Non-homogeneous stochastic birth and death processes

with applications to epidemic outbreak data

Jan van den Broek

ISBN 978-90-393-5835-1
Printed by Ridderprint, Ridderkerk, The Netherlands
cover: Anjolieke Dertien, Multimedia, Faculty of Veterinary Medicine, Utrecht University. (Figure 3.5 chapter 3)

Non-homogeneous stochastic birth and death processes

with applications to epidemic outbreak data

Niet-homogene stochastische geboorte en sterfte processen

met toepassingen voor epidemische uitbraak data
(met een samenvatting in het Nederlands)

Proefschrift
ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op donderdag 27 september 2012 des middags te 12.45 uur

door
Jan van den Broek
geboren op 1 juni 1957 te Ermelo

Promotor: Prof.dr. J.A.P. Heesterbeek

Contents

Chapter 1

General introduction

1.1 Introduction

The subject of this thesis is the non-homogeneous birth-death process with some of its special cases and notably its use in modeling epidemic (outbreak) data. This model describes changes in the size of a population. New population members can appear with a certain rate, called the birth rate or the reproductive power, and members can leave the population with a rate called the death rate. These rates are taken to be non-homogeneous, i.e. they can change over time.

The model is used in the context of interpreting epidemic (outbreak) data, so the population is taken to be the number of infected individuals at a certain point in time. Epidemic data has some important characteristics that have to be taken into account. An important aspect is for instance, that infected individuals can reproduce only during some limited time interval. In section 1 of this introduction some general aspects of modeling epidemic data are discussed. One of these points is that in practice one often has to choose between the complexity of the description of the expected value of the process as a function of time and the description of the distributional part of the model.

The models developed in this thesis are probabilistic in nature and are applied to epidemic (outbreak) data: the Dutch avian influenza outbreak in 2003, the foot-and-mouth disease outbreak in Great Britain and in the republic of Ireland in 2001, the classical swine fever outbreak in the Netherlands in 1997-1998, data of avian influenza in poultry and wild birds in Thailand and data on MRSA in British hospitals. This means that the modeling is statistical and one has to determine how the data can be used as evidence. Therefore in section 2 some statistical modeling issues are discussed.

First two general criteria for choosing a model are mentioned and then an issue from statistical inference is discussed, namely: what can be used as a measure of evidence? As will be argued in that section, underdeterimination plays an important role.

In the next section of this introduction the stochastic (linear) birth-death model is introduced. This model can be used as a compromise between the complexity of the description of the pattern in the data and the description of the distributional part of the data. Techniques are described that can be used in obtaining the stochastic differential equations for this model and from these the partial differential equation for the probability generating function and its solution. Then it is shown how the Lagrange transform might be used to invert the probability generating function into the probability mass function. These techniques are used in later chapters to find the probability mass function of the non-homogeneous birth-death process.

1.2 Modeling epidemic data.

Modeling the spread of an infection in a population is often done by means of a dynamic model, describing the changes in the size of compartments over time. These compartments describe the state in which an individual is with respect to the disease. A famous example of an deterministic epidemic model is the SIR model.

In a simple setting one can think of a population in which an infectious disease occurs, as consisting of individuals in three different states. The susceptible population (S) represents individuals who have not been infected yet but may experience infection in future. The second is a population of infectious individuals (I), which consists of those who have been infected and are infectious to others. The last group consists of removed or recovered individuals (R) who are no longer infectious and are immune. The SIR (Susceptible-Infected-Recovered) model (Diekmann and Heesterbeek, 2000) describes the transition between these 3 states. Let $x(t)$, $y(t)$ and $z(t)$ be the susceptible, infectious and removed fractions of the population at time t, respectively, the transitions according to the SIR-model are:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}x(t) &= -\beta\, y(t)\, x(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}y(t) &= \beta\, y(t)\, x(t) - \mu\, y(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}z(t) &= \mu\, y(t) \end{aligned} \tag{1.1}$$

In this model β, can be interpreted as the number of contacts per unit time, multiplied by the probability that transition occurs given a contact between a susceptible and

an infectious individual (Diekmann and Heesterbeek, 2000) or, roughly, the number of effective contacts per unit time. The force of infection is defined as $\beta y(t)$, it is the probability per unit of time for a susceptible to become infected. This force of infection depends on the number of infectious individuals and on the effective contacts per unit time. In the formulation of (1.1) random mixing is assumed so the probability that two individuals make contact is the same for all pairs of individuals.

If one writes $\lambda(t) = \beta x(t)$, where $\lambda(t)$ can be called the reproductive power, the equation for a deterministic birth-death model is obtained where the birth rate is non-homogeneous over time:

$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) = \lambda(t)\, y(t) - \mu\, y(t) \tag{1.2}$$

The reproductive power depends on the number of susceptibles and can change over time. This means that the reproductive power can change, for instance because the way in which infected individuals and susceptibles mix changes. This means that no homogeneous mixing assumption is needed.

Instead of viewing the process deterministically, i.e. or viewing the equations as describing what happens on average in a large population, one can view the outbreak of an infectious disease as a process that is stochastic, just like the throwing of a dice: one can predict an outcome only with a certain probability. An additional advantage to model infection process stochastically, is that one can explicitly define the probability of transmission, rather than deterministically stating whether or not the transmission happens (Andersson and Britton, 2000).

A stochastic model can be thought of at two levels. First it can describe the quantitative patterns of the observations as a function of time by means of their expected values. This part describes what happens on average over time. The second part is the distributional part on which the formulas for the standard deviations and standard errors depend. Change of this distribution to, for instance, one with heavier tails, changes the formula for the standard deviation. Modeling the standard deviation can be very important, as is explained in chapter 3. A deterministic model with a more or less ad-hoc choice of the distribution for the number of infected per unit of time, gives ad-hoc formulas for the standard deviations and thus only allows only ad-hoc inference about the process.

Another important point of a stochastic model is that the likelihood function can be derived, which will be useful for statistical inference of parameters and critical assessments of the modeling method. Moreover, a stochastic process can model the number of infected over time as being dependent.

There can be a trade-off between the complexity of the description of the ex-

pected value as a function of time and the description of the distributional part of the model. One can use a more complex model to capture the pattern, e.g. recognizing many individual states, and then use something simple for the distributional part like (weighted) least squares or a poisson model, and thus use a more or less ad hoc method for estimating the variance of the data and the variance of the parameter estimates. But since neglecting the distributional part of the model leads to ad hoc inference, one can use a more complex distribution to be able to deal with the variance in a more appropriate way, and then use a relative simple model for the pattern. One approach might be more accurate in estimating the pattern in the data, the other might be more accurate in estimating the variance. Ideally, one searches for a balance that captures the most important part of the pattern over time in the counts of the infectious disease, but also use a stochastic part general enough to be able to describe the variances in the data accurately. If one chooses to increase the accuracy of one part, that in the other might be decreased. In this thesis the non-homogeneous birth-death model is taken as a reasonable compromise between a description of the pattern over time and of the distributional part of the model.

Because these models are stochastic, and intended to describe data, one has to deal with some statistical modeling issues, which I discuss next.

1.3 Statistical modeling

1.3.1 Choosing a model

One might choose a model based on the characteristics of the data: does one have count data or continuous data, are the data independent, are there other covariates or groups to compare etc? Such models can be called 'data driven'models. So for outbreak data one might consider the log of the weekly counts as normally distributed data and model the dependence with an auto correlation process. The time effects could be modeled with orthogonal polynomials or even with a non-parametric local regression form like a lowess curve. This model might give a good description of the observed data, but it explains little. This can sometimes be a good approach, for instance if one compares the efficacy of medication with respect to the number of patients cured. Not only does this approach offer little explanation, there might be a lot of other models, according to the same data considerations, that also give acceptable descriptions of the data. This refers to a problem in philosophy of science which is called (contrastive) underdetermination which states that, given the body of evidence, more than one model (theory) will be supported by this evidence. For a

detailed discussion of underdetermination see (Curd and Cover, 1998, chapter 3).

In science one usually wants models to have explanatory power for the problem at hand. Comparing models can give insight in explanatory power of a current theory. This means that the available theory (biological, medical, etc) dominates the choice of the models. This approach might be called 'theory driven'modeling. With this approach the problem of underdetermination can be limited since there is an extra restriction on the number of possible models. An example of such a model is the SIR model. This model aims to describe the rates at which immunizing infection spreads through a population, while making (model)assumptions about how this model describes the process of spread of the disease, explicit. Usually more than one candidate model is applicable. For instance, two models might be the same except for an age effect on the rate by which the disease spreads.

1.3.2 Statistical inference

Statistical inference asks what conclusion can be drawn about a (null)hypothesis, given the data and given the model used. Usually a statistic (a function of the available data) is evaluated under the (null)hypothesis being tested. This is often done by means of a p-value: the probability of the observed outcome of a test statistic, or a more extreme outcome, given that the null-hypothesis is true. If models are to be compared, then very often the likelihood ratio test statistic is used. The probability of the data, the likelihood, under the null-hypothesis given the model and the available data, is calculated and compared to the likelihood under the alternative hypothesis. The likelihood ratio test statistic then is:

$$2 \log \left(\frac{L_1}{L_0} \right)$$

where L_0 is the likelihood evaluated under the null-hypothesis and L_1 the likelihood evaluated under the alternative. In essence this test statistic measures how much more likely the data is under the alternative hypothesis (given the model and the data) as compared to the null-hypothesis (given the model and the data). A p-value is calculated using the asymptotic chi-square distribution of this statistic under the null-hypothesis (generalized likelihood ratio test). Not only likelihood ratio test statistics are accompanied by p-values. Other ways of calculating test statistics are the Wald test (the estimate of the parameter of interest divided by its standard error) and the score test. For an example of the latter see Van den Broek (1995).

But, there are some concerns with the use of p-values. Fisher saw a small p-value as evidence against the null-hypothesis, with smaller p-values giving stronger

evidence. But actually the p-value measures the degree of (in)consistency of the observed data with the null hypothesis. According to Royall (Royall, 1997) the issue is that the data at hand should be interpreted as evidence and the question is how the strength of this evidence can be measured. Using the p-value as a measure of that evidence is problematic. See for instance Anderson, Burnham and Thompson (2000) for a more elaborate discussion. (For a discussion of p-value problems with multiple testing and optional stopping see Royal (1997), chapter 5.): Below, I elaborate some of the problems with p-values as discussed in the literature:

- Improper interpretation of the p-value.
- The p-value depends on the sampling being random.
- The p-value depends on data not observed.
- The p-value depends on intentions of the researcher.

As will be argued below I do not agree that the last two points are p-value concerns. Rather, they are problems with the data being inconclusive about which model is best.

To start with the first point: the interpretation of the p-value is 'given that H_0 is true what is the probability of these or more extreme observations'. The interpretation for the p-value, often seen in literature is: 'given these data what is the probability that H_0 is true?'(Cohen, 1994). This is known as the p-value fallacy. As Cohen (1994) illustrates, this is a misapplication of deductive syllogistic reasoning. To see this, start with the correct reasoning (modens tollens):

> *If the null-hypothesis is correct, then these observations can not occur.*
> *These observations have occurred*
> *Therefore the null hypothesis is false*

Now, modify the language so that the reasoning becomes probabilistic:

> *If the null-hypothesis is correct, then these observations are highly unlikely.*
> *These observations have occurred*
> *Then the null hypothesis is highly unlikely*

This reasoning is incorrect as the following example illustrates:

> *If a person is an American then he is probably not a member of Congress*
> *This person is a member of Congress*
> *Therefore he is probably not an American*

This illustrates that the statement that a p-value tells you how likely the null hypothesis is, is wrong. It should be interpreted as the probability of the observations (outcome test statistic), or more extreme data under the null hypothesis.

A problem in (medical or biological) practice is that it is common to have a sample that is not drawn randomly. The p-value depends heavily on the randomness of the sample. To see this, think of a test statistic the distribution of which under the null hypothesis is (approximately) known. The observed value of the test statistic under the null hypothesis can only be seen as a random drawing from this distribution if the sample is a random sample. If the sample is not random the test statistic can be biased downwards or upwards. The p-value then is meaningless. One can make the assumption that the sample behaves as if it was a random sample with respect to the models under consideration, but such an assumption is dangerous, as is illustrated by an article as 'Epidemiology faces its limits' (*Science*, 269, 1995). In that article, and in many others, epidemiologists argue that in order for their scientific findings to be valid, the experiments/trials should be repeated. This has its influence on the inductive logic that is going to be applied. A p-value for a non random sample can be used if there is randomization, e.g. of a treatment option. The p-value is then with respect to all possible randomizations.

The third point — the p-value depends on data not observed — is due to the fact that a p-value calculates the probability on the outcome of the test statistic or a more extreme one. Suppose the test statistic used, has an outcome x. The criticism is that two different sampling distributions can give different p-values although the probability of a value x – or of a value in a small neighborhood of x – is the same. This means that the two sample distributions have different probabilities for values larger than x (tail probabilities). So the p-value depends on outcomes of the test statistic which are not observed.

The sampling distribution, however, depends on the models used for the data. Using a different distribution for the data can give a test statistic distribution with a heavier tail as compared to other distributions for the data, and can thus lead to other p-values. This means that the difference in tail probabilities is due to the difference in models used, and that it is important to be able to discriminate between the two models. If this is not possible then the problem is underdetermined. In science it is important to determine which model is appropriate, since a model not necessary predicts but primarily describes the process at hand, and thus may have explanatory power. If a theory generates two different models that might both be plausible but one of these fits the data better, then this has implications for the theory. So in science it is important to use the data to see if there is evidence in favor of some models with respect to others.

That the result can depend on the intention of the researcher, the fourth concren with p-values, is often illustrated with the use of the binomial vs the negative binomial model. Researcher 1 might do a fixed number of trials (Binomial) in order to gain information about a fraction of 'successes', while a second researcher continues until a predetermined number of 'successes 'occurs (Negative binomial). Now, it can happen that if the first researcher observes k successes out of n trials and second researcher observes k successes and $n-k$ failures, so the same data is observed, the p-value is different because the distributions are different (Lindsey, 1996). Also here, a distinction needs to be made between the two models used. In the infectious disease context where there are susceptibles and infecteds, the stochastic death-process model leads to a binomial distribution for the reduction in susceptibles at a certain time point. This model needs a well defined population of susceptibles and it has the force of infection as its parameter. This parameter can be interpreted as describing 'how good susceptibles are in avoiding infection'. Since one starts with a well defined population (susceptibles) and models how the size of this population reduces in time, this model can also be used for other events than infectious disease events. The susceptible population then is called the 'population at risk'. The stochastic birth-process leads to a negative binomial distribution for the number infected at a certain time point. This distribution has the reproductive power as its parameter, which can be interpreted as describing 'how good an infected is in reproducing '. One does not need a well-defined population of susceptibles – except that there should be enough of them available in order for the infected to be able to reproduce– but the disease needs to be communicable, i.e. the number of infected at a certain time point depends on the number of infected before that time point. These models use the data differently, use different assumptions and have different interpretations. The same problem as above occurs: if the observed data is not able to make a distinction between two models, the problem is underdetermined, and either more data need to be gathered, or one can choose a model on theoretical grounds and on the assumptions both models need. So the fourth concern about p-value is not a p-value problem but is a consequence of underdetermination in the research.

The above makes clear that evidence in the observed data is related to the model (and thus to the theory) used. Thompson (Thompson, 2007) makes the more general statement that we need to think in terms of evidential meaning of the observations (y) from an experiment (E) in the context of some theory (T), $E_V[E, T, y]$, and that an important part of this theory is also the inductive logic being employed. Examples of inductive logic are Bayesian analysis, p-values and likelihoods. So the evidence using 'theory 1', $E_V[E, T_1, y]$, need not be equal to the evidence using 'theory 2', $E_V[E, T_2, y]$.

Thus in order to avoid too much underdetermination and to be able to have some explanatory power, models in science can best be theory driven. The question then is: what evidence is there in the observed data for a given model, as compared to another model? In the light of the above discussion, criteria with which to compare models, must be based on the observed data alone and draw conclusion only about what has been observed (unless a truly random sample is taken). Such criteria are best based on the likelihood, since the likelihood describes the probability of the observed data. One measure that compares models using the likelihood is Akaike's Information criteria (AIC), defined as twice the likelihood ratio minus two times the number of parameters that has to be estimated (Lindsey, 1996). This measure penalizes models with a large number of parameters. As a measure for underdetermination one can use Ocham's razor. If two models have approximately the same AIC (the difference between them is less than 2) then the simpler model, in terms of the number of estimated parameters, is chosen with the argument that there is no evidence in the data to choose the more complicated model.

The model used is an attempt to describe the data-generating process that gave rise to the observed data, including the sampling. It is hoped that a model is found that gives a reasonable description of this process, although it is always an approximation.

These points can lead us to consider the stochastic birth-death model as a reasonable candidate for epidemic data. As stated above the reproductive power is a function of the the 'effective number of contacts'-parameter and the fraction of susceptibles. It describes, in a stochastic context, the rate at which infected individuals reproduce and the rate at which they are removed. So it is a model based on a (theoretical) description of the problem at hand and is not data driven. Besides this, with epidemic outbreaks there are no random samples taken but instead data of the whole outbreak is gathered. In practice these outbreak data are often approximate because only the onset of symptoms of the infection is observed rather than the infection event itself. The stochastic birth-death model can take the above mentioned considerations into account. These points makes the likelihood-based approach the most appropriated, because conclusions are drawn about the observed data only. As mentioned above a reasonable model for infectious disease data can be the stochastic birth-death model which model is described next.

1.4 The stochastic (linear) birth-death model

As discussed in section 2 chapter of this, in practice one often has to choose between the complexity of the description of the (time) pattern of the data and the description of the distributional part of the data. Such a choice can be the birth-death model. In a stochastic birth and death process, the probability of a birth or a death will depend on the population size at time t. It is a continuous Markov chain for which transitions from state y can only go to state $y-1$ or to $y+1$. Markov processes are characterized by the condition that the outcome on a future time point only depends on the past through the outcome on the current time point. One can think of a birth-death process by supposing that whenever the population size is y, the time until the next birth is exponentially distributed with rate λ_y, and is independent of the time until the next death, which is exponentially distributed with rate μ_y (Ross, 1983). A birth-death process is called linear if $\lambda_y = \lambda y$ and $\mu_y = \mu y$.

If the process is applied in an infectious disease context, then a birth is taken to be a new infection, and a death a removal from an infectious state.

The discussion below follows Jones and Smith (2010). Suppose the population size at time 0 is y_0, then $P_{y_0}(0) = 1,\ P_y(0) = 0$ for $y > y_0$, where $P_y(t)$ is the probability that the population size at time t is y. The probability of a birth in the time interval $(t, t+\delta t)$ is proportional to δt for small δt. If λ is the birth rate for the stochastic linear birth death model, then the probability of a birth in this time interval if the population size is $y-1$ at time t, is $\lambda(y-1)\delta t$. The probability of two or more births in time interval $(t, t+\delta t)$ is $o(\delta t)$, $o(\delta t)$ representing a term of smaller order than δt.

The probability that an individual dies in a short time interval $(t, t+\delta t)$ is $\mu(y+1)\delta t$ if the population size at time t is $y+1$. The probability of two or more deaths in time interval $(t, t+\delta t)$ is $o(\delta t)$.

The population size y can arise in the small time interval $(t, t+\delta t)$ if :

- the population size at time t is $y-1$ and a birth occurs with probability $\lambda(y-1)\delta t + o(\delta t)$;
- the population size at time t is $y+1$ and a death occurs with probability $\mu(y+1)\delta t + o(\delta t)$;
- the population size at time t is y and nothing happens with probability $[1-(\lambda y + \mu y)\delta t + o(\delta t)]$;

Then, by the law of total probability :

$$\begin{aligned} P_y(t+\delta t) &= [\lambda(y-1)\delta t + o(\delta t)]\, P_{y-1}(t) + [1-(\lambda y+\mu y)\delta t + o(\delta t)]\, P_y(t) \\ &\quad + [\mu(y+1)\delta t + o(\delta t)]\, P_{y+1}(t) \\ P_0(t+\delta t) &= [\mu\delta t + o(\delta t)]\, P_1(t) + [1+o(\delta t)]\, P_0(t) \end{aligned}$$

After dividing by δt and taking limits ($\lim_{\delta t\to 0}$) this can be rewritten as:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}P_y(t) &= \lambda(y-1)P_{y-1}(t) - (\lambda+\mu)yP_y(t) + \mu(y+1)P_{y+1}(t),\ y \geq 1 \\ \frac{\mathrm{d}}{\mathrm{d}t}P_0(t) &= \mu P_1(t) \end{aligned} \tag{1.3}$$

which are the differential equations for the stochastic linear birth-death process.

We now show how to obtain the probability generating function (pgf), which is defined as

$$G(s,t) = \sum_{y=0}^{\infty} P_y(t)s^y$$

We start by multiplying both sides of the equation in (1.3) by s^y and then sum over the appropriate range:

$$\begin{aligned} \sum_{y=0}^{\infty}\frac{\mathrm{d}}{\mathrm{d}t}P_y(t)s^y = \lambda\sum_{y=2}^{\infty}(y-1)P_{y-1}(t)s^y - \lambda\sum_{y=1}^{\infty}yP_y(t)s^y - \mu\sum_{y=1}^{n_0}yP_y(t)s^y + \\ \mu\sum_{y=0}^{n_0-1}(y+1)P_{y+1}(t)s^y \end{aligned} \tag{1.4}$$

If we use $\frac{\delta}{\delta t}G(s,t) = \sum_{y=0}^{\infty}\frac{\mathrm{d}}{\mathrm{d}t}P_y(t)s^y$ and $\frac{\delta}{\delta s}G(s,t) = \sum_{y=0}^{\infty}yP_y(t)s^{y-1}$ it can be seen that:

$$\begin{aligned} \sum_{y=2}^{\infty}(y-1)P_{y-1}(t)s^y &= \sum_{m=1}^{\infty}(m)P_m(t)s^{m+1} = s^2\frac{\delta}{\delta s}G(s,t) \\ \sum_{y=1}^{\infty}yP_y(t)s^y &= s\frac{\delta}{\delta s}G(s,t) \\ \sum_{y=1}^{n_0}yP_y(t)s^y &= s\frac{\delta}{\delta s}G(s,t) \\ \sum_{y=0}^{n_0-1}(y+1)P_{y+1}(t)s^y &= \frac{\delta}{\delta s}G(s,t) \end{aligned}$$

and so (1.4) can be written as :

$$\begin{aligned}\frac{\delta}{\delta t}G(s,t) &= \lambda s^2\frac{\delta}{\delta s}G(s,t) - \lambda s\frac{\delta}{\delta s}G(s,t) - \mu s\frac{\delta}{\delta s}G(s,t) + \mu\frac{\delta}{\delta s}G(s,t)\\ &= \left[\lambda s(s-1) + \mu(1-s)\right]\frac{\delta}{\delta s}G(s,t)\\ &= (\lambda s-\mu)(s-1)\frac{\delta}{\delta s}G(s,t) \qquad (1.5)\end{aligned}$$

This is a partial differential equation for the probability generating function. The initial condition $p_{y_0}(0) = 1$ becomes: $G(s,0) = s^{y_0}$.

There are two cases to consider, the first being $\lambda \neq \mu$.

To remove the term $(\lambda s - \mu)(s-1)$, a change of variables can be used by putting:

$$\frac{\mathrm{d}s}{\mathrm{d}z} = (\lambda s-\mu)(s-1)$$

This separable equation can be integrated to give

$$\begin{aligned}z = \int dz &= \int \frac{ds}{(\lambda s-\mu)(s1-)} = \frac{1}{\lambda}\int \frac{ds}{(\frac{\mu}{\lambda}-s)(1-s)}\\ &= \frac{1}{\lambda-\mu}\int\left[\frac{1}{\frac{\mu}{\lambda}-s} - \frac{1}{1-s}\right]ds\\ &= \frac{1}{\lambda-\mu}\ln\left[\frac{1-s}{\frac{\mu}{\lambda}-s}\right], \quad (0 \leq s \leq min(1,\frac{\mu}{\lambda}) \qquad (1.6)\end{aligned}$$

so that s can be written as:

$$s = \frac{\lambda - \mu e^{(\lambda-\mu)z}}{\lambda - \lambda e^{(\lambda-\mu)z}}$$

Now let $Q(z,t) = G(s,t)$, with s given above, so that $Q(z,t)$ satisfies (using 1.5)

$$\frac{\delta}{\delta z}Q(z,t) = \frac{\delta}{\delta z}G(s,t) = \frac{\delta}{\delta s}G(s,t)\frac{\delta}{\delta z}s = (\lambda s-\mu)(s-1)\frac{\delta}{\delta s}G(s,t) = \frac{\delta}{\delta t}G(s,t) = \frac{\delta}{\delta t}Q(z,t)$$

Finally the partial differential equation for the pgf becomes:

$$\frac{\delta}{\delta z}Q(z,t) = \frac{\delta}{\delta t}Q(z,t)$$

The general solution of this equation is any differentiable function, ν say, of $z+t$, i.e. $Q(z,t) = \nu(z+t)$:

$$\frac{\delta}{\delta z}Q(z,t) = \frac{\delta}{\delta z}\nu(z+t) = \frac{\mathrm{d}}{\mathrm{d}(z+t)}\nu(z+t)\frac{\delta}{\delta z}(z+t) =$$
$$\frac{\mathrm{d}}{\mathrm{d}(z+t)}\nu(z+t) = \frac{\mathrm{d}}{\mathrm{d}(z+t)}\nu(z+t)\frac{\delta}{\delta t}(z+t) = \frac{\delta}{\delta t}Q(z,t)$$

The function ν is determined by the initial conditions. The initial population size is y_0, so by the initial condition

$$G(s,0) = s^{y_0} = \left\{\frac{\lambda - \mu e^{(\lambda-\mu)z}}{\lambda - \lambda e^{(\lambda-\mu)z}}\right\}^{y_0} = \nu(z) = Q(z,0)$$

Hence

$$G(s,t) = Q(z,t) = \nu(z+t) = \left\{\frac{\lambda - \mu e^{(\lambda-\mu)(z+t)}}{\lambda - \lambda e^{(\lambda-\mu)(z+t)}}\right\}^{y_0} = \nu(z)$$

From (1.6), it follows that

$$e^{(\lambda-\mu)z} = \frac{1-s}{\frac{\mu}{\lambda} - s} = \frac{\lambda(1-s)}{\mu - \lambda s}$$

and finally that the probability generating function is

$$G(s,t) = \left\{\frac{\mu(1-s) - (\mu - \lambda s)e^{-(\lambda-\mu)t}}{\lambda(1-s) - (\mu - \lambda s)e^{-(\lambda-\mu)t}}\right\}^{y_0} \tag{1.7}$$

Now we consider the case that the rates are equal: $\lambda = \mu$ In this case (1.5) becomes:

$$\frac{\delta}{\delta t}G(s,t) = \lambda(s^2-1)\frac{\delta}{\delta s}G(s,t)$$

Now $\frac{\mathrm{d}s}{\mathrm{d}z} = \lambda(s^2-1)$, and the change of variables is: $z = \frac{1}{\lambda}\int \frac{\mathrm{d}s}{(s-1)^2} = \frac{-1}{\lambda(s-1)}$ or $s = \frac{\lambda z - 1}{\lambda z}$. It follows that $\nu(z) = s^{y_0} = \left[\frac{\lambda z - 1}{\lambda z}\right]^{y_0}$. Therefore with equal birth and death rate, the probability generating function is

$$G(s,t) = \left[\frac{\lambda(z+t)-1}{\lambda(z+t)}\right]^{y_0} = \left[\frac{1+(\lambda t - 1)(1-s)}{1+\lambda t(1-s)}\right]^{y_0} \tag{1.8}$$

1.5 Lagrange Transform

Once the probability generating function has been found it has to be inverted to find the probability function. Lagrange gave two expansions which can be used to obtain probability distributions. Let $g(z)$ be a infinitely differentiable function such that $g(1) = 1$ and $g(0) \neq 0$. The function $g(z)$ does not need to be a probability generating function. Then the numerically smallest root $z = l(s)$ of the transformation $z = sg(z)$ defines a probability generating function $z = \psi(s)$ with the Lagrange expansion in powers of s as:

$$z = \psi(s) = \sum_{y=1}^{\infty} \frac{s^y}{y!} \left[D^{y-1}[g(z)]^y \right]_{z=0}$$

if $\left[D^{y-1}[g(z)]^y\right]_{z=0} \geq$ for all y and where the symbol D is used as shorthand for the differentiation operator: $\frac{\mathrm{d}^{y-1}}{\mathrm{d}z^{y-1}}$. The probability mass function, called the basic Lagrange function, then is $P(Y = y) = \frac{1}{y!}\left[D^{y-1}[g(z)]^y\right]_{z=0}$ (Consul and Famoye, 2006).

Let $f(z)$ be another successively differentiable function such that $f(1) = 1$ and $f(0) \neq 0$ and $\left[D^{y-1}[g(z)]^y f(\prime z)\right]_{z=0} \geq 0 \quad for\ y \in N$. The probability generating function of the discrete general Lagrange distribution under the transformation $z = sg(s)$ is given by $f(z)$

$$f(z) = f(\psi(s)) = \sum_{y=1}^{\infty} \frac{s^y}{y!} \left[D^{y-1}[g(z)]^y f'(z) \right]_{z=0}$$

The probability mass function then is:

$$\begin{aligned} P(Y = 0) &= f(0) \\ P(Y = y) &= \frac{1}{y!} \left[D^{y-1}[g(z)]^y f'(z) \right]_{z=0} \end{aligned} \tag{1.9}$$

Thus, if one can write a probability generating function as $f(\psi(s))$ with $\psi(s)$ a probability generating function generated by some function $g(z)$ (with $g(1) = 1, g(0) \neq 0$) and with $f(1) = 1$ and $f(0) \neq 0$. One can then use (1.9) to invert the probability generating function and obtain the probability mass function.

Now, if we define $\pi = \frac{\lambda e^{(\mu-\lambda)t} - \lambda}{\mu e^{(\mu-\lambda)t} - \lambda}$ and $\theta = \frac{\mu e^{(\mu-\lambda)t} - \mu}{\lambda e^{(\mu-\lambda)t} - \lambda}$, so $1 - \theta = \frac{\mu - \lambda}{\mu e^{-(\mu-\lambda)t} - \lambda}$ and $1 - \pi - \theta = \frac{\mu e^{(\mu-\lambda)t}}{\mu e^{(\mu-\lambda)t} - \lambda}$, then the probability generating function (1.7) can be written

as $G(t,s) = \left[\frac{\theta+(1-\pi-\theta)s}{1-\pi s}\right]^{y_0}$ since:

$$\begin{aligned} G(t,s) &= \left[\frac{\theta+(1-\pi-\theta)s}{1-\pi s}\right]^{y_0} \\ &= \left[\frac{\mu e^{(\mu-\lambda)t}-\mu+(\mu-\lambda e^{(\mu-\lambda)t})s}{\mu e^{(\mu-\lambda)t}-\lambda+(\lambda-\lambda e^{(\mu-\lambda)t})s}\right]^{y_0} \\ &= \left[\frac{e^{(\mu-\lambda)t}(\lambda s-\mu)-\mu(s-1)}{\lambda-\mu e^{(\mu-\lambda)t}+(\lambda e^{(\mu-\lambda)t}-\lambda)s}\right]^{y_0} \\ &= \left[\frac{e^{(\mu-\lambda)t}(\lambda s-\mu)-\mu(s-1)}{e^{(\mu-\lambda)t}(\lambda s-\mu)-\lambda(s-1)}\right]^{y_0} \end{aligned}$$

wich equals (1.7). The function $G(t,s)$ can be written as

$$\begin{aligned} G(t,s) &= \left[\frac{\theta+(1-\pi-\theta)s}{1-\pi s}\right]^{y_0} \\ &= \left[\theta+(1-\theta)\frac{(1-\pi)s}{1-\pi s}\right]^{y_0} \\ &= [\theta+(1-\theta)\psi(s)]^{y_0} \end{aligned}$$

The function $\psi(s)$ is the probability generating function of the geometric distribution. This can be seen from the basic Lagrange transform by taking $g(z) = 1-\pi+\pi z$, the probability generating function of a Bernoulli distribution.The function $f(z) = (\theta+(1-\theta)z)^{y_0}$ is the probability generating function of the binomial distribution.

We find that the probability generating function of the birth-death process can be written as the probability generating function of the binomial distribution, whose argument is the probability generating function of the geometric distribution. The latter can be obtained from the basic Lagrange by using the probability generating function of the Bernoulli distribution for $g(z)$.

1.6 Content of the thesis

The technique above was explained in some detail because it is used in Chapter 2 to derive the probability distribution function of the more general non-homogeneous birth-death process. The constant birth rate (reproductive power) and constant death rate give rise to exponential survival distributions for the time of reproduction and the time of death. These are replaced with more general survival functions in the non-homogeneous case. A natural parametric form for these distributions is discussed in

Chapter 2 and in Chapter 4. Using the survival distribution in chapter 2, three types of models are examined: a proportional rate model, an accelerated failure time model and a model combining both.The non-homogeneous birth-death model gives rise to the net reproduction ratio which is the rate by which an not yet removed infected reproduces. This net reproduction ratio is used as tool to study the avian influenza outbreak in the Netherlands in 2003.

In Chapter 3, a non-homogeneous martingale model is proposed to model volatility in a stochastic time series of count data wit a constant mean. The approach is derived from a general non-homogeneous birth-and-death process, by taking the reproductive power and the death rate in the non-homogeneous birth-death model to be equal. A model is obtained for which the expected value at a certain time point equals what has been observed at a previous time point. The variance, however, can depend on time. By comparing this model with the non-homogeneous birth-death model, it can be determined whether the observed data is compatible with a drift or that there is volatility with a constant mean in the data. It is shown in that chapter that the net reproduction ratio can be used as a additional tool.These models and procedures are illustrated with MRSA (Methicillin-Resistant Staphylococcus aureus) prevalence data registered since 2001 from three Acute Trusts of hospitals from the National Health Service in Great Britain.

In Chapter 4, two special cases of the non-homogeneous birth-death model are considered: the non-homogeneous death process and the non-homogeneous birth process. In both of these models the end of the epidemic is incorporated through a modification of the survival function with a final size parameter, in the same way as is done in long-term survival models. These models are applied to three outbreaks: The Dutch classical swine fever outbreak from 1997-1998, the foot and mouth disease outbreak in Great Britain from 2001 and the Dutch avian influenza (H7N7) outbreak from 2003.

In Chapter 5 the non-homogeneous birth model is applied to the transmission of avian influenza (H5N1) among poultry in Thailand. To determine the epidemiology of this viral infection and its relation to wild birds outbreaks in Thailand from 2004 through 2007, it was investigated how wild birds play a role in transmission. Areas where there was an outbreak were classified according to whether infected wild bird were found. This definition was then used as an independent variable in the model. A proportional rate regression model was formulated.

In Chapter 6 methods of estimating the reproductive power and the survival function with communicable events are discussed. Using a non-homogeneous birth process, one can estimate the reproductive power and then the survival function and their standard errors directly from the data using the (log-)likelihood, instead of using a

parametric form as is done in the other chapters. It is shown that the standard errors for the estimated reproductive power and for the estimated survival become smaller as time goes on, because with communicable events the amount of information increases with increasing time. Methods are developed to compare empirical estimates in two independent groups by means of the (log) reproduction power ratio. It is shown that the standard error of the log of the reproduction power ratio is usually increasing with increasing time, since this standard error depends on the size of the reproductive power. If these are small, then the standard error is large. These methods are applied to the Dutch avian influenza (H7N7) outbreak from 2003 and to data from avian influenza (H5N1) outbreaks among poultry in Thailand.

-

Chapter 2

Using epidemic prevalence data to jointly estimate reproduction and removal[1]

2.1 Introduction

The data-generating process of an epidemic has special characteristics to which one wants to pay particular attention when modeling these data. First, the observed data of an infectious disease outbreak are limited in the sense that the incidence – expressed as the number of newly infected individuals as a function of time – is usually measured by symptom onset of disease, which is sometimes further accompanied by reporting delay. Thus, all of the observed cases in the reported data represent those who experienced infection at some point in time in the past. Second, epidemic data sets do not usually include information on the number of susceptible individuals as a function of time, but solely records infected (interpreted as symptomatic) individuals. It is therefore unknown if susceptible individuals in the past are still susceptible at a point of time. Third, the susceptible population is usually not well defined at the beginning of an outbreak, and its size may vary with time due to time-dependency in contact behavior and public health countermeasures during the outbreak. In a veterinary context, the countermeasures might include a transportation ban during an infectious disease outbreak on animal farms. Control measures are taken not only

[1]Van Den Broek, J. and Nishiura H. (2009). Using epidemic prevalence data to jointly estimate reproduction and removal.
Annals of Applied Statistics Vol.3, No 4 1505-1520.

to reduce the number of contacts but also to limit the ability of infected individuals to generate secondary cases. For instance, one may think of preemptive culling in the case of an infectious disease outbreak on animal farms. Fourth, since the infection is transmitted from individual to individual, observation of an infected individual is not independent of observing other individuals.

These characteristics lead us to consider developing a method which appropriately captures the dynamics of a directly transmitted infectious disease by modeling the number of infected-and-detected individuals, preferably in discrete time, in order to quantify the reproduction of infected individuals in a non-homogeneous manner. This contrasts with other statistical models which measure the population of susceptibles and model the force of infection at which these susceptibles get infected.

As is usually assumed, one can think of a population in which an epidemic of an infectious disease occurs, as consisting of three groups (or sub-populations) of individuals. The first is the susceptible population which represents individuals who have not been infected yet but may experience infection in future. The second is a population of infectious individuals, which consists of those who have been infected and are infectious to others. The last group consists of removed or recovered individuals who are no longer infectious and may be immune or are removed from the population. The simplest type of the model which describes the transmission dynamics over time is referred to as an SIR (Susceptible-Infected-Recovered) model (Diekmann and Heesterbeek, 2000). Since the present study will concentrate on the number of infected, here we consider their dynamics alone. Let $x(t)$ and $y(t)$ be the susceptible and infectious fractions of the population at time t, respectively, the derivative of $y(t)$ is expressed as:

$$\frac{\mathrm{d}}{\mathrm{d}t}y(t) \;=\; \beta(t)\,y(t)\,x(t) - \mu(t)\,y(t) \tag{2.1}$$

Note that the transmission rate $\beta(t)$ and the removal rate $\mu(t)$ may depend on time. If one rewrites the product of the transmission rate $\beta(t)$ and susceptibles $x(t)$ as $\lambda(t)$ $(= \beta(t)x(t))$, then the equation is rewritten as:

$$\frac{\mathrm{d}}{\mathrm{d}t}y(t) \;=\; \lambda(t)\,y(t) - \mu(t)\,y(t) \tag{2.2}$$

which is the equation of the deterministic non-homogeneous birth-death process. The function $\lambda(t)$ has been referred to as the reproductive power and can be interpreted as the rate at which a single infected individual is able to generate secondary cases (Kendall, 1948). In other words, $\lambda(t)$ is the rate at which an infected individual is

able to reproduce itself. The so-called death rate $\mu(t)$ in a birth-death process is interpreted as the rate at which an infected individual is removed from the sub-population of infected individuals. It should be noted that equation (2.2) relaxed the definition of $y(t)$ compared to that in (2.1). Namely, whereas $y(t)$ in (2.1) has to be infectious to others, we can instead regard $y(t)$ in (2.2) as infected-and-detected individuals (i.e. regardless of infectiousness).

One of the advantages of using this simple equation is that the population of susceptibles is allowed to vary over time. Therefore, the reproductive power varies over time due to two different reasons: (1) the population of susceptibles $x(t)$ varies as a function of time and (2) the transmission rate $\beta(t)$ is non-homogeneous over time. In addition, the removal rate $\mu(t)$ is allowed to vary with time.

It can be an advantage to model infection process stochastically, because one can explicitly define the probability of transmission, rather than deterministicaly stating if the transmission happens (Andersson and Britton, 2000). A stochastic model can describe not only the quantitative patterns of observation with time-dependent expected values, but also offer standard errors of the parameters without making adhoc distributional assumptions. More importantly, the likelihood function can be explicitly derived, which will be useful for statistical inference of parameters and critical assessments of the modeling method. Moreover, such a stochastic process can model the number of infected over time as being dependent.

The present study aims to develop a stochastic model which is based on a non-homogeneous birth-death process. The model is applied to an observed data set of infected-and-detected (but not yet removed) cases, permitting reasonable assessment of the time course of an epidemic. In Section 2, the stochastic version of the non-homogeneous birth-death model is described. A novel analytical solution of the model is obtained with the use of a general Lagrange transformation, derivation of which has not been explicitly discussed to date. In Section 3, we discuss a conditional discrete-time fitting method. The present study is the first to apply the technique to a model where both birth and death rates are non- homogeneous. Depending on the model and the data given, the number of parameters to be estimated can be large and it might be difficult to get stable estimates. Besides, a relationship between the reproductive power and the removal rate might exist. Therefore, more restricted models which employ this relationship, are discussed in section 4. In Section 5, our model is applied to an observed epidemic data set of avian influenza A (H7N7) in the Netherlands, 2003.

2.2 The non-homogeneous birth-death model

The stochastic differential equations for a non-homogeneous birth-death process, with $Y(t)$ and y_0 being the number of infected-and-detected at time t and the initial number of infected-and-detected at time 0, respectively, are:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}p_y(t) &= \lambda(t)(y-1)p_{y-1}(t) + \mu(t)(y+1)p_{y+1}(t) - (\lambda(t)+\mu(t))\, y p_y(t)\\ \frac{\mathrm{d}}{\mathrm{d}t}p_0(t) &= \mu(t)p_1(t)\end{aligned}$$

where $\lambda(t)$ denotes the reproductive power, $\mu(t)$ the death rate and $\Pr(Y(t)=y) = p_y(t)$ the probability that the number of infected-and-detected individuals at time t is y. It should be noted that here we consider $Y(t)$ as the number of infected-and-detected individuals which represents the observed elements of the data and is irrelevant to infectiousness. If we take the probability generating function of the probabilities p, we can derive a partial differential equation for this fraction by multiplying the above differential equations with z^y, summing the result, and taking the derivative. The analytical solution of the partial differential equation is (Kendall, 1948):

$$\Phi(y,u) = \left[\frac{\theta + (1-\pi-\theta)u}{1-\pi u}\right]^{y_0}$$

If we write $\rho(t) = \int_0^t [\mu(\tau) - \lambda(\tau)]\,\mathrm{d}\tau = \log\left[\frac{S_\lambda(t)}{S_\mu(t)}\right]$ and

$$\gamma(t) = \int_0^t e^{\rho(\tau)}\lambda(\tau)\mathrm{d}\tau = -\int_0^t \frac{\mathrm{d}S_\lambda(\tau)}{S_\mu(\tau)} \tag{2.3}$$

we get

$$\theta = 1 - \frac{e^{-\rho(t)}}{1+\gamma(t)e^{-\rho(t)}} = 1 - \frac{S_\mu(t)}{S_\lambda(t)+\gamma(t)S_\mu(t)} \tag{2.4}$$

$$\pi = 1 - \frac{1}{1+\gamma(t)e^{-\rho(t)}} = 1 - \frac{S_\lambda(t)}{S_\lambda(t)+\gamma(t)S_\mu(t)} \tag{2.5}$$

It should be noted that $S_\lambda(t) = e^{-\int_0^t \lambda(\tau)\mathrm{d}\tau}$ is the reproduction survival function and $S_\mu(t) = e^{-\int_0^t \mu(\tau)\mathrm{d}\tau}$ is the removal survival function.

To obtain the probability distribution from $\Phi(y,u)$, the general Lagrange transformation is useful (the details of which can be found elsewhere (Consul and Famoye, 2006)). First, let $\Phi(y,u) = \left[\theta + (1-\theta)\frac{(1-\pi)u}{1-\pi u}\right]^{y_0} = [\theta + (1-\theta)\psi(u)]^{y_0}$ where $\psi(u)$ is the probability generating function (pgf) of the geometric distribution. Second, let $g(z) = 1 - \pi + \pi z$, the pgf of a Bernoulli distribution. Numerically, the smallest root of the transformation $z = ug(z)$ defines a pgf $z = \psi(u) = \frac{(1-\pi)u}{1-\pi u}$ (Consul and Famoye, 2006). Third, additionally considering $f(z) = (\theta + (1-\theta)z)^{y_0}$, the pgf of the discrete general Lagrange probability distribution under the Lagrange transform $z = ug(z)$ is given by $f(z) = f(\psi(u)) = \left[\theta + (1-\theta)\frac{(1-\pi)u}{1-\pi u}\right]^{y_0}$ and, moreover, the probability mass function is a special case of the double-binomial distribution (Consul and Famoye, 2006, page 22-27) which can be referred to as the Bernoulli-binomial Lagrangian distribution in the terminology of (Johnson, Kemp and Kotz, 2005):

$$\begin{aligned} P(Y(t)=0) &= \theta^{y_0} \\ P(Y(t)=y) &= \frac{y_0}{y}\theta^{y_0}\pi^y \sum_{k=0}^{\min(y-1,y_0-1)} \binom{y_0-1}{k}\binom{y}{y-k-1}\left[\frac{(1-\pi)(1-\theta)}{\pi\theta}\right]^{k+1}, \\ & \quad y \geq 1 \end{aligned} \tag{2.6}$$

where θ and π are defined in equations (2.4) and (2.5). It should be noted that $g(z)$ and $f(z)$ are pgf's, and thus, the necessary conditions for Lagrange transformation are satisfied. To the best of our knowledge, detailed derivation of the equation (2.6) has never been discussed in the context of the non-homogeneous birth-death model (see Discussion).

The expectation and the variance of (2.6) are:

$$\begin{aligned} E(y) &= y_0\frac{1-\theta}{1-\pi} \\ &= y_0\frac{S_\mu(t)}{S_\lambda(t)} \\ var(y) &= y_0\frac{(1-\theta)[\theta+\pi(1-\theta-\pi)]}{(1-\pi)^3} \\ &= y_0\frac{S_\mu(t)}{S_\lambda(t)}\left[1+(2\gamma-1)\frac{S_\mu(t)}{S_\lambda(t)}\right] \\ &= y_0R(t)\left[1+(2\gamma-1)R(t)\right] \end{aligned}$$

by using (2.4) and (2.5).

The expected value has two interpretations. The first part is the predicted number

of infected individuals at time t_0 who survived removal (i.e. $y_0 S_\mu(t)$). The second part, $\frac{1}{S_\lambda(t)}$, measures the rate at which a non-removed infected individual reproduces itself. This is similar to an interpretation of a non-homogeneous birth process (chapter 4); the difference of the present study from the previous non-homogeneous birth process is that in the present set up only the predicted non-removed infected-and-detected individuals reproduce. Secondly, the ratio of the survival fractions $\frac{S_\mu(t)}{S_\lambda(t)}$ is the net reproduction ratio with which an infected individual reproduces itself, which is interpreted as the effective reproduction number $R(t)$ as a function of epidemic time t. $R(t)$ in the present study can be regarded as the average number of secondary cases generated by a single primary case at time t. That is, our $R(t)$ is an instantaneous measure of secondary transmissions occurring at time t, whose definition is equivalent to the period total fertility rate in mathematical demography (Nishiura and Chowell, 2009). If $R(t) < 1$, it suggests that the epidemic is in decline and may be regarded as being 'under control' at time t (vice versa, if $R(t) > 1$). It should be noted that the expected value of (2.6) is equivalent to an analytical solution of the deterministic version of non-homogeneous birth-death process (2.2).

The term $\gamma(t)$ in the formula for the variance represents the dependence between the birth and death rate as can be seen from (2.3). As it is clear from the analytical expression for the variance, the variance becomes large if the probability of non-removal is large for an infected individual, if the probability of reproduction is large, or both. This matches intuitive sense. In addition, the variance can be regarded as a type of negative binomial variance, if we rewrite it as $var(y) = E(y)[1 + \frac{2\gamma-1}{y_0}E(y)]$.

2.3 Fitting the model

We have shown how the epidemic data can be generated by a stochastic non- homogeneous birth-death process. Nevertheless, the observed data are, in reality, just one sample path of all the possible sample paths that can arise from such an epidemic process. Considering further that the number of infected-and-detected and individuals at a certain point in time t depends on the number of infected-and-detected individuals at some time point before t, our model is fitted to the data by conditioning on the transmission dynamics which happened before t (Becker, 1989; Becker and Yip, 1989). Moreover, as we briefly discussed in the introduction, another important point in practice is the discrete nature of the time points of observation, say $t_j,\ j = 0, 1, \cdots, n$, where the time unit might typically be days or weeks. Therefore, the number of infecteds at time t_j is modelled conditionally on the number of infecteds by time t_{j-1}.

Since the probability mass function (2.6) is a conditional probability mass function, conditioning being on the number of infecteds at t_0, this can be effectively used as the conditional model for the number of infecteds at time t_j, given the number of infecteds at time t_{j-1}. For this reason, the survival distributions $S_\lambda(t)$ and $S_\mu(t)$ and, of course, $\gamma(t)$, are also conditioned on the past. Let T_λ and T_μ be the stochastic variables that measure the reproduction time and the removal time, respectively. The conditional survival probability for the reproduction time is:

$$\begin{aligned} P(T_\lambda > t_j | T_\lambda > t_{j-1}) &= \frac{S_\lambda(t_j)}{S_\lambda(t_{j-1})} \\ &= 1 - \frac{S_\lambda(t_{j-1}) - S_\lambda(t_j)}{S_\lambda(t_{j-1})} \\ &= 1 - P(T_\lambda \in (t_{j-1}, t_j] | T_\lambda > t_{j-1}) \\ &= 1 - h_\lambda(t_{j-1}) \end{aligned}$$

where $h_\lambda(t_{j-1})$ is interpreted as the discrete reproductive power. Similarly, the conditional survival probability for the removal time is given by

$$P(T_\mu > t_j | T_\mu > t_{j-1}) = 1 - h_\mu(t_{j-1})$$

where $h_\mu(t_{j-1})$ is the discrete removal hazard.

The discrete conditional version of (2.3) in the time interval $(t_{j-1},\, t_j]$ is:

$$\frac{P(T_\lambda \in (t_{j-1}, t_j] | T_\lambda > t_{j-1})}{P(T_\mu > t_j | T_\mu > t_{j-1})} = \frac{h_\lambda(t_{j-1})}{1 - h_\mu(t_{j-1})}$$

When these conditional discrete measurements are considered, θ and π correspond to $h_\mu(t_{j-1})$ and $h_\lambda(t_{j-1})$, respectively. Using these, the conditional probability for (2.6) is expressed as:

$$P(Y(t_j) = 0 | Y(t_{j-1} = y_{t_{j-1}}) = h_\mu(t_{j-1})^{y_0}$$

$$P(Y(t_j) = y_{t_j} | Y(t_{j-1}) = y_{t_{j-1}}) = \frac{y_{t_{j-1}}}{y_{tj}} h_\mu(t_{j-1})^{y_{tj-1}} h_\lambda(t_{j-1})^{y_{tj}} \times$$

$$\sum_{k=0}^{\min(y_{t_j}-1, y_{t_{j-1}}-1)} \binom{y_{t_{j-1}} - 1}{k} \binom{y_{t_j}}{y_{t_j} - k - 1} \left[\frac{(1 - h_\lambda(t_{j-1}))(1 - h_\mu(t_{j-1}))}{h_\lambda(t_{j-1}) h_\mu(t_{j-1})} \right]^{k+1} \quad y \geq 1$$

The expected value of this probability is $y_{t_{j-1}} \frac{1-h_\mu(t_{j-1})}{1-h_\lambda(t_{j-1})}$, which is referred to as sample path profile (Lindsey, 2001). The corresponding conditional measurement of the

effective reproduction number is $R(t_{j-1}) = \frac{1-h_\mu(t_{j-1})}{1-h_\lambda(t_{j-1})}$ with $h_\mu(t_{j-1}) = 1 - \frac{S_\mu(t_j)}{S_\mu(t_{j-1})}$ and $h_\lambda(t_{j-1}) = 1 - \frac{S_\lambda(t_j)}{S_\lambda(t_{j-1})}$.

Let Δ be the vector of parameters from the survival distributions (see section 4). The log-likelihood function is then:

$$l(\Delta) = \sum_{i=1}^{n} \log \left[P(Y(t_j) = y_{t_j} | Y(t_{j-1}) = y_{t_{j-1}}, \Delta) \right]$$

Note that the likelihood is evaluated only for $y_{t_j} > 0$, because zero prevalence is the absorbing state of the process and this state is not observable in reality.

The log-likelihood can be maximized using an optimization procedure, such as the Nelder-Mead method to find the maximum likelihood estimates. In our exercise, the software system R (R Development Core Team, 2010) is used. The information matrix is used to find the standard errors of the parameters, and we use Akaike's information criterion (AIC) to compare model fits.

2.4 The Burr distribution and its special cases

To choose a particular form for the survival functions one might take the early phase of the outbreak into account. The mean of the probability mass function (2.6) depends on these survival functions and is the same as the solution of (2.2). Since (2.2) can be derived from the SIR model, one might look at the deterministic SIR-model to decide the parametric form of the survival function. In the early phase of the outbreak a deterministic SIR-model can be well approximated by an deterministic SI-model since in that phase the number of removals is limited. The dynamic equations for this SI-model hold for the fraction of susceptibles and for the fraction of infected and since the fraction of susceptibles at a point of time is the same as the fraction individuals with infection time larger than that time point, the dynamic equations should also holds for the survival function. The Burr family of distribution functions has precisely this property, see for more details section 4.2.3.

When detection of a symptomatic infected individual occurs, he/she usually will be removed immediately. Thus, it is reasonable to assume that the reproductive power and the removal rate have a similar structure, and follow a similar survival function.

The most well known and useful distribution from the Burr family is the Burr XII, or Singh-Maddala distribution, which is sometimes referred to simply as the

Burr distribution in literature. The survival function is given by:

$$S(t) = \left[1 + \left(\frac{t}{b}\right)^{a}\right]^{-q}, \quad t > 0,\ a, b, q > 0$$

The right tail is governed by the parameters a and q, the left tail by a, and b is the scale parameter (Kleibner and Kotz, 2003, page 198). To reduce the number of parameters to be estimated, one can consider three special cases of the Burr distribution (Kleibner and Kotz, 2003) :

1. The logistic form is obtained for $q = 1$ giving the log-logistic or the Fisk distribution.

2. For $a = 1$, the Burr distribution is reduced to the Lomax (Pareto type II) distribution.

3. The case $a = q$ is also known as the para-logistic distribution.

The Weibull distribution and the Pareto distribution are limiting cases of the Burr distribution (Shao, 2004). An interesting way to arrive at the Burr distribution is to assume that the times follow a Weibull distribution, the scale parameter of which follows an inverse generalized gamma distribution (Kleibner and Kotz, 2003).

Another way to reduce the number of model parameters, and to find a further relationship between the reproductive power and the removal rate, one may rewrite the Burr distribution as a proportional rate or an accelerated event-time distribution (just as in the case of the more famous Weibull distribution). Suppose that the survival function for the reproduction time is a Burr distribution with parameters a, b_1 and q and suppose that the survival function of the removal time is also a Burr distribution with parameters a, b_2 and q. If we replace b_2 by db_1 (where d is a constant), we get

$$S_\mu(t) = \left[1 + \left(\frac{t}{b_2}\right)^{a}\right]^{-q} = \left[1 + \left(\frac{t}{db_1}\right)^{a}\right]^{-q} = \left[1 + \left(\frac{t'}{b_1}\right)^{a}\right]^{-q} = S_\lambda(t')$$

Therefore, the survival distribution for the removal times is exactly the same as that for the reproduction time, except that the removal time is interpreted as accelerated reproduction time.

To employ the proportional rate model, let the survival distribution for the reproduction time be a Burr distribution with parameters a, b and q_1, and suppose that the

survival distribution of the removal times is also a Burr distribution with parameters a, b and q_2. If we replace q_2 by cq_1 (where c is a constant), we get

$$S_\mu(t) = \left[1 + \left(\frac{t}{b}\right)^a\right]^{-q_2} = \left[1 + \left(\frac{t}{b}\right)^a\right]^{-cq_1} = \left\{\left[1 + \left(\frac{t}{b}\right)^a\right]^{-q_1}\right\}^c = \{S_\lambda(t)\}^c$$

indicating that the rates are proportional; i.e., the rate at which an infected-and-detected is removed is proportional to the rate at which an infected-and-detected individual reproduces. Of course, the Burr distribution can also be written as both an accelerated event-time distribution and as a proportional rate distribution; i.e., $S_\mu(t) = [S_\lambda(t')]^c$

All the distributions described above can be written in accelerated event-time form and in proportional hazard form, except for log-logistic distribution which has only an accelerated event-time form. In the next section, we fit all of those models to epidemic data of avian influenza A (H7N7) in the Netherlands, 2003.

2.5 Dutch avian influenza A (H7N7) epidemic in 2003

Here we show an example of our model application to an observed dataset. An epidemic of avian influenza A (H7N7) virus started on February 28, 2003 in the Gelderse Vallei in the Netherlands. In total, 239 flocks experienced infection with known detection date. Control measures taken include movement restrictions, stamping out of infected flocks, and pre-emptive culling of flocks in the neighborhood of infected flocks. As the result, 1,255 commercial flocks and 17,421 flocks of smallholders had to be depopulated, and approximately 25.6 million animals were killed. The virus was also transmitted to humans who had been in close contact with the infected chickens, resulting in one human death. Further details can be found elsewhere (Stegeman et al., 2003).

We examine transmission and detection events between flocks. We regard the detection date of a case (i.e. infected individual) as the date at which there were first signs of infection in a flock. In other words, the detection date of an infected individual is regarded as the birth date in our model. Therefore, the birth date is not the date of infection but the date at which an infected farm is detected, which is used as a surrogate. Moreover, the date of depopulation is regarded as the death date. Consequently, the Dutch data consist of the prevalence of infected-and-detected (but not yet removed) flocks on each epidemic date.

Figure 2.1 shows the temporal distribution of the prevalent cases (representing those who were born and have not been removed yet). As can be seen, the right tail contains gaps and the center of the distribution is not well determined. Usually these make it difficult to fit simple models to the data.

The Burr model with three parameters for birth rate and three for death rate was fitted to the observed data. We refer to it as the full Burr model. To objectively show how this model fits to the data better than other types of model, we compared its likelihood with that of the full inverse Burr (Burr III or Dagum) model. The inverse Burr model is also known as a flexible distribution from the Burr family and can be viewed as a generalized gamma with a scale parameter that follows a inverse Weibull distribution (Kleibner and Kotz, 2003). The inverse Burr yielded an AIC of 363.7, while the full Burr model yielded 342.4. We thus examined the full Burr model (and its special cases) for further analyses. The AICs for these different models are in Table 2.1. The full Lomax model gave the best fit, although the difference in AIC with the full Burr distribution was not particularly large. In other words, the information criterium suggest that both the death rate and the reproductive power may be proportional and that the removal time may be accelerated reproduction time in the observed dataset. Figure 2.2 visually confirms a good fit of the full Lomax model to the observed number of infected-and-detected individuals based on the conditional model in discrete time. The model seems to have well captured the observation, because fitting prevalence, y_{t_j} to the data is conditioned on $y_{t_{j-1}}$. It should be noted that the predicted values in Figure 2.2 reflect the qualitative pattern of the observed data always one or two steps late, which is a general tendency of a conditional fit.

Parameter estimates for the best fitting model are shown in Table 2.2 with their standard errors. The logarithm of the acceleration factor, d, was estimated at 1.47 with a standard error of 0.369, and the logarithm of the proportionality between the reproductive power and the removal rate was 1.54 with a standard error of 0.355. The estimates of mean reproduction and removal times for the Lomax distribution can be calculated from (Kleibner and Kotz, 2003): $E(t) = \frac{b\Gamma(2)\Gamma(q-1)}{\Gamma(q)}$. The mean reproduction time was estimated as 1.81, indicating that it takes on average 1.8 days for a detected case to reproduce another detected case. The mean removal time was estimated as 2.79 indicating that on average it takes 2.8 days for a detected infected case to be removed.

In Figure 2.3, the rate at which a single case survives removal, $(1 - h_\mu(t_{j-1}))$, is shown as a function of epidemic date. In addition, the rate at which a single non-removed case reproduces secondary cases, $(1/(1 - h_\lambda(t_{j-1}))$, is also shown in the figure. The product of these two functions jointly yields $R(t_{j-1})$ (also shown in Figure 2.3). Compared with other modeling results (e.g. (Nishiura and Chowell, 2009)), our

estimates of $R(t_{j-1})$ is smoothed as a function of time owing to our parametric model for the survival functions of reproduction and removal. Nevertheless, it should be noted that our approach does not have to assume that the generation time distribution is known (i.e. common assumption in estimating $R(t)$), because our approach does not have to translate a growth rate of incidence to the reproduction number. As we discussed above, $R(t_{j-1}) < 1$ suggests that the epidemic is in decline at time t_{j-1} (vice versa, if $R(t_{j-1}) > 1$). This can be understood by considering the condition for $R(t_{j-1}) = 1$; i.e., the reproductive power becomes equivalent to the removal rate at time t. In our example, the expected value of $R(t_{j-1})$ declined below unity for the first time on day 23 since the detection of index case, supporting eventual end of the epidemic in the later stage. The sawtooth at the end of the lines is considered to have been caused by zero prevalence during the corresponding time period (i.e. because of our conditional measurement, the survival functions reflect small variations in the observed data). Figure 2.3-B shows the estimated effective reproduction number $R(t_{j-1})$. To get an idea of the statistical uncertainty of $R(t_{j-1})$, 500 sample paths were drawn from the estimated non-homogeneous birth-death model and for each sample path the effective reproduction number was calculated (gray lines). The 95% percentile lines of $R(t_{j-1})$ are also shown.

2.6 Discussion

In the present study, we modelled an epidemic based on the non-homogeneous birth-death process, addressing some of the critical issues which are seen in the observation of directly transmitted infectious diseases. First, we modelled infected-and-detected individuals, which corresponds to observable and countable information in practice (e.g. our model requires neither susceptibles nor infectious individuals). Second, for a similar reason, the application of a birth-death process allowed the population at risk (i.e. the susceptible population) to vary with time. Third, applying the concepts of non-homogeneous birth-death process to epidemic modelling, dependent events (i.e. dependence of a single infected individual on other infected individuals) were addressed in the model. Fourth, our stochastic model offered an explicit likelihood function and yielded standard error of parameters. Lastly, our model allowed estimation of the effective reproduction number, $R(t)$. Although a different probability distribution was given by (Bailey, 1990), the derivation was not given in the literature, and, to the best of our knowledge, equation (2.6) is the first to derive the pdf explicitly.

In chapter 4 a non- homogeneous birth process is applied to epidemic data, in

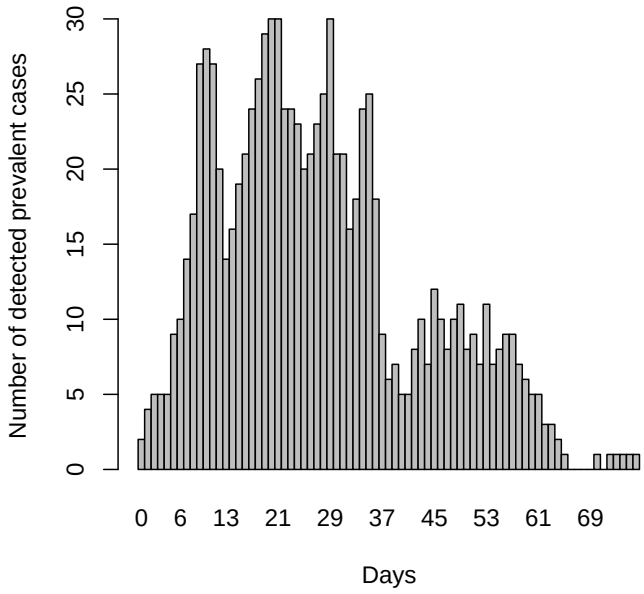

Figure 2.1: Temporal distribution of the prevalent cases of avian influenza A (H7N7) epidemic in the Netherlands, 2003. The index case was reported on February 28, 2003 (and the date is defined as day 0). The prevalent cases represent those who have been infected and detected but not been depopulated yet at day t, which correspond to the expected value of y_t in Section 3. See Stegeman et al. (2003) for further information.

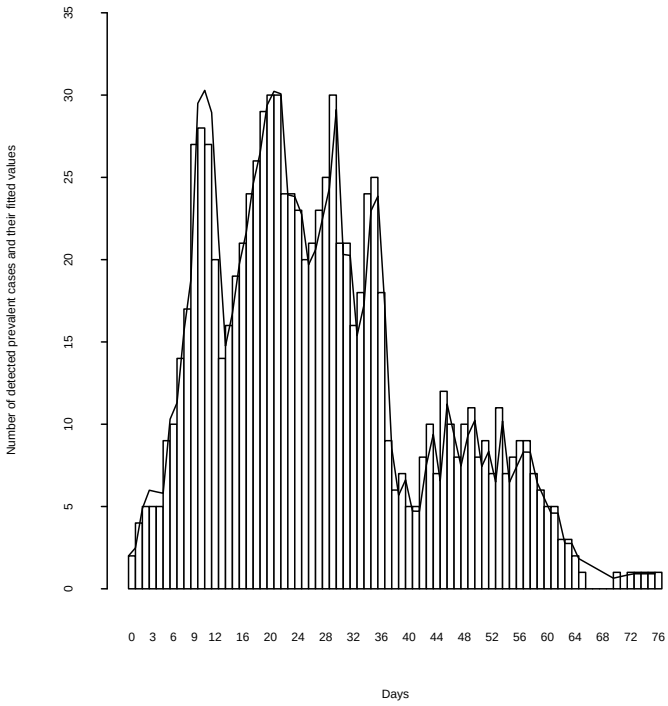

Figure 2.2: Comparison of the observed numbers and the predicted values from the conditional model of prevalent cases of avian influenza A (H7N7) epidemic in the Netherlands, 2003. The index case was reported on February 28, 2003 (and the date is defined as day 0). Observed data (bars) is compared with the predicted number of cases (solid line) based on the full Lomax model. It should be noted that the expectation of prevalence, y_{t_j} is conditioned on $y_{t_{j-1}}$.

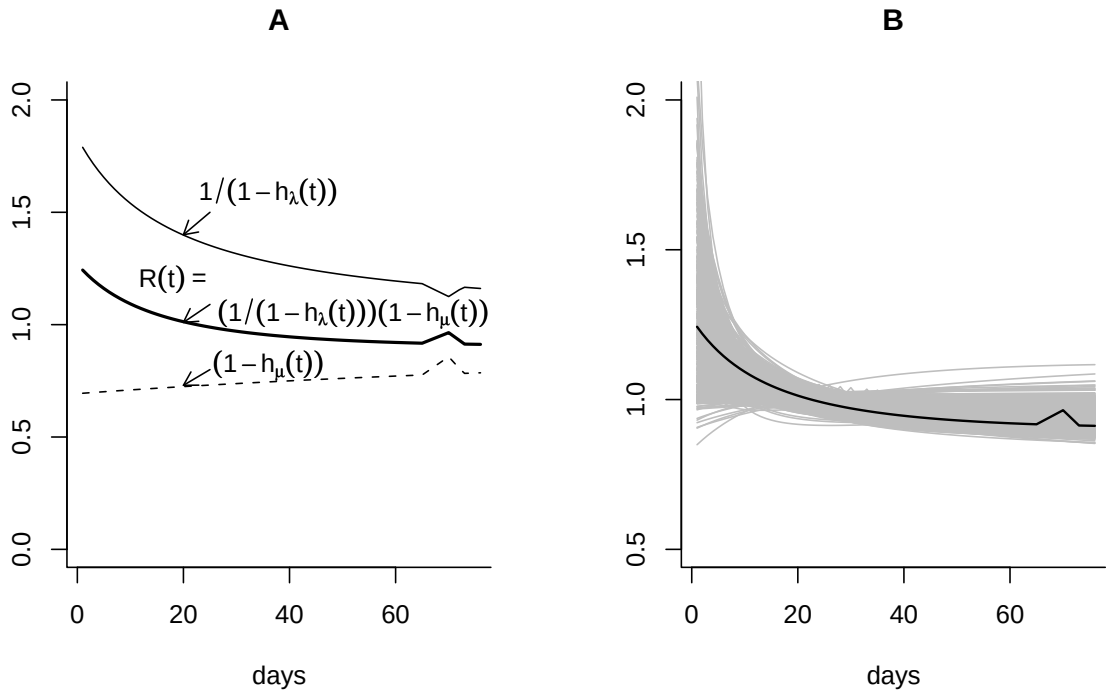

Figure 2.3: Time dependency of birth and death rates which jointly yield the effective reproduction number. **A.**$((1 - h_\mu(t_{j-1}))$ indicates the rate at which an infected-and-detected case escapes removal, whereas $(1/(1 - h_\lambda(t_{j-1})))$ denotes the rate at which a single infected-and-detected (but not yet removed) case reproduces secondary cases. The quotient, $(1 - h_\mu(t_{j-1}))\ /(1 - h_\lambda(t_{j-1}))$ yields the effective reproduction number $R(t)$ as a function of time, which can be interpreted as the average number of secondary cases generated by a single primary case at time t_{j-1}. If $R(t_{j-1}) < 1$, it suggests that the epidemic is in decline. In our example, the expected value of $R(t_{j-1})$ declined below unity on day 23 since the detection of index case. **B.** The effective reproduction number, $R(t_{j-1})$ calculated from sample paths drawn from the underlying estimated process (gray lines) with the estimated value (black line)and the 95% percentile lines (dashed lines).

which the survival-time distribution is modified by a final-size parameter to describe the end of an epidemic (which was influenced by public health countermeasures). The countermeasures would not only reduce the final size of an epidemic but also the reproductive power as a function of time, because secondary transmissions caused by infected individuals are restricted under the control measures. The non-homogeneous birth process in the previous study permitted an explicit assessment of the time variations in the number of newly infected individuals (and thus the reproductive power).

In the present study, the proposed non-homogeneous birth-death process further improved our understanding of time dependency by explicitly adding the non-homogeneous removal rate. Whereas the reproductive power changes as a function of time due to variations in susceptible individuals or in the transmission rate, the removal rate is also non-homogeneous when infected individuals are likely to be removed upon detection. Thus, adding non-homogeneous death to a non-homogeneous birth process enabled us to separately consider the effectiveness of countermeasures as a reduction in reproductive power (e.g. reduction in infectious contacts) and an increase in removal rate of infected individuals (e.g. culling of infected farms). In this way, the fading out of an epidemic was modeled in a smoother way, as compared to the previous model based on a non-homogeneous birth process alone.

Moreover, it should be noted that our model does not necessarily require a homogeneous mixing assumption to describe contacts, because our assumptions of the time-dependent rates implicitly include those non-homogeneities. For example, our model allows time variations in susceptible individuals. Nevertheless, our model only accounts for the non-homogeneity with respect to time in an explicit manner, and understanding other heterogeneous aspects of transmission requires further information.

Of course there are many possible candidate distributions to model the reproduction and the removal times. We have selected the Burr distribution for three reasons:

1. As noted in section 4, the Burr family coincides with our analytical understanding of the epidemic modeling, especially at the early stage of an epidemic.

2. Essentially, the Burr distribution is flexible and has special and limiting cases. For instance, the distribution can be regarded as a Weibull distribution with a random scale parameter.

3. In the context of the present study, the proportional hazard rate and the accelerated event time interpretation may be very helpful in model reduction and futher interpretation of the data.

Table 2.1: Akaike's information criterion (AIC) for different models.

	Full model	Acc. event time[1]	Prop. rate[2]	Both acc. event and prop rate
Burr	342.4	345.90	346.4	342.7
Log-logistic	371.3	369.8	–	–
Lomax	341.9	344.8	344.9	same as full model
Para-logistic	351.8	354.0	357.5	same as full model

As an example, the non-homogeneous birth-death model was applied to epidemic data of avian influenza A (H7N7) in the Netherlands, 2003, showing that the model fitted to the data very well. Indeed, since the dataset has information on both birth and death events for each individual case, the Dutch data appeared very useful for fitting prevalence data and applying our modelling method. Even in the presence of gaps in the right tail of the epidemic curve, and even though the centre was not well determined, our model reasonably described the time-course of the observed epidemic. In particular, our model permitted an estimation of the effective reproduction number $R(t)$ as a function of time without imposing a specific distribution of the generation time.

As is often the case with natural outbreaks, a single observation represents just one sample path from the process for which the above-mentioned model is imposed as the generator. There is no random sampling of infectious disease outbreaks, and a repeated sampling interpretation for the resulting model fit might be difficult. In other words, the description and conclusions arising from analysis of a single outbreak data set is valid only for that outbreak (see section 1.3.2 for a discussion). To find some general disease-specific conclusions from such an exercise, we stress that it is important to analyze several different outbreaks for the same disease. For such a purpose, one may use our model to accumulate the experience of applying our method to several outbreaks.

[1] Accelerated event-time model
[2] Proportional rate model

Table 2.2: Parameter estimates for the Lomax distribution.

Parameter	Estimate	St.Error
$\ln(b_1)$	3.235	0.5998
$\ln(q_1)$	2.712	0.3611
$\ln(b_2)$	4.987	1.0396
$\ln(q_2)$	3.980	0.8480

Chapter 3

Modeling volatility using a non-homogeneous martingale model for processes with constant mean on count data[1]

3.1 Introduction

Count data, such as the population size of some biological species, or the number infected by some infectious disease in a population, are often regulated by rich underlying mechanisms, and as a consequence can reveal complex behavior. There are, for example, situations in which it is difficult to determine in a time series of counts whether there is upward or downward change in the mean, or whether there is just some random fluctuation around a constant value, which only seems to show a trend. In situations were there appears to be curvature, in the case of population data for example because of density dependence, distinguishing between the situations might be even more difficult. This 'turbidity' can occur because the variance of the process might not be constant over time so that, for instance, an increase in variance might generate several relatively high or low values in a row, suggesting a upward or downward change of the mean. This variability in a measurement of interest over time is often, especially in the economic literature, referred to as volatility.

Volatility is used in the economic literature for the constant diffusion coefficient in a geometric Brownian motion when modeling the (log) outcome of a financial in-

[1] submitted/ in revision

strument over time, such as with the famous Black-Scholes model. However, the idea of constant volatility in such models is usually replaced by one where volatility can change in time, since in many cases this was what the data suggests (Lewis, 2000). Stochastic volatility models were developed that deal with this changing volatility. So volatility in this literature refers to the variability, usually measured in terms of standard deviation, of time series that have a continuous state space. Modeling volatility is seen in the economic literature as being of vital importance to obtain reasonable models for financial instruments.

Volatility is also defined as the occurrence of very extreme values of a certain outcome over a period of time (Lindsey, 2004). The term then refers to a change in variance over time. This might be modeled with heavy tail distributions and by allowing the dispersion to depend on the previous values of the variability.

This 'increase in variance' interpretation of volatility is also seen in the environmental literature, for instance in climate research. There it is used as an indication that more extreme outcomes in climate measurements occur (Ahmed, 2009). In climate studies it has also been recognized, for some time now, that volatility plays an important role in understanding climate change (Katz and Brown, 1992).

Modeling volatility usually means modeling the change in variability in time. For count models with a time changing mean, the model implicitly accounts for changing volatility, since with these models there is a relationship between the mean and the variance, and if the mean is changing in time (e.g. drift), then so is the variance, and thus the volatility. A good example is the much used Poisson model for count data, for which the standard deviation is $\sqrt{\mu_t}$, were μ_t is the mean at time t. In such cases, when the change in volatility comes with the change in time of the mean of the process, the question might be whether or not the model describes the change in volatility adequately. If not, a version of the model can be used that is either over- or under dispersed.

This leaves open the case for stochastic processes that have counts as their state space, were the mean of the process is constant, but were there might nevertheless be a change in volatility. The case of a model with only a time trend for the mean and not for the variance is usually not a valid case for count data.

In foregoing modeling studies of data from such a stochastic process with constant mean, the population size or the number of infected-and-immune individuals in a population in the case of infectious diseases, tended to be described by deterministic models and, to estimate the parameters of the model, rather ad-hoc distributional assumption are made. But, it is not always realized that an ad-hoc choice of a distribution also means an ad hoc choice for the formula describing the variance of the data. This paper proposes a new non-homogeneous martingale model for (changing)

volatility in count data with constant mean. Volatility is used here for the (homogeneous) standard deviation of the outcomes of a stochastic process. In section 3.2 the non-homogeneous birth-and-death process is used to derive this non-homogeneous martingale model. A description of a likelihood-based approach to fit a conditional version of the model is explained in Section 3.3.

It might be important to detect a change in volatility for processed data with constant mean because this change in volatility might be a signal that in the near future a change in the mean might occur. An increase in variance as a signal that there is going to be a change of state is known in complex dynamical systems. As discussed in (Scheffer et al., 2009), in such systems these changes can be early warning signals that the system is approaching a state transition. As discussed there, a system approaching critical points become increasingly slow in recovering from small perturbations. This critical slowing down leads to three possible early warning signals, among which an increase in variance. Examples that an increase in variance can be an early warning signal can be observed in such diverse areas as climate change, ecosystems, asthma attacks and epileptic seizures (Scheffer et al., 2009).

In section 3.4 it is described how a change in variance in a model with constant mean might be observed. Also the choice of a survival curve is explained there. In section 3.5 the models are applied to Methicillin resistant staphylococcus aureus (MRSA) data obtained from the National Health Service (NHS) Acute Trusts of hospitals from Great Britain.

3.2 Non-homogeneous martingale process

As time evolves, a stochastic process can generate count data from a distribution, with the same mean and variance. Such a process can be called a stable process. A stable process can be 'destabilized' by a change in volatility only, in the absence of a change in the mean: the counts in time remain scattered around a certain level but the amount of scatter might increase or decrease with time. As explained above, this kind of disturbance might be a signal that in the near future the mean might change over time. This change in the mean can manifests itself as a smooth – possibly nonlinear – change in the mean over time, or it might be a jump in the average count, upwards or downwards, after which the process might 'stabilize' again at a new level. To study destabilization, the model of the process that generates the count data must be able to distinguish between the state in which only the volatility changes and the state in which there is a change in the mean over time, and thus also a change in volatility. In order to achieve this, the 'change-in-volatility-only' model should be a

special case of the 'change-in-mean and change-in-volatility model'. Besides this, the model should take the dependence between the counts in different time points into account.

In the non-homogeneous birth-death models the rates are allowed to change over time. This can be very useful in modeling the dynamics of a process, for example in infectious disease models. A possible interpretation for the time-varying rates is given in (Becker and Yip, 1989). The non-homogeneous birth-death model with counts as its state-space can be reduced to a model that models volatility only, as will be shown below by taking the birth-rate – or reproductive power – and the death-rate equal. In the non-homogeneous birth-death model the probability of a birth or a death at a given point in time will depend on the population size at the previous point in time only, i.e. it is a Markov model.

So, the general non-homogeneous birth-and-death process is an example of a stochastic process that is suitable for our purpose. It is used in this paper to derive a new model that only describes the variance as a function of time and that is shown to be a non-homogeneous martingale process. We will use the language of infectious disease epidemiology as our motivating practical example.

First a brief discussion of the non-homogeneous birth-death process is given. Details can be found in chapter 2. A random count variable, measuring population size at calendar time t, or the number of infected-and-detected individuals of an infectious disease at time t, is denoted by $Y(t)$. Let y_0 be the count at the time point of the start of the observation period. The non-homogeneous birth-and-death process is governed by the stochastic differential equations which can be written a linear partial differential equation for the probability generating function, the solution of which is given by (Kendall, 1948):

$$\phi(z,t) = \left[\frac{\theta(t) + (1-\pi(t)-\theta(t))z}{1-\pi(t)z}\right]^{y_0},$$

where $\theta(t)$ and $\pi(t)$ in the above solution are written as

$$\theta(t) = 1 - \frac{e^{-\rho(t)}}{1+\gamma(t)e^{-\rho(t)}} = 1 - \frac{S_\mu(t)}{S_\lambda(t)+\gamma(t)S_\mu(t)}, \tag{3.1}$$

and

$$\pi(t) = 1 - \frac{1}{1+\gamma(t)e^{-\rho(t)}} = 1 - \frac{S_\lambda(t)}{S_\lambda(t)+\gamma(t)S_\mu(t)}, \tag{3.2}$$

and

$$\gamma(t) = \int_0^t e^{\rho(\tau)}\lambda(t)\mathrm{d}\tau = -\int_0^t \frac{\mathrm{d}S_\lambda(\tau)}{S_\mu(\tau)}, \tag{3.3}$$

$\rho(t) = \int_0^t [\mu(\tau) - \lambda(\tau)]\,\mathrm{d}\tau = \ln\left[\frac{S_\lambda(t)}{S_\mu(t)}\right]$, $S_\lambda(t) = e^{-\int_0^t \lambda(\tau)\mathrm{d}\tau}$ is the survival function for reproduction time (birth process), and similarly, $S_\mu(t) = e^{-\int_0^t \mu(\tau)\mathrm{d}\tau}$ is the survival function for removal time (death process).

The non-homogeneous martingale process is a special case of the above mentioned non-homogeneous birth-and-death process. Because a population with constant mean indicates that the birth and death rates are equal (i.e. $\lambda(t) = \mu(t)$ (say $\eta(t)$)), it follows that $\rho(t) = 0$ and thus $\theta(t) = \pi(t)$ (say $\delta(t)$). The probability distribution then is given by

$$\begin{aligned} P(Y(t) = 0) &= \delta^{y_0}(t), \qquad (3.4)\\ P(Y(t) = y) &= \frac{y_0}{y}\delta^{y+y_0}(t) \sum_{k=0}^{\min(y-1,y_0-1)} \binom{y_0-1}{k}\binom{y}{y-k-1}\left[\frac{(1-\delta(t))}{\delta(t)}\right]^{2(k+1)} \quad y \geq 1. \end{aligned}$$

Since $\eta(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log(S_\eta(t))$, it follows that $\gamma(t) = \int_0^t \eta(\tau)\mathrm{d}\tau$, is the integrated rate function, similar to the integrated hazard function in survival analysis. By using $\eta(t)$ (instead of $\lambda(t)$ or $\mu(t)$) and (3.1) or (3.2), it can be seen that $\delta(t) = \frac{\ln(S_\eta^{-1}(t))}{1+\ln(S_\eta^{-1}(t))} = \frac{\gamma(t)}{1+\gamma(t)}$.

The expected value of the population size at time t derived from (3.4) is

$$E(Y(t)) = y_0.$$

That is, the expected value of the population size for any time point t, after time zero is equal to the population size at time zero and thus the process is a martingale process. The variance of the population size is

$$var(Y(t)) = 2y_0 \ln(S_\eta^{-1}(t)) = 2y_0 \int_0^t \eta(\tau)\mathrm{d}\tau \qquad (3.5)$$

which depends on time through the integrated rate function. In a simpler case of a homogeneous birth-and-death process (i.e. with equal and constant birth and death rates, η), the variance is $2y_0\eta t$, and therefore the variance increases linearly as a function of time t, see (Bailey, 1990, page 94 - 96) for details. Note that this is the variance of a stable process and that if η is large this process can handle a large variability. In that case a couple of large or small observations in a row is not unlikely. This shows that it can be hard to distinguish this model from a model which describes a change in mean and thus in volatility or a changing volatility model. In the non-homogeneous case, the derivative of the variance is positive, and thus the variance is an increasing function with time.

3.3 Fitting the model

According to the model described in section 3.2, the data are generated by the stochastic non-homogeneous birth-and-death process where the rate at which a birth (new case, for the 'infectious-disease interpretation') occurs is equal to the rate at which a death (recovery/removal) occurs. To fit the model to empirical data, it can be noted that the observed data represent just a single sample path of all possible sample paths that can be generated by such a model. The population size at time t depends on the population size before t (Markov property), and the model fitting procedure needs to account for this conditionality, see chapter 2 and 4.

In addition, empirical observations of such a process are usually made in discrete time t_j $(j = 0, 1, \cdots, n)$ where the interval of observation might be for instance days, weeks or months. Accordingly, the population size at time t_j is conditioned on population size at time point t_{j-1}. in other words, the model fitting is achieved by considering only those sample paths of the process that go through the point $(t_{j-1}, y_{t_{j-1}})$.

Because the probability mass function (3.4) is actually a conditional probability (i.e. conditioned on the observed population size at t_0), it can readily be used as a conditional model for the population size at time t_j given the population size at t_{j-1}. Because $\eta(t)$ (and thus $S_\eta(t)$) varies as a function of time, the conditional form of $S_\eta(t)$ must be taken into account. Let T be a stochastic variable that measures the time at which an event (i.e. birth or death) occurs. The conditional survival probability is:

$$P(T > t_j | T > t_{j-1}) \;=\; \frac{S(t_j)}{S(t_{j-1})} = 1 - h(t_{j-1})$$

were $h(t_{j-1})$ is the birth rate or the death rate in discrete time, depending on the event considered (see chapter 2 section 2.3).

Using $h(t_{j-1})$, and defining $c(t_{j-1})$ as the discrete version of $\delta(t)$ in the non-homogeneous martingale model: $c(t_{j-1}) = \frac{\ln(\frac{1}{1-h(t_{j-1})})}{1-\ln(\frac{1}{1-h(t_{j-1})})}$, the conditional version of (3.4) is

$$P(Y(t_j) = 0 | Y(t_{j-1}) = y_{t_{j-1}}) = c(t_{j-1})^{y_{t_{j-1}}}, \tag{3.6}$$

$$P(Y(t_j) = y_{t_j} | Y(t_{j-1}) = y_{t_{j-1}}) = \frac{y_{t_{j-1}}}{y_{t_j}} c(t_{j-1})^{y_{t_j} + y_{t_{j-1}}} \times$$

$$\sum_{k=0}^{\min(y_{t_j}-1, y_{t_{j-1}}-1)} \binom{y_{t_{j-1}} - 1}{k} \binom{y_{t_j}}{y_{t_j} - k - 1} \left[\frac{(1 - c(t_{j-1}))}{c(t_{j-1})}\right]^{2(k+1)} \quad y_{t_j} \geq 1.$$

In this process, the expected value of the population size at time t_j is

$$E(Y(t_j)) = y_{t_{j-1}},$$

and this can be referred to as sample path profile (Lindsey, 2001). The variance of the population size at time t_j, given that the process passes through the point $(t_{j-1}, y_{t_{j-1}})$ is:

$$var(Y(t_j)) = 2y_{t_{j-1}} \ln\left(\frac{1}{1-h(t_{j-1})}\right). \tag{3.7}$$

In this variance, $h(t_{j-1}) = 1 - \frac{S(t_j)}{S(t_{j-1})}$. Depending on the choice of this survival function, let Δ represent the vector of its parameters; the log-likelihood as a function of Δ is then written as:

$$l(\Delta) = \sum_{i=1}^{n} \ln\left[P(Y(t_j) = y_{t_j} | Y(t_{j-1}) = y_{t_{j-1}}, \Delta)\right],$$

for $y_{t_j} > 0$. It should be noted that $y_{t_j} = 0$ is an absorbing state of the process. The log-likelihood can be maximized using an optimization procedure such as the Nelder-Mead simplex method followed by the Newton-Raphson method to find maximum likelihood estimates. In the present study, we used the software system R for the iterations (R Development Core Team, 2010). The information matrix is used for computing the standard errors of the parameters and we use the Akaike Information Criterion (AIC) to compare models.

3.4 Modeling Issues

3.4.1 Using the net reproduction ratio

To compare the martingale model derived in sections 3.2 and 3.3 with a model with a change in the mean, the general non-homogeneous birth-death process seems the most obvious candidate since it has the martingale model as a special case. The process is described by the stochastic non-homogeneous birth-and-death differential equations, shown in section two, the solution and properties of which is given elsewhere in chapter 2. The process involves time-dependent variation both in the mean and in the variance of the population size. The expectation and the variance of the population size are:

$$\begin{aligned} E(Y(t)) &= y_0 \frac{S_\mu(t)}{S_\lambda(t)} = y_0 R(t), \\ var(Y(t)) &= y_0 R(t)\left[1 + (2\gamma - 1)R(t)\right], \end{aligned}$$

respectively, where $R(t) = \frac{S_\mu(t)}{S_\lambda(t)}$ is referred to as the net reproduction ratio for this model at time t (chapter 2) since it is the expected number of secondary cases per primary case at time t (Nishiura and Chowell, 2009; Diekmann and Heesterbeek, 2000). The functions $S_\lambda(t)$ and $S_\mu(t)$ are the survival functions of the reproduction times and removal times, respectively. If $R(t) < 1$, it suggests that the population is in decline, whereas it is growing if $R(t) > 1$. Moreover, $R(t)$ measures steepness of the trend. Because of non-homogeneity, the model is able to reflect time-dependency in the trend.

The mean population count in this model can be thought of as consisting of two parts. The first is the expected number present at t_0 which have survived until time t (i.e. $y_0 S_\mu(t)$, those available on time zero that are not removed up to t), and the second part is $\frac{1}{S_\lambda(t)}$, which measures the rate at which a non-removed individual reproduces itself (which is the number of individuals needed to have at least one individual who is still reproducing). This interpretation is similar to the interpretation of the effective reproduction number in the sense that only those individuals that survive removal can reproduce (chapter 2). In other words, if non-removed individuals reproduce faster then the speed at which they are removed, $R(t) > 1$ and the population size increases (vice versa for a decreasing population). It can also be seen from the formula for $var(Y(t))$ above that increase or decrease in the population size directly influences the variance, because the variance depends on time t and $R(t)$.

Note that the expected value of the process obeys the differential equation for the deterministic birth-death process, that is $\frac{\mathrm{d}}{\mathrm{d}t}E(Y(t)) = \lambda(t)E(Y(t)) - \mu(t)E(Y(t))$. In a SIR-model – a compartmaent model with compartments susceptible, infectives and removed – interpretation the reproductive power can be written as the product of the infection rate parameter and the time dependent fraction of susceptibles: $\lambda(t) = \beta s(t)$. So in this interpretation the reproduction power depends on the fraction of susceptibles and the non-homogeneous character of the reproductive power is inherited from the fraction of susceptibles. In other words the reproductive power is changing in time because the fraction of susceptibles is.

The rates in this model, the birth rate $\lambda(t)$ and the death rate $\mu(t)$, both depend on time. These rates can also be taken constant in time. This is the same as using the exponential distribution as the survival distribution for the reproduction times and the removal times, that is by using $S_\lambda(t) = e^{-\lambda t}$ and $S_\mu(t) = e^{-\mu t}$. This process yields time-dependent variations in the mean population size, and the model is a special case of the non-homogeneous process with a change in the mean in that the birth and death rates are independent of time. Therefore, the model is able to capture a change in population counts over time, but the birth and death rates are kept constant. This

homogeneous birth-death model has been frequently used in literature and has been described in detail elsewhere (e.g. (Allen, 2003)).

Since in the present work a model with changing volatility is compared to a model with a changing mean and thus also changing volatility, model comparison might be difficult using only AIC. For instance, in the first part of the observation period the martingale model may seems to fit well whereas in the last part it may not. In that case the net reproduction ratio might be an additional tool to compare models because the net reproduction ratio for the martingale model is one, since there the birth rate equals the death rate. For the non-homogeneous birth-death model, the net reproduction ratio can be an increasing or a decreasing function or both. In this way it can be used to decide which model is better: if the net reproduction ratio is constant around one, then the martingale model might be appropriate, if not, the non-homogeneous model is preferred. But it can also show the changing behavior mentioned: in the first part the net reproduction ratio can be constant around one but it can increase or decrease in a latter part of the observation period or the other way around. That is, it might show a change when time evolves.

The conditional version of the net reproduction ratio is: $R(t_{j-1}) = \frac{1-h_\mu(t_{j-1})}{1-h_\lambda(t_{j-1})}$ with $h_\mu(t_{j-1}) = 1 - \frac{S_\mu(t_j)}{S_\mu(t_{j-1})}$, the discrete hazard rate of removal, and $h_\lambda(t_{j-1}) = 1 - \frac{S_\lambda(t_j)}{S_\lambda(t_{j-1})}$, the discrete rate of reproduction. This net reproduction ratio is calculated using the birth-death model with the survival functions that fits the data best according to the AIC. From the estimated non-homogeneous birth-death process that fits the data best, sample paths can be drawn and for each sample path the net reproduction ratio can be calculated for each time point. From these the 95% percentile lines can be inspected to see how the line $R(t) = 1$ behaves with respect to these 95% lines. This technique is illustrated in section 5.

3.4.2 Survival function

Given that the population dynamics involve complex mechanisms, on the one hand the family of survival functions used in the model needs to be flexible. On the other hand, one generally should aim to use the simplest model, e.g. exponentially distributed survival functions for homogeneous processes, where rates are constant. The generalized gamma is a rich family of distributions that include the exponential as a special case but also the Weibull and the Gamma distribution. As a consequence, by checking the estimated parameters of the generalized gamma one can check if the data supports constant rates.

The survival function is then given by:

$$S(t) = \frac{1}{\Gamma(p)} \int_0^{(\frac{t}{b})^a} t^{p-1} e^{-t} \mathrm{d}t$$

where $p, b > 0$. If $a < 0$, the distribution is referred to as inverse generalized gamma distribution. Special cases of the generalized gamma distribution are:

- Gamma distributions for $a = 1$ and the inverse gamma for $a = -1$.
- Weibull distribution for $p = 1, a > 0$ and the inverse Weibull (log-Gompertz) for $p = 1, a < 0$.
- Exponential distribution for $a = p = 1$ and the inverse exponential distribution for $a = -1, p = 1$.
- Log-normal, Pareto and power function distributions for appropriate limits.

Further details of the generalized gamma distribution are given elsewhere (Kleibner and Kotz, 2003).

3.5 Prevalence of MRSA in three NHS Trust in Great Britain

Methicillin-Resistant *Staphylococcus aureus* (MRSA) is a bacterial infection that causes infections in different parts of the body and is resistant to several antibiotics such as methicillin, amoxicillin, penicillin and oxacillin. MRSA infections are an important risk for people who have weakened immune systems and are in hospital intensive care units, nursing homes, and other health care centers. There are many risk factors for acquiring MRSA like surgery, duration of hospitalization, compliance with hand disinfection procedures and antibiotic exposure, among others (Tacconelli, et al., 2008).

Because hospital patients have an increased risk of being infected with MRSA it is important for hospitals to monitor their MRSA-prevalence and to take measures to prevent spread. Are the (precautionary) measures taken effective and is the prevalence decreasing in time, or is the volatility increasing as a consequence of the measures? It is also possible that the measures taken, hardly have influence and that the prevalence of MRSA is randomly fluctuating around some fixed level. For a discussion of the problems that can arise with MRSA infection data, including the

variability of the rates, see (Spiegelhalter , 2005).

Other stochastic models for hospital outbreak data have been used. For instance (Pelupessy et al., 2009) describe a model for different colonization routes of pathogens within a hospital. They derive a stationary probability distribution for the colonized patients that takes the different routes of colonization into account and need data on these routes. In contrast the present paper describes a model that uses a probability distribution for which the expected value is not changing over time but which is nevertheless depending on time through the variance of the process and only relays on infection prevalence data.

An important point is that the rate at which MRSA reproduces, depends on the duration of hospitalization (Beyersman et al., 2011). The above model deals with this by taking the reproduction power to be time dependent. As is explained by (Becker and Yip, 1989) a rate parameter might behave in a time dependent manner because there is a difference in susceptibility among the susceptibles. Those with higher susceptibility tend to be infected earlier while the low-risk susceptibles will tend to remain susceptible longer. Hence heterogeneity among susceptibles (related to duration of hospitalization) can make the rate behave in a time dependent manner. Furthermore, as explained, in a SIR interpretation, the reproductive power can be thought of as a product of the infection rate parameter (β) and the time dependent fraction susceptibles. This fraction is a parameter in the model. So if the fraction susceptibles is changing because the discharge rate is changing over time, then this is reflected in the reproductive power through the fraction of susceptibles. Besides this, as said, the model discussed here is for the case where only counts per unit time are available although it still depends implicitly on the time-dependent fraction of susceptibles.

Because of the importance to monitor the MRSA prevalence, there is a surveillance of MRSA among NHS Acute Trust hospitals in Great Britain. This surveillance has been mandatory since April 2004. Quarterly data from April-June 2001 until January-march 2010 and monthly tables from February 2010 until February 2011 can be obtained from the web site of the Health Protection Agency:

$http://www.hpa.org.uk/webw/HPAweb\&HPAwebStandard/HPAweb_C/1233906819629?p=1191942169773$

Positive blood cultures from the same patient within 14 days of the initial culture were considered to be part of a single infected episode. Duplicate reports, more than 14 days apart are considered to be separate episodes of infection of the same patient. That is, the data consist of 14-day-episodes prevalence data for patients who are MRSA positive. New cases in a hospital are reproduced from existing cases usually through health care workers. The rate at which this happens is the reproductive power in the model. After discovery of MRSA a case is usually removed or isolated

so that it is not possible for this person to reproduce any longer. The rate at which this removal happens is the death rate in the model.

The models and methods described in this paper are used for the data of three of the NHS Trust hospitals to illustrate the following specific data issues:

- The data from the Leeds Teaching Hospitals Trust show that the non-homogeneous martingale model fits the data best, based on Akaike's Information Criterium. There is a change in volatility.

- The data from the Guy's & St. Thomas' Trust does not show a clear choice between some birth-death models and a martingale model, based on Akaike's Information Criterium. Inspection of the curve of the net reproduction ratio shows that there is evidence for a birth-death model and thus a change in mean.

- Also the data for the King's College Hospital Trust is not conclusive between some birth-death models and the martingale model, based on Akaike's Information Criterium. Here, however, the net reproduction ratio does not show evidence of a change in mean, indicating that the martingale model is to be preferred. The data also shows that there is a constant rate at which an event (new MRSA infection or MRSA removal) occurs, meaning that volatility is changing linearly in time.

We assumed that the characteristics of the process that generates the data within a Trust is approximately the same for all hospitals of that Trust and that thus the data within a Trust can be aggregated.

Using the data of these three Trusts, 14 different models were fitted. Four martingale models were fitted in which the survival distribution (of the reproduction and removal times) is taken to be the Generalized Gamma, the Gamma, the Weibull and the exponential distribution. Ten birth-death models were fitted: birth-death models where the reproduction times and the removal times had different Generalized Gamma, Gamma, Weibull and exponential distributions, birth-death models where the reproduction times had an exponential distribution- constant birth rate- and the removal times had an generalized Gamma, Gamma or a Weibull distribution and finally birth-death models where the removal times had an exponential distribution (constant death rate) and the reproduction times had an generalized Gamma, Gamma or a Weibull distribution. Table 3.1 compares AIC values between the different models for each of the three Trusts mentioned above.

Figure 3.1 shows the data (upper part) for Leeds Teaching Hospital. The number of cases per quarter seems to stay reasonably stable for a long period (about 30

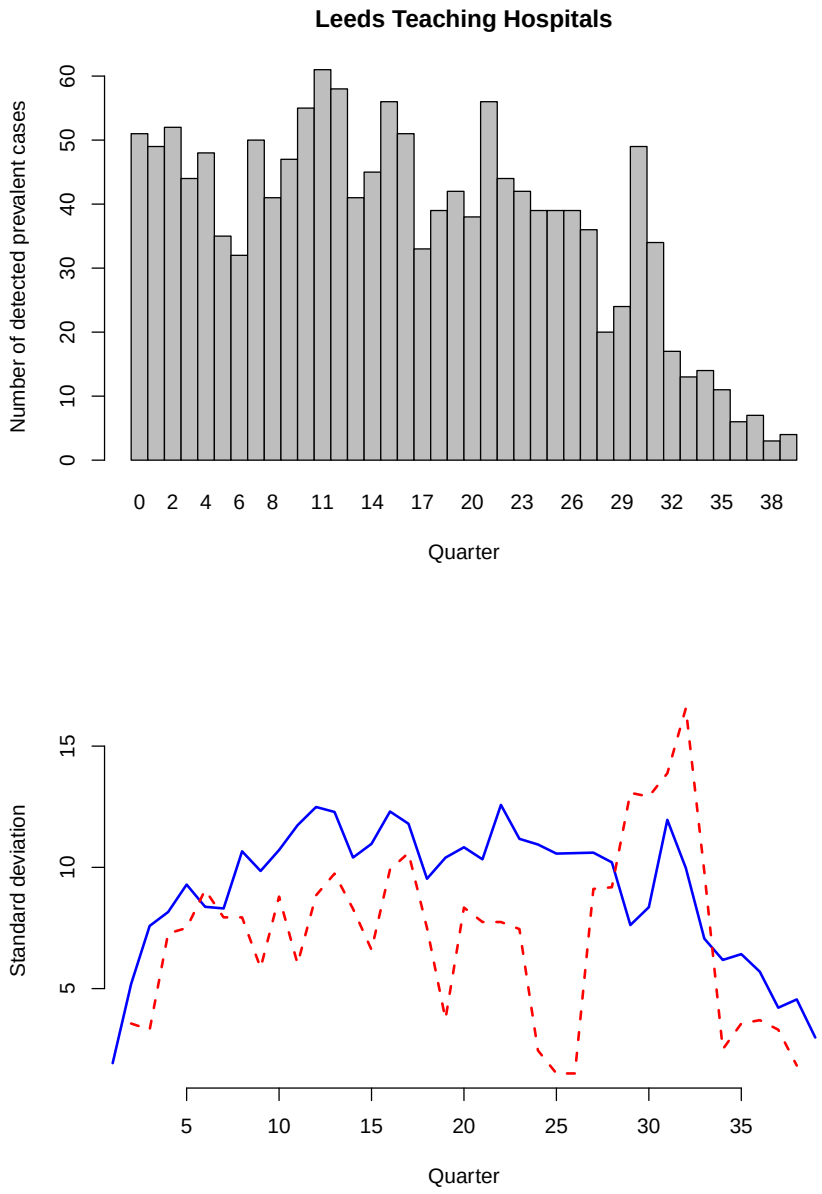

Figure 3.1: Leeds Teaching Hospital. Upper: the MRSA prevalence. Lower: the four-period moving standard deviation (dashed line) and the standard deviation from the conditional model (solid line).

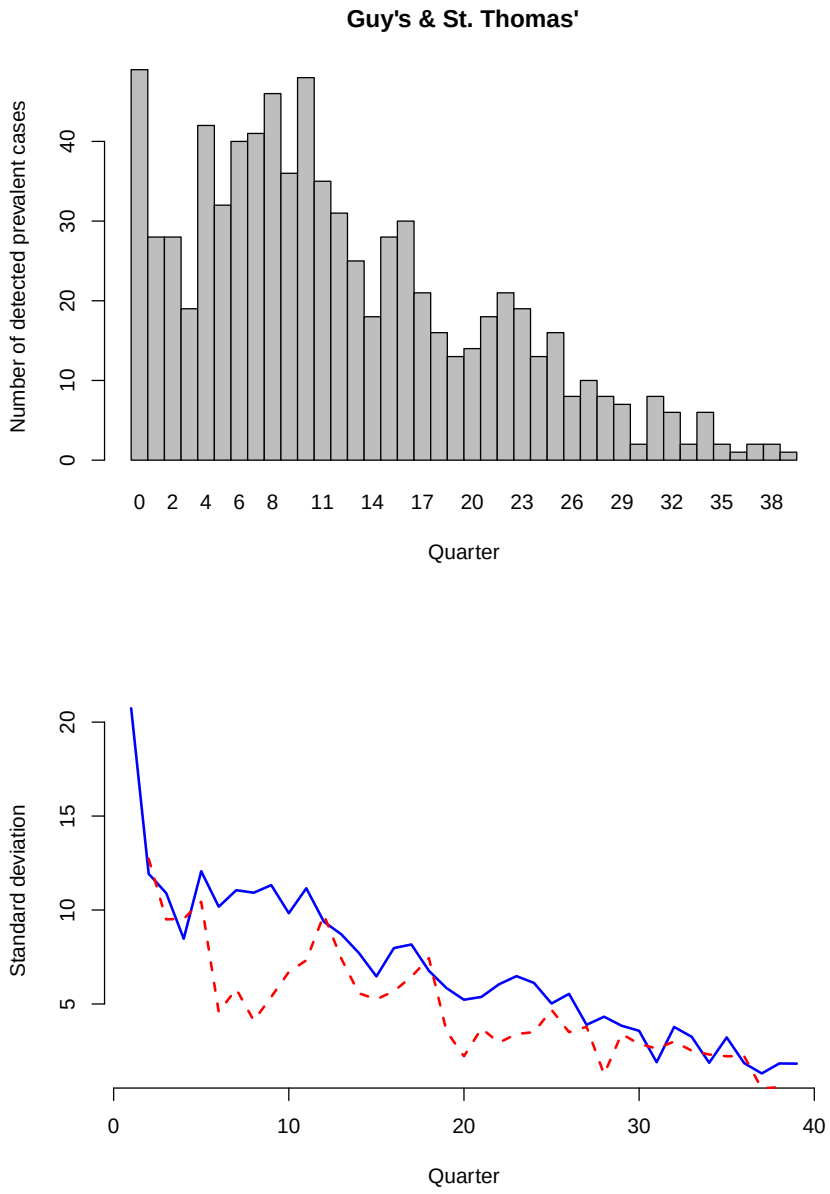

Figure 3.2: Guy's & St.Thomas'. Upper: the MRSA prevalence. Lower: the four-period moving standard deviation (dashed line) and the standard deviation from the conditional model (solid line).

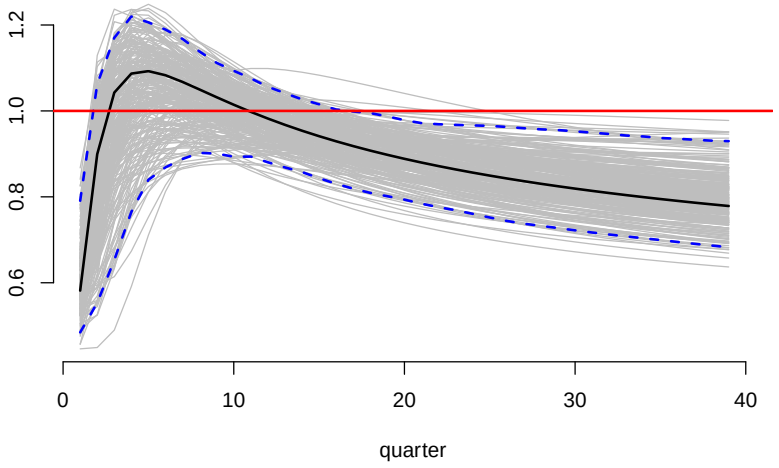

Figure 3.3: Guy's & St.Thomas'. Net reproduction ratio as a function of time (black solid line) and the net reproduction ratio calculated from sample paths drawn from the underlying estimated process (gray lines) with the 95 % percentile lines (dashed lines).

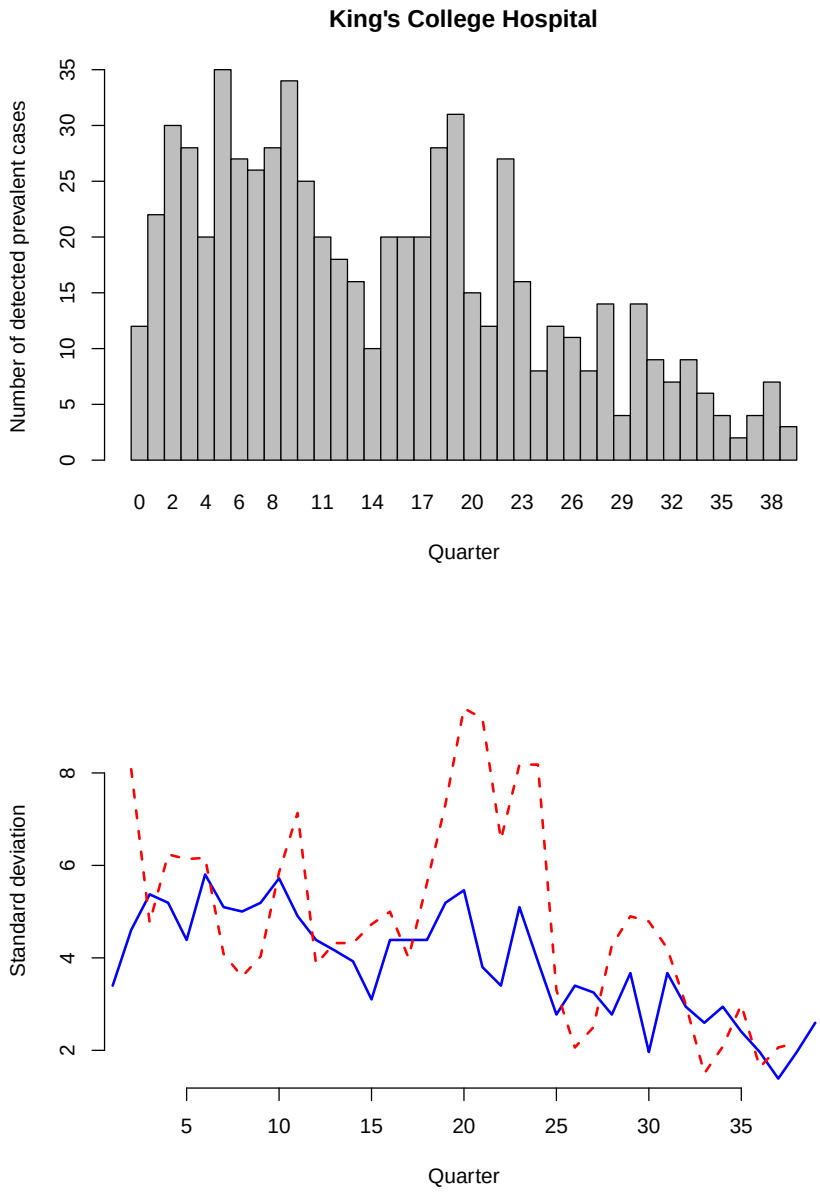

Figure 3.4: King's College Hospital. Upper: the MRSA prevalence. Lower: the four-period moving standard deviation (dashed line) and the standard deviation from the conditional model (solid line).

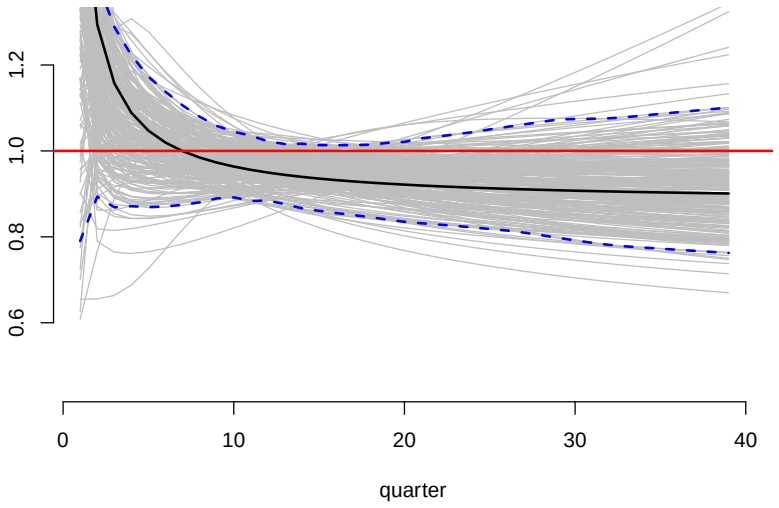

Figure 3.5: King's College Hospital. Net Reproduction ratio as a function of time (black solid line) and the Net reproduction ratio calculated from sample paths drawn from the underlying estimated process (gray lines) with the 95 % percentile lines (dashed lines).

Table 3.1: AIC's for 3 Trusts using 14 different models.

Model	Survival distribution	Leeds Teaching	Guy's & St.Thomas'	King's College
Martingale	Generalized Gamma	281.85	249.76	262.33
	Gamma	279.87	248.55	260.73
	Weibull	281.52	247.81	260.89
	Exponential	282.11	249.60	258.90
Birth-death	Generalized Gamma	285.88	250.56	265.12
	Gamma	283.22	251.03	261.10
	Weibull	285.03	250.85	262.60
	Exponential	283.50	250.35	260.86
Birth-death with birth rate exponential	Generalized Gamma	286.63	249.97	260.88
	Gamma	285.41	249.43	259.13
	Weibull	285.07	251.12	260.62
Birth-death with death rate exponential	Generalized Gamma	287.05	247.24	262.52
	Gamma	285.50	250.84	260.53
	Weibull	285.42	252.34	261.07

quarters). Only at the end of the observation period there seems to be a disturbance in volatility or in the mean (and thus also in volatility). As can be seen from Table 3.1 the martingale model with the Gamma distribution gives the best fit according to AIC, indicating that the volatility of the process is changing. This martingale model has an AIC 3.25 smaller the the best fitted birth-death model so the choice for a martingale model as compared to a non-homogeneous birth-death model, is clearly indicated by the AIC. This is not the case for the choice of the distribution since for the martingale model, the Weibull distribution gives a fit that is reasonably close (according to AIC) and has the same number of parameters. The (log)parameter estimates and their standard errors for the martingale model with the Gamma distribution are: $\ln(p) = 1.533(0.2599); \ln(b) = -0.455(0.2248)$. From these estimates the standard deviations of the conditional process can be estimated using (3.7). These estimated conditional standard deviations are shown in Figure 3.1 in the lower-part (the solid line). As can be seen, these standard deviations increase in the early observation period, after which they stay reasonably stable, until in the last part of the observation period where they seem to decrease. After time point 30, there seems to be an increase in variance. A couple of relatively low counts are produced and the variance process adapts itself to these low values, and as a result more low counts

are observed. If this is continued (the production of low counts), then there might be evidence for a changing mean in the data.

In this figure, the line for the four-period moving standard deviation is also shown (dashed line), as an empirical measure of the standard deviation of this process. This is of course different from the conditional standard deviation of the martingale model, but shows approximately the same pattern, although in a more 'zig-zag'-style: an increase in the beginning and a decrease in the end of the observation period.

If the choice for the martingale model in the case of the Leeds Teaching Hospital seems a clear one, such is not the case with Guy's & St. Thomas. As can be seen from Table 3.1, the martingale model with a Weibull and the birth-death model with a exponential distribution for the removal times (a constant death rate) and a generalized gamma distribution for the reproduction times are very close. Figure 3.2 shows a barplot of the data in the upper part and in the lower part the four-period moving standard deviation (dashed line) together with the conditional standard deviation (solid line) from the martingale model with the Weibull distribution. These are both decreasing.

One can then pick the martingale model because its number of parameters is smaller but one may also try to see if there is some other information available. As mentioned previously, Akaike's information criterion judges the fit over the whole observation range. There can be evidence in the data, however, that the birth and death rates are approximately equal in the first part of the observation period but not for a later part and one can then, as discussed earlier, use the net reproduction ratio $R(t)$ for comparison. Figure 3.3 shows the conditional version of the net reproduction ratio $R(t_{j-1})$. This net reproduction ratio is calculated using the birth-death model with an exponential distribution for the removal times and a generalized gamma distribution for the reproduction times (being the best fitting birth-death model according to the AIC). From this model, 250 sample paths are drawn and for each sample path the net reproduction ratio is calculated (gray lines); from these the 95% percentile lines are determined (dashed lines). From this plot it can be seen that the net reproduction ratio does not differ much from one in the first part of the observation period (about the first 18 quarters), indicating that the birth and death rates are approximately equal, but is less than one in the second part. So by judging the net reproduction ratio over the observation period, the conclusion can be drawn that the removal rate is larger that the reproduction rate and that thus the number of cases is decreasing after approximately 18 quarters. In other words, the net reproduction ratio and its 95% percentile lines, indicate that there is evidence in the data for a decreasing mean of the average number of MRSA cases in the second part of the observation period. This also causes a decrease in the standard deviation.

Similar observations can be made for King's College Hospital. Two models are very close in fit as can be seen from Table 3.1: the martingale model with exponential distribution and the birth-death model with exponential reproduction, times and gamma-distributed removal times (although there are other birth-death models giving a fit reasonable close). Figure 3.4 shows the data in the top graph and in the lower graph the four-period moving standard deviation (dashed line), with the conditional standard deviation from the martingale model with the exponential distribution (solid line). Again, one can have a look at the net reproduction ratio which is shown in Figure 3.5, together with the net reproduction ratio's from 250 sample paths and their 95% percentile lines. Here, the line $R(t) = 1$ lies between the 95% percentile lines, and there thus seems no evidence in the data that the birth rates and death rates are different. So, in this case one can take the martingale model with the exponential distribution as the one that best fitted the data, and conclude that there is no evidence in the data for a change in mean or in volatility.

3.6 Discussion

Modeling volatility is becoming an important issue, not only in the economic literature where the Black-Scholes model has drawn much attention, but also in areas as ecology, and more recently in the environmental sciences were climate change is thought to cause an increase in volatility. Modeling volatility can be done separately from modeling the mean of a process when the normal distribution is used. When models for counts are used, things are slightly more complicated since there a change in mean has a change in variance as consequence because of the relationship between the mean and the variance. If there is more (or less) variance as there must be according to the model, then models for over (or under) dispersion can be used.

This leaves the important case were there is no change in the mean but there is a change in volatility. It is important to be able to distinguish this case from the 'changing-mean' case, because a change in volatility can give the erroneous impression that there is a change in the mean. Consider for instance the case that the variance is increasing, then by coincidence a few large (or small) values in a row might be observed, suggesting a change in the mean. A change in volatility only can also indicate a disturbance in a stable process, after which a change in the mean can occur. This change can be smooth, or it can manifest as a jump, after which the process can stabilize again.

This paper proposes a new non-homogeneous martingale model for count data, derived from a non-homogeneous birth-death process, to study the changes in volatility

without change in the mean. This model is capable of modeling a change in volatility while the mean is staying the same, but also can deal with a process that shows no change at all, a stable process.

The net reproduction ratio plays an important role in choosing between a volatility-only model and a model with different non-homogeneous birth and death rates. It might be that in a part of the observation period the net reproduction ratio is approximately one, indicating an equal birth and death rate and thus pointing to the non-homogeneous martingale model, where in another part there might be evidence from the data that this is not the case.

As an illustration the models were fitted to MRSA prevalence data from three Trusts in Great Britain: Leeds Teaching Hospital, Guy's & St. Thomas and King's college Hospital. The procedure described above was able to reveal different aspects of the data: a changing volatility only (Leeds Teaching Hospital), evidence using AIC and the net reproduction ratio that there was a change in mean in the later part of the observation period (Guy's & St. Thomas) and no evidence for change, not in volatility and not in mean (King's college Hospital). For the hospital Guy's & St. Thomas the data showed that the net reproduction ratio dropped below one, indicating that the MRSA-measures taken in that hospital are effective. For King's college the data did not show that the net reproduction ratio was different from one indicating that MRSA is persisting there.

As is clear from the models used here and the data of the three hospitals, it might not at all be obvious whether there is a changing mean in the data or change in volatility or that the time series as a whole stays reasonable stable. This is especially the case in data that shows a large variance. This would lead to the policy implication for these kind of surveillance that rather long time series are needed in order to be able to identify a changing mean.

Chapter 4

Non-homogeneous birth and death models for epidemic outbreak data [1]

4.1 Introduction

Modelling the outbreak of an infectious disease, such as the foot and mouth disease epidemic in Great Britain in 2001, is not a straightforward task because it is not only a biological process that generates the data but also control measures, which are often rigorous, being taken to control and finally end the outbreak. This gives an unnatural ending of the process which is totally different from the outbreak process. Therefore a model describing the epidemic outbreak might only be appropriate for the first stage of the outbreak. Besides this there are other aspects of the data-generating process that one wants to take into account. Firstly, infected individuals arise in time which makes the data dependent: what happens on a certain time point depends on what happened before that. Secondly, these infecteds are, for various reasons, not immediately observed at the time of infection but after some detection time. The prevalence of the infecteds since the start of the outbreak at a certain time point can be related to the distribution function of the infection times, by noting that the fraction of infecteds since the outbreak is the same as the fraction of individuals with a infection time less than or equal to that time. So the infected fraction since the

[1] Van Den Broek, J. and Heesterbeek, J.A.P. (2007). Nonhomogeneous birth and death models for epidemic outbreak data.
Biostatistics 8, 453-467.

outbreak is the same as the fraction of left-censored infection times. In the same way the susceptible fraction at a certain time point is the same as the fraction of individuals with an infection time larger than that time point, i.e. it is the same as the fraction right-censored infection times. This kind of data is often referred to as current status data.

As the third point we have that infecteds are registered by the symptoms of the disease, usually with some delay. This means that it is hard to model susceptibles because an individual who is not a registered infected is not automatically a susceptible. It might be a not-yet-detected infected. Besides this it is often not known precisely how many susceptibles there are at the start of the outbreak.

Although it is not a purely biological process that generates the data, a good starting point might nevertheless be the SIR model. We assume throughout that this model is a valid description of the underlying outbreak process. In an SIR model only three classes of individuals are recognized: Susceptibles, Infectives (i.e. those individuals that have been infected and are infectious to others) and Removed/Recovered individuals (i.e. individuals that are no longer infectious and are immune). In the context of the model we will therefore refer to currently infected individuals as infect*ives* (because there is no distinction); we will use infect*ed* in more general descriptions (because in reality an infected individual could be in a latency phase and not yet infectious to others). In this Chapter attention is focused on the total number of infecteds, or the total reduction of susceptibles caused by the outbreak. So we shall be concerned with the first part of the SIR model only. By writing $x(t)$ for the susceptible fraction at time t and $y(t)$ for the infective fraction at time t, the first part of the SIR model for a closed population is:

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) \;=\; -\beta(t)\,y(t)\,x(t) \tag{4.1}$$

See, for example, Diekmann and Heesterbeek (2000). This equation states that the change in susceptibles at time t depends on the fraction of susceptibles, the fraction of infectives and an infection rate parameter which in this case depends on t.

In Section 2 two models derived from (4.1) are described, the force of infection model and the reproductive power model, together with their non-homogeneous stochastic versions. Also conditional models are discussed in order to fit a single outbreak. The stochastic models depend on the distribution of the infection times. The Burr family of distribution functions are proposed for this distribution and a parameter for the final size is introduced in the model. In Section 3 the models are fitted to three epidemic outbreaks: the Dutch classical swine fever outbreak from 1997-1998, the foot and mouth disease outbreak in Great Britain from 2001 and the

Dutch avian influenza (H7N7) outbreak from 2003.

4.2 The models

4.2.1 Force of Infection.

Assume disease is irreversible and concentrate on monitoring the total number of susceptibles. Equation (4.1) describes the reduction of susceptibles as a function of time where time is measured from the start of the epidemic until infection. By taking $\lambda(t) = \beta(t)y(t)$ this can be written as:

$$\frac{\mathrm{d}}{\mathrm{dt}}x(t) = -\lambda(t)x(t) \tag{4.2}$$

The force of infection, $\lambda(t) = \beta(t)y(t)$ is the rate at which susceptibles become infected.

The change in the proportion of susceptibles at time t is $\lambda(t)$ times the proportion of the susceptibles already there. The rate $\lambda(t) = \beta(t)y(t)$ itself depends on the fraction infectives. The force of infection model makes the following approximations:

1. x_0 is reasonably accurately known. This means that the area in which the epidemic might spread is approximately known.

2. Since it is usually not determined at time t that a susceptible is still a susceptible, x_t is approximately equal to $x_0-\{$the total number infected at time $t\}$. This approximation is reasonable if an infected is detected relatively fast, i.e. when the incubation period is short.

The well known solution of (4.2) is

$$x(t) = \exp\left[\int_0^t -\lambda(\tau)\mathrm{d}\tau\right] = S_\lambda(t)$$

with $S_\lambda(t)$ the survival function of the susceptibles. This is used in the literature together with the independent observations assumption — observations are implicitly assumed independent or are treated as such (Becker, 1989) — to form the currently used models. To relax the independence assumption one can use the stochastic version of (4.2) which is the differential equation of the non-homogeneous death process with

$X(t)$ the number of susceptibles at time t, $X(0) = x_0$ at time 0 and with $p_x(t)$ the probability that the number of susceptibles at time t is x:

$$\frac{\mathrm{d}}{\mathrm{d}t}p_x(t) = (x+1)\lambda(t)p_{x+1}(t) - x\lambda(t)p_x(t).$$

Its solution can be derived by the probability generating function given by Kendall (1948) as

$$\Phi(z,t) = \left[1 - e^{-\rho(t)} + ze^{-\rho(t)}\right]^{x_0}$$

and thus:

$$P(X(t) = x_t) = \binom{x_0}{x_t} \left[e^{-\rho(t)}\right]^{x_t} \left[1 - e^{-\rho(t)}\right]^{x_0 - x_t}$$

the binomial distribution with $e^{-\rho(t)} = \exp\left[-\int_0^t \lambda(\tau)\mathrm{d}\tau\right] = S_\lambda(t)$ and $\lambda(t)$ the hazard rate. Note that the expected value of $X(t)$ is the same as the solution of the deterministic equation (4.1) for the number of susceptibles. The expected value for the number of susceptibles at time t is $x_0 S_\lambda(t)$. This can be called the underlying profile of the process (Lindsey, 2001).

4.2.2 Reproductive power.

With force of infection models the data are modeled from the perspective of the susceptibles. This is often not desirable with an epidemic outbreak because not the number susceptibles but the number of detected infecteds is measured over time. It may, moreover, take a while before an infected is detected. This means that if an infected is detected at time t all that is known is that the time of infection is smaller than or equal to t.

Usually the end of the epidemic is not only determined by the removals or the population size but also by other measures taken like stamping-out, a transportation ban and all kinds of hygiene measures. So in that case only the first part of the outbreak can be reasonably modeled with an epidemic model such as the SIR model.

In the first stage of the outbreak the number of removals is limited and since the number of susceptibles is large, $y(t)$, the fraction of infectives, can approximately taken to be the fraction of infected at time t, and hence $x(t) + y(t) \approx 1$. In this case one can write (4.1) as a non-homogeneous birth process:

$$\frac{\mathrm{d}}{\mathrm{d}t}y(t) = \mu(t)y(t) \tag{4.3}$$

where $\mu(t) = \beta(t)\,[1 - y(t)]$ can be called the reproductive power (Kendall, 1948). The change in the fraction of cases at time t is $\mu(t)$ times the fraction of cases available at time t. This reproductive power itself is dependent on the proportion of susceptibles. In the situation described above the force of infection and the reproductive power are related: $\lambda(t) = \frac{y(t)}{1-y(t)}\mu(t)$, In other words: the force of infection is the odds of the prevalence times the reproductive power or the prevalence odds is the ratio between the force of infection and the reproductive power. The reproductive power model makes the following approximations:

1. The control measures are not immediately fully operational, so that the SIR description is in principle valid.

2. The number of removals is limited. This is reasonable in the first phase of the outbreak since in that phase an infected might be detected relatively late and large-scale control measures are only beginning to be implemented.

The solution of (4.3) is:

$$y(t) = \exp\left[\int_0^t \mu(\tau)\mathrm{d}\tau\right] = S_\mu^{-1}(t)$$

with $S_\mu(t)$ the survival function.

Using a the stochastic version of the non-homogeneous birth process gives with $Y(t)$ the total number of infected at time t and y_0 the number of infected at time 0, we have the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}p_y(t) = (y-1)\mu(t)p_{y-1}(t) - y\mu(t)p_y(t)$$

with $p_y(t)$ the probability that the number of detected infected at time t is y. The solution has probability generating function (Kendall, 1948):

$$\Phi(z,t) = \left[\frac{ze^{\rho(t)}}{1-\left[1-e^{\rho(t)}\right]z}\right]^{y_0}$$

and thus:

$$P(Y(t) = y_t) = \binom{y_t - 1}{y_0 - 1}\left[e^{\rho(t)}\right]^{y_0}\left[1 - e^{\rho(t)}\right]^{y_t - y_0} y_t = y_0, y_0 + 1, \cdots$$

with $e^{\rho(t)} = \exp\left[-\int_0^t \mu(\tau)\mathrm{d}\tau\right] = S_\mu(t)$. This is a shifted negative binomial distribution. It is the probability of obtaining $y_t - y_0$ infectives at time t in an epidemic that started with y_0 infectives at time zero.

Two comments are in order. Firstly, if for $S_\mu(t)$ the exponential distribution is taken then this model is the same as that of Langberg (Langberg, 1980). He showed that for a simple stochastic epidemic the distribution of the number of infectives converges in distribution to a negative binomial, where the inter-infection times are independently exponentially distributed.

Secondly, the expected value of $Y(t)$ is the same as the solution of the deterministic equation (4.3) for the number of infectives (instead of the fraction of infectives). The expected value for the number of infected at time t is $\frac{y_0}{S_\mu(t)}$, the underlying profile of the process.

4.2.3 The distribution function of the infection times

In order to choose a distribution function, or a survival function, for the infection time one might take equation (4.1) into account. We have assumed that this deterministic equation holds approximately in the first stage of the outbreak for the susceptible fraction and for the infected fraction. Because the susceptible fraction at time t is the same as the fraction of individuals with an infection time larger than t, and similar for the infected fraction, this should hold also for the distribution function and the survival function. Thus the first equation of the SIR model (4.1) should also hold for the survival function and the distribution function because of this equivalence. In that case the expected value of the non-homogeneous stochastic process is modeled in the same way as the solution of the deterministic equation. The distribution functions of the Burr family have precisely this property, that is, these distribution functions satisfy the differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}F(t) = \beta(t)\,F(t)\,(1-F(t)) \text{ or } \frac{\mathrm{d}}{\mathrm{d}t}S(t) = -\beta(t)\,S(t)\,(1-S(t)) \tag{4.4}$$

for some non-negative function $\beta(\cdot)$, (Kleibner and Kotz, 2003, page 52). This function $\beta(\cdot)$ can also depend on $F(t)$. When it does only depend on t the solution of (4.4) reduces to the logistic form: $F(t) = \left[1+\exp\left(-\int_0^t \beta(\tau)\mathrm{d}\tau\right)\right]^{-1}$ (Burr, 1942).

There are 12 known distributions in this family, denoted Burr I to Burr XII. The Burr I is the uniform distribution and a special case of the Burr II is the logistic distribution.

The logistic distribution occurs frequently in epidemic modelling since it is the well known solution of (4.1) in terms of fractions, see, for instance, Renshaw (1991). The Burr family can thus been seen as a generalization of the logistic distribution. The distribution function of the Burr II is: $F(t) = \left(1+e^{-\frac{t}{b}}\right)^{-p}$, which reduces to

the logistic distribution function for $p = 1$. This can also be seen by writing (4.4) for the Burr II in terms of the logistic distribution function. If $F_L(t)$ denotes the logistic distribution function then (4.4) for the Burr II can be written as $\frac{\mathrm{d}}{\mathrm{d}t}F(t) = \frac{p}{b}\left[1 - F_L(t)\right]\left[F_L(t)\right]^p$, which reduces to the differential equation for the logistic case for $p = 1$ with constant $\beta(t) = 1/b$. This also shows that, compared to the logistic case, in the Burr II case the infected fraction ($F_L(t)$) is weighted and can receive more or less weight. Note that the Burr II and the logistic distribution are defined on the whole real line and therefore can not be interpreted as the distribution function of the infection times. This can be remedied by incorporating $\ln\left(\frac{F(0)}{1-F(0)}\right)$, the log-odds of the infected fraction at time 0, as an offset. If this fraction is very small then the infection-time distribution is approximately a distribution function for positive values. In order to put this offset in the model one has to know the number of susceptibles at time 0, so this model can only be used in the non-homogeneous death process case.

Only three distributions are defined on positive values, which is necessary because they are used to model time. Two of these are the most useful ones, the Burr III and the Burr XII.

The Burr XII is known as simply the Burr distribution or the Singh-Maddala distribution (Kleibner and Kotz, 2003). It has distribution function:

$$F(t) = 1 - \left[1 + \left(\frac{t}{b}\right)^a\right]^{-q}, \quad t > 0,\ a, b, q > 0$$

The right tail is governed by the parameters a and q, the left tail by a and b is the scale parameter (Kleibner and Kotz, 2003, page 198).

The logistic form is obtained for $q = 1$ giving the log-logistic or the Fisk distribution (Kleibner and Kotz, 2003): $F(t) = \frac{\left(\frac{t}{b}\right)^a}{1+\left(\frac{t}{b}\right)^a} = \frac{\exp(-a\ln(b)+a\ln(t))}{1+\exp(-a\ln(b)+a\ln(t))}$. Equation (4.4) for the Burr XII distribution can also be written in the logistic form. Denote by $F_{LL}(t)$ the log-logistic distribution function then: $\frac{\mathrm{d}}{\mathrm{d}t}F(t) = \frac{qa}{t}\left[1 - F_{LL}(t)\right]^q\left[F_{LL}(t)\right]$. As compared to the log-logistic case the susceptible fraction is weighted up or down.

For $a = 1$ the Burr distribution reduces to the Lomax (Pareto type II) distribution. The case $a = q$ is also known as the para-logistic distribution. The Weibull distribution and the Pareto distribution are limiting cases of the Burr distribution (Shao, 2004). Another interesting way to arrive at the Burr distribution is by assuming the infection times have the Weibull distribution the scale parameter of which follows an inverse generalized gamma distribution.

The Burr III is also known as the Dagum distribution or as the inverse Burr. This last name is not surprising since if X has a Burr distribution then $1/X$ has the inverse

Burr distribution. The distribution function of the inverse Burr is:

$$F(t) = \left[1 + \left(\frac{t}{b}\right)^{-a}\right]^{-p}, \quad t > 0,\ a, b, p > 0$$

The inverse Burr has a more flexible shape than the Burr and the roles of the parameters are reversed: there is one parameter for the upper tail and two for the region where the largest part of the data is situated (often around the origin) (Kleibner, 1996) (Kleibner and Kotz, 2003).

Equation (4.4) written in the logistic form becomes: $\frac{\mathrm{d}}{\mathrm{d}t}F(t) = \frac{pa}{t}\left[1 - F_{LL}(t)\right]\left[F_{LL}(t)\right]^p$ from which the log-logistic distribution is obtained by taking $p = 1$, see also Shao (2000). As compared to the log-logistic the infected fractions are weighted up or down.

Another way to obtain the inverse Burr for our application is to assume that the infection times have the flexible generalized gamma distribution the scale parameter of which follows an inverse Weibull distribution.

In short: The log-logistic is the logistic distribution on log-time and, roughly, from this log-logistic the Burr distribution is obtained by weighting the survival function (the fraction susceptibles) up or down, and roughly the inverse Burr is obtained by weighting the distribution function (the infected fraction) up or down.

4.2.4 The end of the epidemic

When no measures are taken during an outbreak the epidemic could end by a biological process, for example because the size of the population of susceptibles has (locally) decreased too much to prevent fade out. But usually measures are taken to influence the final size of the epidemic in the sense that one wants the final size to be as low as possible. This makes the end of the outbreak very different from the start; the start can be described well with an epidemic model, for the end this is much harder. To incorporate at least part of this in the model the survival function could be modified in such a way that the expected value of the number of susceptibles goes to a final size value in the long run, i.e. $S(t) = 1 - \pi F(t)$ where $F(t)$ is the Burr or the inverse Burr. So, when the distribution function goes to 1 there will be a fraction $1 - \pi$ still not infected. This is called long-term survival in the literature (Shao and Zhou, 2004).

For the non-homogeneous death process one could take $\pi = 1 - \frac{\theta}{x_0}$ where θ is the expected final size of the outbreak, i.e. given the control measures imposed, and x_0 is the number of susceptibles at the start of the outbreak. Because here the number of

susceptibles at the start is known, a long-term survival interpretation makes sense: a fraction $1-\pi$ of the population never got infected.

For the non-homogeneous birth process things are different. If $\pi = 1 - \frac{y_0}{\theta}$, with y_0 being the number of infected that starts the epidemic, then the expected value of the total number infected is θ. A long term survival interpretation here makes no sense since nothing is assumed known about the population size. One can interpret $1-\pi$ as that fraction of the total number of infected that were generated by the y_0 infected that started the epidemic.

4.2.5 Fitting the models

In general one observes just one outbreak which can be interpreted as a sample path if the model outlined above is imposed as the generator of the process. This means that the model should be fitted conditionally on the past. Furthermore, in practice the outbreak is observed in discrete time (Becker, 1989, page 108) with, say, $t_j, j = 0, ..., n$ as the observation times. So at time point t_j one models the number of susceptibles or the number of infectives conditionally on what was observed at time point t_{j-1}.

With the non-homogeneous death model the process starts at t_0 with x_0 susceptibles. At t_1 the number of susceptibles has a binomial distribution with x_0 and probability $S(t_0)$. We introduce the random variable T which measures the time from the start of the epidemic until infection. Suppose at time point t_{j-1} there were $x_{t_{j-1}}$ susceptibles observed, then the number of susceptibles at time point t_j, conditional on the number of susceptibles on time t_{j-1}, has a binomial distribution with $x_{t_{j-1}}$ and with probability: $P(T > t_j | T > t_{j-1}) = \frac{S(t_j)}{S(t_{j-1})} = 1 - h(t_{j-1})$, where $h(t)$ is the discrete-time hazard rate (see also section 2.3). So at time t_j the distribution of the number of susceptibles conditional on the number of susceptibles at t_{j-1} is:

$$P(X(t_j) = x_{t_j} | X(t_{j-1}) = x_{t_{j-1}}) = \binom{x_{t_{j-1}}}{x_{t_j}} [1 - h(t_{j-1})]^{x_{t_j}} [h(t_{j-1})]^{x_{t_{j-1}} - x_{t_j}} \quad (4.5)$$

The expected value on time point t_j is $x_{t_{j-1}}(1 - h(t_j)), j = 0, ..., n$ and might be called the individual or sample path profile (Lindsey, 2001). The diffrence between the individual and the underlying profile is that the former relates to the conditional model above whereas the later relates to the unconditional model. This model is very similar to the model of Becker in chapter 6, page 109 Becker (1989).

If δ is the vector of parameters for the Burr distribution, then from (4.5) the

log-likelihood $l(\delta)$ can be written as:

$$l(\delta) = \sum_{j=1}^{n} \log \binom{x_{t_{j-1}}}{x_{t_j}} + x_{t_j} \log\left[1 - h(t_{j-1})\right] + (x_{t_{j-1}} - x_{t_j}) \log\left[h(t_{j-1})\right]$$

where $h(t_{j-1}) = 1 - \frac{S(t_j)}{S(t_{j-1})}$, $S(t) = 1 - \pi F(t)$ and $F(t)$ a distribution from the Burr family. This log-likelihood can be maximized using an optimization procedure like the quasi-Newton method to find the maximum likelihood estimates. For this purpose the package R can be used (R Development Core Team, 2010). The information matrix can be used to find the standard errors.

With the non-homogeneous birth model the same arguments lead to the distribution of the number of infected at time t_j conditionally on the number of infectives at time t_{j-1}. The process takes off at time 0 with y_0 infectives. So the number of infected at time t_1 has a shifted negative binomial with y_0 and probability $S(t_0)$. The probability of having y_{t_j} infectives at time t_j, given that there were $y_{t_{j-1}}$ infectives at time t_{j-1}, is then given by

$$\begin{aligned} P(Y(t_j) &= y_{t_j} | Y(t_{j-1}) = y_{t_{j-1}}) = \binom{y_{t_j} - 1}{y_{t_{j-1}} - 1} \left[1 - h(t_{j-1})\right]^{y_{t_{j-1}}} \left[h(t_{j-1})\right]^{y_{t_j} - y_{t_{j-1}}} \\ & y_{t_j} = y_{t_{j-1}}, y_{t_{j-1}} + 1, y_{t_{j-1}} + 2, \cdots . \end{aligned}$$

The expected value for time point t_j is $\frac{y_{t_{j-1}}}{1 - h(t_{j-1})}$, the sample path profile.

This model can be fitted as an ordinary negative binomial by taking $Z(t_j) = Y(t_j) - Y(t_{j-1})$. Then

$$\begin{aligned} P(Z(t_j) &= z_{t_j} | Y(t_{j-1}) = y_{t_{j-1}}) = \binom{z_{t_j} + y_{t_{j-1}} - 1}{y_{t_{j-1}} - 1} \left[1 - h(t_{j-1})\right]^{y_{t_{j-1}}} \left[h(t_{j-1})\right]^{z_{t_j}} \\ & z_{t_j} = 0, 1, \cdots . \end{aligned} \tag{4.6}$$

So, conditioning on the previous observed number of cases makes the process start again in the same way as it was started at time 0. The differences are that there are more infected to start the process and that it starts at another time point.

If δ is the vector of parameters for the Burr distribution, then from (4.6) the log-likelihood $l(\delta)$ can be written as:

$$l(\delta) = \sum_{j=1}^{n} \log \binom{z_{t_j} + y_{t_{j-1}} - 1}{y_{t_{j-1}} - 1} + y_{t_{j-1}} \log\left[1 - h(t_{j-1})\right] + z_{t_j} \log\left[h(t_{j-1})\right]$$

Table 4.1: Parameter estimates for the classical swine fever outbreak

Burr distribution			Inverse Burr distribution		
Parameter	Estimate	St.Error	Parameter	Estimate	St.Error
$\ln(b)$	7.654	11.320	$\ln(b)$	2.477	0.107
$\ln(a)$	-0.271	0.147	$\ln(a)$	1.330	0.074
$\ln(q)$	5.124	8.090	$\ln(p)$	-2.672	0.370
$\ln(\theta)$	6.079	0.319	$\ln(\theta)$	5.993	0.319

where $h(t_{j-1}) = 1 - \frac{S(t_j)}{S(t_{j-1})}$, $S(t) = 1 - \pi F(t)$ and $F(t)$ a distribution from the Burr family. This log-likelihood can be maximized in the same way as the log-likelihood of the non-homogeneous death process giving the maximum likelihood estimates and, by using the information matrix, the standard errors can also be obtained.

4.3 Three outbreaks

4.3.1 The classical swine fever outbreak in the Netherlands in 1997-1998

Classical swine fever is an infectious viral disease that occurs in domestic pigs and wild boar under natural conditions. An outbreak of the disease can lead to huge financial losses in the pig-production industry. In the Dutch classical swine fever outbreak during the years 1997-1998, the first case was detected on February 4, 1997 on a mixed sow and finishing pig farm. The epidemic lasted 57 weeks and there were in total 427 herds infected and approximately 700,000 pigs from these herds were slaughtered. The number of new outbreaks per week is given in Figure 4.1. The first week with nine outbreaks is considered to be time point zero. For more information about this outbreak the reader might want to consult a special issue on this subject of *Preventive veterinary medicine 42, 1999.*

Since it is not completely clear how large the number of susceptible farms was at the start of the outbreak and because mainly (detected) infecteds are registered, the data are fitted with the non-homogeneous birth process.

With the Burr distribution the parameter q only affects the right tail. As can be seen from Figure 4.1 the right tail does not contain much information. There is a relatively large number of weeks with no new cases. This is why the standard

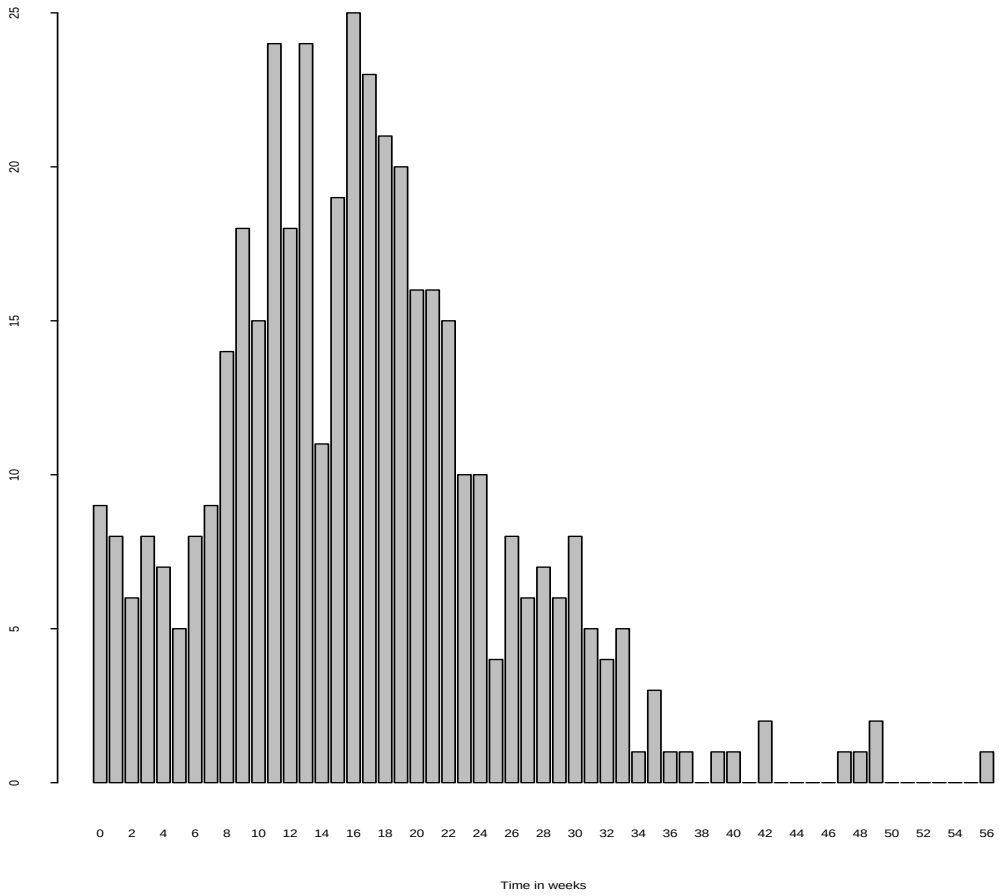

Figure 4.1: New cases in the classical swine fever outbreak in The Netherlands, 1997-1998 by week

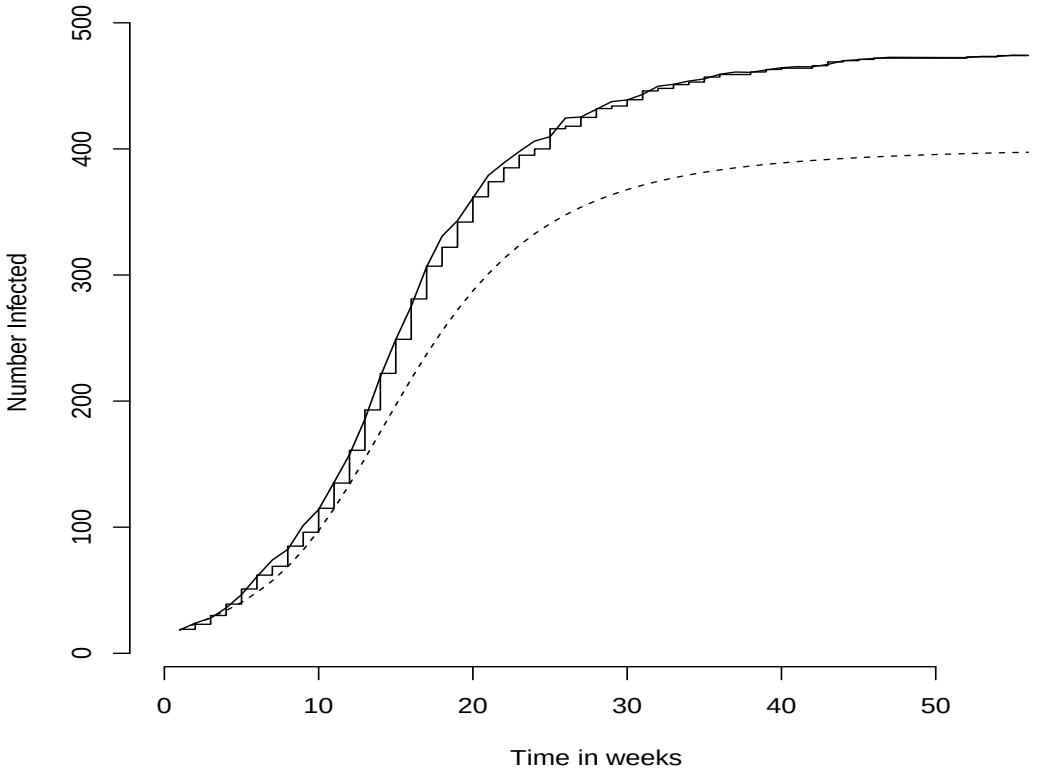

Figure 4.2: The classical swine fever outbreak data (step line)by time in weeks. The individual profile (solid line) and the underlying profile (dashed line) for the inverse Burr

error of the logarithm of this parameter is very large, see Table 4.1. From the same Figure one can see that the distribution of the new cases is not very peaked so the standard error of the logarithm of the parameter b is also very large. These large standard errors indicate a close to singular information matrix. The AIC for the Burr distribution is 222.9. With the inverse Burr things are reversed, now two parameters are used to describe the body of the data and one parameter for both tails which is in this case an advantage. The AIC is 221.9 and the estimates for the logarithm of the parameters are in Table 4.1. The individual profile and the underlying profile for the inverse Burr are in Figure 4.2. The individual profile shows that this model fitted the data well. The expected total number infected of the underlying profile is estimated with the inverse Burr as 400.8 while the observed final size was 427.

The reproductive power function for the inverse Burr is shown in Figure 4.3. Stegeman et al. in (Stegeman et al., 1999) distinguish several phases of the outbreak according to the measures taken to control the spread the epidemic. The first phase is the detection of the first case. In the second phase measures like stamping out, establishing a protection zone and a transportation ban were taken. This phase lasted until April 20, 1997, which is in week 11 after the start of the outbreak. Then the third phase began with additional measures. The fourth phase started around mid-June and the last phase was the last part of the outbreak. The reproductive power function for the inverse Burr seems to reflect these first four phases pretty well: it starts off in the neighbourhood of 0.3, then decreases quickly to a value just below 0.2 were it stays approximately until week 10-11 and then drops again until about week 20 after which it still decreases but less fast.

4.3.2 The avian influenza (H7N7) outbreak in the Netherlands in 2003

On February 28, 2003 an epidemic of avian influenza (H7N7) started in the Gelderse Vallei in the Netherlands, spreading to adjacent areas and to the province of Limburg. In total 239 flocks were infected with known detection date. The epidemic was controlled by movement restrictions, stamping out of infected flocks, and pre-emptive culling of flocks in the neighbourhood of infected flocks. In total 1,255 commercial flocks and 17,421 flocks of smallholders had to be depopulated. Approximately 25.6 million animals were killed. The virus was also transmitted to humans that had been in close contact with the infected poultry, leading to one human death. For more information on this outbreak see Stegeman *et al.* (2003).

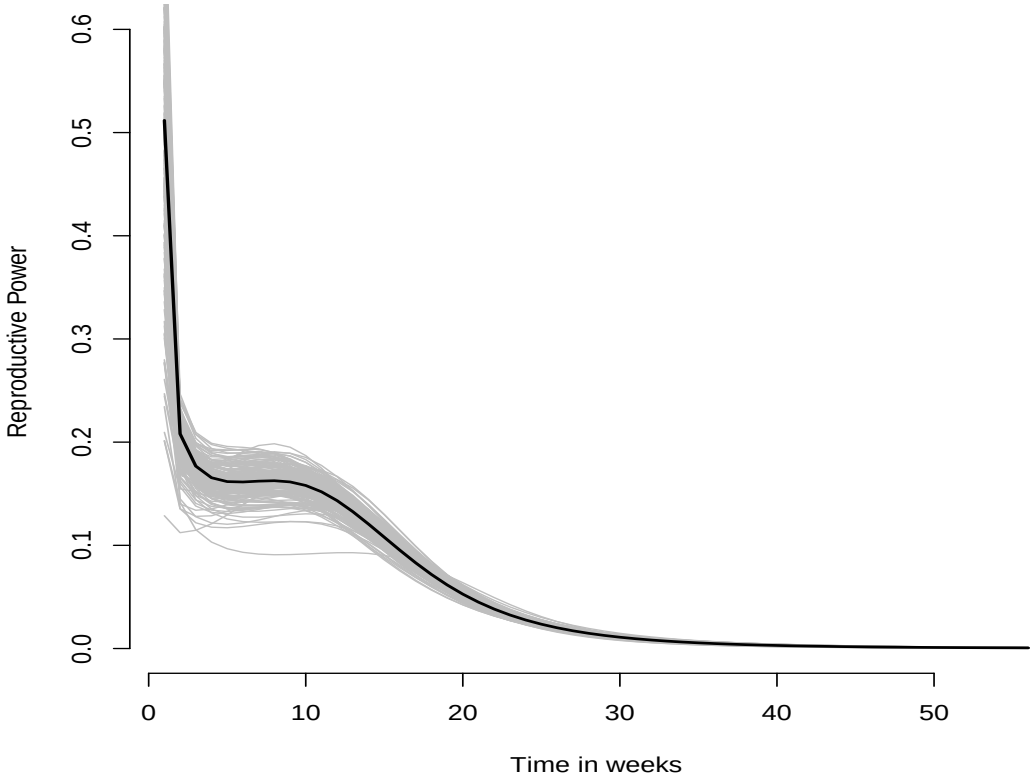

Figure 4.3: The reproductive power the inverse Burr for the classical swine fever outbreak data against time in weeks calculated from 200 sample paths drawn from the underlying estimated process (gray lines) and the estimated value (black line).

Table 4.2: Parameter estimates for the avian influenza (H7N7) outbreak

Burr distribution			Inverse Burr distribution		
Parameter	Estimate	St.Error	Parameter	Estimate	St.Error
$\ln(b)$	11.720	28.271	$\ln(b)$	2.440	0.350
$\ln(a)$	-0.569	0.136	$\ln(a)$	0.850	0.151
$\ln(q)$	6.261	15.705	$\ln(p)$	-2.232	0.612
$\ln(\theta)$	5.439	0.430	$\ln(\theta)$	5.418	0.434

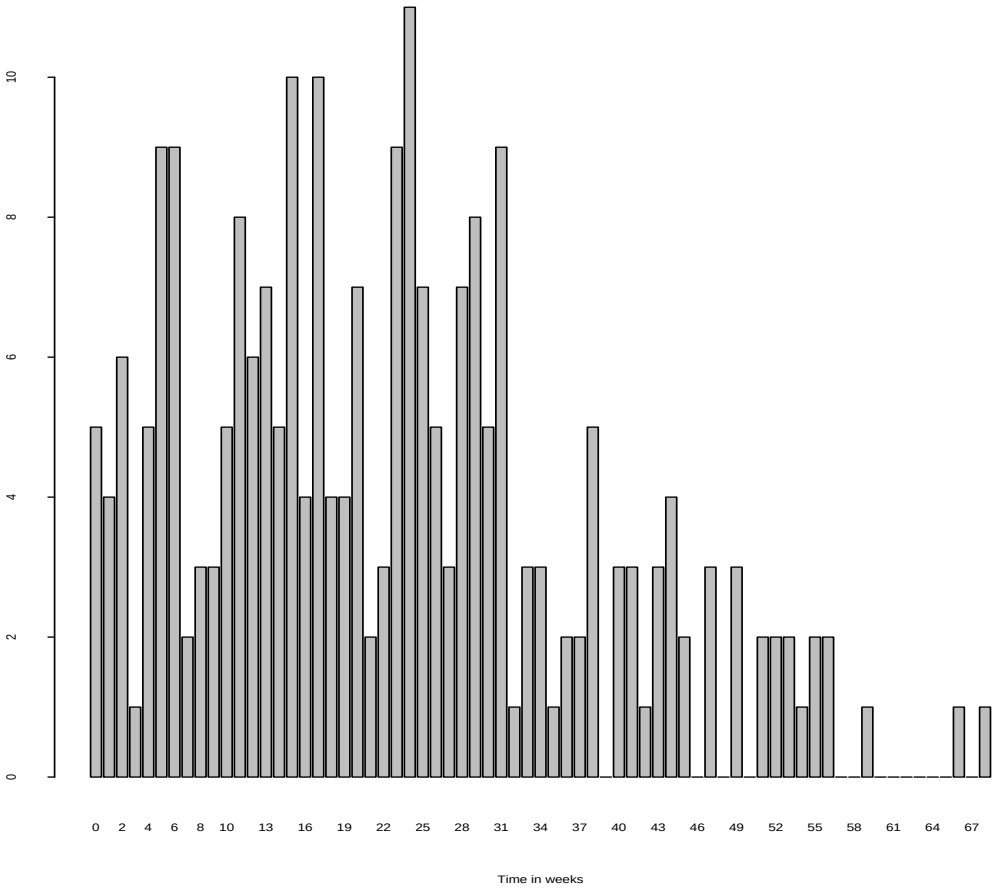

Figure 4.4: New cases in the avian influenza (H7N7) out break by day

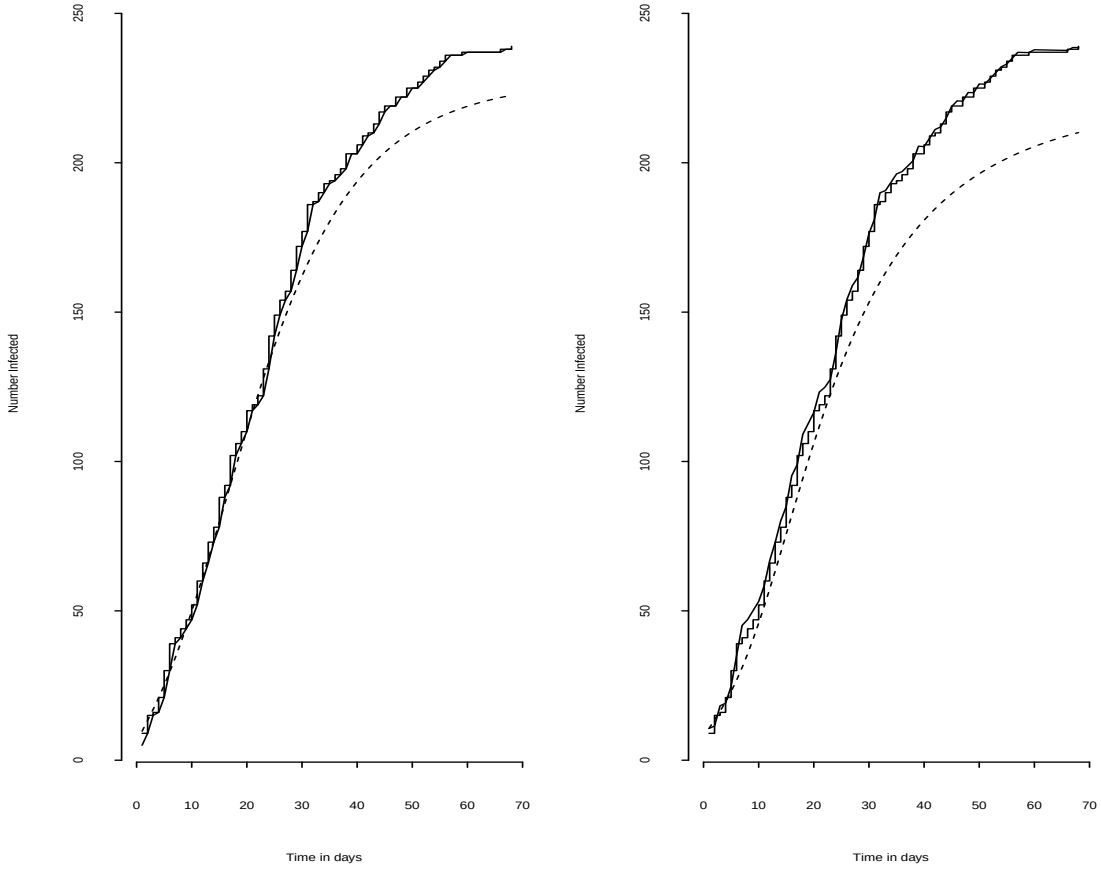

Figure 4.5: The avian influenza (H7N7) outbreak data (step line)against time in days. The individual profile (solid line) and the underlying profile (dashed line). Left for the Burr and right for the inverse Burr

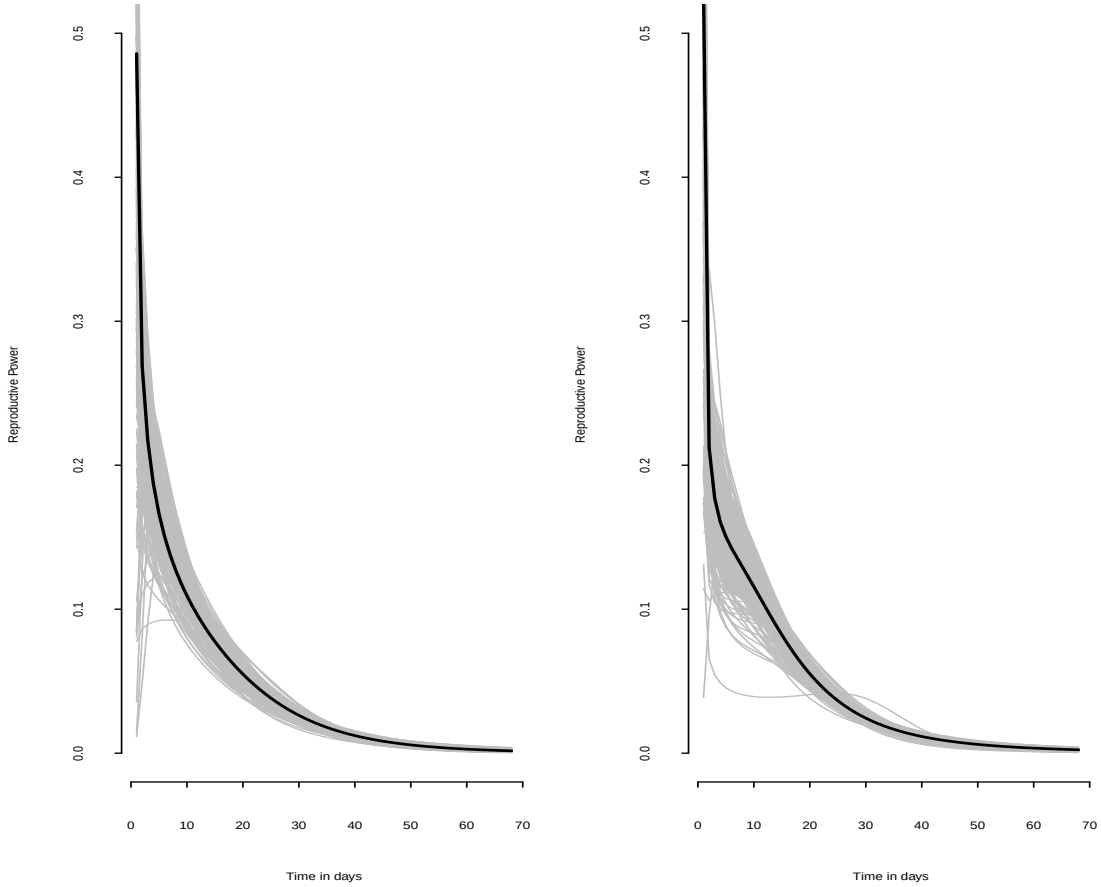

Figure 4.6: The reproductive power for the Burr (left) and the inverse Burr (right) for the avian influenza (H7N7) outbreak data against time in days calculated from 200 sample paths drawn from the underlying estimated process (gray lines) and the estimated value (black line).

For the same reasons as for the classical swine fever outbreak, the non-homogeneous birth model is more appropriate here. As compared to the classical swine fever outbreak things seem even worse here. The right tail contains many gaps indicating days with no new cases and the distribution seems to be almost flat. Not surprising the logarithm of the tail parameter and the logarithm of the scale parameter have very large standard errors. So also here the information matrix is probably close to singular. The AIC for this fit is 270.1. The fit of the inverse Burr does not show these large standard errors. The AIC though has a value of 274.8 indicating that the Burr give a better fit. The logarithms of the estimates are in table 4.2 and Figure 4.5 shows the underlying profile and the individual profile, the latter showing that the model gave a good fit. Note that the parameter estimates are quite similar to those of the classical swine fever outbreak which might indicate a similar outbreak process. The expected total number of infecteds of the underlying profile was estimated as 230.2 for the Burr and as 225.4 for the inverse Burr.

Stegeman *et al.* (2003), in an epidemiological description of the outbreak, distinguish roughly three periods. In the first week the number of diagnosed outbreaks increased. Subsequently, the number of outbreaks per day fluctuated between 2 and 11 until the end of March. During this period outbreaks were only detected in the Gelderse Vallei. By the end of March, the virus had escaped to the Southern part of The Netherlands; between 0 and 5 outbreaks per day were observed throughout most of April. The plot of the reproductive power shown in Figure 4.6 seems to approximately reproduce these periods. For the inverse Burr case the reproductive power starts off at 0.3, decreases sharply until about 5 days, where it bends and afterwards decreases less sharply. Then after about 30 days the decrease levels off. The curve for the Burr case is approximately the same except that it is smoother in the first 10 days.

4.3.3 The foot and mouth disease outbreak in Great Britain and the republic of Ireland in 2001

Foot and mouth disease is a highly transmissible viral infection, which can spread very rapidly. An outbreak of foot and mouth disease began in Great Britain in February 2001, 34 years since the last major outbreak. The primary infection was a pig herd in Northumberland in early February. From there the disease spread rapidly via long distance animal movements, and locally via contact and wind borne transmissions (Ferguson et al., 2001). For more information on foot and mouth disease see the website of the British Department of environment foot and rural affairs:

Table 4.3: Parameter estimates for the foot and mouth disease outbreak

Burr distribution			Inverse Burr distribution		
Parameter	Estimate	St.Error	Parameter	Estimate	St.Error
$\ln(b)$	-1.007	0.120	$\ln(b)$	-0.189	0.334
$\ln(a)$	0.667	0.234	$\ln(a)$	0.727	0.028
$\ln(q)$	0.039	0.246	$\ln(p)$	-1.581	0.688
$\ln(\theta)$	7.654	0.378	$\ln(\theta)$	7.663	0.320

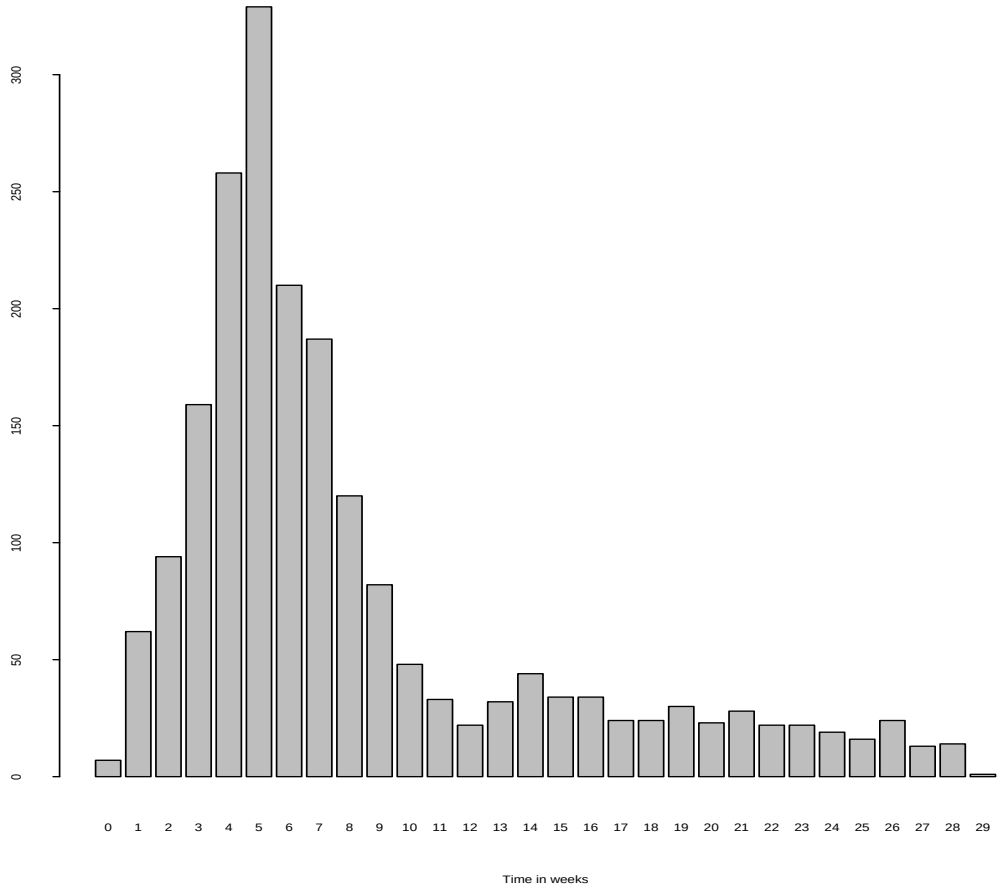

Figure 4.7: New cases in the foot and mouth disease outbreak by week

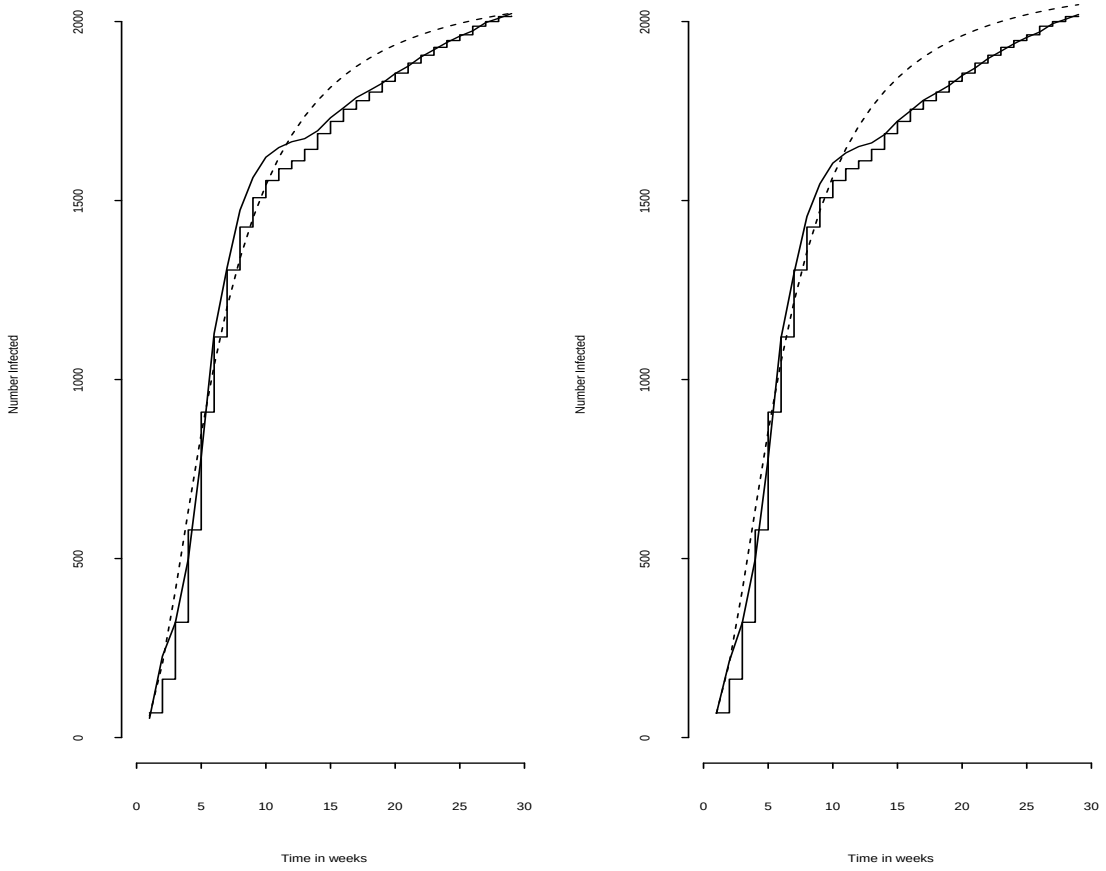

Figure 4.8: The foot and mouth disease outbreak data (step line). The individual profile (solid line) and the underlying profile (dashed line). Left for the Burr and right for the inverse Burr

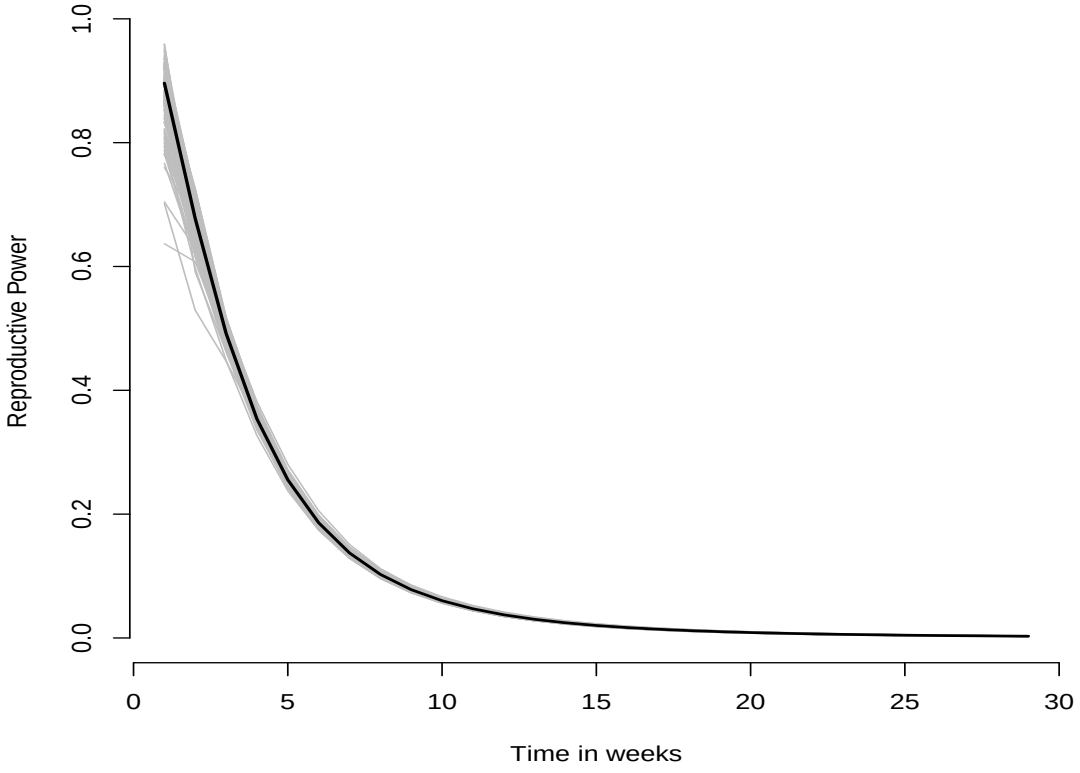

Figure 4.9: The reproductive power for the Burr for the avian influenza (H7N7) outbreak data against time in weeks calculated from 200 sample paths drawn from the underlying estimated process (gray lines) and the estimated value (black line).

$http://www.defra.gov.uk/footandmouth$. There were in total 2015 detected farms registered.

As can be seen from Figure 4.7 this outbreak is quite different from the former two. It has a clear peak around 5 weeks and has a strikingly long tail with no zero gaps. In this case the Burr distribution for the infection times, with its special parameter for the right tail, fits the data well (AIC=451.2). Nevertheless the inverse Burr fits the data slightly better (AIC=449.2). The parameter estimates are in Table 4.3, the underlying profile and the individual profile for both the distributions are in Figure 4.8. When comparing the parameter estimates of the inverse Burr with the outbreaks discussed above one will note that they are quit different, suggesting that the foot and mouth disease outbreak is different from the former two. After 5 weeks the number of new outbreaks drops fast until about 10 weeks were after it stabilizes on a value around 25. This can also be seen from the individual profiles in Figure 4.8 where both the Burr and the inverse Burr have a difficulty to follow the sharp bend in the curve. The reproductive power of both distributions is given in Figure 4.9; they are hardly distinguishable. They both drop fast until about 5 weeks then decrease more slowly until about 10 weeks, after which the decrease levels off. The expected total number of infecteds of the underlying profile is estimated as 2109.1 by the Burr and as 2128.1 by the inverse Burr.

4.4 Discussion

The generalized non-linear models described in this Chapter incorporate some issues met when an epidemic disease outbreak is studied. The dependence of the data is handled with a non-homogeneous death or a non-homogeneous birth process. The prevalence is related to the censored infection times in such a way that the distribution function and thus the survival function agrees with the first equation of the SIR model for infection spread, incorporating the fact that cases are reported with some delay. The non-homogeneous birth process handles the fact that in practice (detected) infecteds are registered, rather than susceptibles. While the first phase of the outbreak might be well described with an epidemic model, the end of the outbreak is a different process influenced by the measures taken. This is incorporated in our analysis by modifying the survival function with a final size parameter in the same way as is done in long-term survival modelling. The three examples show that the approach with the inverse Burr distribution is very capable of fitting these outbreak data even in a "messy" case such as the Dutch avian influenza (H7N7) outbreak. Furthermore, one can differentiate between different kinds of outbreak processes. This was clearly the

case with the foot and mouth disease as compared to the other two. The model with the Burr distribution gave a reasonable fit with the foot and mouth disease outbreak, whereas in the other two cases it gave problems with the information matrix. In addition, the parameter estimates of the inverse Burr were clearly different with the foot and mouth disease outbreak. The parameter estimates of the classical swine fever outbreak and the avian influenza (H7N7) outbreak were reasonably similar indicating similar outbreak processes. The classical swine fever outbreak was a single outbreak in a relatively small area. The same is true for the avian influenza outbreak (H7N7) except that at the end there was a small outbreak in another area. Things were different for the foot and mouth disease outbreak. Early in the outbreak infected sheep were moved all over the country which led to the almost simultaneous introduction of foot and mouth disease into different areas. This led to several distinct outbreaks in different counties not starting all at the same point in time. The features of the distinct outbreaks are lost when the data for these separate outbreaks are pooled.

This chapter focusses on the application of the non-homogeneous birth process since with spontaneous outbreaks this seems the most appropriate. The non-homogeneous death process however will be applicable in a more controlled setting such as transmission experiments. In that case the number of susceptibles at the start of the outbreak is known by design and it can also be determined wether a susceptible is still a susceptible at a certain point in time.

It seems that the reproductive power function is an interesting concept because it appears to mimic the various phases during the outbreaks when different sets of control measures were active. This might be explained by noting that the reproductive power $\mu(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\ln[1-\pi F(t)]$ depends on the final size parameter through π and on the fraction of detected infecteds. The final size parameter dominates the final phase of the outbreak which is mainly caused by the measures taken. This does not mean that the epidemic process itself is not affected by the imposed measures. Measures taken relatively early restrict the population at risk of being infected. This means that the infectives at time t are less capable of generating new cases. This influences the fraction of infecteds after time t and thus it influences $F(t)$ after time t.

In the above sense the reproductive power could be called an outbreak signature. If one wishes to track the effectiveness of (sets of) control measures over time, such a function might play a valuable role and merits closer scrutiny.

As can be seen from the formula for $\mu(t)$, estimating, for example, standard errors and confidence bands for the reproductive power function does not seem to be an easy task. Besides this one may also carefully think about the interpretation of such bands since with epidemic outbreaks one usually does not have a sample so the repeated sampling interpretation might be difficult. One just has one sample path as

an observation.

Chapter 5

Wild Birds and Increased Transmission of Highly Pathogenic Avian Influenza (H5N1) among Poultry, Thailand [1]

5.1 Introduction

Avian influenza is a viral disease of poultry and is distributed worldwide. The virus is classified based on 2 surface proteins, the hemagglutinin (HA) protein (H1–H16) and the neuraminidase (NA) protein (N1–N9), which can be found in numerous combinations (Alexander, 2007). All H and N subtypes can be found as low pathogenic avian influenza virus strains in aquatic wild birds, which are assumed to be the main reservoirs outside poultry (Fouchier et al., 2005; Webster et al., 1992). Occasionally, low pathogenic avian influenza virus strains are introduced into domestic poultry flocks with no clinical signs or only mild clinical consequences, but strains carrying the H5 or H7 gene can mutate into highly pathogenic avian influenza (HPAI) strains that cause high death rates in domestic poultry (Alexander, 2000) and, occasionally, in migratory birds (Gauthier-Clerc, Lebarbenchon and Thomas, 2007; Olsen et al., 2006). Because of the devastating effect of HPAI outbreaks in commercial poultry,

[1]Extented version of: Keawcharoen, J., Van den Broek, J., Bouma, A., Tiensin, T., Osterhaus, A.D.M.E. and Heesterbeek, H. (2011). Wild Birds and Increased Transmission of Highly Pathogenic Avian Influenza (H5N1) among Poultry, Thailand.
Emerging Infectious Diseases 17, No. 6, 1016-1022.

all outbreaks caused by H5 and H7 subtypes are notifiable (World Organisation for Animal Health, 2010).

Currently, a HPAI virus strain of subtype H5N1 is circulating in many countries in Eurasia and Africa, causing high death rates in poultry, substantial economic losses, and human deaths. The strain was first identified in Southeast Asia in 1996 and has since spread to 63 countries in Asia, Europe, Africa, and the Middle East (World Organisation for Animal Health, 2010). Millions of domestic poultry died from the effects of the disease or from culling efforts to control the spread of the virus (Alexander, 2007; Fouchier et al., 2005; Claas, 1998; Peiris,Jong and Guan, 2007). The spread of the HPAI (H5N1) virus from Southeast Asia to Russia, Europe, and Africa was assumed to originate from a virus source at Qinghai Lake, People's Republic of China (Olsen et al., 2006; Spackman, 2008). Therefore, migratory birds were considered to be responsible for long distance dispersal of the virus (Boyce et.al., 2009; Kilpatrick et al., 2006; Liu et al., 2005).

In Thailand, 7 waves of HPAI (H5N1) virus outbreaks have occurred since January 2004. Poultry and wild bird populations in 1,417 villages in 60 of the 76 provinces were affected, and $>$ 62 million birds died or were culled to prevent further transmission (Siengsanan et al., 2009; Songserm et al., 2006; Tiensin et al., 2009). Introduction of the virus into poultry flocks is considered to be possible through infected wild birds. Additional insight on the basis of quantitative data into the role of wild birds would be necessary to further develop control measures and surveillance programs.

Relatively little effort has been made to quantify the association between infection in wild birds and outbreaks in poultry flocks, most likely because of the lack of data on infection in wild birds. Recently, a preliminary study was carried out that analyzed the prevalence of HPAI (H5N1) infection in wild birds in Thailand (Siengsanan et al., 2009). In that study, 6,263 pooled surveillance samples from wild birds in Thailand, collected from January 2004 through December 2007, were tested for evidence of infection. Testing indicated that prevalence patterns in wild birds mirrored outbreaks among poultry; however, the association was not proven or quantified. We studied extensive data on 24,712 wild birds, sampled and analyzed from 2004 through 2007 in Thailand, to quantify the possible effect of infection in wild birds on the spread of the infection among poultry flocks.

5.2 Materials and Methods

5.2.1 Data Collection

Data about subtype H5N1 infections in wild bird populations were provided by the National Institute of Animal Health of Thailand, Regional Veterinary Research and Development Centers, the Veterinary Science faculty of Mahidol University, and the Department of Livestock Development, Thailand. A total of 24,712 wild bird samples were collected from January 2004 through December 2007. During 2004–2006, sampling was part of a general countrywide surveillance program; in 2007, sampling was targeted specifically at areas where outbreaks in poultry had occurred.

Sampling methods have been described previously (Siengsanan et al., 2009; Tiensin et al., 2009, 2005). Wild birds were either trapped by using baited traps, hand nets, or mist nets, or shot. Tracheal/oropharyngeal swabs and cloacal swabs of live birds and bird carcasses were collected from active surveillance (sampling of healthy wild birds) and passive surveillance (sampling of sick or dead birds). Swab samples were collected in viral transport media, stored at 4^oC, and shipped to the laboratory, where they were stored at -80^oC until further analysis could be done.

5.2.2 Virus Detection

Methods used for antigen detection have been described by Tiensin et al. (Tiensin et al., 2009) and Siengsanan et al. (Siengsanan et al., 2009). Supernatants from homogenized tissue and swab samples were filtrated and inoculated in 11-day-old embryonated chicken eggs or MDCK cell cultures. After incubation at 37^oC for 3 days, allantoic fluid was harvested. The inoculated MDCK cell culture was observed daily for cytopathic effect, and supernatant fluid was harvested by day 4, even if no cytopathic effect was observed. Viruses were initially identified in allantoic fluids or culture supernatants by the HA assay according to World Health Organization recommendations (Siengsanan et al., 2009). Negative samples were inoculated 2 additional times in embryonated chicken eggs before specimens were confirmed as negative. RNA from positive samples acquired from virus culture was extracted by using a viral RNA extraction kit (QIAGEN, Valencia, California, USA), according to the manufacturer's instructions. Reverse transcription PCR (RT-PCR) was performed by using a 1-step RT- PCR ki (QIAGEN) to identify the subtype, according to the manufacturer's instructions. Primers for RT of viral genome and all HA, NA, and matrix (M) genes for virus subtype and influenza A virus identification have been published elsewhere (Siengsanan et al., 2009; Tiensin et al., 2005; Uchida et al., 2008;

Lee et al., 2001). PCR products were processed with 1% agarose gel electrophoresis and were purified by QIAquick PCR purification kit (OIAGEN). Sequencing was performed by using the H5 and N1 specific primers, and sequence data were edited following methods previously described (Siengsanan et al., 2009; Tiensin et al., 2005; Uchida et al., 2008).

5.2.3 Statistical Analysis

Data considerations

For each identified bird species, geographic location and season were recorded. Bird species were divided into 3 groups: 1) resident birds (nonmigratory populations), present year-round in Thailand; 2) migratory (visitor) birds, bird populations moving between Russia or China to Thailand during September/October and March/April; and 3) breeding visitor birds, which migrate to Thailand for breeding in different periods of the year. To study the relevance between the regions and subtype H5N1 outbreaks in wild birds, we divided Thailand into 4 major geographic regions (northern, northeastern, central, and southern) on the basis of the former administrative region grouping system used by the Ministry of Interior, Thailand. Because of the high number of outbreaks in the Central region (Siengsanan et al., 2009; Tiensin et al., 2005; Gilbert et al., 2006), it was further divided into 6 parts: central-northwest, central-north, central-central, central-east, central-southeast, and central-southwest.

Data on outbreaks among poultry were taken from (Tiensin et al., 2009). We used their definition of poultry, which encompasses all farmed avian species in Thailand, including backyard chickens and ducks. Different species or types of production systems were not differentiated in the data. Using a nonhomogeneous birth model (chapter 4), we investigated the association between subtype H5N1 presence in infected poultry flocks and wild birds. Prevalence data from the 9 different regions were modeled independently and conditioned on the number of infected birds during the first month of detected infection for each region. Time lapse was measured in months from the first month infection was detected. To analyze the association between presence of subtype H5N1 in wild birds and outbreaks in poultry, we pooled data for the 3 wild bird groups (resident birds, migratory visitor birds, and breeding visitor birds) to increase power. In most regions, sampling among wild birds was only done systematically after a poultry outbreak in that region, except in the central-northwest, central-north, and central-central regions (regions 2,3 and 4). We could therefore only use the latter 3 regions to investigate whether the presence of infected wild birds was related to the poultry outbreak.

The model

The stochastic version of the non-homogeneous birth process with $Y(t)$ the total number of infected (and detected) individuals at time t and y_0 the number of infected (and detected) individuals at time 0, has the following probability mass function (see chapter 4 for details):

$$P(Y(t) = y_t) = \binom{y_t - 1}{y_0 - 1} \left[e^{\rho(t)}\right]^{y_0} \left[1 - e^{\rho(t)}\right]^{y_t - y_0} \; y_t = y_0, y_0 + 1, \cdots$$

with $e^{\rho(t)} = \exp\left[-\int_0^t \lambda(\tau)\mathrm{d}\tau\right] = S_\lambda(t)$ and with y_t the number of infected. This is a shifted negative binomial distribution. It is the probability of obtaining $y_t - y_0$ infected (and detected) at time t in an epidemic that started with y_0 infected (and detected) at time zero.

The expected value of $Y(t)$ is the same as the solution of the deterministic equation for the number of infectived individuals. The expected value for the number of infected at time t is

$$E(Y(t)) = \frac{y_0}{S_\lambda(t)} \tag{5.1}$$

and the variance is:

$$var(Y(t)) = y_0 \left[\frac{1 - S_\lambda(t)}{S_\lambda^2(t)}\right] \tag{5.2}$$

The non-homogeneous birth model depends on the reproductive power λ: the rate at which infected individuals are capable of reproducing. We will call the time points at which such reproduction occurs: 'reproduction times' and study their distribution. In chapter 4, arguments are given to use a distribution from the Burr family. The most well known and useful distribution from the Burr family is the Burr XII, or Singh-Maddala distribution, which is sometimes referred to simply as the Burr distribution. The survival function is given by:

$$S(t) = \left[1 + \left(\frac{t}{b}\right)^a\right]^{-q}, \;\; t > 0, \; a, b, q > 0$$

The right tail is governed by the parameters a and q, the left tail by a, and b is the scale parameter (Kleibner and Kotz, 2003, page 198). To reduce the number of parameters to be estimated, one can consider three special cases of the Burr distribution (Kleibner and Kotz, 2003) :

1. The logistic form is obtained for $q = 1$ giving the log-logistic or the Fisk distribution.

2. For $a = 1$, the Burr distribution is reduced to the Lomax (Pareto type II) distribution.

3. The case $a = q$ is also known as the para-logistic distribution.

The Burr III distribution is also known as the Dagum distribution or as the inverse Burr. This last name is not surprising since if X has a Burr distribution then $1/X$ has the inverse Burr distribution. The distribution function of the inverse Burr is:

$$S(t) = 1 - \left[1 + \left(\frac{t}{b}\right)^{-a}\right]^{-p}, \quad t > 0,\ a, b, p > 0$$

Because in practice only one sample path of the outbreak is observed and usually in discrete time we use a conditional fitting procedure in which the survival distribution is replaced by the discrete hazard function $h(t_{j-1} = 1 - \frac{S(t_j)}{S(t_{j-1})}$ with $t_j, j = 0, 1, \cdots, n$ the n time points (see chapter 4). The log-likelihood for this conditional shifted negative binomial then is:

$$\begin{aligned} l(\cdot) &= \sum_{j=1}^{n} \log \binom{y_{t_j} - 1}{y_{t_{j-1}} - 1} + y_{t_{j-1}} \log\left[1 - h(t_{j-1})\right] + (y_{t_j} - y_{t_{j-1}}) \log\left[h(t_{j-1})\right] \quad (5.3) \\ & y_{t_j} = y_{t_{j-1}}, y_{t_{j-1}} + 1, y_{t_{j-1}} + 2, \cdots . \end{aligned}$$

Just like the more famous Weibull distribution the Burr XII (Singh-Maddala) distribution can be written as an accelerated time model and as a proportional rate model yielding two ways of incorporating covariates in the survival function. Here we use the proportional rate model. To employ the proportional rate model, let the survival distribution for the reproduction time of a non-wildbird detected month S_{nwm}, be a Burr distribution with parameters a, b and q_1, and suppose that the survival distribution of the reproduction times wildbird detetected month S_{wm}, is also a Burr distribution with parameters a, b and q_2. If we replace q_2 by cq_1 (where c is a constant), we get

$$S_{wm}(t) = \left[1 + \left(\frac{t}{b}\right)^{a}\right]^{-q_2} = \left[1 + \left(\frac{t}{b}\right)^{a}\right]^{-cq_1} = \left\{\left[1 + \left(\frac{t}{b}\right)^{a}\right]^{-q_1}\right\}^{c} = \{S_{nwm}(t)\}^{c}$$

indicating that the rates are proportional; i.e., the rate at which an infected-and-detected reproduces for a wildbird detected month is proportional to the rate at which an infected-and-detected individual reproduces in a non wildbird detected month.

The Burr XII distribution can also be written as both an accelerated event-time distribution and as a proportional rate distribution; i.e., $S_{wm}(t) = S_{nwm}(t')^c$. The Burr III is not a proportional rate model although the parameter p from the equation above can be model as a proportional rate parameter giving $F_{wm}(t) = \{F_{nwm}(t)\}^c$ were F is the distribution function instead of the survival function.

To model the effect of the independent variables the log of the parameter c is modeled linearly in the independent variables area (indicates area 2, 3 or 4) and wildbird infected month indicator (wb): $\ln(c) = lc1 + lc2 \times area + lc3 \times wb$

Both the Burr III and the Burr XII are used in the data analysis and it is checked in the next session, which of these two fits the data best.

5.3 Results

5.3.1 Descriptive Statistics

Infected poultry flocks and wild birds were found in all 9 regions during the study period. In Figure 5.1, we present the numbers of wild birds sampled per month for each of the 9 regions and outbreak data of subtype H5N1 in poultry flocks. A total of 24,712 wild birds were sampled, consisting of 303 species, 64 families, and 20 orders (online Table 1, www.cdc.gov/EID/content/17/6/ 1016-appT1.htm). Of these, 192 samples were positive for subtype H5N1, resulting in an overall prevalence of 0.78%. A relatively high number of wild birds positive for subtype H5N1 were detected from January 2004 through May 2004, before the poultry outbreaks in June 2004. Infections in wild birds were consistently detected after the poultry outbreaks had ended, except during April and May in 2005, 2006, and 2007. The spatial distribution and size classes of infected poultry flocks, as well as numbers of infected wild birds detected, are shown in Figure 5.2. In 2004 and 2005, infected wild birds were reported in the same locations where infected poultry flocks were found, especially in the central region. No infected poultry flocks were found in 2006 and 2007 in these areas.

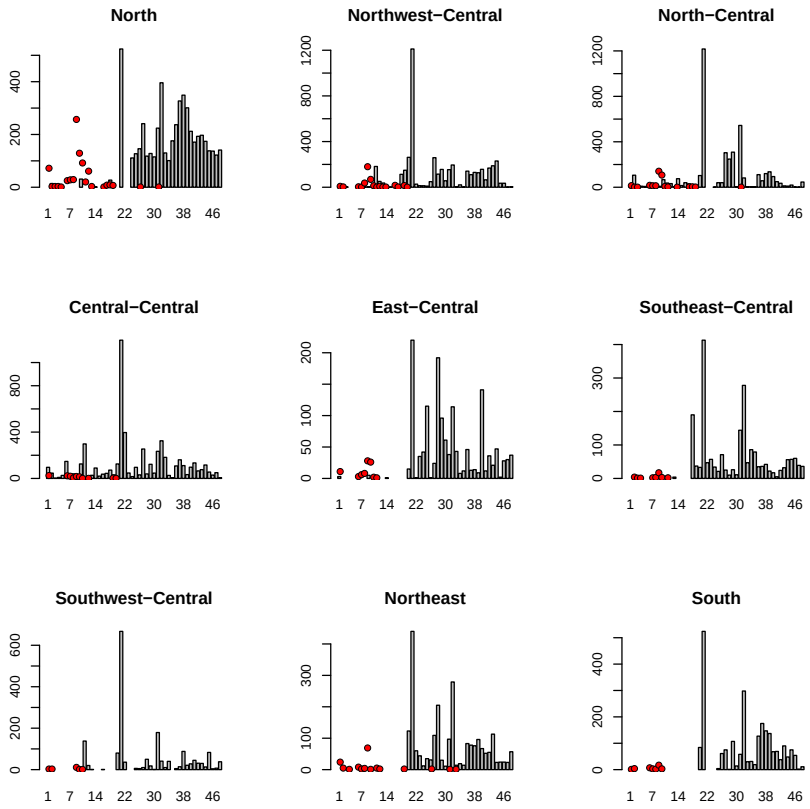

Figure 5.1: The numbers of wild birds sampled (vertical axes) per month (horizontal axes) for each of the 9 regions and outbreak data of subtype H5N1 in poultry flocks. Bullets represent the number of infected poultry

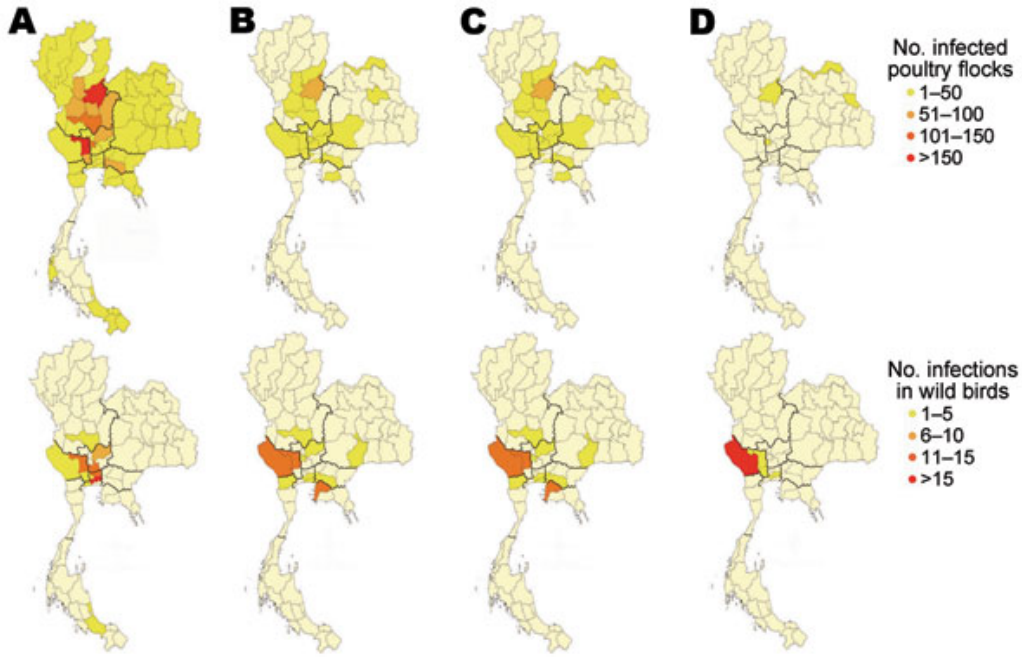

Figure 5.2: The spatial distribution and size classes of infected poultry flocks, as well as numbers of infected wild birds detected.**A**2004,**B**2005,**C**2006,**D**2007

Table 5.1: Parameter estimates for the Burr XII model

Parameter	Estimate	Standard Error
$\ln(b_1)$	0.772	0.0777
$\ln(a)$	1.142	0.0627
$\ln(c_1)$	-1.574	0.0746
$\ln(c_2)$	-0.045	0.0627
$\ln(c_3)$	0.523	0.073

5.3.2 Association between Outbreaks in Poultry and Infection in Wild Birds

The Burr XII and Burr III distributions each have 5 parameters. These distributions were used to model the observed poultry outbreak data for each of the 9 regions, taking into account wild-bird infection. The AIC, when we used the Burr XII model to fit the observed data, was 5,628.6, substantially lower than that for the Burr III distribution, which gave an AIC of 5,829.8. We therefore chose the Burr XII distribution to model the data. Figure 5.3 gives the fit to the data for all 9 regions. The model fits the data rather well. The Burr XII has standard 3 parameters together with the proportional rate factor for the detected wild-bird infected month and the effect of area's 2,3 and 4 versus others gives 5 parameters. To see whether the proportionality factor for the wild-bird infected months versus not wild-bird infected months, was needed in the model a Burr XII model without this factor was fitted, giving an AIC of 5677.7 So the data clearly show that the reproductive power of poultry flocks in wild-bird infected months was higher than in non wild-bird infected months. Parameter estimates for the Burr XII model are shown in the Table 5.1. The log of the proportionality (ln[c3]) is 0.523, corresponding to a proportionality factor of 1.67, indicating that the reproductive power in wild-bird infected months is 1.7 times higher than that in non wild-bird infected months (Figure 5.4, where we give the reproductive power for the associated period). In Figure 5.4, we have also plotted the reproductive power for the 6 regions for which we could not do the wild-bird related comparison (regions 1,5–9). The reproductive power as a function of time was almost indistinguishable from the curve for the non wild- bird infected months in regions 2, 3, and 4.

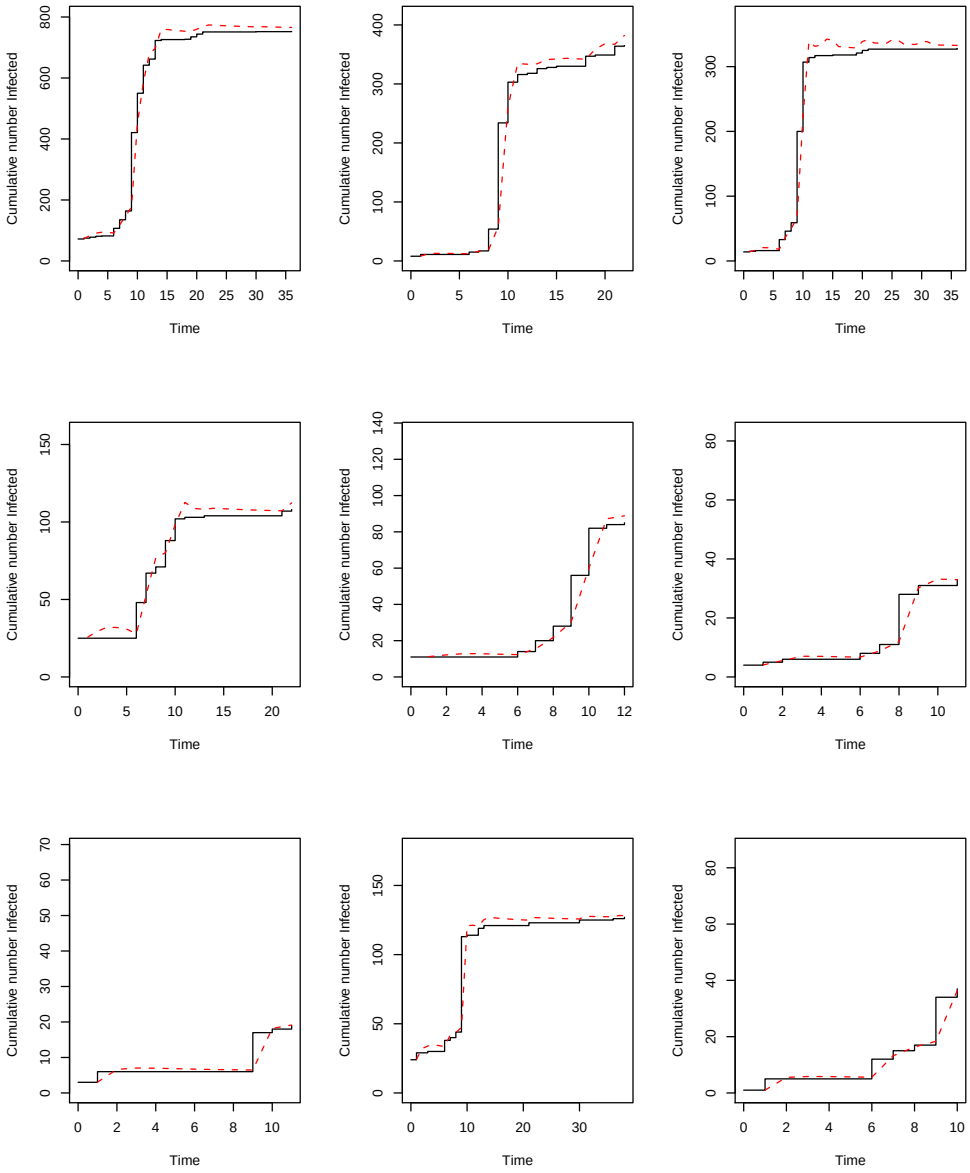

Figure 5.3: The fit of the Burr XII to the data for the 9 regions

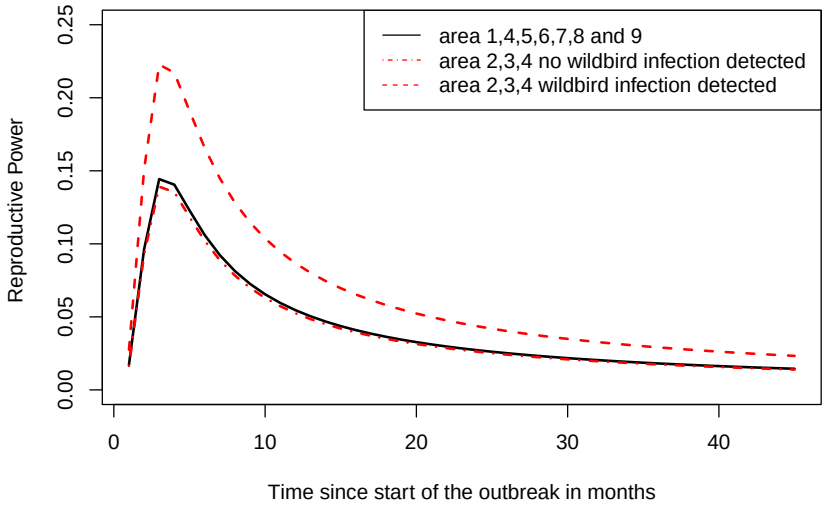

Figure 5.4: The reproductive powers for the Burr XII model

5.4 Discussion

We analyzed one of the largest datasets available of wild birds sampled for HPAI (H5N1) infection in Thailand, a country where several outbreaks of the disease have occurred in poultry flocks. Our aim was to determine the prevalence and distribution of HPAI (H5N1) in wild birds and to determine whether an association exists between outbreaks in poultry flocks and in wild birds within different regions in Thailand. We calculated the reproductive power in poultry flocks, a measure for the ability of a poultry flock to infect other susceptible poultry flocks. Notably, reproductive power was 1.7 times higher in so-called wild-bird infected months, compared with poultry outbreaks in non wild-bird infected months, suggesting a strong association of spread among poultry flocks and the presence of the infection in wild birds. Poultry flocks in this study represent several avian species, which were considered as a single group with equal infectiousness, susceptibility, and other characteristics, in the absence of more precise information. Domestic ducks, which normally manifest a subtype H5N1 infection subclinically, were included in the poultry group. Ducks were not sampled according to criteria related to clinical signs. Available data do not allow a more differentiated analysis. To quantify the association with outbreaks in poultry, we

regarded wild birds as 1 group. We can therefore not differentiate the quantification of interaction to the level of specific wild-bird groups. In our additional analyses, however, most cases of HPAI (H5N1) infection in wild birds were found in resident birds, as compared with migratory and breeding visitor birds. Therefore, resident wild birds may be responsible for the association that we quantified. Our results can possibly be explained by the difference in exposure time of the wild birds. We partially confirmed, but more importantly expanded and added detail to, the conclusions reached by (Siengsanan et al., 2009), on the basis of pooled samples for a smaller part of the database. Bird species seemed to differ in susceptibility for infection. In our study, H5N1 virus infection was detected in many resident bird species, but we did not have a sufficient number of birds to differentiate in the quantitative analysis between different species. Species do differ, however, in terms of potential contact to poultry, especially birds considered to be peridomestic species of the Columbiformes, Cuculiformes, and Passeriformes orders, which are commonly associated with poultry environments. Transmission of subtype H5N1 to poultry populations by this group of resident bird species is more likely than transmission by other resident birds, including those belonging to the Galliformes, Gruiformes, Piciformes, Psittaciformes, and Struthioniformes orders. The habitats of these birds are not located near poultry areas. Previous experimental studies have shown that infected individuals of peridomestic species such as sparrow and starling can shed subtype H5N1 after infection, but they die quickly (Boon et al., 2007; Brown et al., 2009). Therefore, these birds are unlikely to be long-term reservoirs but may be a higher risk to poultry than other resident bird species. Pigeons were found to be less susceptible to severe neurologic signs and death from HPAI (H5N1) infection (Brown et al., 2009). Infected pigeons appeared to shed low amounts of virus, thereby limiting virus transmission to sentinel birds (Boon et al., 2007; Brown et al., 2009; Jia et al., 2008; Klopfleisch et al., 2006; Liu et al., 2007; Werner et al., 2007; Bavinck et al., 2009). Our data showed a relatively high prevalence of HPAI (H5N1) in herons and storks (commonly known as scavengers and hunters of juvenile aquatic birds), which suggests that these birds are predominantly infected by contact with infected poultry flocks. The prevalence of HPAI (H5N1) infections in resident birds was higher in areas with poultry flocks. We could not determine whether wild birds became infected because of spillover from poultry flocks or whether wild birds were the origin of outbreaks in poultry flocks. The association we found is not necessarily one of cause and effect. The 2 populations may have been affected by the same factors that increase transmission between flocks, e.g., contaminated water, movement between poultry flocks, or even increased transmission through fomites. Even though data results are from the largest sampling effort available, the lack of a clear sampling strategy in the collection of wild-bird

data precludes a definite answer to whether poultry flocks were infected with HPAI (H5N1) from infected wild birds or vice versa. (Siengsanan et al., 2009) suggested that poultry outbreaks precede detection of the infection in wild birds, but we have found no evidence either for or against that claim, again because of the sampling strategy used. One could argue the fact that infected poultry flocks produce massive amounts of virus, which supports the view that infection in wild birds is mostly seeded from poultry. A study carried out by (Bavinck et al., 2009) suggested that small backyard flocks did not contribute to the spread of subtype H7N7 infection in the Netherlands during 2003. Seasonal bird migration, as well as enhanced movement and trade of poultry in the winter period caused by major social events occurring at the end of the year, may play a role in virus spread (Leslie and Ian, 2008). Our data show increased prevalence among wild birds in all winter periods, with the exception of 2007 in which neither poultry farm outbreaks nor wild bird infections were detected. The actual sources of new introductions of HPAI (H5N1) into the commercial poultry flocks in Thailand could not be elucidated by our analysis. From January through October 2004, a relatively small number of wild-bird samples was collected, compared with the number of samples collected from November 2004 to December 2007. Selection bias may have occurred during this period. Despite a bias in sampling numbers, HPAI (H5N1) infected wild birds were detected during April May 2004 just before the onset of the 2004 outbreak, but were not observed in that same period during 2005–2007 despite larger sampling numbers. Variation in geographic distribution of HPAI (H5N1) infections in wild birds was observed over different areas. The central region of Thailand with dense poultry populations and large populations of birds living in the surrounding wetlands can be considered a hotspot for HPAI (H5N1) outbreaks. Our dataset shows high prevalence rates of the virus in the central region, corresponding with previous studies of HPAI (H5N1) surveillance in wild birds (Siengsanan et al., 2009), in poultry flocks during 2005–2005 (Tiensin et al., 2009, 2005), and in cases of HPAI (H5N1) infection among humans during 2004 (Areechokchai, 2006). Associating these observations to our statistical model is interesting, because the reproductive power of poultry flocks in regions 1, 5, 6, 7, 8, and 9 was almost identical to that in regions 2, 3, and 4 during non wild-bird infected months (Figure 5.4); regions 1, 5, 6, 7, 8, and 9 experienced no outbreaks in wild birds. It is however impossible to conclude from the current data that absolutely no wild birds were infected because, in these regions, relatively few samples were collected during the appropriate periods (Figure 5.1). By determining the reproductive power in poultry, which is the ability of infected poultry flocks to spread infection to susceptible poultry flocks, we quantified the association between wild bird infection and outbreaks in poultry. We also attempted to take the reproductive power in wild birds, during poultry-infected

months, as our starting point. However, too few infected wild birds were available for a reliable analysis.

Acknowledgement. We thank the Department of National Park Wildlife and Plant Conservation, Thailand; and the following members of Wild Bird Study Group, Thailand, who contributed data to our study: Thaweesak Songserm, Kridsada Chaichoun, Parntep Rattanakorn, Surapong Wongkasemjit, and Tuangthong Patchimasiri. Dr Keawcharoen is a veterinarian in the Virology Unit, Department of Pathology, Faculty of Veterinary Science, Chulalongkorn University, Bangkok, Thailand. Her research interests include the epidemiology of avian influenza virus in wild birds.

Chapter 6

Estimating survival from communicable-events data [1]

6.1 Introduction

In this chapter methods of estimating the reproductive power and the accompanied survival function with communicable events, e.g. an infectious disease, are discussed. Using a non-homogeneous birth process, the survival distribution appears in the distribution for the number of events, in a natural way. This survival function can be estimated by assuming a particular form which depends on some unknown parameters the value of which is then estimated. One can also estimate the reproductive power and the survival function and their standard errors directly from the data using the (log-)likelihood. It is shown that the standard errors for the estimated reproductive power and for the estimated survival become smaller as time goes on, because with communicable events the amount of information tends to increase with increasing time.
Methods are developed to compare empirical estimates in two independent groups by means of the (log) reproduction power rate. It is shown that the standard error of log of the reproduction power rate is usually increasing since this standard error depends on the size of the reproductive power. Are these small then the standard error is large.
These methods are applied to the Dutch avian influenza (H7N7) outbreak from 2003 and on data from the avian influenza (H5N1) among poultry in Thailand.

[1] in preparation

6.2 Estimating Survival.

In survival analysis the population at risk plays a key role and should be well defined. This is not always straight forward. If, for instance, one is interested in the survival of patients who underwent a particular operation, one very often deals with patients taken from different hospitals. This artificial population can give biased results in many cases (catchment populations). Besides the definition of the population at time zero one also has to take censoring into account in order to define the population at risk at a certain time point. One has to make assumptions about the censoring mechanism in order to avoid bias. Only in the case that the population at risk is well defined can one calculate a characteristic, such as the hazard rate for a disease which is the instantaneous risk of contracting the disease at a certain time point, given absence of disease before that time point. Since one starts with a well defined population (susceptibles) and models how the size of this population reduces in time, this model can also be used for other events than infectious disease events.

This is different with communicable events. Communicable events are events that occur in the spread of an infectious disease, the spread of a rumor and the spread of some kind of behavior through a population and so on. With communicable events like the spread of an infectious disease such a definition of the population at risk is not necessary. Of course, in order to be able to spread there must be enough people without the disease in the population and there must be some kind of contact pattern, but a complete specification of the population at risk at each time point is not necessary. This is due to the communicability of the disease. With such a disease the risk of obtaining it at a certain time point is characterized by the ability of current infectious individuals to infect, the so called reproductive power. In large enough populations this characterization depends mainly on how many infectious individuals there are. If this reproductive power can be taken to be non-homogeneous in time, there is also no need for a homogeneous mixing assumption, since in that case there might be periods in which the diseased infected individuals mix well with other individuals and there might be periods in which this is less the case.

A model that can be used to describe a reproducing disease is the birth model whose parameter, the reproductive power (birth rate), can depend on time. For the hazard function estimators are available in the literature. This is not the case for the reproductive power. The aim of this chapter is to estimate the reproductive power in two ways. The first way is by using a parametric form for the survival function and thus also for the reproductive power, and estimate the parameters of this survival function. The second way is to estimate the reproductive power directly from the data and use this estimated reproductive power to estimate the survival function.

The stochastic version of the non-homogeneous birth process with the stochastic variable $Y(t)$ representing the total number of infected at time t and y_0 the number of infected at time 0, has the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}p_{y(t)} = (y-1)\lambda(t)p_{y-1}(t) - y\lambda(t)p_y(t)$$

with $p_y(t)$ the probability that the number of detected infected at time t is y. The solution has probability mass function (Kendall, 1948):

$$P(Y(t)=y) = \binom{y-1}{y_0-1}\left[S_\lambda(t)\right]^{y_0}\left[1-S_\lambda(t)\right]^{y-y_0} \quad , y = y_0, y_0+1, \cdots$$

with $S_\lambda(t) = \exp\left[-\int_0^t \lambda(\tau)\mathrm{d}\tau\right]$ the survival function and with y_t the number of infected. This is a shifted negative binomial distribution. It is the probability of obtaining $y - y_0$ infected at time t in an epidemic that started with y_0 infected at time zero.

The expected value of $Y(t)$ is the same as the solution of the deterministic equation for the number of infectives. The expected value for the number of infected at time t is

$$E(Y(t)) = \frac{y_0}{S_\lambda(t)} \tag{6.1}$$

which is the underlying profile of the process, and the variance is:

$$var(Y(t)) = y_0\left[\frac{1-S_\lambda(t)}{S_\lambda^2(t)}\right] \tag{6.2}$$

In general one observes just one outbreak, which can be interpreted as a sample path if the model outlined above is imposed as the generator of the process. This means that the model is fitted conditionally on the past. Furthermore, in practice the outbreak is observed in discrete time (Becker, 1989, page 108) with, say, $t_j, j = 0, ..., n$ as the observation times. So at time point t_j one models the number of infected conditionally on what was observed at time point t_{j-1}. A consequence of this is that the survival distribution is evaluated conditionally:

$$\begin{aligned} P(T > t_j | T > t_{j-1}) &= \frac{S(t_j)}{S(t_{j-1})} \\ &= 1 - h(t_{j-1}) \end{aligned} \tag{6.3}$$

where $h(t)$ is the discrete-time hazard rate (**??**). The probability of having count outcome y_{t_j} at time t_j, given that there were $y_{t_{j-1}}$ counted at time t_{j-1}, is then given by

$$\begin{aligned} P(Y(t_j) &= y_{t_j}|Y(t_{j-1}) = y_{t_{j-1}}) = \binom{y_{t_j}-1}{y_{t_{j-1}}-1} [1-h(t_{j-1})]^{y_{t_{j-1}}} [h(t_{j-1})]^{y_{t_j}-y_{t_{j-1}}} \\ & y_{t_j} = y_{t_{j-1}}, y_{t_{j-1}}+1, y_{t_{j-1}}+2, \cdots . \end{aligned} \tag{6.4}$$

which is the shifted negative binomial distribution. Note that $Y(t_j)$ is the stochastic variable representing the population count at time t and y_{t_j} is its observed outcome. The expected value for time point t_j is :

$$E(y_{t_j}) = \frac{y_{t_{j-1}}}{1-h(t_{j-1})},$$

the sample path profile and the variance is:

$$var(y_{t_j}) = y_{t_{j-1}} \frac{h(t_{j-1})}{[1-h(t_{j-1})]^2}$$

The log-likelihood for this conditional shifted negative binomial then is:

$$\begin{aligned} l(\cdot) &= \sum_{j=1}^{n} \log \binom{y_{t_j}-1}{y_{t_{j-1}}-1} + y_{t_{j-1}} \log [1-h(t_{j-1})] + (y_{t_j}-y_{t_{j-1}}) \log [h(t_{j-1})] \\ & y_{t_j} = y_{t_{j-1}}, y_{t_{j-1}}+1, y_{t_{j-1}}+2, \cdots . \end{aligned} \tag{6.5}$$

In the next section, the above mentioned two ways of estimating the reproductive power and thus the survival function, are discussed. First, two parametric families for the reproductive power function – and thus also for the survival function –are described (section 6.3.1) and then a discussion follows on estimating the reproductive power function directly from the data by maximum likelihood without using a formulation of the survival function (section 6.3.2). With this estimate and its standard error one can check whether the parametric survival function used are supported by the data. This is illustrated with the avian influenza (H7N7) outbreak in the Netherlands in 2003 in section 6.4.

This method of estimating the survival directly from the data – empirical estimate– can also be used to compare to independent groups with respect to their reproductive power. This is explained in section 6.3.3 and illustrated with data on the transmission of avian influenza (H5N1) among poultry in Thailand in section 3.2.

6.3 Estimating the survival function.

6.3.1 Using parametric distribution functions

Burr family

The non-homogeneous birth model depends on the reproductive power $\lambda(t)$: the rate at which infected individuals are capable of reproducing themselves. This rate can also be used in a survival distribution which gives the probability of escaping infection produced by the communicable disease by reproduction. Several parametric forms might be chosen for this survival distribution, among them distributions from the Burr family (see chapter 4 for details). The most well known and useful distribution from the Burr family is the Burr XII, or Singh-Maddala distribution, which is sometimes referred to simply as the Burr distribution in literature. The survival function is given by:

$$S(t) = \left[1 + \left(\frac{t}{b}\right)^{a}\right]^{-q}, \quad t > 0,\ a, b, q > 0$$

The right tail is governed by the parameters a and q, the left tail by a, and b is the scale parameter (Kleibner and Kotz, 2003, page 198). To reduce the number of parameters to be estimated, one can consider three special cases of the Burr distribution (Kleibner and Kotz, 2003) :

1. The logistic form is obtained for $q = 1$ giving the log-logistic or the Fisk distribution.

2. For $a = 1$, the Burr distribution is reduced to the Lomax (Pareto type II) distribution.

3. The case $a = q$ is also known as the para-logistic distribution.

The Weibull distribution and the Pareto distribution are limiting cases of the Burr distribution (Shao, 2004). An interesting way to arrive at the Burr distribution is to assume that the times follow a Weibull distribution, the scale parameter of which follows an inverse generalized gamma distribution (Kleibner and Kotz, 2003).
The Burr III is also known as the Dagum distribution or as the inverse Burr distribution. This last name is not surprising since if X has a Burr distribution then $1/X$ has the inverse Burr distribution. The distribution function of the inverse Burr is:

$$S(t) = 1 - \left[1 + \left(\frac{t}{b}\right)^{-a}\right]^{-p}, \quad t > 0,\ a, b, p > 0$$

The Generalized Gamma

The generalized gamma is also a rich family of distributions that include the exponential as a special case but also the Weibull and the Gamma distribution. As a consequence by checking the estimated parameters of the generalized gamma one can check if the data supports constant rates.

The survival function is then given by:

$$S(t) = \frac{1}{\Gamma(p)} \int_0^{(\frac{t}{b})^a} t^{p-1} e^{-t} \mathrm{d}t$$

where $p, b > 0$. If $a < 0$, the distribution is referred to as inverse generalized gamma distribution. Special cases of the generalized gamma distribution are:

- Gamma distributions for $a = 1$ and the inverse gamma for $a = -1$.
- Weibull distribution for $p = 1, a > 0$ and the inverse Weibull (log-Gompertz) for $p = 1, a < 0$.
- Exponential distribution for $a = p = 1$ and the inverse exponential distribution for $a = -1, p = 1$.
- Log-normal, Pareto and power function distributions for appropriate limits.

Further details of the generalized gamma distribution are given elsewhere (Kleibner and Kotz, 2003).

6.3.2 Estimation of the reproductive survival function directly from the data

The negative binomial log-likelihood (6.5) is proportional to the binomial log-likelihood except that now the total is the dependent variable. The derivative of (6.5) w.r.t $h(t_{j-1})$, $j = 1, \cdots, n$ is a vector with elements:

$$\frac{\mathrm{d}l(\cdot))}{\mathrm{d}h(t_{j-1})} = \left[\frac{-y_{tj-1}}{1 - h(t_{j-1})} + \frac{y_{t_j} - y_{t_{j-1}}}{h(t_{j-1})} \right] , \; j = 1, \cdots, n$$

where y_{t_j} is the observed value of the stochastic variable $Y(t_j)$. If one equates to zero and solves for $h(t_{j-1})$, one finds the maximum likelihood estimator for the reproductive power:

$$\hat{h}(t_{j-1}) = 1 - \frac{y_{t_{j-1}}}{y_{t_j}}\,,\; j = 1, \cdots, n$$

Minus the second derivative, the negative of the Hessian, is an $n \times n$ diagonal matrix with elements $\left[\frac{y_{tj-1}}{\{1-h(t_{j-1})\}^2} + \frac{y_{t_j}-y_{t_{j-1}}}{h(t_{j-1})^2}\right]$, $j = 1, \cdots, n$, which evaluated at the maximum likelihood estimate $\hat{h}(t_{j-1})\,,\; j = 1, \cdots, n$, become $\frac{y_{t_j}^2}{y_{t_{j-1}}} + \frac{y_{t_j}^2}{y_{t_j}-y_{t_{j-1}}} = \frac{y_{t_j}^3}{y_{t_{j-1}}(y_{t_j}-y_{t_{j-1}})}\,,\; j = 1, \cdots, n$. The inverse is the estimated variance-covariance matrix, which is also diagonal, and has elements:

$$\begin{aligned} \hat{var}(\hat{h}(t_{j-1}) &= \frac{y_{t_{j-1}}(y_{t_j} - y_{t_{j-1}})}{y_{t_j}^3} \\ &= \frac{1}{y_{t_j}}\hat{h}(t_{j-1})[1 - \hat{h}(t_{j-1})]\,,\; j = 1, \cdots, n \end{aligned} \tag{6.6}$$

and so

$$SE(\hat{h}(t_{j-1})) = \sqrt{\frac{\hat{h}(t_{j-1})[1 - \hat{h}(t_{j-1})]}{y_{t_j}}}\,,\; j = 1, \cdots, n$$

Since the conditional negative binomial log-likelihood (6.5) is a log-likelihood of a generalized linear model, the parameter estimates are approximately normally distributed so that the estimated reproductive power $\pm$ the standard error is an approximately 95 % confidence interval, i.e. given the outcome of the stochastic process under consideration up until a point in time, what can be expected about the reproductive power, with a probability of 0.95. This empirical estimate with its confidence interval can be compared with the estimates obtained from assuming a particular form for the survival distribution. One then gets as an estimator for the survival, using (6.3)

and with y_{t_j} being the observed value of the stochastic variable $Y(t_j)$:

$$\begin{aligned}
\hat{S}(t_j) &= \hat{S}(t_{j-1})\left[1-\hat{h}(t_{j-1})\right] \\
&= \hat{S}(t_{j-2})\left[1-\hat{h}(t_{j-2})\right]\left[1-\hat{h}(t_{j-1})\right] \\
&= \cdots \\
&= \prod_{k=1}^{j}\left[1-\hat{h}(t_{k-1})\right] \\
&= \frac{y_{t_0}}{y_{t_j}}
\end{aligned}$$

From (6.6) it can be seen that $\hat{var}\left(\frac{1}{Y(t_j)}\right) = \frac{1}{y_{t_j} y_{t_{j-1}}^2}\hat{h}(t_{j-1})[1-\hat{h}(t_{j-1})]$ and thus

$$\begin{aligned}
\hat{var}\left(\hat{S}(t_j)\right) &= y_{t_0}^2 \hat{var}\left(\frac{1}{Y(t_j)}\right) \\
&= \frac{y_{t_0}^2}{y_{t_{j-1}}^2}\frac{\hat{h}(t_{j-1})[1-\hat{h}(t_{j-1})]}{y_{t_j}}
\end{aligned}$$

and so

$$SE\left(\hat{S}(t_j)\right) = \frac{y_{t_0}}{y_{t_{j-1}}}\sqrt{\frac{\hat{h}(t_{j-1})[1-\hat{h}(t_{j-1})]}{y_{t_j}}}$$

This standard error can also be derived using the delta method. See appendix for details.

6.3.3 Comparing two groups

Sometimes it is of interest when data of two independent outbreaks are available to see if the two outbreaks have the same reproductive power, i.e. if the rate at which the infection reproduces is the same in both groups. Especially in an experimental setting this might be valuable. It might be important to let the group effect depend on time since it can sometimes be interesting to see when groups start to differ and how this difference develops.

In order to model this one can take the log of the hazard at time t_j as a linear function of the groups, that is the hazard in goup i at time t_j is taken as

$$h_i(t_{j-1}) = e^{\alpha_j + \beta_j x_{i,j}}, \quad i = 1, 2$$

where x_{ij} is zero for group $i = 1$ and one for group $i = 2$. The log-likelihood then is:

$$\begin{aligned} l(\alpha_j, \beta_j) &= \sum_{i=1}^{2}\sum_{j=1}^{n} \log \binom{y_{it_j} - 1}{y_{it_{j-1}} - 1} + y_{it_{j-1}} \log\left[1 - h_i(t_{j-1})\right] + (y_{it_j} - y_{it_{j-1}}) \log\left[h_i(t_{j-1})\right] \\ & y_{it_j} = y_{it_{j-1}}, y_{it_{j-1}} + 1, y_{it_{j-1}} + 2, \cdots . \end{aligned} \tag{6.7}$$

The maximum likelihood estimates are given by the first derivative w.r.t. the vector of parameters (α_j, β_j), $j = 1, \cdots, n$. Since $\frac{\mathrm{d}l(.)}{\mathrm{d}h_i(t_{j-1})} = \sum_i \left\{ \frac{-y_{it_{j-1}}}{1-h_i(t_{j-1})} + \frac{y_{it_j} - y_{it_{j-1}}}{h_i(t_{j-1})} \right\}$, $i = 1, 2$, one has:

$$\frac{\mathrm{d}l(.)}{\mathrm{d}\alpha_j} = \sum_i \left\{ \frac{-y_{it_{j-1}}}{1 - h_i(t_{j-1})} + \frac{y_{it_j} - y_{it_{j-1}}}{h_i(t_{j-1})} \right\} h_i(t_{j-1}) \tag{6.8}$$

and

$$\frac{\mathrm{d}l(.)}{\mathrm{d}\beta_j} = \sum_i \left\{ \frac{-y_{it_{j-1}}}{1 - h_i(t_{j-1})} + \frac{y_{it_j} - y_{it_{j-1}}}{h_i(t_{j-1})} \right\} h_i(t_{j-1}) x_{ij} \tag{6.9}$$

Putting these to zero and using (6.9) one obtains: $\hat{h}_2(t_{j-1}) = 1 - \frac{y_{2t_{j-1}}}{y_{2t_j}}$, $j = 1, \cdots, n$ and $\hat{\alpha}_j + \hat{\beta}_j = \log[\hat{h}_2(t_{j-1})] = \log\left[1 - \frac{y_{2t_{j-1}}}{y_{2t_j}}\right]$. Using this together with (6.8) gives: $\hat{h}_1(t_{j-1}) = e^{\hat{\alpha}_j} = 1 - \frac{y_{1t_{j-1}}}{y_{1t_j}}$, $j = 1, \cdots, n$. And so:

$$\hat{\alpha}_j = \log\left(1 - \frac{y_{1t_{j-1}}}{y_{1t_j}}\right)$$

and

$$\begin{aligned} \hat{\beta}_j &= \log\left(1 - \frac{y_{2t_{j-1}}}{y_{2t_j}}\right) - \log\left(1 - \frac{y_{1t_{j-1}}}{y_{1t_j}}\right) \\ &= \log\left(\frac{\hat{h}_2(t_{j-1})}{\hat{h}_1(t_{j-1})}\right) \end{aligned} \tag{6.10}$$

So, $\hat{\alpha}_j$ is the estimated log-reproductive power in group 1 at time t_j and $\hat{\beta}_j$ is the differences in estimated log-reproductive powers of the two groups or it is the estimated log-reproductive power ratio at time t_j, which means that $e^{\hat{\beta}_j}$ is the estimated reproductive power ratio at time t_j

The standard errors are obtained from the observed information matrix, the negative of the Hessian matrix. Because the vector of parameters is $(\alpha_1, \beta_1, \cdots, \alpha_n, \beta_n)$, the Hessian matrix is block-diagonal, with block j containing the second derivatives with respect to α_j, β_j and with respect to α_j and β_j, which shows that the observations are conditionally independent over time. The entry's for block j of the Hessian are:

$$\begin{aligned}
\frac{\mathrm{d}^2 l(\cdot)}{\mathrm{d}\alpha_j^2} &= \sum_i -y_{it_{j-1}} \frac{h_i(t_{j-1})}{[1-h_i(t_{j-1}]^2} \\
\frac{\mathrm{d}^2 l(\cdot)}{\mathrm{d}\beta_j^2} &= \sum_i -y_{it_{j-1}} x_{ij}^2 \frac{h_i(t_{j-1})}{[1-h_i(t_{j-1}]^2} \\
\frac{\mathrm{d}}{\mathrm{d}\alpha_j}\frac{\mathrm{d}l(\cdot)}{\mathrm{d}\beta_j} = \frac{\mathrm{d}}{\mathrm{d}\beta_j}\frac{\mathrm{d}l(\cdot)}{\mathrm{d}\alpha_j} &= \sum_i -y_{it_{j-1}} x_{ij} \frac{h_i(t_{j-1})}{[1-h_i(t_{j-1}]^2}
\end{aligned}$$

The observed information matrix – minus the Hessian evaluated at the maximum likelihood estimates – then is block-diagonal with entries for block j given by:

$$\begin{aligned}
-\frac{\mathrm{d}^2 l(\cdot)}{\mathrm{d}\alpha_j^2} &= \sum_i \frac{y_{it_j}}{y_{it_{j-1}}} \left(y_{it_j} - y_{it_{j-1}}\right) \\
-\frac{\mathrm{d}^2 l(\cdot)}{\mathrm{d}\beta_j^2} &= \sum_i x_{ij}^2 \frac{y_{it_j}}{y_{it_{j-1}}} \left(y_{it_j} - y_{it_{j-1}}\right) \\
&= \frac{y_{2t_j}}{y_{2t_{j-1}}} \left(y_{2t_j} - y_{2t_{j-1}}\right) \\
-\frac{\mathrm{d}}{\mathrm{d}\alpha_j}\frac{\mathrm{d}l(\cdot)}{\mathrm{d}\beta_j} &= \sum_i x_{ij} \frac{y_{it_j}}{y_{it_{j-1}}} \left(y_{it_j} - y_{it_{j-1}}\right) \\
&= \frac{y_{2t_j}}{y_{2t_{j-1}}} \left(y_{2t_j} - y_{2t_{j-1}}\right)
\end{aligned}$$

The inverse of the information matrix, the variance-covariance matrix of the parameters, then has the following block stucture at time t_j:

$$\begin{pmatrix}
\frac{1}{\frac{y_{1t_j}}{y_{1t_{j-1}}}(y_{1t_j}-y_{1t_{j-1}})} & \frac{-1}{\frac{y_{1t_j}}{y_{1t_{j-1}}}(y_{1t_j}-y_{1t_{j-1}})} \\
\frac{-1}{\frac{y_{1t_j}}{y_{1t_{j-1}}}(y_{1t_j}-y_{1t_{j-1}})} & \frac{1}{\frac{y_{1t_j}}{y_{1t_{j-1}}}(y_{1t_j}-y_{1t_{j-1}})} + \frac{1}{\frac{y_{12t_j}}{y_{2t_{j-1}}}(y_{2t_j}-y_{2t_{j-1}})}
\end{pmatrix}$$

This is equal to:

$$\begin{pmatrix} \frac{1}{y_{1t_j}\frac{\hat{h}_1(t_{j-1})}{1-\hat{h}_1(t_{j-1})}} & \frac{-1}{y_{1t_j}\frac{\hat{h}_1(t_{j-1})}{1-\hat{h}_1(t_{j-1})}} \\ \frac{-1}{y_{1t_j}\frac{\hat{h}_1(t_{j-1})}{1-\hat{h}_1(t_{j-1})}} & \frac{1}{y_{1t_j}\frac{\hat{h}_1(t_{j-1})}{1-\hat{h}_1(t_{j-1})}} + \frac{1}{y_{2t_j}\frac{\hat{h}_2(t_{j-1})}{1-\hat{h}_2(t_{j-1})}} \end{pmatrix}$$

So, the reproductive power ratio e^{β_j} is estimated as the ratio of the fraction of new cases during $t_{j-1}$and t_j, in both groups $e^{\hat{\beta}_j} = \frac{\frac{y_{2t_j}-y_{2t_{j-1}}}{y_{2t_j}}}{\frac{y_{1t_j}-y_{1t_{j-1}}}{y_{1t_j}}}$. The standard error of the log of this ratio is:

$$se(\hat{\beta}_j) = \sqrt{\frac{1}{y_{1t_j}\frac{\hat{h}_1(t_{j-1})}{1-\hat{h}_1(t_{j-1})}} + \frac{1}{y_{2t_j}\frac{\hat{h}_2(t_{j-1})}{1-\hat{h}_2(t_{j-1})}}} \tag{6.11}$$

Since the specific formulation of the negative binomial model used here, is a generalized linear model, the parameter estimators are approximately normally distributed, so $\hat{\beta}_j \pm 1.96 \times se(\hat{\beta}_j)$ is an approximate 95% confidence interval for the parameter β_j, again with the interpretation that it gives what can be expected about the reproductive power ratio – given the outcome of the stochastic process under consideration up until a point in time – with a probability of 0.95. The log reproductive power in the first group is $\hat{\alpha}_j = \log[\hat{h}_1(t_{j-1})]$ and $se\left(\log[\hat{h}_1(t_{j-1})]\right) = \sqrt{\frac{1}{y_{1t_j}\frac{\hat{h}_1(t_{j-1})}{1-\hat{h}_1(t_{j-1})}}}$. Further, $\log[\hat{h}_2(t_{j-1})] = \hat{\alpha_j} + \hat{\beta}_j$ so $se\left(\log[\hat{h}_2(t_{j-1})]\right) = \sqrt{\frac{1}{y_{2t_j}\frac{\hat{h}_2(t_{j-1})}{1-\hat{h}_2(t_{j-1})}}}$. These, together with the approximate normality of the parameters, can be used to calculate approximate confidence intervals for the reproductive powers.

6.4 Some applications

6.4.1 The avian influenza (H7N7) outbreak in the Netherlands in 2003

On February 28, 2003 an epidemic of avian influenza (H7N7) started in the Gelderse Vallei in the Netherlands, spreading to adjacent areas and to the province of Limburg. In total 239 flocks were infected with known detection date. The epidemic was controlled by movement restrictions, stamping out of infected flocks, and pre-emptive

Table 6.1: Parameter estimates for the Lomax distribution

Parameter	Estimate	St.Error
$\ln(b)$	-1.114	0.6221
$\ln(c)$	-0.394	0.0689

culling of flocks in the neighborhood of infected flocks. In total 1,255 commercial flocks and 17,421 flocks of smallholders had to be depopulated. Approximately 25.6 million animals were killed, see Stegeman et al.(2003) for more details.

The data of the detected new asian influenza cases per day are in given figure (6.1). As one can see the right tail contains gaps and the center of the distribution is not well determined. This usually makes model fitting difficult. The AIC for the model with the Burr XII distribution was 339.08, using the Burr III distribution it was 342.92 and for the model with the generalized gamma distribution 386.42, indicating the Burr XII fitted the data best. The parameters a and q govern the tails of the distribution. Due the the many gaps in the right tail these parameters have large standard deviations, indicating that there is a lack of information. The Lomax distribution ($a = 1$) has an AIC of 340.32 and the log-logistic ($q = 1$) has an AIC of 344.30. This shows that the model with the Lomax distribution fits the data almost as good as the Burr XII and it has not the large standard errors as can be seen in table (6.1). Figure (6.2) shows the reproductive powers of the used models. The reproductive power for the Burr III distribution and the Lomax distribution are almost indistinguishable (solid line). The dashed line shows the reproductive power for the generalized gamma distribution. The step line shows the empirical reproductive power with its confidence region. As the figure shows, in the early phase of the outbreak, the lomax seems to have better agreement with the confidence region of the empirical estimate. Furthermore the figure shows that the confidence limits are getting smaller as time progresses since information accumulates with these models.

6.4.2 Transmission of avian influenza (H5N1) among poultry in Thailand

On January 23, 2004, the Ministry of Public Health in Thailand informed the World Health Organization of an avian influenza A (H5N1) outbreak. To determine the epidemiology of this viral infection and its relation to poultry outbreaks in Thailand

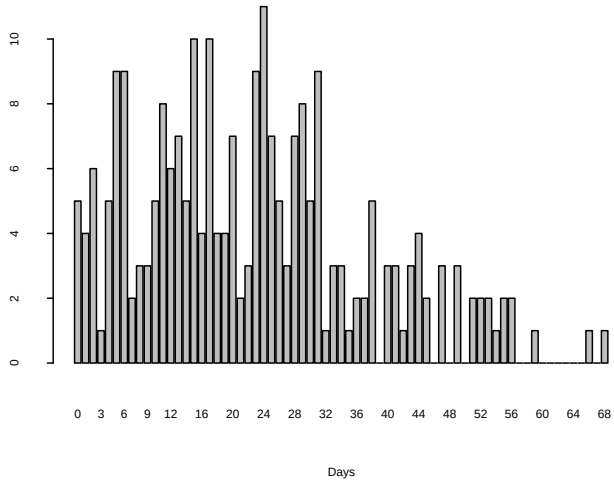

Figure 6.1: Detected new cases in the avian influenza (H7N7) outbreak in the Netherlands 2003 by day.

from 2004 through 2007, it was investigated how wild birds play a role in transmission (Keawcharoen et al., 2011). A total of 24,712 swab samples were collected from migratory and resident wild birds. Reverse transcription PCR showed a 0.7% HPAI (H5N1) prevalence. The highest prevalence was observed during January–February 2004 and March–June 2004, predominantly in central Thailand, which harbors most of the country's poultry flocks. Interest was in the relationship between poultry and wild bird outbreaks. Does for instance the joint presence of infected wild birds and poultry increase spread among poultry flock.

For each bird species identified, geographic location and season were recorded. To study the effect of regions on the subtype H5N1 outbreaks in wild birds, Thailand was divided into 4 major geographic regions (northern, northeastern, central, and southern) on the basis of the former administrative region grouping system used by the Ministry of Interior, Thailand. Because of the high number of outbreaks in the Central region, this was further divided into six parts: central-northwest, central-north, central-central, central-east, central-southeast, and central-southwest.

Time was measured in months from the first month that infection was detected. In most regions, sampling among wild birds was only done systematically after a poultry outbreak in that region, except in the central-northwest, central-north, and central-central regions.

As an illustration of the methods introduced in the present chapter, the outbreak

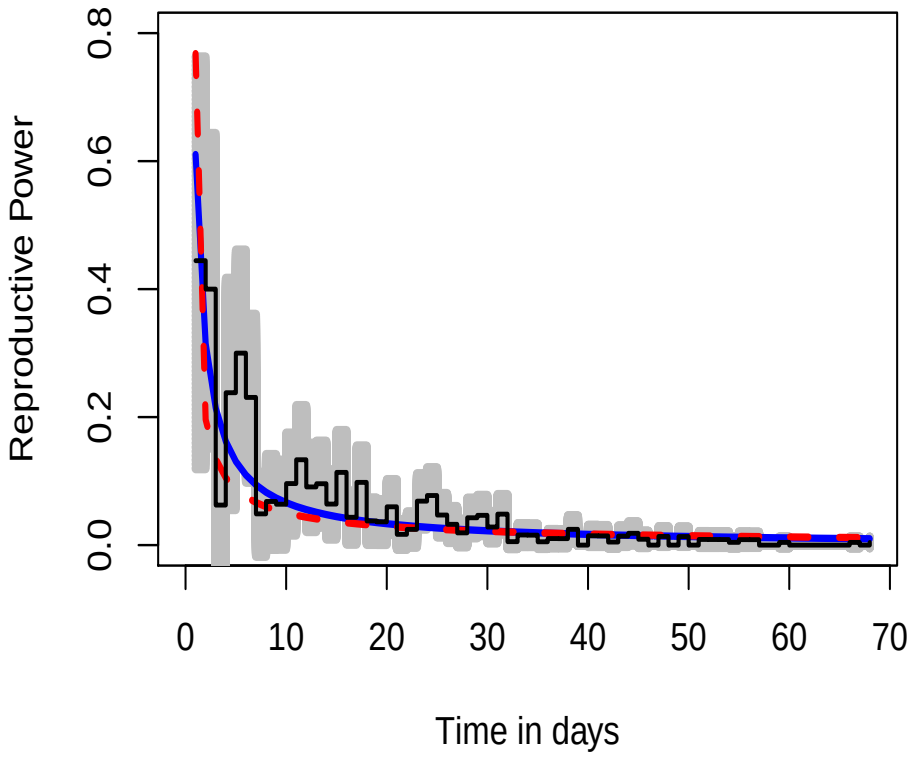

Figure 6.2: The reproductive power for the Burr III and the Lomax distribution (solid line, they are practically the same) and for the generalized gamma distribution(dashed line). The step function is the empirical reproductive power and the gray area indicates the 95 % confidence interval.

Table 6.2: Estimates of the log reproductive power ratio β, their standard errors and the approximate 95 % confidence limits by time (time 0 is January 2004).

Time	Estimate (β)	Stander Error	lower limit	upper limit
6	0.877	0.5389	-0.179	1.933
7	-0.060	0.4145	-0.872	0.753
8	-0.260	0.3863	-1.017	0.497
9	0.344	0.1412	0.067	0.620
10	0.095	0.1799	-0.258	0.447
11	-0.066	0.7923	-1.6188	1.487
12	-0.218	1.1482	-2.4682	2.033

among poultry in the region central-north is compared to that in the region central-east. Central-north is taken as a region where infection was detected in wild birds, whereas central-east is a region with no detected infected wild birds. The cumulative number of cases among poultry for both regions is in figure (6.3). In table (6.2), estimates of the log of the reproductive power ratio (6.10), their standard errors (6.11) and the 95 % confidence limits are given. From this table it can be seen that only in month 9 after the start of outbreak (October 2004) there seemed to be a difference between the regions. In region north-central (wild bird infected region), the number of detected infected poultry is $e^{0.344} = 1.4$ times higher as compared to the region east-central (no infected wild bird detected). Figure (6.4) shows the log reproductive power ratio's ($\hat{beta}$'s) (solid line) and their 95% confidence area (gray area). Due to very sparse data on other time points, the the log reproductive power ratio's could not be calculated since there the reproductive power in one of the groups was zero. Note that the confidence interval is becoming wider as time progresses because the reproductive power is decreasing, leading to an increase in the standard error as can be seen from (6.11).

6.5 Discussion

In this chapter, methods for estimating and comparing survival functions for communicable events are discussed. Using a non-homogeneous birth process the survival distribution appears in the distribution for the number of communicable events in a natural way. This survival function can be estimated by assuming a particular form

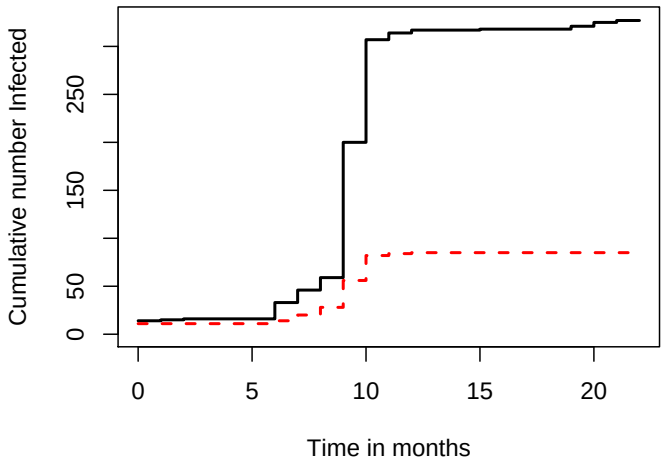

Figure 6.3: Cumulative number of prevalence cases for the wild bird infected region (solid step line) and for the non-wildbird infected region (dashed step line).

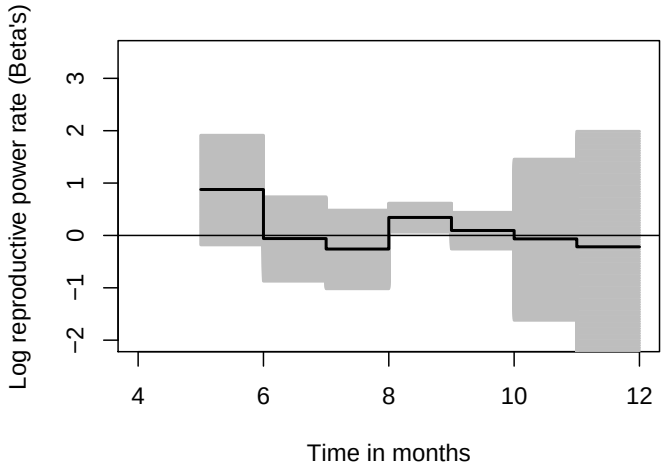

Figure 6.4: The log reproductive power rates (solid line) and their 95% confindence area (grey area).

for this distribution and then estimating the value of the unknown parameters .

One can also first estimate the reproductive power, and then the survival function and their standard errors directly from the data using the likelihood. From the expressions derived in section 6.3.2, it can be seen that the standard errors for the estimated reproductive power and for the estimated survival become smaller as time goes on. That is, information on the reproductive power and on survival is accumulating as time evolves. This is in contrast to 'ordinary' survival. There, the population at risk is decreasing and the estimate of the survival has an increasing standard error. This empirical estimate with its confidence interval can be compared to the estimates obtained from assuming a particular form for the survival distribution.

In the case of comparing two groups, things are slightly different. Although the standard error for the reproductive power in both group is decreasing due to accumulating information, the standard error of log of the reproduction power ratio is increasing since this standard error depends on the size of the reproductive powers. Are these small then the standard error is large as can be seen from (6.11).

Appendix: Using the delta method to obtain the standard error for the survival function

To obtain the standard error for $\hat{S}(t)$ one can use the delta method and (6.1) and (6.2):

Refer to $S'\left(E(Y(t_j))\right)$ as the derivative of $S(t)$ seen as a function of $Y(t_j)$ evaluated at its expected value. Then:

$$\begin{aligned} var\left[\hat{S}(t_j)\right] &= \left\{\hat{S}'\left(E(Y(t_j))\right)\right\}^2 var(Y(t_j)) \\ &= \frac{y_{t_0}^2}{(E(Y(t_j))^4} y_{t_{j-1}} \frac{h(t_{j-1})}{1-h(t_{j-1})} \\ &= \frac{y_{t_0}^2}{y_{t_{j-1}}^4}[1-h(t_{j-1})]^4 y_{t_{j-1}} \frac{h(t_{j-1})}{[1-h(t_{j-1})]^2} \\ &= \frac{y_{t_0}^2}{y_{t_{j-1}}^3}[1-h(t_{j-1})]^2 h(t_{j-1}), \quad j=1,\cdots,n \end{aligned}$$

and so

$$\begin{aligned} \hat{var}\left(\hat{S}(t_j)\right) &= \frac{y_{t_0}^2}{y_{t_{j-1}}^3}\frac{y_{t_{j-1}}}{Y(t_j)}[1-\hat{h}(t_{j-1})]\hat{h}(t_{j-1}) \\ &= \frac{y_{t_0}^2}{y_{t_{j-1}}^2}\frac{\hat{h}(t_{j-1})[1-\hat{h}(t_{j-1})]}{Y(t_j)}, \quad j=1,\cdots,n \end{aligned}$$

and

$$SE\left(\hat{S}(t_j)\right) = \frac{y_{t_0}}{y_{t_{j-1}}}\sqrt{\frac{\hat{h}(t_{j-1})[1-\hat{h}(t_{j-1})]}{Y(t_j)}}, \quad j=1,\cdots,n$$

Bibliography

Bibliography

Ahmed, S. A., Diffenbaugh, N.S. and Hertel, T.W. (2009) Climate volatility deepens poverty vulnerability in developing countries. *Environmental Research Letters* **4** 1–8.

Alexander D.J. (2007) An overview of the epidemiology of avian influenza. *Vaccine* **25** 5637–5644

Alexander D.J. (2000) A review of avian influenza in different bird species. *Microbiol* **74** 3–13

Allen, L.J.S. (2003). *Stochastic processes with applications to biology.* Pearson Education,Inc.

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis.* Springer Verlag, Berlin.

Anderson, D.R. , Burnham, K.P. and Thompson, W.L. (2000). Null hypothesis testing: problems, prevalence and an alternative. *Journal of wildlife management* **64**,(4), 912-923

Centers for Disease Control and Prevention. Areechokchai D., Jira-phongsa C., Laosiritaworn Y., Hanshaoworakul W. and OReilly M. (2006) Investigation of avian influenza (H5N1) outbreak in humans –Thailand, 2004. *MMWR* **44(SUP01)**3–6

Bailey, N.T.J. (1990). *The elements of stochastic processes with applications to the natural sciences.* Reprint edition, Wiley-interscience.

Bavinck V., Bouma A., van Boven M., Bos M.E., Stassen E., Stegeman J.A. (2009) The role of backyard poultry flocks in the epidemic of highly pathogenic avian influenza virus (H7N7) in the Netherlands in 2003. *Prev. Vet. Med.***88** 247–254

Becker,N.G. (1989). *Analysis of Infectious disease data.* London,New York: Chapman and Hall.

Becker, N.G. and Yip, P. (1989). Analysis of variations in an infection rate. *Australian Journal of Statistics* **31**,42–52.

Beyersman, J., Wolkewitz, M., Alligol, A.,Grambauer, N. and Schumacher, M. (2011) Application of multistate models in hospital epidemiology: advantages and challenges. *Biometrical Journal* **53** 332–350

Billard, L. and Dayananda P.W.A. (1995). Epidemic Plant Diseases: a Stochastic Model of leaf and stem lesion growth. In: *Epidemic Models: their Structure and Relation to data*, Mollison, D. (editor).

Boon A.C., Sandbulte M.R., Seiler P., Webby R.J., Songserm T., Guan Y., et al. (2007) Role of terrestrial wild birds in ecology of influenza A virus(H5N1). *Emerg. Infect. Dis.***13** 1720–172

Boyce W.M., Sandrock C., Kreuder-Johnson C., Kelly T. and Cardona C. (2009) Avian influenza viruses in wild birds: a moving target. *Comp.Immunol. Microbiol. Infect. Dis.* **32** 275–286

Brown J.D., Stallknecht D.E., Berghaus R.D., Swayne D.E.(2009) Infectious and lethal doses of H5N1 highly pathogenic avian influenza virus for house sparrows (Passer domesticus) and rock pigeons (Columbia livia). *J. Vet. Diagn. Invest.* **21** 437-445

Burr,I.W. (1942) Cumulative frequency functions. *Annals of Mathematical Statistics* **13**, 215–232.

Claas E.C., Osterhaus A.D., van Beek R., De Jong J.C., Rimmelzwaan G.F., Senne D.A., et al. (1998) Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet.* **351** 472–477

Cohen,J. The World is Round ($p < .05$). *American Psychologist* **49**, (12), 997-1003

Consul, P.C. and Famoye, F. (2006). *Lagrange probability distributions.* Boston: Birkhäuser

Curd, M. and Cover, J.A. (1998). *Philosophy of science, the central issues.* New York: W.W. Norton & Company.

Diekmann, O and Heesterbeek, J.A.P. (2000). *Mathematical Epidemiology of Infectious Diseases. Model Building, Analysis and Interpretation.* Chichester: John Wiley & Sons.

Ferguson, N., Donnely, C. and Anderson, R. (2001). The Foot-and-Mouth Epidemic in Great Britain: Pattern of Spread and Impact of interventions. *Science* **292**, 1155–1160.

Fouchier RAM, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, et.al. (2005) Characterization of a novel influenza A virus hemag-glutinin subtype (H16) obtained from black-headed gulls. *J Virol.* **79** 2814–2822

Gauthier-Clerc M., Lebarbenchon C., Thomas F. (2007) Recent expansion of highly pathogenic avian influenza H5N1: a critical review. *IBIS* **149** 202–214

Gilbert M., Chaitaweesub P., Parakamawongsa T., Premashthira S., Tiensin T., Kalpravidh W., et al. (2006) Free-grazing ducks and highly pathogenic avian influenza, Thailand. *Emerg. Infect. Dis.* **12** 227–234

Jia B., Shi J., Li Y., Shinya K., Muramoto Y., Zeng X., et al. (2008) Pathogenicity of Chinese H5N1 highly pathogenic avian influenza viruses in pigeons. *Arch. Virol.* **153** 1821–1826

Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate discrete distributions*, Third edition. John Wiley & Sons

Jones, P.W., and Smith, P. (2010). *Stochastic processes. An introduction.* London: CRC press.

Katz, R.W. and Brown, B.G.(1992) Extreme events in a changing climate: variablity is more important than averages. *Climate Change* **21**, 289–302

Keawcharoen, J., Van den Broek, J., Bouma, A., Tiensin, T., Osterhaus, A.D.M.E. and Heesterbeek, H. (2011). Wild Birds and Increased Transmission of Highly Pathogenic Avian Influenza (H5N1) among Poultry, Thailand. *Emerging Infectious Diseases* **17**, No. 6, 1016–1022

Kendall, D.G. (1948). On the generalized "birth-and-death" process. *Annals of Mathematical Statistics* **19**, 1–15

Kilpatrick A.M., Chmura A.A., Gibbons D.W., Fleischer R.C., Marra P.P. and Daszak P. (2006) Predicting the global spread of H5N1 avian influenza. *Proc. Natl. Acad. Sci. USA.* **103** 19368–19373

Kleibner, C. (1996). Dagum vs. Singh-Maddala income distribution. *Economics Letters* **53**, 265–268.

Kleibner, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences.* Hoboken,New Jersey: John Wiley.

Klopfleisch R., Werner O., Mundt E., Harder T. and Teifke J.P. (2006) Neurotropism of highly pathogenic avian influenza virus A/chicken/Indonesia/2003 (H5N1) in experimentally infected pigeons (Columbia livia f. domestica). *Vet. Pathol.***43** 463–470

Langberg, N.A. (1980). The convergence in Distribution of Some Simple Epidemics. *Mathematical Biosciences* **50**, 273–284.

Lee M.S., Chang P.C., Shien J.H., Cheng M.C., Shieh H.K. (2001) Identification and subtyping of avian influenza viruses by reverse transcription-PCR. *J. Virol. Methods* **97** 13–22

Leslie D.S. and Ian H.B. (2008) Multicontinental epidemic of H5N1 HPAI virus (1996–2007). In: Swayne DE, editor. *Avian influenza.* Ames (IA): Blackwell Publishing, 2008. p. 269.

Lewis, A.L. (2000). *Option Valuation under Stochastic Volatility: with Matematica Code.* Newport Beach, California, USA: Finance Press.

Lindsey, J.K. ,(1996). *Parametric Statistical Inference.* Oxford: Clarendon Press.

Lindsey, J.K. (2001). *Nonlinear Models in Medical Statistics.* Oxford: Oxford University press.

Lindsey, J.K. (2004). *Statistical Analysis of Stochastic Processes in Time.* Cambridge, UK: Cambridge University press.

Liu J., Xiao H., Lei F., Zhu Q., Qin K., Zhang X.W., et al. (2005) Highly pathogenic H5N1 influenza virus infection in migratory birds. *Science.* **309**

Liu Y., Zhou J., Yang H., Yao W., Bu W., Yang B., et al. (2007) Susceptibility and transmissibility of pigeons to Asian lineage highly pathogenic avian influenza virus subtype H5N1. *Avian Pathol.***36** 461–465

Nishiura, H., and Chowell, G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In: *Mathematical and Statistical Estimation Approaches in Epidemiology* (Edited by: Chowell, G., Hyman, J.M., Bettencourt, L.M.A. and Castillo-Chavez, C.), Springer, New York, pp. 103-121.

Olsen, B., Munster, V., Wallensten, A., Waldenstrom, J., Osterhaus, A. and Fouchier, R. (2006) Global patterns of influenza A virus in wild birds. *Science* **312** 384–388

Peiris, J.S.M, de Jong, M.D. and Guan, Y. (2007) Avian influenza virus (H5N1): a threat to human health. *Microbiol Rev.* **20** 243– 267

Pelupessy, I., Bonten, M.J.M. and Diekman, O. (2009). How to assess the relative importance of different colonization routes of pathogens within hospital settings. *PNAS* **99**, 5601–5605

R Development core team (2010). R: A language and environment for statistical computing, R Foundation for Statistical Computing:Vienna, Austria, 2010, ISBN 3-900051-07-0, http://www.R-project.org

Rensaw, E. (1991). *Modelling Biological Populations in Space and Time.* Camebridge: Cambridge University press.

Ross, S.M. (1983). *Stochastic processes.* New York: John Wiley & sons.

Royall, R. (1997). *Statistical Evidence, a likelihood paradigma.* London: Chapman & Hall.

Scheffer, M., Bascompte, J., Brock, W.A., Brovkin, V., Carpenter, R.C., Dakos, V., Held, H., Van Nes,E.H., Rietkerk, M. and Sugihara, G.(2009) Early-warning signals for critical transitions. *Nature* (2009) **461**, 53–59

Shao, Q. (2000) Estimation for hazardous concentrations based on NOEC toxicity data: An alternative approach. *Environmetrics* **11**, 583–595.

Shao, Q. (2004). Notes on the maximum likelihood estimation for the three parameter Burr XII distribution. *Computational Statistics* & Data Analysis **45**,675–687.

Shoa, Q. AND ZHOU, X. (2004). A new parametric model for survival data with long-term survivors. *Statistics in Medicine* **23**,3525–3543.

Shkedy, Z. (2003). *Flexible Statistical Modelling: Application to Infectious Diseases and Astronomical Data.* PhD.thesis 2003; Limburgs Univeritair Centrum, Diepenbeek, Belgium.

Siengsanan, J., Chaichoune, K., Phonaknguen, R., , L., Prompiram, P., Kocharin, W., et al. (2009) Comparison of outbreaks of H5N1 highly pathogenic avian influenza in wild birds and poultry in Thailand. *J.Wildl.Dis.* **45** 740–747

Songserm, T., Jam-on, R., Sae-Heng, N., Meemak, N., Hulse-Post, D.J., Sturm-Ramirez, K.M., et al. (2006) Domestic ducks and H5N1 influenza epidemic, Thailand. *Emerg. Infect. Dis.* **12** 575–581

Spackman (2008) A brief introduction to the avian influenza virus. *Methods Mol. Biol.* **436** 1–6

Spiegelhalter,D.J.(2005) Problems in assessing the rates of infection with methicillin resistant Staphylococcus aureus. *BMJ* **331**, 1013–1015

Stegeman, J.A., Bouma, A., Elbers, A.R.W., Van Boven, M., De Jong, M.C.M., Nodelijk, G., De Klerk, F. and Koch,G. (2003) Avian influenza a virus (H7N7) epidemic in The Netherlands in 2003: course of the epidemic and effectiveness of control measures. *The Journal of Infectious Diseases.* **190**, 2088–2095

Stegeman, J.A., Elbers, A.R.W., Smak, J. and De Jong, M.C.M. (1999). Quantification of the transmission of classical swine fever virus between herds during the 1997-1998 epidemic in The Netherlands. *Preventive Veterinary Medicine* **42**, 219–234.

Tacconelli, E.,De Angelis, G., Cataldo, M.A., Pozzi, E. and Cauda, R. (2008) Does antibiotic exposure increase the risk of methicilin-resistant Stahylococcus aurus (MRSA) isolation? A systemetic review and meta-analysis. *Journal of Antimicrobial Chemotherapy* **61**, 26–38

Tiensin, T., Ahmed Syed, S.U., Rojanasthien, S., Songserm, T., Ratanakorn, P., Chaichoun, K., et al. (2009) Ecologic risk factor investigation of clusters of avian influenza A (H5N1) virus infection in Thailand. *J. Infect. Dis.* **199** 1735–1743

Tiensin, T., Chaitaweesub, P., Songserm, T., Chaisingh, A., Hoonsuwan, W., Buranathai, C., et al. (2005) Highly pathogenic avian influenza H5N1, Thailand, 2004. *Emerg. Infect. Dis.***11** 1664–1672

Thomson, B. (2007) *The nature of statistical evidence.* Springer, New York

Uchida, Y., Chaichoune, K., Wiriyarat, W., Watanabe, C., Hayashi, T., Patchimasiri, T., et al. (2008) Molecular epidemiological analysis of highly pathogenic avian influenza H5N1 subtype isolated from poultry and wild bird in Thailand. *Virus Res.* **138** 70–80

Van Den Broek, J. (1995) A score test for zero inflation in a Poisson distribution. *Biometrics* **51**, 738–743

Webster RG, Bean WJ, Gorman OT, Chambers TM and Kawaoka Y. (1992) Evolution and ecology of influenza A viruses. *Microbiol Rev.* **56** 152–179

Werner O., Starick E., Teifke J., Klopfleisch R., Prajitno T.Y., Beer M., et al. (2007) Minute excretion of highly pathogenic avian influenza virus A/ chicken/Indonesia/2003 (H5N1) from experimentally infected do- mestic pigeons (Columbia livia) and lack of transmission to sentinel chickens. *J Gen Virol.***88** 3089–3093

World Organisation for Animal Health. (2010) 63 countries report H5N1 avian influenza in domestic poultry wildlife. 2003-2010

`http://www.oie.int/eng/info_ev/en_AI_factoids_2.htm`

Samenvatting

Het onderwerp van dit proefschrift is het niet-homogene geboorte-sterfte proces met enkele van zijn speciale gevallen en de toepassing ervan in het modelleren van epidemische uitbraak data. Dit model beschrijft veranderingen in populatie groottes. Nieuwe populatie elementen kunnen ontstaan met een bepaald aantal per tijdseenheid, welk aantal de geboorte rate of het reproducerende vermogen (reproductive power) wordt genoemd en elementen kunnen verdwijnen met een bepaald aantal per tijdseenheid, wat de sterfte rate wordt genoemd. Deze rates zijn niet tijds homogeen dat wil zeggen ze kunnen veranderen in de tijd.

Omdat het model wordt gebruikt voor het modelleren van epidemische data is de populatie grootte in dit proefschrift het aantal geinfecteerde individuen op een bepaald moment in de tijd. Omdat de modellen worden toegepast op uitbraak data worden in de introductie een korte beschrijving gegeven van wat algemene aspecten van epidemische modellen. Bovendien wordt er enkele onderwerpen besproken die te maken hebben met de manier waarop data als bewijs kan worden gebruikt. Er worden ook technieken besproken aan de hand van het homogene geboorte-sterfte proces, om de stochastische differentiaal vergelijkingen op te stellen en met behulp van deze de differentiaal vergelijking voor de kans genererende functie. Dan wordt besproken hoe de Lagrange transformatie kan worden gebruikt om de kans genererende functie om te schrijven in een kansverdeling.

De technieken besproken in het eerste hoofdstuk worden in het tweede hoofdstuk gebruikt om de verdelingsfunctie van het meer algemene niet-homogene geboorte-sterfte proces af te leiden. Als het reproducerend vermogen en de sterfte rate constant zijn dan zijn de reproductie tijd en de sterfte tijd exponentieel verdeeld. Deze verdelingen hebben in het geval van de niet homogene rates een algemene vorm. Een natuurlijke keuze voor deze verdelingen en argumenten daarvoor wordt besproken in hoodfstuk 2 en 4. Voor wat betreft de survival functie worden in hoofdstuk 2, 3 modellen besproken: het "proportional rate" model, het "accelerated failure time" model en een combinatie van beide. Het niet homogene geboorte- sterfte proces genereert

een parameter die de netto reproductie ratio (net reproduction ratio)genoemd kan worden en die het uitbraak-proces karakteriseert. Deze ratio wordt gebruikt om de vogelgriep (aviaire influenza H7N7) uitbraak in Nederland in 2003 te bestuderen.

In hoofdstuk 3 wordt een niet-homogeen martingaal model besproken om volatiliteit– zeg maar verandering in variatie in een tijdreeks – te modelleren. Dit model wordt afgeleid uit het niet homogene geboorte-sterfte proces door het reproducerend vermogen gelijk te stellen aan de sterfte rate maar wel afhankelijk van de tijd te houden. Zo wordt een model verkregen waarvan de verwachting op een bepaald tijdstip gelijk is aan de populatie grootte (aantal geinfecteerden) op een vorig tijdstip. De variantie echter is afhankelijk van de tijd. Door dit model te vergelijken met het niet-homgene geboorte-sterfte proces, kan worden bepaald of er een tijd trend aanwezig is in de geobserveerde data (wat bij data die bestaat uit tellingen ook betekent dat de variantie afhankelijk is van de tijd) of dat er alleen sprake is van volatiliteit met een constant gemiddelde. In dit hoofdstuk wordt geillustreerd dat de netto reproductie hiervoor goed als hulpmiddel gebruikt kan worden. Deze modellen worden toegepast op MRSA (Methicillin-Resistant Staphylococcus aureus) data van 3 "Acute Trust" ziekenhuizen van het "National Health Service" in Groot-Brittannie.

In hoofdstuk 4 worden twee speciale gevallen van het niet-homogene geboorte-sterfte proces besproken: het niet-homogene geboorte proces en het niet-homogene sterfte proces. De overlevingsfunctie wordt aangepast door er een "final size" parameter in op te nemen op dezelfde manier als dat gebeurt bij "long term survival" modellen. Deze modellen worden toegepast op 3 uitbraken: de Nederlandse varkenspest uitbraak in 1997-1998, de mont en klauwzeer uitbraak in Groot-Brittannie en de Nederlandse vogelgriep (aviaire influenza H7N7) uitbraak in 2003.

In hoofdstuk 5 wordt het niet-homogene geboorte proces toegepast op de transmissie van aviaire influenza (H5N1) onder pluimvee in Thailand. Om de epidemiologie van deze virale infectie te bepalen en de relatie met een uitbraak onder wilde vogels in Thailand van 2004 tot 2007, werd er onderzocht wat de rol was van wilde vogels in de transmissie. Gebieden waar er een uitbraak was, werden geclassificeerd als een gebied met een uitbraak onder wilde vogels of als een gebied waar geen uitbraak was onder wilde vogels. Deze definitie werd gebruikt als onafhankelijke variabele in een "proportional rate" formulering van het niet-homogene geboorte model.

In hoofdstuk 6 wordt het schatten van het reproducerend vermogen en de overlevingsfunctie met overdraagbare gebeurtenissen besproken. Met het niet-homogene geboorteproces kan men het reproducerend vermogen en de overlevingsfunctie en hun standaard fouten direct uit de data schatten met behulp van de log-likelihood in plaats van een parametrische vorm voor de overlevingsfunctie te kiezen zoals dat in de andere hoofdstukken gedaan is. Er wordt in dit hoofdstuk aangetoond dat de standaard

fouten van het geschatte reproducerend vermogen en van de geschatte overlevingsfunctie kleiner worden met toenemende tijd, omdat met overdraagbare gebeurtenissen de hoeveelheid informatie toeneemt met toenemende tijd. Er wordt een methode ontwikkeld om empirische schattingen van twee onafhankelijke groepen met elkaar te vergelijken door middel van de reproducerend vermogen ratio. Er wordt aangetoond dat de standaard fout van de logaritme van de ratio van reproducerend vermogens groter wordt met toenemende tijd omdat de deze standaard fout afhangt van het reproducerend vermogen. Deze methode worden toegepast op de Nederlandse vogelgriep (aviaire influenza H7N7) uitbraak in 2003 en op de aviaire influenza (H5N1) uitbraak onder pluimvee in Thailand.

Acknowledgment

First of all I want to thank my promotor Hans Heesterbeek. Hans bedankt voor de vrijheid die ik kreeg om mijn eigen gang te kunnen gaan bij de keuze van het onderwerp. Bedankt voor de discussie en het meedenken en voor je geduld met mijn spelling.

I would like to thank Hiroshi Nishiura for his cooperation with chapter two and I would Jim Lindsey for comments on a early draft of chapter 4.

Curriculem Vitae

Jan van den Broek was born on June 1st in Ermelo (The Netherlands). Since 1984 he works as a statistician, first at the National Institute for Public Health and the Environment in Bilthoven and at TNO in Soesterberg and since 1988 at the Faculty of Veterinary Medicine of the Utrecht University. There he teaches statistics and is involved in veterinary research projects, and provides statistical consulting. Besides this he does statistical research. In 1995 he received his Master in Biostatistics from the Limburgs Universitair Centrum in Diepenbeek (Belgium). In 2003 he started, for one day a week, Phd research resulting in this thesis.