\alpha(G^{-1})_{L(1,0)_cL(1,1)} \\ &= -p_1\beta(G^{-1})_{3,4} - p_1\alpha(G^{-1})_{2,4} \end{aligned}$$

$$\begin{aligned} m_{13} &= -p_1\beta(G^{-1})_{L(1,0)L(1,3)} - p_1\alpha(G^{-1})_{L(1,0)_cL(1,3)} \\ &= -p_1\beta(G^{-1})_{3,6} - p_1\alpha(G^{-1})_{2,6} \end{aligned}$$

$$\begin{aligned} m_{31} &= -p_3\beta(G^{-1})_{L(3,0)L(1,1)} - p_3\alpha(G^{-1})_{L(3,0)_cL(1,1)} \\ &= -p_3\beta(G^{-1})_{15,4} - p_3\alpha(G^{-1})_{14,4} \end{aligned}$$

$$\begin{aligned} m_{33} &= -p_3\beta(G^{-1})_{L(3,0)L(1,3)} - p_3\alpha(G^{-1})_{L(3,0)_cL(1,3)} \\ &= -p_3\beta(G^{-1})_{15,6} - p_3\alpha(G^{-1})_{14,6} \end{aligned}$$

Finally we find that

$$R_0 = \frac{(m_{11} + m_{33}) + \sqrt{(m_{11} - m_{33})^2 + 4m_{31}m_{13}}}{2}.$$

$\diamondsuit$

In the general case G will be a $n(n+3) \times n(n+3)$ matrix of the following form (where the A_i, D_i, and 0 are $(n+3) \times (n+3)$ versions of their namesakes from example 3.1)

$$G = \begin{pmatrix} A_1 & 0 & \dots & \dots & 0 \\ D_1 & A_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & D_{n-1} & A_n \end{pmatrix}.$$

One checks easily that

$$G^{-1} = \begin{pmatrix} A_1^{-1} & 0 & \dots & \dots & 0 \\ & A_2^{-1} & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & B_{rs} & & \ddots & 0 \\ & & & & A_n^{-1} \end{pmatrix},$$

where the r, s^{th}-matrix below the diagonal, B_{rs} $(r > s)$, is given by

$$B_{rs} = (-1)^{r+s} D_s D_{s+1} \cdots D_{r-1} A_r^{-1} A_{r-1}^{-1} \cdots A_s^{-1}.$$

Note that we only need to know the inverse of every A_i in order to calculate G^{-1}.

Remark 1. As we have seen in example 3.1 the analysis is simplified if some of the p_i's are zero. If m of the p_i's are non-zero then M' becomes a $m \times m$ matrix.

$\diamondsuit$

Example 3.2. We elaborate our result for the case of one disease-state and then compare it with the expression given in Dietz (1989). For $n = 1$ we have

$$G = \begin{pmatrix} -\mu_1 - \rho & \mu_0 + \sigma_c & \mu_0 + \sigma & \mu_1 + \sigma \\ \rho & -\mu_0 - \mu_1 - \sigma_c - \alpha & 0 & 0 \\ 0 & (1-p)\alpha & -\mu_0 - \mu_1 - \sigma - p\beta & 0 \\ 0 & p\alpha & p\beta & -2\mu_1 - \sigma \end{pmatrix}.$$

and R_0 is given by

$$\begin{aligned} R_0 &= p\beta \int_0^\infty P_{L(1,0)L(1,1)}(\tau) d\tau + p\alpha \int_0^\infty P_{L(1,0)_c L(1,1)}(\tau) d\tau \\ &= -p\beta (G^{-1})_{3,4} - p\alpha (G^{-1})_{2,4}. \end{aligned}$$

Explicitly we find

$$R_0 = \frac{p\beta\alpha(\mu_1 + \sigma)(1-p)\rho + p\rho\alpha(\mu_1 + \mu_0 + \sigma + p\beta)(\mu_1 + \sigma)}{\mu_1(x+y)}. \qquad (3.2.3)$$

with

$$\begin{aligned} x &= (\mu_0 + \mu_1 + \sigma + p\beta)(2\mu_1 + \sigma)(\mu_0 + \mu_1 + \rho + \sigma_c + \alpha), \\ y &= \rho\alpha(p\mu_0 + (2-p)\mu_1 + \sigma + p\beta). \end{aligned}$$

In Dietz (1989) the courtship period is infinitely short. If we let the rate α of entering the sexually active phase tend to infinity (i.e. the average length of a courtship period tends to zero), we get from (2.3)

$$R_0 = \frac{p\rho(\mu_1 + \sigma)(\mu_0 + \mu_1 + \sigma + \beta)}{\mu_1(\mu_0 + \mu_1 + \sigma + p\beta)(2\mu_1 + \rho + \sigma) + \mu_1(1-p)\rho(\mu_1 - \mu_0)} \qquad (3.2.4)$$

which is exactly expression (12) from Dietz (1989) (with appropriate renaming of parameters).

$\diamondsuit$

Remark 2. If we choose $\alpha = \beta$ and $\sigma_c = \sigma$ and we 'lump' the types $(i,0)_c$ and $(i,0)$, for each $i \in \{1,...,n\}$, we are in a situation similar to the one described in Dietz (1989) but with the difference that a pair does not necessarily start with a sexual contact. Let us consider the case $n = 1$. Then G is a 3×3-matrix given by

$$G = \begin{pmatrix} -\mu_1 - \rho & \mu_0 + \sigma & \mu_1 + \sigma \\ \rho & -\mu_0 - \mu_1 - \sigma - p\beta & 0 \\ 0 & p\beta & -2\mu_1 - \sigma \end{pmatrix},$$

and $R_0 = -p\beta(G^{-1})_{2,3}$, or explicitly

$$R_0 = \frac{p\beta\rho(\mu_1 + \sigma)}{\mu_1(\mu_0 + \mu_1 + \sigma + \rho + p\beta)(2\mu_1 + \sigma) + \rho p\beta\mu_1}. \tag{3.2.5}$$

$\diamondsuit$

3.3. Incorporating heterogeneity in susceptibility

If we want to incorporate heterogeneity among the individuals in the population we have to specify not only the characteristics (called h-state) of an individual itself but also those of its current partner (if the individual is not single) because the h-state of the partner can influence the death-rate and in this way the probability that the original individual becomes single. The characteristics can take discrete or continuous values and the h-state of an individual can be constant in time or dynamic. Among the most important characteristics to be incorporated in the context of sexually transmitted diseases are age, sexual activity level, gender, homo-/bi-/heterosexuality, and behavioural traits such as condom use.

Let a variable ξ represent the heterogeneity characteristics of an individual; ξ is assumed to take values in some set Ω. The type of a sexually active individual in a pair is now represented by

$$(i, j; \xi_i, \xi_j), \quad i \in \{1,...,n\}, \quad j \in \{0,...,n\}, \quad \xi_i, \xi_j \in \Omega,$$

while $(i, -1; \xi_i)$ denotes a single infected individual, and $(i, 0; \xi_i, \xi_0)_c$ describes the relevant types of infected individuals in the courtship phase of a partnership. Suppose we have an individual, say x, that was infected by an individual with disease-state j and h-state ν. Suppose x itself had h-state η at the moment of infection. Then x was 'born' with type $(1, j; \eta, \nu)$. As time progresses, assuming

that x stays alive, x will become separated from its original partner, the h-state of x will change to, say, θ, and x will form a pair with a susceptible with h-state, say, ξ. Analogously to the case without heterogeneity in section 3.2, we want to evaluate the expected number of type-transitions

$$(i,0;\theta,\xi)_c \longrightarrow (i,1;\theta,\xi)$$
$$(i,0;\theta,\xi) \longrightarrow (i,1;\theta,\xi),$$

i.e. partner infections, of our individual x during its entire remaining life. In analogy with the notation of section 3.2 we write $m_{ij}(\xi,\theta;\eta,\nu)$ for this number. Let the current infectivity of x towards its partner be described by the probability $p_i(\xi,\theta)$, the rate of entering the sexually active phase by $\alpha(\xi,\theta)$, and let $\beta(\xi,\theta)$ be the number of sexual contacts per unit of time within the sexually active phase of a partnership. Generalising the expressions from section 3.2 we find

$$m_{ij}(\xi,\theta;\eta,\nu) = p_i(\xi,\theta)\beta(\xi,\theta)\int_0^\infty P_{L(i,0)L(1,j)}(\tau,\xi,\theta;\eta,\nu)d\tau$$
$$+ p_i(\xi,\theta)\alpha(\xi,\theta)\int_0^\infty P_{L(i,0)_cL(1,j)}(\tau,\xi,\theta;\eta,\nu)d\tau,$$

where $P_{L(i,0)L(1,j)}(\tau,\xi,\theta;\eta,\nu)$ denotes the probability that at time τ after x became infected as type $(1,j;\eta,\nu)$ it is still alive and has currently type $(i,0;\theta,\xi)$. The analogous quantity with index c has a similar interpretation.

Let $\phi = \phi(j;\eta,\nu)$ be the distribution of just infected individuals over the space $\{1,...,n\}\times\Omega\times\Omega$. We call this a generation. The next generation consists of all individuals that are infected by the members of this generation ϕ. When one individual 'born' with type $(j;\eta,\nu)$ infects $m_{ij}(\xi,\theta;\eta,\nu)$ partners of h-state ξ while having disease-state i and h-state θ, we obtain, by summing with respect to j,η and ν, for the next generation

$$(K\phi)(i;\xi,\theta) = \sum_{j=1}^{n}\int_{\Omega\times\Omega} m_{ij}(\xi,\theta;\eta,\nu)\phi(j;\eta,\nu)d\eta d\nu \qquad (3.3.1)$$

cases that are 'born' with type $(i;\xi,\theta)$. The formula (3.3.1) defines the next-generation operator K which tells us both how many secondary cases arise and how they are distributed with respect to type at 'birth'.

We regard the next-generation operator K as an operator mapping the space $L_1(\{1,...,n\}\times\Omega\times\Omega)$ into itself. As shown in chapter 1, R_0 is the spectral radius of the operator K.

Remark 3. It could prove to be no more than an academic exercise to work at this level of generality because it will be rather involved to determine analytically the probabilities P in the case of a dynamic continuous h-state like for example *age*. Instead of solving a coupled set of ODE's, which is basically

what happens in section 3.2, one has to solve a coupled system of PDE's. See Knolle (1990) for a different approach in the case of age as h-state, however with much more restrictive assumptions.

$\diamondsuit$

Example 3.3. We discuss the simplest example: let $\Omega = \{1, 2\}$ where '1' represents females, '2' represents males, and take only heterosexual contacts into account. The next-generation operator K is in this case a $2n \times 2n$ matrix of the following form

$$K = \begin{pmatrix} 0 & K_1 \\ K_2 & 0 \end{pmatrix},$$

where $K_1 = (m_{ij}(1, 2; 2, 1))_{1 \leq i,j \leq n}$ and $K_2 = (m_{ij}(2, 1; 1, 2))_{1 \leq i,j \leq n}$. For the spectral radius $r(A)$ of an operator A we have that $r(A^k) = r(A)^k, k \geq 1$; furthermore, if B is a second operator, then $r(AB) = r(BA)$. Since

$$K^2 = \begin{pmatrix} K_1 K_2 & 0 \\ 0 & K_2 K_1 \end{pmatrix}$$

we find

$$R_0 = \sqrt{r(K_1 K_2)}.$$

Note that, because we only look at heterosexual contacts, the probabilities $P_{L(i,0)L(1,j)}(\tau)$ and $P_{L(i,0)_c L(1,j)}(\tau)$ can be calculated in a way that is completely analogous to the example in section 3.2, with the only difference that death-rates, the infectivities in each disease-state, and the rates of change in disease-state are allowed to depend on the h-state of the individuals in the pair. Let us treat the case of $n = 1$ in somewhat more detail. Let the male parameter-set be given by $\{\mu_1, \mu_0, p, \beta, \sigma, \rho, \alpha\}$ and the female set by $\{\mu_1', \mu_0', p', \beta, \sigma', \rho', \alpha\}$, where p is the probability that a male infects a female. Note that there will be consistency requirements involving the pair formation parameters.

The transition matrix G is given by

$$G = \begin{pmatrix} G^{12} & 0 \\ 0 & G^{21} \end{pmatrix},$$

where G^{12} (G^{21}), which describes how the types of a female (male) individual change, is essentially the matrix from example 3.2 with appropriate placing of accents. We find

$$K_1 = m_{11}(1, 2; 2, 1) = -p\beta(G^{21})^{-1}_{3,4} - p\alpha(G^{21})^{-1}_{2,4}$$
$$K_2 = m_{11}(2, 1; 1, 2) = -p'\beta(G^{12})^{-1}_{3,4} - p'\alpha(G^{12})^{-1}_{2,4}$$

and $R_0 = \sqrt{K_1 K_2}$. In section 3.5 we discuss the case $n = 4$ in detail.

$\diamondsuit$

For arbitrary heterogeneity one can derive an explicit formula for R_0 in the very special case that the next-generation operator K has a one-dimensional range. If we assume

$$m_{ij}(\xi, \theta; \eta, \nu) = a_i(\xi, \theta) b_j(\eta, \nu) \tag{3.3.3}$$

then the only eigenvalue of K is

$$R_0 = \sum_{j=1}^{n} \int_{\Omega \times \Omega} b_j(\eta, \nu) a_j(\eta, \nu) d\eta d\nu.$$

In chapter 1 assumption (3.3.3) is called separable infectivity and susceptibility, or separable mixing. If the functions a and b are equal up to a multiplicative constant, then the assumption is known as proportionate mixing, see chapter 1. Somewhat less restrictive than separable mixing is the assumption

$$m_{ij}(\xi, \theta; \eta, \nu) = a_i(\xi, \theta) b_{ij}(\eta, \nu) \tag{3.3.4}$$

which leads to a next-generation operator with n-dimensional range, where n is the number of disease-states. We then have that the eigenvalues of K are equal to those of an $n \times n$ matrix $E = (e_{ij})$ with

$$e_{ij} = \int_{\Omega \times \Omega} b_{ij}(\eta, \nu) a_j(\eta, \nu) d\eta d\nu.$$

R_0 is then the dominant eigenvalue of E. Assumption (3.3.4) is called local separable infectivity and susceptibility, or local separable mixing, in chapter 1.

3.4. Various limit procedures

In this short section we show how various limit procedures can lead to interesting expressions for R_0. We restrict ourselves to the case of one disease-state. The R_0 for this case is explicitly given in equation (3.2.3) in example 3.2. First we 'collapse' the sexually active period of a partnership to a point event. We write $\beta = k\sigma + O(1)$ and let $\sigma \to \infty$. Then $1 + k$ is the average number of sexual contacts during one partnership with a sexually active phase and (3.2.3) simplifies to

$$R_0 = \frac{\rho \alpha p (1 + k)}{\mu_1 (1 + pk)(\mu_0 + \mu_1 + \rho + \sigma_c + \alpha)}. \tag{3.4.1}$$

The interpretation of (3.4.1) is as follows: α is the rate of becoming sexually active, given that one is in the courtship phase; $\rho/(\mu_1(\mu_0 + \mu_1 + \rho + \sigma_c + \alpha))$ is

the expected time that an infected individual will have a susceptible partner, i.e. the expected time spent in courtship (the product of this term with α gives the expected number of sexual partners); $p(1+k)/(1+pk)$ is the success-ratio per sexual partner (this identical to the formula on page 405 of Dietz (1988)). Note that (3.4.1) also covers the case where β remains bounded (simply put $k=0$, this means that there are only 'first contacts').

To completely eliminate pair formation from our model we still have to let $\alpha \to \infty$ in (3.4.1), or, in other words, we have to let the length of the courtship period tends to zero. We then find

$$R_0 = \frac{\rho p(1+k)}{\mu_1(1+pk)}. \tag{3.4.2}$$

This gives $R_0 = p\rho/\mu_1$ for $k=0$, which can be found immediately from the appropriate non pair formation model by looking at the interpretation of the parameters.

If, instead of α, we let $\rho \to \infty$ in (3.4.1), we are in the situation where the individual is constantly in the courtship phase,

$$R_0 = \frac{\alpha p(1+k)}{\mu_1(1+pk)}.$$

.

Remark 4. In the case without pair formation the formal route to R_0 would be to specify the infectivity A as a function of disease-age τ and calculate, see chapter 0, $R_0 = \int_0^\infty A(\tau)d\tau$. Under our assumptions, listed in section 3.2, $A(\tau)$ has a special form $A(\tau) = p\rho e^{-\mu_1\tau}$, and this leads once more to (3.4.2) with $k=0$. In the case of n possible disease-states, $A(\tau)$ involves an *expectation* for an infected individual to have a certain disease-state. The limit procedure to eliminate the pair formation completely is essentially the same as above: let $\sigma \to \infty$, $\alpha \to \infty$ and take $k=0$.

In the heterogeneous case we conjecture that a similar limit argument collapses the spectral radius of the next-generation operator (3.3.1) into the spectral radius of the operator

$$(K\phi)(\xi) = \int_\Omega \int_0^\infty A(\tau,\xi,\eta)d\tau\phi(\eta)d\eta$$

from chapter 1, but with a special form for the infectivity kernel A. As an illustration we look at the case where we recognise male and female individuals and allow only heterosexual contacts (example 3.3). In the situation without pair formation we have, May and Anderson (1988),

$$R_0 = \sqrt{\frac{pp'\rho\rho'}{\mu_1\mu_1'}}. \tag{3.4.3}$$

R_0 for the pair formation case, from example 3.3, transforms into (3.4.3) if we let σ, σ' and α tend to infinity.

$\Diamond$

3.5. Application to HIV

In this section and the next one, for the largest part based on Dietz, Heesterbeek and Tudor (1992), we apply our method of calculation to study the effects of variable infectivity on the spread of HIV. In addition we obtain some preliminary results on the effects of behaviour changes by the individuals.

From the many types of behaviour change that could be important, we have chosen two. One is the use of condoms, or some other measure of protection, during intercourse, and the other is the possibility for an individual to become permanently sexually inactive after a partnership with a partner suffering from AIDS breaks up. At that time these individuals will become aware that they are, with high probability, themselves infected, and could opt not to have sexual relations again. We could, in our model formulation, easily incorporate both a permanent and a temporary inactive period, which would be more realistic still. For example, a temporary inactive period could reflect a period of mourning after a partnership has come to an end through the death of the partner. However, in order not to complicate matters here, we concentrate on permanent sexual inactivation.

We assume that for the first sexual contact with a new partner there is a certain probability of adequate protection. This quantity could, for example, be the product of the probability of condom use with the probability that a condom actually protects. For all further sexual contacts between the members of a pair, we assume a constant, but possibly different, probability of adequate protection per contact. We can then study the effect of the advice to use condoms during each first sexual contact with a new partner.

We start by indicating which modifications, and concretisations of the model assumptions in section 3.2 were carried out for the present section.

Originally every pair initially passed through a courtship period characterised by the absence of sexual contacts. A courtship period can come to an end if the partners separate before a sexual contact has occured, or, alternatively, after the first sexual contact. For the sake of simplicity, we neglect the coursthip period in the present chapter, i.e. we take it to be infinitely short. This implies that a pair is established by a sexual contact.

We take $n = 4$ in our application. The infection-states 1 to 4 are assumed to correspond to the possible phases in the development of HIV-infection: an initial burst of infectiousness, followed by a, possibly long, period of virtual non-infectiousness, then a new infectious period in a pre-AIDS phase, and finally the disease-phase of so-called full-blown AIDS. We do not include a latency period to simplify matters, and because we are, in the present paper, not interested in determining the influence of such a period on the spread of HIV (see Watts and May (1992) for a discussion of the influence of the latency period). (The model, however, can easily incorporate a latency period; one just adjusts the number of possible states.)

We extend our notation by taking the male/female dichotomy into account. We indicate the sex of an individual by $k \in \{1,2\}$, where '1' indicates 'male', and '2' indicates 'female'. We restrict our attention to heterosexual contacts. The type of an *infected* individual x is then written as

$$(i,j;k), \quad i \in \{1,..,4\},\ j \in \{-1,0,..,4\},\ k \in \{1,2\},$$

where k is the sex of x, i its infection-state, and j its partnership-state. The sex of x's partner is, of course, fixed by the sex of x. Singles are of type $(i,-1;k)$.

We list the assumptions that are modified:

1'. The infection states are passed through in natural order. In particular, a newly infected individual of sex k starts infected life with type $(1,j;k)$ for some $j \in \{1,...,4\}$.

3'. The infectivity of an individual of sex k in infection-state i is described by the probability $p_i(k)$ that an unprotected sexual contact with a susceptible of the opposite sex leads to transmission.

4'. μ_0 is the death-rate of susceptible males and females, μ_i is the death-rate of infected male or female individuals with infection-state i.

5'. Every single individual of sex k with infection state unequal to 4, has a constant probability per unit of time $\rho_(k)$ of acquiring a new (sexual) partner. Individuals in infection state 4 are assumed not to establish new sexual relationships. The divorce rate is σ. We assume that each time an infected individual in infection state i, $i \in \{1,2,3,4\}$, with a partner in state j, $j \in \{0,...,4\}$, becomes single, either by divorce or by the partner's death, there is a probability s_{ij} that this individual stays sexually active (i.e. a probability $1 - s_{ij}$ of becoming permanently sexually inactive).

6'. By definition a partnership starts with one sexual contact. The probability of adequate protection at the first contact is q_0.

7'. Following the initiating contact, there are β sexual contacts per unit of time during partnerships if non of the partners has infection-state 4. If one of the partners has infection-state 4 we assume that there are no sexual contacts. The probability of efficient protection per post-initial contact is q.

Note that there will be consistency requirements involving the pair formation rates for males and females. These rates must satisfy the condition $\rho(1)x(1) = \rho(2)x(2)$, where $x(k)$ denotes the number of susceptibles of sex k in the population. In other words, the ratio of the partner acquisition rates equals the reciprocal of the sex ratio in the susceptible population.

Let $\Lambda := \{(k;i,j) | k \in \{1,2\},\ 1 \leq i \leq 4,\ -1 \leq j \leq 4\}$ be the set of all possible types. Then $|\Lambda| = 48$ and consequently our type-space is $\mathbb{R}^{48}$. Let $\Sigma := \{1,2,...,48\}$. Let a matrix $G : \mathbb{R}^{48} \to \mathbb{R}^{48}$ describe the transition probabilities per unit of time between the states, i.e. g_{rs} gives the rate of leaving state $s \in \Sigma$ to go to state $r \in \Sigma$. On the basis of the assumptions

stated above, G is given by

$$G = \begin{pmatrix} G(1) & 0 \\ 0 & G(2) \end{pmatrix},$$

where $G(k) : \mathbb{R}^{24} \to \mathbb{R}^{24}$ describes how the type of an infected individual of sex k changes; '0' denotes the 24×24 zero-matrix (the sex of an individual is usually fixed).

From the assumptions it follows that there are only six types with which an infected individual can be 'born' (from the point of view of the infection): $(1;1,j)$ for males, and $(2;1,j)$ for females, $1 \leq j \leq 3$. To calculate R_0, for each of the possible types at birth, we follow the remainder of the life of the corresponding individuals and count how many individuals of the different birth-types it produces on average. R_0 is then given by the dominant eigenvalue of the matrix $M : \mathbb{R}^6 \to \mathbb{R}^6$

$$\begin{pmatrix} 0 & M_2 \\ M_1 & 0 \end{pmatrix},$$

where the $M_k = (m_{ij}(k))_{1 \leq i,j \leq 3}$, $k \in \{1,2\}$, describe the average number of new cases caused by the different male (female) types amongst the different female (male) types, respectively. For these we have expressions

$$\begin{aligned} m_{ij}(1) = &-(1-q)p_i(1)\beta(G(1)^{-1})_{L(1;i,0)L(1;1,j)} \\ &- (1-q_0)p_i(1)\rho(1)(G(1)^{-1})_{L(1;i,-1)L(1;1,j)} \end{aligned}$$

and

$$\begin{aligned} m_{ij}(2) = &-(1-q)p_i(2)\beta(G(2)^{-1})_{L(2;i,0)L(2;1,j)} \\ &- (1-q_0)p_i(2)\rho(2)(G(2)^{-1})_{L(2;i,-1)L(2;1,j)} \end{aligned}$$

for $1 \leq i,j \leq 3$. These formulas have the following interpretation: the second term of $m_{ij}(1)$ describes the expected number of first contacts (pair establishing), where the infection is successfully transmitted by an infected male, 'born' with type $(1;1,j)$, while it is of type $(1;i,-1)$. The first term of $m_{ij}(1)$ describes the average number of females that become infected while being a partner of an infected male with infection-state i, that was 'born' with type $(1;1,j)$.

Next let us describe $G(k)$, $k \in \{1,2\}$,

$$G(k) = \begin{pmatrix} A_1 & 0 & 0 & 0 \\ D_1 & A_2 & 0 & 0 \\ 0 & D_2 & A_3 & 0 \\ 0 & 0 & D_3 & A_4 \end{pmatrix}$$

where '0' is the 6×6 zero-matrix, $D_j = diag(\theta_j)$, and A_i, $i \in \{1,2,3,4\}$ is given by

$$\begin{pmatrix} a_1(i) & b_0 & b_1 & b_2 & b_3 & b_4 \\ (1-(1-q_0)p_i(k))\rho(k) & a_2(i) & 0 & 0 & 0 & 0 \\ (1-q_0)p_i(k)\rho(k) & p_i(k)(1-q)\beta & a_3(i) & 0 & 0 & 0 \\ 0 & 0 & \theta_1 & a_4(i) & 0 & 0 \\ 0 & 0 & 0 & \theta_2 & a_5(i) & 0 \\ 0 & 0 & 0 & 0 & \theta_3 & a_6(i) \end{pmatrix}$$

where $a_1(i) = -\mu_i - \theta_i - \rho(k)$; $a_2(i) = -\mu_i - \mu_0 - \theta_i - \sigma - p_i(k)(1-q)\beta$; $a_3(i) = -\mu_i - \mu_1 - \theta_i - \theta_1 - \sigma$; $a_4(i) = -\mu_i - \mu_2 - \theta_i - \theta_2 - \sigma$; $a_5(i) = -\mu_i - \mu_3 - \theta_i - \theta_3 - \sigma$; $a_6(i) = -\mu_i - \mu_4 - \theta_i - \sigma$, and where $b_k = s_{ik}(\mu_k + \sigma)$, $k \in \{0, \ldots, 4\}$. For A_4 we have to read zero's for β and $\rho(k)$.

As already stated, $G(k)^{-1}$ can be expressed in the 6×6 matrices that constitute $G(k)$. $G(k)^{-1}$ is given by

$$\begin{pmatrix} {A_1}^{-1} & 0 & 0 & 0 \\ -D_1{A_2}^{-1}{A_1}^{-1} & {A_2}^{-1} & 0 & 0 \\ D_1D_2{A_3}^{-1}{A_2}^{-1}{A_1}^{-1} & -D_2{A_3}^{-1}{A_2}^{-1} & {A_3}^{-1} & 0 \\ B_{41} & D_2D_3{A_4}^{-1}{A_3}^{-1}{A_2}^{-1} & -D_3{A_4}^{-1}{A_3}^{-1} & {A_4}^{-1} \end{pmatrix},$$

with $B_{41} = -D_1D_2D_3{A_4}^{-1}{A_3}^{-1}{A_2}^{-1}{A_1}^{-1}$. The matrix M can now be determined. We consider the special case $p_2(k) = 0$, for $k \in \{1,2\}$. We find that

$$M_k = \begin{pmatrix} m_{11}(k) & 0 & m_{13}(k) \\ 0 & 0 & 0 \\ m_{31}(k) & 0 & m_{33}(k) \end{pmatrix}.$$

In terms of the elements of $G(k)^{-1}$ we can write

$$\begin{aligned} m_{11}(k) &= -(1-q)p_1(k)\beta(G(k)^{-1})_{2,3} - (1-q_0)p_1(k)\rho(k)(G(k)^{-1})_{1,3} \\ m_{13}(k) &= -(1-q)p_1(k)\beta\{G(k)^{-1})_{2,5} - (1-q_0)p_1(k)\rho(k)(G(k)^{-1})_{1,5} \\ m_{31}(k) &= -(1-q)p_3(k)\beta(G(k)^{-1})_{14,3} - (1-q_0)p_3(k)\rho(k)(G(k)^{-1})_{13,3} \\ m_{33}(k) &= -(1-q)p_3(k)\beta(G(k)^{-1})_{14,5} - (1-q_0)p_3(k)\rho(k)(G(k)^{-1})_{13,5}. \end{aligned}$$

Then the spectral radius $r(M) = \sqrt{r(M_1M_2)} = \sqrt{r(M'_1M'_2)}$, where

$$M'_k = \begin{pmatrix} m_{11}(k) & m_{13}(k) \\ m_{31}(k) & m_{33}(k) \end{pmatrix}, \quad k \in \{1,2\},$$

see example 3.3.

Finally we find that

$$R_0 = \frac{(a+d) + \sqrt{(a-d)^2 + 4cb}}{2},$$

with

$$\begin{aligned} a &:= m_{11}(1)m_{11}(2) + m_{13}(1)m_{31}(2) \\ b &:= m_{11}(1)m_{13}(2) + m_{13}(1)m_{33}(2) \\ c &:= m_{31}(1)m_{11}(2) + m_{33}(1)m_{31}(2) \\ d &:= m_{31}(1)m_{13}(2) + m_{33}(1)m_{33}(2). \end{aligned}$$

Example 3.4 (Dynamic protection strategy). As an interlude we show how to calculate R_0 if we allow for switching between protected and unprotected contacts by the partners of a pair. In order not to complicate the notation we regard the easiest case only, without adding the further complications of inactivation and the female/male dichotomy. Furthermore we give the calculation only in the case of a single infection-state. The 'full' model, with four infection-states/male-female-distinction/inactivation, is an easy generalisation, along the lines laid out in this section.

We retain the relevant assumptions made above. If the members of a pair use condoms, or any other means of protection against disease transmission, we indicate so by attaching a subscript 'p' to the type: $(1,0)_p$. For pairs where both members are infected it does not matter whether they use protection or not. For the R_0-calculation there is no reason to distinguish between protected and unprotected pairs of infecteds. So, as far as our calculations are concerned, infected individuals with an infected partner are denoted by the type $(1,1)$, irrespective of the use of protection. The upshot of all this is that there is only one type at birth, $(1,1)$, and we can write down R_0 explicitly.

All pairs start with a protected phase. With probability per unit of time γ, the protected phase is left. With probability per unit of time γ^* the couple can have second thougths and re-enter the protected phase.

We order the different types lexicographically with the additional requirement that a protected type directly precedes the corresponding unprotected type.

The matrix G, containing the transition probabilities per unit of time is given by

$$G = \begin{pmatrix} a_1 & \mu_0 + \sigma & \mu_0 + \sigma & \mu_1 + \sigma \\ (1-(1-q_0)p_1)\rho & a_2 & \gamma^* & 0 \\ 0 & \gamma & a_3 & 0 \\ (1-q_0)p_1\rho & (1-q)p_1\beta & p_1\beta & a_4 \end{pmatrix}$$

where $a_1 = -\mu_1 - \rho$; $a_2 = -\mu_1 - \mu_0 - \sigma - (1-q)p_1\beta - \gamma$; $a_3 = -\mu_1 - \mu_0 - p_1\beta - \gamma^* - \sigma$; $a_4 = -2\mu_1 - \sigma$. From previous considerations it then follows that R_0 is given by

$$R_0 = -(1-q_0)p_1\rho(G^{-1})_{1,4} - (1-q)p_1\beta(G^{-1})_{2,4} - p_1\beta(G^{-1})_{3,4}.$$

$\Diamond$

3.6. Results for the model of section 3.5

We define N to be the average number of new sexual partners that a newly infected individual will have in the remainder of its life, and C to be the average number of sexual contacts that a newly infected individual will have in the remainder of its life.

Let r be the right-eigenvector of matrix M corresponding to the dominant eigenvalue R_0: $Mr = R_0 r$. Then r gives the distribution of infected individuals over the six possible birth types $(k;1,j)$, $j \in \{1,2,3\}$, $k \in \{1,2\}$ (which we number 1 to 6). We normalise r to $\bar{r} = (\frac{r_1}{|r|}, .., \frac{r_6}{|r|})^T$ then $\bar{r}_\ell$ gives the probability that an individual will be 'born' with birth type $\ell \in \{1, .., 6\}$.

Denote by N_ℓ the average number of new sexual partners that a just infected individual born with birth type $\ell \in \{1, ..6\}$ will have during the rest of its life. Then the correct way to weigh the six numbers N_ℓ to determine N is to let

$$N = \sum_{\ell=1}^{6} N_\ell \bar{r}_\ell .$$

The numbers N_ℓ are given by

$$N_{(k;1,j)} = -\sum_{i=1}^{3} \rho(k)(G(k)^{-1})_{L(k;i,-1)L(k;1,j)}.$$

To calculate C let C_ℓ denote the average total number of sexual contacts the individual will have after becoming infected as birth type $\ell \in \{1, .., 6\}$. Then C_ℓ is given by

$$C_\ell = N_\ell + \beta T_\ell$$

being the sum of the number of 'first contacts' and the average future time T_ℓ a newly infected individual with birth type ℓ will spend with a partner multiplied by the frequency β of sexual contacts within a partnership. T_ℓ can be written as

$$T_{(k;1,j)} = -\sum_{i=1}^{3} \Big\{ (G(k)^{-1})_{L(k;i,0)L(k;1,j)} + (G(k)^{-1})_{L(k;i,1)L(k;1,j)} + (G(k)^{-1})_{L(k;i,2)L(k;1,j)} + (G(k)^{-1})_{L(k;i,3)L(k;1,j)} \Big\}.$$

The quantity C is then given by the sum of the six C_ℓ's weighted according to birth type

$$C = \sum_{\ell=1}^{6} C_\ell \bar{r}_\ell .$$

We want to investigate R_0 as a function of N. Of particular interest is the threshold value N for which $R_0 = 1$. If we want to assure uniqueness of this threshold value, to make the results comparable, we are not free to choose any combination of the pair formation parameters ρ and σ. If σ is allowed to increase while all other parameters are held constant, then the graph $R_0 = f(N)$ need not be monotonically increasing, and therefore the threshold value may not be unique. The reason for this lack of monotonicity is that for increasing separation rate σ, other parameters remaining constant, the infected individual spends less and less time in partnerships. Hence, the average total number of contacts will decrease, which in turn results in a smaller value of R_0. Therefore, to assure monotonicity (and uniqueness of the threshold value), we keep the total number of contacts after infection C constant when we vary N.

It is worthwile to spend some lines explaining how the graphs of R_0 as a function of N are made. In the numerical calculation below we only regard the case of a sex ratio of one in the susceptible population. In that case, $\rho(1) = \rho(2) = \rho$ because of the consistency requirement mentioned before. We want to plot $R_0 = f(N)$, for $N \in [1, 20]$ say, with $C = \bar{c}$ a pre-chosen constant. Both N and C are functions of σ and ρ, $N = g_1(\sigma, \rho)$ and $C = g_2(\sigma, \rho)$ say. If $\rho = \bar{\rho}$ has been chosen, then we can determine $\sigma = \bar{\sigma}$ such that $g_2(\bar{\sigma}, \bar{\rho}) = \bar{c}$ (formally write $\bar{\sigma} = h(\bar{c}, \bar{\rho})$. Now calculate $R_0 = f(g_1(h(\bar{c}, \bar{\rho}), \bar{\rho}))$, plot the result against $g_1(\bar{\sigma}, \bar{\rho})$ and choose a new $\bar{\rho}$.

We discuss here only two sets of graphs derived for the model as an illustration. The numerical calculations were performed by David Tudor. In both figures 3.2 and 3.3, we chose $\bar{c} = 500$. For $\rho \in [0.75, 5.25]$ per year, the corresponding σ's then ranged over $[0.18, 4.35]$ per year. The rest of the parameters (all 'per year') were chosen as follows: $\theta_1 = 4.0, \theta_2 = \theta_3 = 0.2, \theta_4 = 0$; $\mu_i = 0.02 (i \neq 4), \mu_4 = 0.5$; $\beta = 100$. We assume that $s_{ij} = 1$ for $i \in \{1, 2, 3\}$, $j \in \{0, 1, 2, 3\}$ because in our model we disregard testing for sero-positivity and therefore infected individuals that do not have AIDS, and whose current partner also does not have AIDS, have no way of knowing that they are infected. Things are different if the current partner of our index case (the infected individual we follow) has AIDS. Then, if this partnership breaks up, our index case will know that it has a high probability of being infected. So, we choose $s_{i4} =: s$, $i \in \{1, 2, 3\}$ different from one. The remaining s_{4j}'s, $j \in \{0, 1, 2, 3\}$, do not require special treatment. In these cases the index case itself has AIDS, and because of our assumptions this individual will then not have any sexual contacts or new sexual partners for the remaining time of its infectious period. The behaviour of these individuals, and the value of s_{4j}, will not influence R_0, nor does it influence the calculation of N for the population. It would perhaps be natural to choose $s_{4j} = 0$, but in our calculations below we have set $s4j = 1$. One could argue that the value of s_{ij} can depend on the reason (separation or death of the partner) that a partnership of the index case breaks up. However, as all these influences are unknown there is really no point to such generality.

In a first series of four graphs, we evaluate the effect of the dependence

of the infection probabilities per sexual contact on time and sex. We take $q_0 = q = s = 0$ for this set of graphs. We regard four situations (the numbers correspond to the graphs in figure 3.2):

(1) time dependent and sex dependent p_i's.
(2) time dependent and sex independent p_i's.
(3) time independent and sex dependent p_i's.
(4) time independent and sex independent p_i's.

We have to be careful in gauging the four situations if we want to compare the results. We calibrate situations 2 through 4 using our choice for the most general situation 1.

For situation 1 we chose $p_1(1) = 0.05, p_2(1) = 0.001, p_3(1) = 0.01$ for the 'male to female' probabilities and $p_1(2) = 0.025, p_2(2) = 0.0005, p_3(2) = 0.005$ for the 'female to male' probabilities ($p_4(k) = 0, k = 1, 2$).

For situations 3 and 4 we first calculate the expected duration of infectivity D

$$D = \frac{1}{\mu_1 + \theta_1} + \frac{\theta_1}{(\mu_1 + \theta_1)(\mu_2 + \theta_2)} + \frac{\theta_1\theta_2}{(\mu_1 + \theta_1)(\mu_2 + \theta_2)(\mu_3 + \theta_3)}.$$

Then, the mean infectivity $\bar{p}$ is

$$\bar{p} = \frac{1}{D}\Big(\frac{p_1}{\mu_1 + \theta_1} + \frac{p_2\theta_1}{(\mu_1 + \theta_1)(\mu_2 + \theta_2)} + \frac{p_3\theta_1\theta_2}{(\mu_1 + \theta_1)(\mu_2 + \theta_2)(\mu_3 + \theta_3)}\Big). \quad (3.6.1)$$

We can use (3.6.1), with appropriate placing of k's, to calculate, from the original $p_i(k)$'s, a $\bar{p}(1)$ for males and a $\bar{p}(2)$ for females. This leads to $\bar{p}(1) = 0.00654$ and $\bar{p}(2) = 0.00327$. We use these in situation 3. For situation 4 we take the geometric mean $\sqrt{\bar{p}(1)\bar{p}(2)} = 0.00462$.

Finally in situation 2, we take for $i = 1, 2$ the geometric mean over the sexes, $p_i = \sqrt{p_i(1)p_i(2)}$. This leads to $p_1 = 0.0354, p_2 = 0.0007$. Now to gauge all situations we demand that p_1, p_2 and p_3 are such that $\bar{p}$ from (3.6.1) is equal to the geometric mean of $\bar{p}(1)$ and $\bar{p}(2)$. This requirement leads to $p_3 = 0.0071$.

We make two remarks about the graphs in figure 3.2. The ordering of the graphs is:

$$R_0(1) \leq R_0(2) \leq R_0(3) \leq R_0(4).$$

Remarks

(5) An interesting observation is that the sex-dependence does not appear to have a marked influence on R_0. One could therefore argue that the often made distinction between 'male to female' and 'female to male' infection probabilities, is irrelevant for questions involving the basic reproduction ratio.

(6) Time-dependence has a marked influence on R_0 and the use of a time independent infection probability would lead to an overestimation of R_0. The effect of time-dependence on R_0 is larger for higher values of N.

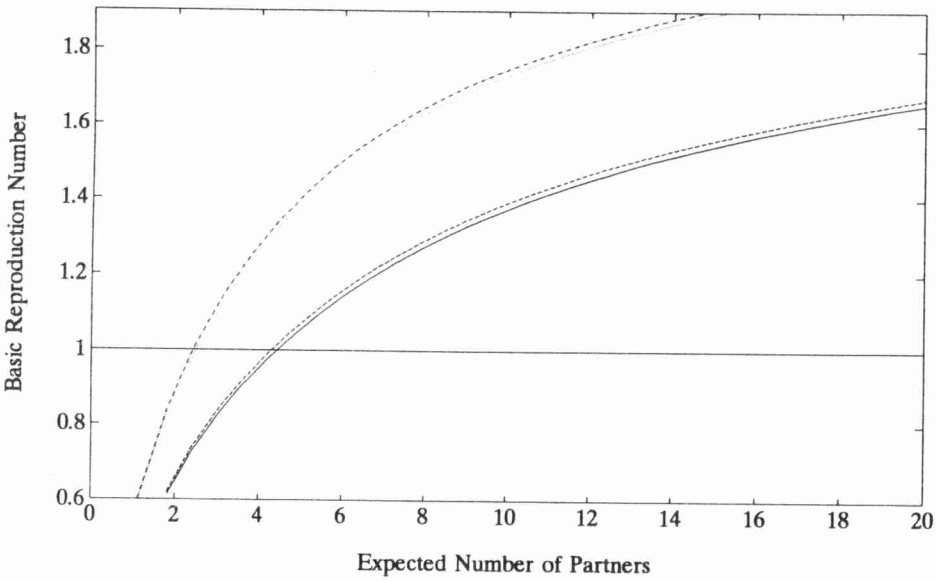

Figure 3.2 R_0 as a function of N, for four situations of variability in infectivity, see text for details

In any case, the numerical computations suggest that the introduction of variance in infectivity results in a decrease in the predicted R_0-value for a given value of N if the average infectivity is kept constant. Repeating the calculations with different distributions of the infectivity over the infection states, while keeping the (weighted) integral over the entire infectious period fixed, did not lead to qualitatively different results.

The threshold values N^* for the number of partners during the infectious period of $D = 8.88$ years, satisfy the inequalities:

$$N^*(4) < N^*(3) < N^*(2) < N^*(1)$$

where $N^*(4) = 2.42$; $N^*(3) = 2.49$; $N^*(2) = 4.28$; and $N^*(1) = 4.42$ (note that $N^*(1)$ is 83% larger than $N^*(4)$). If we let N tend to the maximum value which corresponds to the condition $C = 500$, in other words, if we let $N = 500$ while letting $\sigma \to \infty$, we obtain the same value for all four curves: $R_0 = 2.31$. The curves differ primarily for intermediate values of N. This is easy to explain. For large values of N they agree because the number of contacts per partner tends to one, so that the infection probability per partner equals the infection probability per contact and the average infection probabilities all agree. Here we note that with respect to sex differences one has to take the geometric average, and with respect to time differences the intergral over the infectious

period. For small values of N the curves agree because the number of contacts per partner is large, so that the infection probability per partner tends to one in all four cases.

For a heuristic argument that makes a decrease in the infection probability per partner in the case of increased variability in infectivity plausible, see Dietz, Heesterbeek and Metz (1992).

In a second series of four graphs we take a preliminary look at the effects of different values for the protection parameters q_0, q and s on R_0. For the infection probabilities we take the sex- and time dependent values from figure 2.3. In figure 3.3 the numbers 1 to 4 correspond to the following situations:

(1) $q_0 = 0, q = 0.10, s = 0$;
(2) $q_0 = 0.25, q = 0, s = 0$;
(3) $q_0 = 0, q = 0, s = 0$;
(4) $q_0 = 0, q = 0, s = 0.95$.

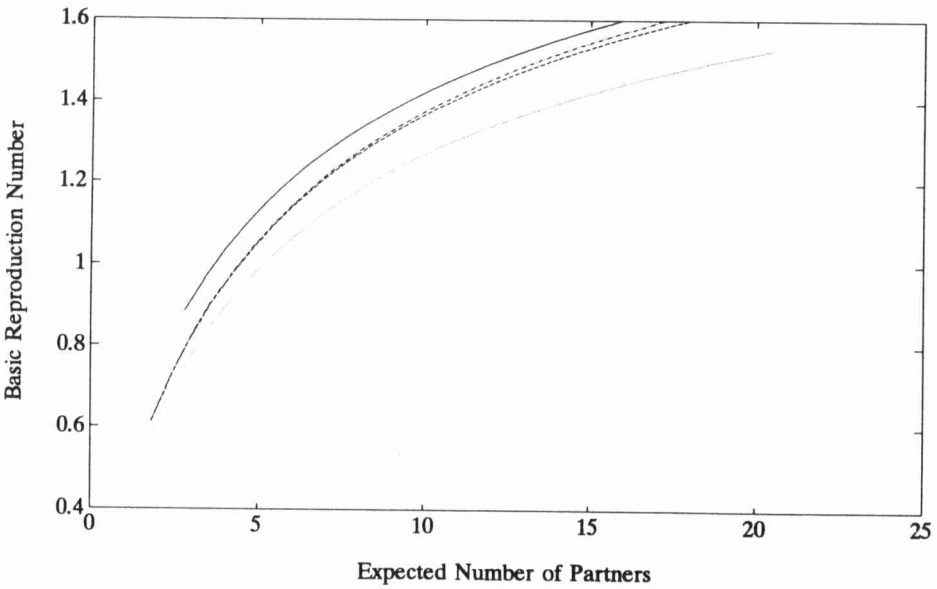

Figure 3.3 R_0 as a function of N, for four combinations of protection against infection, see text for details

We make three remarks about the graphs in figure 3.3. The ordering of the graphs is:

$$R_0(1) \leq R_0(2) \leq R_0(3) \leq R_0(4).$$

Remarks.

(7) The effect of using protection with probability $q = 0.1$ after the initiating contact, as opposed to no protection at all, is more pronounced at higher values of N.

(8) Variation in 'withdrawal' probability s does not seem to have a major influence on R_0. Furthermore, this influence decreases with increasing N.

(9) It does not make much difference if protection is used during the initiating contact. This influence increases as N becomes larger, but that is to expected since the number of 'first contacts' rises linearly with N. The degree of influence does not suggest that the, for some groups of individuals theoretically not unrealistic, protection policy that is based on advising couples to use protection at least for their first sexual contact, would be succesful. One would, however, aim such a policy at individuals who have a large number of 'one night stands' (pardon the expression), and therefore a large value of N. Therefore, the effect of q_0 has to be studied in a much wider N-range.

3.7. References

Blythe, S.P. & R.M. Anderson (1988): Variable infectiousness in HIV transmission models. *IMA J. Math. Appl. Med. Biol.* **5**: 181-200.

Dietz, K. (1988): On the transmission dynamics of HIV. *Math. Biosc.* **90**: 397-414.

Dietz, K. (1989): The role of pair formation in the transmission dynamics of HIV. In: *Sesquicentennial Invited Paper Sessions (M.H. Gail and N.L. Johnson, eds.)* American Statistical Association, Alexandria, Virginia, 609-621.

Dietz, K. & K.P. Hadeler (1988): Epidemiological models for sexually transmitted diseases. *J. Math. Biol.* **26**: 1-25.

Dietz, K., Heesterbeek, J.A.P. & D.W. Tudor (1992): The basic reproduction ratio for sexually transmitted diseases, Part 2: Effects of variable HIV infectivity. Subm. to *Math. Biosc.*

Kermack, W.O. & A.G. McKendrick (1927): Contributions to the mathematical theory of epidemics. *Proc. Roy. Soc. A* **115**: 700-721.

Knolle, H. (1990): Age preference in sexual choice and the basic reproduction number of HIV/AIDS. *Biom.J.* **32**: 243-256.

May, R.M. & R.M. Anderson (1988): The transmission dynamics of human immunodeficiency virus (HIV). *Phil. Trans. R. Soc. Lond. B* **321**: 565-607.

Watts, C.H. & R.M. May (1992): The influence of concurrent partnerships on the dynamics of HIV/AIDS. *Math. Biosc.* **108**: 89-104.

Chapter 4

The saturating contact rate

In this chapter we show how to derive, by a mechanistic argument, an expression for the saturating contact rate of individual contacts in a population that mixes randomly. The main assumption is that the individual interaction times are typically short as compared to the time-scale of changes in, for example, individual-type, but that the interactions yet make up a considerable fraction of the limited time-budget of an individual. In special cases an explicit formula for the contact rate is obtained. The result is applied to mathematical epidemiology and marriage models.

4.1. Introduction

A problem that has been around for a long time in mathematical epidemiology, is that of giving a mechanistic description of the saturation in the number of (re)new(ed) contacts that an individual can make per unit of time, given that the time that an individual has available for these contacts is limited. Of epidemiological importance is of course the number of contacts between infected and susceptible individuals, which determines the possible number of new infections per unit of time (see e.g. Anderson and May (1991)). The same problem occurs in marriage models, where one needs to model the number of 'steady relationships' that are established per unit of time. The idea here is that individuals have a number of short lasting contacts per unit of time within a limited time available, and that steady relationships ('marriages') may result from these brief encounters. Frequently a Holling-type argument, borrowed from predator-prey systems, is thought to be the solution to the problem. However, as we explain below, on closer examination the application of the usual Holling argument to epidemic- and marriage models cannot be justified. In this chapter we give a mechanistically based, answer to the 'contact-rate problem' in sections 4.3 (epidemic models) and 4.4 (marriage models).

Suppose we have a well-mixed closed population, divided into n different types of individuals (think e.g. of infected or non-infected males and females), of total density $N(t)$ at time t. Suppose furthermore that two individuals can together establish a temporary complex, and that these complexes are formed according to mass-action kinetics. The rate constants of complex formation and dissociation are allowed to depend on the types of the individuals involved. Our key assumption is that the complexes are of short duration, as compared to the time-scale on which the individuals change their type (where the latter also comprises the formation of more permanent relationships), or the various type-densities change by births or deaths.

Instead of 'saturating contact rate' one could also speak of 'the functional response' in the number of individual contacts, but this formulation is traditionally reserved for predator-prey systems. There we can distinguish, for example, two types (prey individuals, and predators) and one complex (predators that are busy handling prey). The Holling argument then gives an expression for the number of prey caught by a predator, taking into account that a proportion of the predators' available time is spent handling the already caught prey. A prerequisite for the Holling argument is that on the time-scale of predator type change (searching $\leftrightarrow$ busy with prey) the environment of the predators (in this case the density of the prey population) stays constant. This is not the case if interactions also occur between individuals of comparable type (think for example of predators fighting with predators). For those types of individuals the differential equations are no longer linear in the type density itself, but contain quadratic interaction terms. This then makes it impossible to study the problem by following one individual and describing the possible type-changes by a continuous time (semi)-Markov chain (which is one fruitful way of looking at the Holling-type problems). In the case of self interaction the time-scale on which the density of individuals of a particular type changes is the same as that on which their environment changes. Therefore the usual Holling-type arguments are not applicable to the description of social interactions such as marriages and contacts between infected and susceptible individuals where interactions between comparable types are important. Our argument in section 4.2 serves as a replacement. In section 4.5 we discuss a slightly more involved Holling argument that does work in our situation. However, we could only deduce this argument from the results of the mathematically correct calculations in the intervening sections.

4.2. Main result

Let $X_i(t)$ denote the density of free individuals of type i, $i \in \{1, ..., n\}$, at time t. In this paper we assume that complexes are formed between two individuals and we therefore disregard larger groups. Let $K_{ii}(t)$ denote the

density of complexes at time t involving two i-type individuals; $2K_{ij}(t)$ denotes the density of complexes at time t involving one i- and one j-type individual ($K_{ij} = K_{ji}$). There are then $\frac{1}{2}n(n+1)$ different complexes.

The rate constant for the formation of an (i,j)-complex is denoted by r_{ij}, the dissociation rate constant by s_{ij} (we assume $r_{ij} = r_{ji}$, and $s_{ij} = s_{ji}$). Type-changes of free individuals of type i are described by a function F_i, type-changes by individuals that are part of a complex ij are described by a function G_{ij} (describing, for example, infections or marriages). Birth and death of individuals are included in these functions as well. The F_i and G_{ij} will, in general, be functions of $(X_1, ..., X_n; K_{11}, ..., K_{nn})$ and we assume that they are Lipschitz continuous in all variables (in many applications they are actually linear). Our main assumption is that the time-scale of type-change is much longer than that of complex formation and dissociation.

We can write down differential equations for the changes in X_i and K_{ij}. For $1 \leq i, j \leq n$ we have

$$\frac{dX_i}{dt} = -X_i \sum_{j=1}^{n} r_{ij} X_j + 2 \sum_{j=1}^{n} s_{ij} K_{ij} + F_i, \tag{S1}$$

$$\frac{dK_{ij}}{dt} = \frac{1}{2} r_{ij} X_i X_j - s_{ij} K_{ij} + G_{ij} \tag{S2}$$

In order to correctly apply a time-scale argument, we rewrite (S) as a singular perturbation problem. If the processes of type-change occur with rate constants expressed per unit of t-time, then the processes of complex formation and dissociation occur, by assumption, with rate constants (r_{ij} and s_{ij}) expressed per unit of εt-time, for $\varepsilon \ll 1$. If we re-scale by writing $r_{ij} = \rho_{ij}/\varepsilon$ and $s_{ij} = \sigma_{ij}/\varepsilon$, then all processes involved in our system are on the same time-scale, and (S) turns into

$$\varepsilon \frac{dX_i}{dt} = -X_i \sum_{j=1}^{n} \rho_{ij} X_j + 2 \sum_{j=1}^{n} \sigma_{ij} K_{ij} + \varepsilon F_i, \tag{S_ε1}$$

$$\varepsilon \frac{dK_{ij}}{dt} = \frac{1}{2} \rho_{ij} X_i X_j - \sigma_{ij} K_{ij} + \varepsilon G_{ij}. \tag{S_ε2}$$

Define $\theta_{ij} := \frac{\rho_{ij}}{\sigma_{ij}} = \frac{r_{ij}}{s_{ij}}$ and, for $i \in \{1, ..., n\}$, let

$$\xi_i(t) := X_i(t) + 2 \sum_{j=1}^{n} K_{ij}(t),$$

be the total density of i-type individuals in the population at time t. Then $\sum_{i=1}^{n} \xi_i(t) = N(t)$.

Theorem 4.1 *For the solution $X_{\varepsilon i}(t)$, $K_{\varepsilon ij}(t)$ of system (S_ε), we have*

$$\lim_{\varepsilon \downarrow 0} X_{\varepsilon i}(t) = X_i^* \tag{4.1.1}$$

$$\lim_{\varepsilon \downarrow 0} K_{\varepsilon ij}(t) = K_{ij}^*, \tag{4.1.2}$$

where the convergence is uniform on bounded intervals bounded away from zero. Here $(X_1^*, .., X_n^*)$ *is the unique positive solution of the system*

$$X_i + X_i \sum_{j=1}^{n} \theta_{ij} X_j = \xi_i \tag{4.1.3}$$

$i \in \{1, ..., n\}$, *with* $\xi_i \geq 0$,

$$K_{ij}^* = \frac{1}{2} \theta_{ij} X_i^* X_j^* \tag{4.1.4}$$

and ξ_i *is the solution of*

$$\frac{d\xi_i}{dt} = F_i(X_1^*, .., X_n^*, K_{11}^*, .., K_{nn}^*) + 2 \sum_{j=1}^{n} G_{ij}(X_1^*, .., X_n^*, K_{11}^*, .., K_{nn}^*). \tag{4.1.5}$$

Proof: We use a standard singular perturbation argument (see, e.g., Tikhonov, Vasil'eva and Sveshnikov (1985)) for system (S_ε). We start by regarding system (S_ε) on the 'fast time-scale' by first substituting $\tau := t/\varepsilon$, and then taking the limit $\varepsilon \downarrow 0$. The substitution leads to

$$\frac{dX_i}{d\tau} = -X_i \sum_{j=1}^{n} \rho_{ij} X_j + 2 \sum_{j=1}^{n} \sigma_{ij} K_{ij} + \varepsilon F_i, \tag{4.1.6}$$

$$\frac{dK_{ij}}{d\tau} = \frac{1}{2} \rho_{ij} X_i X_j - \sigma_{ij} K_{ij} + \varepsilon G_{ij}, \tag{4.1.7}$$

and we obtain the following system of equations for the quasi-steady state

$$0 = -X_i \sum_{j=1}^{n} \rho_{ij} X_j + 2 \sum_{j=1}^{n} \sigma_{ij} K_{ij}, \tag{$S_0 1$}$$

$$0 = \frac{1}{2} \rho_{ij} X_i X_j - \sigma_{ij} K_{ij} \tag{$S_0 2$}$$

$1 \leq i, j \leq n$. Write $K_{ij} = \frac{1}{2} \theta_{ij} X_i X_j$, then both $(S_0 1)$ and $(S_0 2)$ are satisfied. On the fast time-scale we have an additional relation between K_{ij}, X_i and X_j. On that time scale N and the ξ_i's do not change:

$$\frac{d\xi_i}{d\tau} = \varepsilon F_i + 2\varepsilon \sum_{j=1}^{n} G_{ij}$$

and therefore we have $\frac{d\xi_i}{d\tau} = 0$, which implies that we have to solve (S_0) within the manifolds $\xi_i = \textit{constant}$. This leads to the conservation equations $X_i + 2 \sum_{j=1}^{n} K_{ij} = \xi_i = \textit{constant}$ or

$$X_i + X_i \sum_{j=1}^{n} \theta_{ij} X_j = \xi_i = \textit{constant} \tag{4.1.8}$$

$i \in \{1, ..., n\}$.

In the appendix we show that system (4.1.8) has a unique positive solution $(X_1^*, ..., X_n^*)$, and that $(X_1^*, ..., X_n^*, K_{11}^*, ..., K_{nn}^*)$, the corresponding positive solution of (S_0), is an asymptotically stable steady state of (4.1.6)-(4.1.7) at $\varepsilon = 0$.

Application of the singular perturbation theorem from Tikhonov, Vasil'eva and Sveshnikov (1985) now gives that, for $\varepsilon \downarrow 0$, the solution to system (S_ε) converges to the steady state $(X_1^*, ..X_n^*, K_{11}^*, .., K_{nn}^*)$ uniformly on intervals bounded away from zero and infinity. Remember that the X_i^*'s and the K_{ij}^*'s are functions of time. They change on the 'slow time-scale' because on that time-scale N and the ξ_i's change. We have that ξ_i, $(1 \leq i \leq n)$, is the solution of

$$\frac{d\xi_i}{dt} = F_i(X_1^*, .., X_n^*, K_{11}^*, .., K_{nn}^*) + 2\sum_{j=1}^{n} G_{ij}(X_1^*, .., X_n^*, K_{11}^*, .., K_{nn}^*), \quad (4.1.9)$$

which describes the changes in the density of i-type individuals on the slow time-scale.

$\diamondsuit$

In some special cases one can explicitly solve (4.1.3) in terms of the ξ_i.

Corollary 4.2 *Let $\theta_{ij} = \theta$ for all $1 \leq i, j \leq n$. Then*

$$X_i^* = \frac{-1 + \sqrt{1 + 4\theta N}}{2\theta N}\xi_i \quad (4.1.10)$$

$$K_{ij}^* = \frac{1 + 2\theta N - \sqrt{1 + 4\theta N}}{2\theta N^2}\xi_i\xi_j \quad (4.1.11)$$

where ξ_i is the solution of (4.1.5), $1 \leq i, j \leq n$.

Proof: Let $X := \sum_{i=1}^{n} X_i$, and $K := \sum_{ij} K_{ij}$. Then system (4.1.3) can be written as

$$X + \theta X^2 = N. \quad (4.1.12)$$

Its unique positive solution is

$$X^* = \frac{-1 + \sqrt{1 + 4\theta N}}{2\theta}.$$

From the relation $K^* = \frac{1}{2}\theta X^{*2}$ we then obtain

$$K^* = \frac{1 + 2\theta N - \sqrt{1 + 4\theta N}}{4\theta}.$$

We express K_{ij}^* in terms of K^*. If we substitute (4.1.4) in the definition of ξ_i, then we can write X_i^* as

$$X_i^* = \frac{\xi_i}{1 + \theta X^*} = \frac{\xi_i}{N}X^*$$

where, in the last equality, we have used equation (4.1.12). Then

$$K_{ij}^* = \frac{\xi_i \xi_j}{N^2} K^*.$$

◊

The following special case was suggested by sexual activity models in the AIDS-literature.

Corollary 4.3 *Let* $\theta_{ij} = \alpha_i \alpha_j$ *and define* $A := \sum_{j=1}^n \alpha_j X_j^*$. *Then*

$$X_i^* = \frac{\xi_i}{1 + \alpha_i A}$$
$$K_{ij}^* = \frac{\alpha_i \alpha_j}{(1 + \alpha_i A)(1 + \alpha_j A)} \xi_i \xi_j$$

where A *is the unique positive solution of*

$$A = \sum_{j=1}^{n} \frac{\alpha_i \xi_i}{1 + \alpha_i A} \tag{4.1.13}$$

and where ξ_i *is the solution of (4.1.5).*

Proof: From (4.1.3) we obtain $X_i + \alpha_i X_i A = \xi_i$ which gives

$$X_i^* = \frac{\xi_i}{1 + \alpha_i A}$$

from which the result follows by using (4.1.4). Substitution of X_i^* in the defining relation for A gives (4.1.13), which can easily be seen to have a unique positive solution.

◊

4.3. Application to mathematical epidemiology

In non-pair formation models for sexually transmitted diseases it has been argued (see e.g. Thiema and Castillo-Chavez (1989)) that the number of new cases of the infection arising per unit of time, should be written as

$$\beta C(N) S \frac{I}{N}$$

where β is the average probability of transmision of the infection between two individuals taking part in a meeting (during which they are allowed more than

one contact); $C(N)$ is the 'unknown' probability per unit of time for an infected individual to take part in a meeting (this is the fraction of the time that an infected individual is meeting with another individual); S and I are the densities of the non-infected and infected populations, respectively. Some reasonable demands on $C(N)$ are that it should be a non-decreasing function of N and that $C(N)/N$ should be a non-increasing function of N. Furthermore, the function should behave linearly in N, for small N, and it should be independent of N, for N large. Of course, many functional forms can, and have been, suggested that have these properties, but a mechanistically derived form was lacking.

Let X_1 and X_2 denote the densities of susceptible and infected *singles*, respectively. The infection process constitutes a change from $K_{12} \rightarrow K_{22}$, which we assume to happen with some probability per unit of time β. If we now write $S := X_1 + 2K_{11} + 2K_{12}$ and $I := N - S$ for ξ_1 and ξ_2, respectively, then the number of new cases of the disease appearing per unit of time is

$$\beta K_{12}^* = \beta \frac{1 + 2\theta N - \sqrt{1 + 4\theta N}}{2\theta N^2} SI,$$

if we regard the simplest case where the disease has no influence on the propensity to form complexes and the time spent in a complex. We find therefore, the following expression for $C(N)$ in the simplest case

$$C(N) = \frac{1 + 2\theta N - \sqrt{1 + 4\theta N}}{2\theta N}.$$

This expression has the four properties mentioned above. If we multiply both the numerator and the denominator by $1 + 2\theta N + \sqrt{1 + 4\theta N}$ to obtain

$$C(N) = \frac{2\theta N}{1 + 2\theta N + \sqrt{1 + 4\theta N}},$$

then we see that, for N small, $C(N) \sim 2\theta N$, whereas for N large, $C(N) \sim 1$. Furthermore, $C(N)$ is non-decreasing and $C(N)/N$ is non-increasing.

Remark. If we equate 'time-scale of the infection process' with the length of the infectious period, and 'time-scale of the complex' with the inverse of the dissociation rate constant, then for most infectious diseases in many different populations our assumption that the infection processes and the formation of complexes are on two different time-scales, is reasonable. This not only applies to sexually transmitted diseases. For example, regard influenza or measles where contacts are often on the scale of minutes or hours, whereas the infectious period is on the scale of days. The problem with the present approach for these non-sexually transmitted diseases however, is that complexes in these cases may sometimes consist of more than two individuals. In order to describe the contact process for these cases more realistically, the approach in Theorem 4.1 should then in theory be generalised to allow for larger complexes. The problem is that such a generalisation immediately leads to an almost limitless proliferation of

rate constants. Moreover, one should be very careful about what one considers to be complexes in this generalisation, and which variability in contacts one could better take care of by extending the number of individual types. For example, classmates in school, fellow passengers on a commuter train, and colleagues at work are not met randomly but over and over again. This means that your schoolclass, the commuter train you take, and the place where you work, all should be made part of your individual type. In a certain sense (to be discussed more fully in section 4.5) our present model is the simplest mechanistical model that can account for the fact that individuals have limited time-budgets for their social interactions, as well as variability among types.

$\diamondsuit$

4.4. Application to marriage models

The calculations in this paper can also be applied to marriage models. In that case, one would let X_1 and X_2 be single females and males, respectively. The complexes would signify the brief encounters between singles, for example in a bar, in a train, on the street, etcetera. 'Brief' should here be interpreted as short relative to the time-scale of 'steady partnerships' between two individuals (for convenience called: marriages). One introduces a new group within the population, the *married couples*. The saturating contact rate then determines, in randomly mixing population (a well-stirred society), the possible number of new marriages formed per unit of time.

To illustrate the use of the approach in section 4.2, we will now derive a well-known simple marriage model, based on our mechanistic principles. Let the index '1' refer to female individuals, and '2' to male individuals. The longer lasting relationships are the married couples (p), that are assumed to be exclusively heterosexual; the shorter lasting encounters are the complexes $K_{11}, K_{12}(= K_{21})$, and K_{22}. The idea is that a considerable fraction of the time that an individual has available to find 'the one-and-only', is wasted by brief encounters with both sexes. We want to determine K_{12}^*.

Let μ denote the per capita death rate, b the constant birth rate, γ the probability per unit of time for a complex consisting of a male and a female to get married, and α the divorce rate for married couples. The other parameters are as in section 2.

System (S) now reads

$$\begin{aligned}\frac{dX_1}{dt} &= -r_{11}X_1^2 - r_{12}X_1X_2 + 2s_{11}K_{11} + 2s_{12}K_{12} + b - \mu X_1 \\ &\qquad + 2\mu K_{11} + \mu K_{12} + (\mu + \alpha)p \\ \frac{dX_2}{dt} &= -r_{12}X_2X_1 - r_{22}X_2^2 + 2s_{22}K_{22} + 2s_{12}K_{12} + b - \mu X_2 \\ &\qquad + 2\mu K_{22} + \mu K_{12} + (\mu + \alpha)p\end{aligned}$$

$$\frac{dK_{11}}{dt} = \frac{1}{2}r_{11}{X_1}^2 - s_{11}K_{11} - 2\mu K_{11}$$
$$\frac{dK_{12}}{dt} = \frac{1}{2}r_{12}X_1X_2 - s_{12}K_{12} - 2\mu K_{12} - \gamma K_{12}$$
$$\frac{dK_{22}}{dt} = \frac{1}{2}r_{22}{X_2}^2 - s_{22}K_{22} - 2\mu K_{22}$$
$$\frac{dp}{dt} = \gamma K_{12} - (2\mu + \alpha)p.$$

If we carry through our time-scale argument from section 4.2, we can collapse the above system into a standard system of three differential equations describing heterosexual pair formation in a two-sex population (see for example Hadeler, Waldstätter and Wörz-Busekros (1988)). In the process we get a specific form of the marriage function. Define $x := X_1 + 2K_{11} + 2K_{12}$, and $y := X_2 + 2K_{22} + 2K_{12}$ as the total density of females and males that are not members of a married couple, respectively. Then we arrive at

$$\frac{dx}{dt} = b - \mu x + (\mu + \alpha)p - \gamma K_{12}^*$$
$$\frac{dy}{dt} = b - \mu y + (\mu + \alpha)p - \gamma K_{12}^*$$
$$\frac{dp}{dt} = \gamma K_{12}^* - (2\mu + \alpha)p.$$

Here

$$K_{12}^* = \frac{1}{2}\frac{r_{12}}{s_{12}}X_1^*X_2^*$$

and (X_1^*, X_2^*) is the unique positive solution, in terms of x and y, of the system

$$X_1 + \theta_{11}X_1^2 + \theta_{12}X_1X_2 = x$$
$$X_2 + \theta_{22}X_2^2 + \theta_{12}X_1X_2 = y.$$

This system reduces to an equation in one unknown of degree 4, and can therefore be solved explicitly. In the not unreasonable special case that $x = y =: z$ (equal sex-ratio in the population) and $\theta_{11} = \theta_{22} =: \theta$, we easily obtain the unique positive solution

$$X_1^* = X_2^* = \frac{-1 + \sqrt{1 + 4\zeta z}}{2\zeta},$$

where $\zeta := \theta + \theta_{12}$.

4.5. Encore: Holling squared

In the introduction we argued that the usual Holling argument could not be applied to the 'contact rate problem' because *both* individuals that are involved in a contact are time-limited. However, we can adapt the Holling argument to this situation, and we christen this adaptation 'Holling squared'.

For convenience only, we regard the easiest case with only one type of single individual. This is not a restriction on the method, it is similar for the general setting of Theorem 4.1. Let Z denote the number of (re)new(ed) contacts in time interval T by a given individual. Let $Y = Z/T$. With $\tau = s^{-1}$ we denote the mean contact duration, with r we denote the complex formation rate constant among singles. We let N be the density of individuals, and K and X the density of complexes and singles, respectively. We have $K = \frac{1}{2}NY\tau$, and $X + 2K = N$. The usual Holling argument would be

$$Z = rX(T - Z\tau) \Rightarrow Y = rX(1 - Y\tau) \Rightarrow Y = \frac{rX}{1 + \theta X}$$

with $\theta = r\tau$. In our case however, the singles are time limited, and the available singles are given by

$$X = N(\frac{T - Z\tau}{T}) \Rightarrow X = n(1 - Y\tau).$$

Inserting this in the equation for Y above, we find

$$Y = rN(1 - Y\tau)^2$$

which leads, with $K = \frac{1}{2}NY\tau$, to

$$K = \frac{1}{2}\theta X^2$$

and this is exactly the condition found in Theorem 4.1, for this particular special case. This shows that the Holling squared leads to the same saturating contact rate expression as Theorem 4.1. Analogously one shows that the general case of Holling squared leads to the conditions (4.1.3)-(4.1.4).

The fact that the equilibrium conditions are the same for both the heuristic Holling squared and the rigorous mechanistic Theorem 4.1, raises an important point. In the Holling argument one does not use the fact that τ comes from any particular probability distribution. This suggests that, in the general approach, we can replace the exponential distribution by an arbitrary distribution and take for our parameter s, the inverse of the mean duration of the complex time period.

4.6. References:

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

Dugundji, J. (1966): *Topology.* Allyn and Bacon, Boston.

Hadeler, K.P., Waldstätter, R., Wörz-Busekros, A. (1988): Models for pair formation in bisexual populations. *J. Math. Biol.* **26**, 635-649.

Higgins, J. (1968): Some remarks on Shear's Liapunov function for systems of chemical reactions. *J. Theoret. Biol.* **21**: 293-304.

Horn, F., Jackson, R. (1972): General mass action kinetics. *Arch. Rational Mech. Anal.* **47**: 81-116.

Shear, D. (1967): An analog of the Bolzmann H-Theorem (a Lyapunov function) for systems of coupled chemical reactions. *J. Theoret. Biol.* **16**: 212-228.

Thieme, H.R., Castillo-Chavez, C. (1989): On the role of variable infectivity in the dynamics of the human immunodeficiency virus epidemic. In: *Mathematical and Statistical Approaches to AIDS Epidemiology*, C. Castillo-Chavez (ed.). Lect. Notes in Biomath. Vol. 83, Springer Verlag, Berlin.

Tikhonov, A.N., Vasil'eva, A.B., Sveshnikov, A.G. (1985): *Differential Equations.* Springer Verlag, Berlin.

4.7. Appendix

In this appendix we prove that the system

$$\frac{dX_i}{d\tau} = -X_i \sum_{j=1}^{n} \rho_{ij} X_j + 2 \sum_{j=1}^{n} \sigma_{ij} K_{ij}, \tag{A1}$$

$$\frac{dK_{ij}}{d\tau} = \frac{1}{2} \rho_{ij} X_i X_j - \sigma_{ij} K_{ij}, \tag{A2}$$

together with the conservation equations

$$X_i + 2 \sum_{j=1}^{n} K_{ij} = \xi_i, \quad 1 \leq i \leq n, \tag{A3}$$

has a unique asymptotically stable positive equilibrium.

We first prove existence of positive solutions $(X_1^*, ..., X_n^*)$ to system (4.1.3). Solutions to (4.1.3) correspond to equilibria of (A1-A3) by letting $K_{ij}^* =$

$\frac{1}{2}\theta_{ij}X_i^*X_j^*$. For given $\xi_1, ..., \xi_n > 0$ and $\theta_{ij} \geq 0$, $(i, j \in \{1, ..., n\})$ define the map $A : \mathbb{R}^n \to \mathbb{R}^n$ acting on a vector $X = (X_1, ..., X_n)^T$ by

$$(X_1, ..., X_n) \mapsto \big(\frac{\xi_1}{1 + a_1(X)}, ..., \frac{\xi_n}{1 + a_n(X)}\big),$$

where $a_i(X) := \sum_{j=1}^n \theta_{ij}X_j$. Note that positive fixed points of A correspond to positive solutions to system (4.1.3). The operator A is continuous and maps the bounded, convex and closed set $\Lambda := \{X \in \mathbb{R}^n : 0 \leq X_i \leq \xi_i\}$ into itself. Therefore, there exists at least one $X^* \in \Lambda$ with $AX^* = X^*$, by the Brouwer fixed point theorem (see, e.g. Dugundji (1966)).

Lemma A1 *The solutions to system (4.1.3) are isolated.*

Proof: Define a map $F : \mathbb{R}^n \to \mathbb{R}^n$ acting on a vector $X = (X_1, ..., X_n)^T$ by

$$(X_1, ..., X_n) \mapsto \big(X_1(1 + a_1(X)) - \xi_1, ..., X_n(1 + a_n(X)) - \xi_n\big),$$

whith $a_i(X) := \sum_{j=1}^n \theta_{ij}X_j$. Then F is differentiable and a solution to (4.1.3) corresponds to a zero of F. Let $X \geq 0$ satisfy $F(X) = 0$, and regard the derivative $D := DF(X) = (d_{ij})_{1 \leq i,j \leq n}$ of the map F at the point X in $\mathbb{R}^n$,

$$D = \begin{pmatrix} 1 + \theta_{11}X_1 + a_1(X) & \theta_{12}X_1 & \cdots & \theta_{1n}X_1 \\ \theta_{12}X_2 & 1 + \theta_{22}X_2 + a_2(X) & & \vdots \\ \vdots & & \ddots & \vdots \\ \theta_{1n}X_n & \cdots & \cdots & 1 + \theta_{nn}X_n + a_n(X) \end{pmatrix}.$$

Regard the transpose D^T of D. Note that for D^T we have

$$d_{ii} > \sum_{j \neq i} d_{ij}. \tag{A4}$$

It is elementary that such a matrix is non-singular by the following well-known argument. Suppose to the contrary that there exists a non-trivial solution z of $D^T z = 0$. Let k be one of the indices with maximal $|z_k|$ and consider the k'th equation of $D^T z = 0$. Rearranging this equation and taking absolute values leads to the estimate

$$d_{kk}|z_k| \leq \sum_{j \neq k} d_{kj}|z_j| \leq |z_k| \sum_{j \neq k} d_{kj}$$

which is a contradiction to (A4).

We can now apply the inverse function theorem to F at X. This asserts that F is a homeomorphism in some neighbourhood of X. The zeros of F are therefore isolated.

$\diamondsuit$

We now consider system (A1-A3). Let $m = \frac{1}{2}n(n+1)$, and write $c = (c_1, ..., c_m)$ for $(X_1, ..., X_n, K_{11}, ..., K_{nn})$ (where we order the components of the latter lexicographically, starting with the X_i's). We are only concerned with c's in the positive cone of $\mathbb{R}^m$. Let $c^* = (c_1^*, ..., c_m^*)$ be a given positive equilibrium of (A1-A3) and define $H_{c^*} : \mathbb{R}^m \to \mathbb{R}$ as

$$H_{c^*}(c) = \sum_{i=1}^{m} (c_i \ln \frac{c_i}{c_i^*} - c_i + c_i^*).$$

In Shear (1967), it is shown that H_{c^*} is a Lyapunov function for closed mass-action chemical reaction systems, of which (A1-A3) is an example. Specifically, the following holds: 1) $H_{c^*}(c) \geq 0$ for all positive $c \in \mathbb{R}^m$, and $H_{c^*}(c) = 0 \Leftrightarrow c = c^*$; 2) $\frac{dH_{c^*}}{dt}(c) \leq 0$ for all positive $c \in \mathbb{R}^m$, and $\frac{dH_{c^*}}{dt}(c) = 0 \Leftrightarrow \frac{dc}{dt} = 0$; 3) $\frac{\partial H_{c^*}}{\partial c_i} = \ln \frac{c_i}{c_i^*}$. The Lyapunov function H_{c^*} assures that in the case of a unique positive equilibrium c^*, this equilibrium is asymptotically stable. What remains to be shown is that we indeed have a unique positive equilibrium c^* of (A1-A3).

Lemma A2 *There exists only one positive equilibrium c^* to (A1-A3).*

Proof: By lemma A1, the equilibria are isolated. So assume, without loss of generality, that there are two, $r, s \in \mathbb{R}^m$, with $r \neq s$. Construct the Lyapunov function H_r based on the equilibrium r. Then $H_r(s) =: h_s > 0$. The graph of the function $H_r(c_1, ..., c_m)$ in $\mathbb{R}^{m+1}$ has a single zero for $c = r$, and in every coordinate direction i, H_r strictly decreases when $c_i < r_i$ and strictly increases for $c_i > r_i$. This implies that levelsets of fixed values of H_r have dimension $m - 1$. Regard the levelset $\{c \in \mathbb{R}^m : H_r(c) = h_s\}$. Because this set is of dimension less than m, it follows that in every neighbourhood of s there is a point $z \in \mathbb{R}^m$ such that $H_r(z) < h_s$. But H_r decreases along the trajectories of system (A1-A3), so the equilibrium s cannot be stable. Now construct the Lyapunov function H_s based on the equilibrium s. By the theorem of Lyapunov it follows that s is stable, and we have found a contradiction.

♢

Remark. For completeness we mention that the claim in Shear (1967) that the function H_{c^*} is also a Lyapunov function for open mass-action chemical reaction systems, was proved false in Higgins (1968). Furthermore, there is an alternative way to show asymptotic stability of the equilibrium of (A1-A3). In Horn and Jackson (1972), a general theory for mass-action kinetics is developed. One can show that our system (A1-A3) is, in the terminology of Horn and Jackson, so-called *complex-balanced*, and therefore *quasi-thermodynamic*. According to Lemma 4c in Horn and Jackson (1972), this is sufficient to assure asymptotic stability of the unique equilibrium.

♢

Samenvatting

In dit proefschrift worden twee aspecten van de verspreiding van besmettelijke ziekten vanuit de wiskunde benaderd. Het eerste aspect gaat over de vraag hoe men zou kunnen beslissen of een besmettelijke ziekte die (bijvoorbeeld) in Nederland 'binnenkomt', een epidemie zal gaan veroorzaken of niet. Het blijkt dat een bepaald getal, dat we met R_0 (R-nul) aangeven, het mogelijk maakt deze vraag te beantwoorden. De definitie en de berekening van dit getal R_0 worden behandeld in de hoofdstukken 1, 2 en 3. Het tweede aspect wordt behandeld in hoofdstuk 4 en heeft te maken met de vraag hoeveel ontmoetingen met andere mensen iemand per dag kan hebben. Ik zal in deze samenvatting proberen beide aspecten toe te lichten aan de hand van de ziekte mazelen. De theorie in het proefschrift wordt echter in grote algemeenheid behandeld en heeft betrekking op willekeurige besmettelijke ziekten bij mensen, andere dieren of planten. Ik geef eerst een korte inleiding, vertel daarna kort iets over de inhoud van hoofdstuk 4 en vervolgens ga ik wat uitgebreider in op het onderwerp van de hoofdstukken 1, 2 en 3.

Stel we bekijken een groep mensen, bijvoorbeeld de inwoners van IJsland, die allemaal vatbaar zijn voor mazelen, terwijl de ziekte niet in de groep voorkomt. Als op een gegeven moment de ziekte in de groep geïntroduceerd wordt, bijvoorbeeld doordat een met mazelen besmette matroos in een IJslandse haven aan wal gaat, dan kan men zich afvragen of er in de groep een epidemie van mazelen zal ontstaan (een 'lawine' van nieuwe gevallen) of dat het met een sisser zal aflopen.

Voor de beantwoording van deze vraag is het van belang te bepalen welke factoren van invloed zouden kunnen zijn op de verspreiding van mazelen. Om te beginnen is belangrijk hoe lang een zieke gemiddeld besmettelijk is. Bovendien is de vraag *hoe* besmettelijk de zieke gemiddeld is. Met andere woorden, wat is de kans dat tijdens een 'ontmoeting' tussen een besmet persoon en een vatbare, die vatbare persoon ook met mazelen besmet raakt. Een 'ontmoeting' moet geïnterpreteerd worden als een zodanig in contact komen met een besmet persoon dat daarbij de ziekte in kwestie kan worden overgebracht. Bijvoorbeeld, 'in het gezicht hoesten' is geen ontmoeting voor het overbrengen van 'AIDS', maar wel voor het overbrengen van 'griep'. Als we de besmettingskans per ontmoeting weten is het vervolgens van belang hoeveel ontmoetingen een besmet persoon per dag gemiddeld heeft met vatbaren.

Over het laatst genoemde aspect gaat hoofdstuk 4 van het proefschrift. Daar wordt besproken hoe men het aantal 'ontmoetingen' kan bepalen voor het geval dat steeds precies twee personen met elkaar in contact komen. Een van de bijzonderheden van 'ontmoetingen' is dat beide deelnemers in hun tijd beperkt zijn; ontmoetingen nemen een bepaalde tijd in beslag en men kan dus niet willekeurig veel ontmoetingen per dag hebben. Het was lange tijd een onopgelost probleem in met name de wiskundige epidemiologie, om een preciese afleiding te geven van een 'ontmoetingsfunctie' die alle eigenschappen heeft die men van een dergelijke functie zou verwachten. In hoofdstuk 4 wordt een oplossing van dit probleem gegeven.

Stel nu dat een zieke gemiddeld b dagen besmettelijk is en gemiddeld o ontmoetingen met vatbaren per dag heeft. Laten we de kans dat bij zo'n ontmoeting de ziekte wordt overgedragen met p aangeven. Stel dat we, bijvoorbeeld door experimenteel onderzoek, getallen kunnen vinden die we voor b, o en p kunnen invullen (voor mazelen is bijvoorbeeld $b = 7$, daarna herstelt men en wordt de persoon immuun). We verwachten dat een zieke *op* vatbaren per dag aansteekt (o ontmoetingen maal een 'succeskans' p). Omdat de zieke hier b dagen mee doorgaat, verwachten we dus dat hij in totaal *bop* nieuwe gevallen van mazelen zal veroorzaken.

Wat levert dit op voor ons oorspronkelijke probleem of er al dan niet een epidemie van mazelen op gang komt? Welnu, als het produkt van de getallen b, o en p groter dan één is, dus als elke zieke gemiddeld meer dan één vatbare aansteekt, dan zal het aantal zieken groeien (epidemie). Maar als het produkt kleiner is dan één, dus als iedere zieke gemiddeld niet eens een vervanger voor zichzelf kan produceren, dan zal het aantal zieken in de groep niet groeien en krijgen we geen epidemie. We hebben onze vraag, of er een epidemie komt of niet, dus opnieuw geformuleerd in: is een bepaald getal, *bop* in dit geval, groter dan wel kleiner dan één. We noemen zo'n getal een 'drempelgetal'.

Het idee dat er een drempelgetal is dat onze vraag kan beantwoorden, is afkomstig van Sir Ronald Ross (ongeveer 1909). Hij 'ontdekte' dit tijdens het analyseren van een eenvoudig wiskundig model voor de verspreiding van malaria. Tegenwoordig wordt in de wiskundige epidemiologie het drempelgetal met het symbool R_0 aangegeven. We zouden het ruwweg als volgt kunnen definiëren: *R_0 is het gemiddeld aantal nieuwe gevallen van de ziekte dat veroorzaakt wordt door één besmettelijk persoon.* De redenering van de vorige alinea suggereert dat voor mazelen zal gelden

$$R_0 = bop.$$

Als de wereld zo eenvoudig zou zijn dan was dit proefschrift overbodig. We hebben bij de bepaling van de R_0 voor mazelen impliciet enkele vereenvoudigende aannamen gedaan. De belangrijkste hiervan is dat we geen onderscheid hebben gemaakt in vatbaarheid tussen personen (we veronderstelden een homogene groep mensen). In werkelijkheid vertonen mensen (dieren, planten) onderling verschillen in vatbaarheid en besmettelijkheid. Natuurlijk zijn lang

niet alle verschillen van invloed op de verspreiding van een besmettelijke ziekte, maar vele kenmerken, zoals bijvoorbeeld 'leeftijd' en 'geslacht', zijn dit vaak wel. Als verschillen van invloed zijn op de vatbaarheid, het verloop van de ziekte en het bewegingspatroon van zieken en vatbaren ('mensen met welke kenmerken ontmoeten welke andere mensen hoe vaak?'), dan moeten we daar rekening mee houden als we R_0 nauwkeurig willen berekenen voor een bepaalde ziekte.

Als we alleen verschillen in besmettelijkheid toelaten (bijvoorbeeld dat de hevigheid van de besmetting van de leeftijd van de zieke afhangt), dan kunnen we R_0 nog uitrekenen door gewoonweg het gemiddelde te nemen van alle mogelijke verschillende 'typen' zieken (zieken van alle leeftijden bijvoorbeeld). Als we bovendien verschillen in vatbaarheid toestaan kunnen we nog eenvoudig gemiddelden nemen, maar alleen als de vatbaarheid en de besmettelijkheid elkaar niet beïnvloeden. Laten we 'leeftijd' als voorbeeld nemen. Dan bedoel ik met 'elkaar niet beïnvloeden', dat de kans dat een zieke en een vatbare elkaar ontmoeten, niet af mag hangen van hun respectievelijke leeftijden. We komen pas in de problemen als vatbaarheid en besmettelijkheid elkaar wel beïnvloeden. Bijvoorbeeld bij mazelen is het zo dat de zieken voornamelijk schoolkinderen zijn die klasgenootjes (en dus leeftijdsgenootjes) aansteken, en met die klasgenootjes ook veel meer ontmoetingen hebben dan met mensen van andere leeftijden. Als we met mogelijke verschillen en hun interacties rekening houden dan is het niet meteen duidelijk hoe we het drempelgetal R_0 moeten definiëren, laat staan dat het duidelijk is hoe we het zouden kunnen uitrekenen. Toch willen we voor heterogene groepen de epidemie/geen epidemie-vraag met een soortgelijk drempelgetal beantwoorden.

Gezien het belang van R_0 (ik kom hier nog op terug) en gezien het feit dat R_0 de laatste tien jaar in zeer veel wetenschappelijke artikelen op het gebied van de wiskundige epidemiologie figureert, is het verbazingwekkend dat nog niet eerder een algemene wiskundige theorie voor R_0 is gepubliceerd. (Dit is echter voor mij weer prettig, want het ontwikkelen van die wiskundige theorie voor willekeurig heterogene groepen is precies het onderwerp van dit proefschrift.) In hoofdstuk 1 wordt behandeld hoe men R_0 in grote algemeenheid wiskundig kan definiëren (men kan laten zien dat deze definitie een drempelgetal oplevert) en er onder speciale voorwaarden een formule voor kan geven. In hoofdstuk 3 gebeurt iets dergelijks voor sexueel overdraagbare ziekten zoals 'AIDS' waar rekening gehouden wordt met het gegeven dat mensen met een vaste partner een paar vormen (en dan vervolgens doorgaans alleen met die partner sexuele contacten hebben totdat het tot een scheiding komt). In hoofdstuk 2 wordt door middel van enkele voorbeelden uitgebreid getoond dat ideeën uit de theorie van hoofdstuk 1 het mogelijk maken om in schijnbaar ingewikkelde situaties R_0 toch uit te kunnen rekenen. Een van de voorbeelden betreft het berekenen van R_0 voor de verspreiding van 'AIDS' onder heterosexuelen, als gegeven is dat een tweede sexueel overgedragen ziekte (zoals syfilis) in de groep aanwezig is. Het is bekend dat zowel de vatbaarheid voor 'AIDS' voor een persoon met syfilis, als de besmettelijkheid van een seropositief persoon die ook met syfilis besmet is,

toeneemt. Als men de vraag wil beantwoorden of deze verhoogde vatbaarheid en besmettelijkheid er toe zouden kunnen bijdragen dat er, bij een toename van het aantal syfilis gevallen, een epidemie van 'AIDS' onder heterosexuelen ontstaat, dan moet men eerst in staat zijn om R_0 uit te kunnen rekenen voor deze situatie. Een ander voorbeeld betreft de verspreiding van een virus van varkens (dat bijna overal in Nederland voorkomt).

Er is een belangrijk verschil tussen de twee genoemde voorbeelden. In het eerste geval gaat het inderdaad om de vraag of een ziekte in een bepaalde groep een epidemie zou kunnen veroorzaken, maar in het tweede voorbeeld gaan we uit van een situatie waarbij de ziekte al aanwezig is in de groep. Dit is een mooie aanleiding om iets meer te vertellen over een van de belangrijkste toepassingen van het drempelgetal R_0. Stel dat een ziekte al in de groep aanwezig is, bijvoorbeeld mazelen in Nederland, en stel dat we de beschikking hebben over een vaccin tegen die ziekte (of een andere 'bestrijdingsmaatregel'). We weten dat als R_0 kleiner dan één is, de ziekte zich niet kan handhaven in de groep, en uitsterft. Een voor de hand liggende vraag is dan: is er een zodanige vaccinatie strategie (of toepassing van de bestrijdingsmaatregelen) dat de ziekte na verloop van tijd uitsterft? Bijvoorbeeld, welk percentage van de bevolking moeten we minimaal inenten om mazelen uit Nederland te verdrijven? Men heeft uitgerekend dat dit ongeveer 95% is. Hierbij kan men handig gebruik maken van R_0. Noem de fractie van de bevolking die we moeten inenten f. Het komt erop neer dat we vervolgens R_0 gaan uitrekenen alsof de ziekte nog niet aanwezig is maar nu voor een groep vatbaren die een deel $(1-f)$ is van de oorspronkelijke groep, namelijk diegenenen die niet zijn ingeënt. De drempel R_0 zal dan van het getal f afhangen. We proberen nu f zo te kiezen dat R_0 onder de drempel één terecht komt. Als we er dan in slagen om constant minimaal een fractie f van de bevolking ingeënt te houden, dan zal de ziekte zich niet kunnen handhaven. Voor een ziekte als 'AIDS' kan men zich bijvoorbeeld afvragen: stel er wordt een vaccin gevonden, welke groepen mensen dienen dan met voorrang te worden ingeënt om het meest effectief de verspreiding tegen te gaan?

Het getal R_0 is van groot belang bij de evaluatie van bestrijdingsmaatregelen. Het is dan wel zaak dat men in staat is om R_0 voor ingewikkelde en 'realistische' situaties uit te rekenen. Precies hierover handelt dit proefschrift.

Curriculum vitae

Ik werd geboren op 6 december 1960 in Roermond. Mijn wetenschappelijk onderwijs begon toen ik aan de Landbouw Universiteit in Wageningen plantenziektekunde ging studeren (oude stijl). Op 23 juni 1986 behaalde ik het doctoraal diploma (met lof) met als afstudeervakken: Fythopathologie en Wiskunde. Het vak Fytopathologie bestond uit twee onderzoeken (beide van ongeveer een half jaar): het eerste was een theoretisch onderzoek naar de modellering van continentale epidemieën (o.l.v. prof.dr. J.C. Zadoks), het tweede was praktisch van aard en betrof een taxonomisch onderzoek naar de schimmelsoorten die voorkomen in potgrond (> 60) en de successie van deze soorten na een behandeling van de grond met stoom (o.l.v. dr. G.J. Bollen). Het vak Wiskunde bestond uit vrijwel alle wiskunde vakken die er op de LU te doen waren aangevuld met het bestuderen van een boek over functionaal-analyse (o.l.v. prof.dr. B. van Rootselaar). Mede dankzij stimulerende invloed van deze laatste ben ik na afloop wiskunde gaan studeren. Eveneens van invloed hierbij was de stage van een half jaar die ik tijdens mijn studie in Wageningen deed bij O. Diekmann bij het Centrum voor Wiskunde en Informatica in Amsterdam. Op 31 augustus 1988 behaalde ik het doctoraal diploma wiskunde (cum laude) aan de Universiteit van Amsterdam met als afstudeerrichting Topologie.

Van 1 oktober 1988 tot en met 31 december 1992 was ik Onderzoeker in Opleiding bij de afdeling AM (Analyse, Algebra en Meetkunde) van het CWI te Amsterdam, binnen het project Niet-Lineaire Analyse en Biomathematica. Het onderzoeksgebied was de wiskundige beschrijving van aspecten van de verspreiding van besmettelijke ziekten in gestructureerde populaties. Van 1 juni 1991 tot 1 juni 1992 was ik tevens voor 2 dagen per week als toegevoegd docent verbonden aan de vakgroep Theoretisch Biologie van de Rijksuniversiteit te Leiden om mee te werken aan het vervaardigen van de syllabus 'Wiskundige Procesbeschrijving' voor biologie-studenten.

Ons werk

Modelling pandemics of quarantine pests and diseases: problems and perspectives. *Crop Protection* (1987), **6**: 211-221, (met J.C. ZADOKS). Ook beschikbaar in het Chinees (denk ik).

How to estimate the efficacy of periodic control of an infectious plant disease. *Math. Biosc.* (1989), **93**: 15-29, (met H.R. THIEME).

On sums of remainders and almost perfect numbers. *CWI-Quarterly* (1989), **2**: 15-17, (met H.A.J.M. SCHELLINX)

Building Blocks and Prototypes for Epidemic Models. Lecture Notes, preprint (1989) (met O. DIEKMANN, M. KRETZSCHMAR & J.A.J. METZ)

On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* (1990), **28**: 365-382, (met O. DIEKMANN & J.A.J. METZ).

The nature of fractal geometry. *CWI-Quarterly* (1990), **3**: 137-149, (met J.M.A.M. van NEERVEN, H.A.J.M. SCHELLINX & M. ZWAAN).

Several elementary proofs that $0 = 1$. *Nieuw Archief voor Wiskunde* (1990), **series 4, part 8**: 253-256, (met J.M.A.M. van NEERVEN & H.A.J.M. SCHELLINX).

The influence of certain co-factors on the spread of HIV. In: *AIDS Impact Assessment, Modelling and Scenario-Analysis* J.C. Jager & E.J. Ruitenberg (eds.), Elsevier Science Publishers, (1992), pp. 73-81.

The basic reproduction ratio for sexually transmitted diseases, Part 1: Theoretical considerations. *Math. Biosc.* (1991) **107**: 325-339, (met O.DIEKMANN & K. DIETZ).

The saturating contact rate in marriage- and epidemic models. Te verschijnen in: *J. Math. Biol.* (1992) (met J.A.J. METZ).

A singular perturbation theorem for evolution equations and time-scale arguments for structured population models. Opgestuurd naar: *Can. Appl. Math. Quart.* (1991) (met G. GREINER & J.A.J. METZ).

The basic reproduction ratio for sexually transmitted diseases, Part 2: Effects of variable HIV-infectivity. Opgestuurd naar: *Math. Biosc.* (met K. DIETZ

& D.W. TUDOR).

An algorithm to calculate the basic reproduction ratio for discrete time heterogeneous epidemic models. Manuscript (met M.C.M. De JONG & O. DIEKMANN).

Das Fegefeuer Theorem (De Purgatorio), Verlag Die Libelle, Bottighofen am Bodensee, Litzelstetter Libellen, Abteilung Handbüchlein und Enchiridia, Z.N.-F. (Ziemlich Neue Folge) 2, te verschijnen (1992), (met J.M.A.M. van NEERVEN & H.A.J.M. SCHELLINX).

STELLINGEN

bij het proefschrift 'R_0'
van Hans Heesterbeek

1. Nu de existentie van het Vagevuur (Purgatorio) eindelijk is bewezen, is het grootste open probleem van de Mathematische Theologie (s.l.) de realiteit van de Hel (Inferno) aan te tonen binnen de huidige grenzen van de axiomatische Zondenleer.

   J.A.P. Heesterbeek, J.M.A.M. van Neerven, H.A.J.M. Schellinx (1992): *Das Fegefeuer Theorem (De Purgatorio)*, Libelle Verlag, Bottighofen am Bodensee, Litzelstetter Libellen, Abteilung Handbüchlein und Enchiridia, Z.N.F. (Ziemlich Neue Folge) 2 (ISBN 3-909081-55-x).

2. Het meest opvallende en frustrerende effect van het publiceren van een abstracte wiskundige 'oplossing' in grote algemeenheid van een probleem uit de biologie is, dat menig onderzoeker daarna speciale gevallen nog steeds hardnekkig op een eigen (vaak foute) manier blijft oplossen, terwijl tegelijkertijd het bewuste abstracte artikel (op een veelal onhandige plaats) geciteerd wordt.

3. Beschouw een systeem van evolutie vergelijkingen, die afhangen van een kleine parameter $\varepsilon \in (0, \varepsilon_0]$, van de volgende vorm

$$\dot{\gamma}_\varepsilon(t) = f(\gamma_\varepsilon(t), w_\varepsilon(t), \varepsilon) \tag{1.1a}$$

$$\varepsilon \dot{w}_\varepsilon(t) = A_0 w_\varepsilon(t) + \varepsilon F(\gamma_\varepsilon(t), w_\varepsilon(t), \varepsilon) \tag{1.1b}$$

   met beginconditie

$$\gamma_\varepsilon(0) = \bar{\gamma} \ , \ w_\varepsilon(0) = \bar{w} \ . \tag{1.1c}$$

   De functie γ_ε heeft waarden in $\mathbb{R}^m$, en w_ε heeft waarden in een (oneindig dimensionale) Banach ruimte X. Voor A_0, f en F nemen we aan:

   i). A_0 genereert een C_0-halfgroep $\{T_0(t)\}$ op X die exponentieel stabiel is.

   ii). Zowel $f : \mathbb{R}^m \times X \times (0, \varepsilon_0] \to \mathbb{R}^m$ als $F : \mathbb{R}^m \times X \times (0, \varepsilon_0] \to X$ zijn continu en locaal Lipschitz met betrekking tot de eerste twee variabelen, uniform in ε.

   Dan geldt de volgende singuliere storingsstelling (á la Tykhonov):

   Stelling *Zij* $\gamma_0 : [0, T] \to \mathbb{R}^m$ *een oplossing van* $\dot{\gamma}(t) = f(\gamma(t), 0, 0)$, *met* $\gamma(0) = \bar{\gamma}$. *Dan is er voor elke* $\delta > 0$ *een* $\varepsilon_1 > 0$ *zodat voor* $\varepsilon \in (0, \varepsilon_1]$ *de oplossing* γ_ε, w_ε *van* (1.1) *bestaat op* $[0, T]$ *en voldoet aan de volgende afschattingen*

$$|\gamma_\varepsilon(t) - \gamma_0(t)| \leq \delta \quad \textit{voor alle} \quad t \in [0, T]$$

$$\|w_\varepsilon(t)\| \leq \delta \quad \textit{voor alle} \quad t \in [\delta, T]$$

G. Greiner, J.A.P. Heesterbeek, J.A.J. Metz (1991): A singular perturbation theorem for evolution equations and time-scale arguments for structured population models, subm. to *Can. Appl. Math. Quart.*

4. 'For the *continuing health* of their subject mathematicians *must* become involved with biology. (...) it is *clear* that if mathematicians do not become involved in the biosciences they will *simply* not be part of what are likely to be the most important and exciting discoveries *of all time.*'

   De cursiveringen zijn van mij.

   J. Murray (1989): *Mathematical Biology*, Springer Verlag, Berlin (1^e alinea preface).

5. Met de zin 'Love of deity [is the] effect of organization' in een van zijn *Notebooks (C 166)* bedoelde Darwin dat het geloof in goddelijke machten een evolutionair artefact is van de ingewikkelde structuur van onze hersenen.

   A. Desmond, J. Moore (1991): *Darwin*, Michael Joseph, London.

6. In alle wetenschappelijke publikaties dient een alfabetische auteursvolgorde te worden aangehouden, zoals dit gebruikelijk is in wiskundige publikaties.

7. De door Nederlandse wetenschappers meest gemaakte (ver)taalfout tijdens voordrachten in het Engels is de zin: 'This is how it looks like'.

8. De stelling 'Beter 1 vogel in de hand dan 10 in de lucht', kan verscherpt worden tot 'Beter 1 vogel in de hand dan 10.4319 in de lucht'.

   G.V Decnop, J. Verhoeff (1956): 'n Bewijs van de stelling: 'Beter één vogel in de hand dan tien in de lucht.' In: *Het Tentamen Algebra e.a*, A.B. Paalman de Miranda, B. van Rootselaar e.a (Eds.), Mathematisch Instituut, Universiteit van Amsterdam.

9. Enkele weken na de verschijning van 'Het ontbrokene' overleed de dichter Hans Faverey. De slotregels van het laatste gedicht dat hij schreef:

   > Laat de god die zich in mij verborgen houdt
   > mij willen aanhoren, mij laten uitspreken,
   > voor hij mij met stomheid slaat en mij
   > doodt waar ik bij sta, waar jij bij staat.

   Wat moet ik dan nog schrijven,
   wat moet ik dan nog lezen.

R_0

PROEFSCHRIFT

ter verkrijging van de graad van Doctor
aan de Rijksuniversiteit te Leiden,
op gezag van de Rector Magnificus
Dr. L. Leertouwer,
hoogleraar in de Faculteit der
Godgeleerdheid,
volgens besluit van het College van Dekanen
te verdedigen op donderdag 17 september 1992
te klokke 14.15 uur

door

JOHAN ANDRE PETER
HEESTERBEEK

geboren te Roermond in 1960

1992
Centrum voor Wiskunde en Informatica, Amsterdam

Promotiecommissie:

Promotores:	Prof. dr. O. Diekmann
	Prof. dr. J.A.J. Metz
Referent:	Prof. dr. K.P. Hadeler
Overige leden:	Prof. dr. ir. J. Grasman
	Prof. dr. T.M. Konijn
	Prof. dr. ir. L.A. Peletier
	Prof. dr. ir. M.W. Sabelis

Natura infinita est,
sed qui symbola animadverterit
omnia intelliget
licet non omnino

J.W. von Goethe

The deeper understanding Faust sought
could not from the Devil be bought
but now we are told
by theorists bold
all we need to know is R_0

R.M. May

I'm just giving harmless advice you thought
but the eye of the Devil's been caught
he'll let your theorists bold
not live to grow old
and claims their inventions are naught

The Devil's Advocate

Although the Devil may well rage as Nero,
every theorist he damns is a hero
and one thing behold
will surely grow old
he can't kill ideas like R_0

The Author

Voor Marion, Madelief
Jodokus & Bop Niksaart

Preface

It is, in my opinion, worthwhile devoting
some energy to proving the obvious.

J.B.S. Haldane

This thesis should be filed, if at all, on the shelf of mathematical epidemiology. To be more precise, it treats some aspects of deterministic mathematical modelling of the spread of infectious diseases in heterogeneous populations. To be more honest, it treats just two aspects of this, and one of these only briefly at that.

We are mainly concerned with the methodology of determining whether or not a contagious disease can, upon entering a population of humans, other animals or plants, cause a spreading epidemic in that population. The basic tool for deciding this so-called 'invasion problem' is a quantity we call 'the basic reproduction ratio' which is, in mathematical epidemiology at least, commonly denoted by the symbol R_0 (note that this immediately fully explains the title of this thesis). Almost everything you will find in these pages is somehow concerned with this quantity.

In the last decade R_0 has become, judging by its appearance in many papers as a minor or major aspect (although the subject matter of these papers is mutually very divergent), one of the most important concepts in mathematical epidemiology. It is strange that, notwithstanding all the research effort devoted to this quantity, no general mathematical theory concerning the definition of R_0 and its calculation has been published sooner. Of course, one praises oneself lucky that such has not been the case. Stabs at a mathematical 'treatment' have certainly been taken before, but only with regards to special cases, never as a systematic approach. In retrospect, there were both correct definitions for special cases (R_0 as dominant eigenvalue of a matrix, if only a finite number of different types of individuals are considered, and R_0 as the spectral radius of some operator derived from an age-dependent model), and incorrect definitions that sometimes gave the correct answer only because the assumptions happened to be just right (e.g. R_0 as a simple weighted average). In the following chapters we explain the formal mathematical side of the basic reproduction ratio in a very general context, and make some excursions

into the computational aspects involved. Furthermore, we pay attention to the application of our results to certain 'real-world' problems. Although the title is boldly 'R_0' we can of course not even remotely be all-encompassing. The main viewpoints that are lacking are a stochastic approach and an approach for 'macro-parasitic' diseases.

The idea behind R_0 and some aspects of its history are introduced in section 0.3 of the introductory chapter. The mathematical way of looking at R_0, and the methodology of its calculation are dealt with in chapter 1. This is an extended version of [1]. In chapter 2, we treat two detailed examples of situations where the methods of chapter 1 are applied fruitfully to calculate R_0 for seemingly complicated 'real-world' problems, one concerning the spread of HIV ('causing' AIDS) in a population where the presence of certain other sexually transmitted diseases can enhance HIV's probability of infection (this is an extension of [2]), the other concerning the spread of a virus disease in a population of farm-animals. In addition we regard an example with age-dependence, taken verbatim from [1]. In chapter 3 we explain how to obtain R_0 for sexually transmitted diseases if we take into account the fact that individuals can form relationships for longer periods of time. This is a generalisation of ideas of K. Dietz, and the first four sections are a reprint of [3]. In the last two sections of this chapter, we apply the developed techniques to study, in collaboration with K. Dietz and D.W. Tudor, the effect of variable infectivity on the spread of HIV. In addition we give some preliminary results about the incorporation in our model of certain behaviour changes. The larger part of section 3.5 and the variable infectivity part of section 3.6 are based on [4] (the reader should pardon the unavoidable repetition of ideas from earlier sections).

What remains is non-R_0. The introductory chapter explains the terminology used in later chapters (section 0.1), highlights the basic questions of mathematical epidemiology (section 0.2), and briefly reports on work in progress (with M. Kretzschmar) on a model incorporating superinfection (section 0.4) ending in a discussion on the location, in my opinion, of the heart of mathematical epidemiology. In the final chapter 4 an open problem from mathematical epidemiology is solved. It concerns the problem of determining how many contacts with other individuals (that can possibly lead to disease transmission) an individual is capable of having, given that both individuals that take part in the contact are time-limited. This chapter is a reprint of [5].

Throughout this thesis we will adhere to the principle borrowed from Desmond Morris (*Babywatching*, Jonathan Cape, London, 1991) that all individuals will be referred to as 'it'. The only individual that is excepted from this rule is yours truly. I will be referred to as 'we', in the best of scientific and royal traditions.

[1] On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* (1990), **28**: 365-382, (with O. DIEKMANN & J.A.J. METZ).

[2] The influence of certain co-factors on the spread of HIV. In: *AIDS Impact Assessment, Modelling and Scenario-Analysis* J.C. Jager & E.J. Ruitenberg (eds.), Elsevier Science Publishers, (1992), pp. 73-81.

[3] The basic reproduction ratio for sexually transmitted diseases, Part 1: Theoretical considerations. *Math. Biosc.* (1991) **107**: 325-339, (with O. DIEKMANN & K. DIETZ).

[4] The basic reproduction ratio for sexually transmitted diseases, Part 2: Effects of variable HIV-infectivity. Subm. to *Math. Biosc.* (with K. DIETZ & D.W. TUDOR).

[5] The saturating contact rate in marriage- and epidemic models. *J. Math. Biol.* (to appear) (with J.A.J. METZ).

Contents