Biostatistics (2007), **8**, 2, *pp.* 453–467
doi:10.1093/biostatistics/kxl023
Advance Access publication on September 6, 2006

Nonhomogeneous birth and death models for epidemic outbreak data

JAN VAN DEN BROEK*, HANS HEESTERBEEK
Faculty of Veterinary Medicine, Utrecht University, The Netherlands
j.vandenbroek@vet.uu.nl

SUMMARY

In this paper, generalized nonlinear models are proposed in order to incorporate the following considerations in modeling an epidemic disease outbreak statistically. (1) The dependence of the data is handled with a nonhomogeneous death or a nonhomogeneous birth process. (2) The first stage of the outbreak is described with an epidemic susceptibles-infectives-removed (SIR) model. Soon the control measures taken will dominate the process. These measures are in addition to the natural epidemic removal process. The prevalence is related to the censored infection times in such a way that the distribution function and thus the survival function satisfy approximately the first equation of the SIR model. This leads in a natural way to the Burr family of distributions. (3) The nonhomogeneous birth process handles the fact that in practice, with some delay, infecteds are registered, but not susceptibles. (4) Finally, the ending of the epidemic caused by the measures taken 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 to 1998, the foot-and-mouth disease outbreak in Great Britain from 2001, and the Dutch avian influenza (H7N7) outbreak from 2003.

Keywords: Avian influenza (H7N7); Burr distribution; Classical swine fever; Foot-and-mouth disease; Force of infection; Negative binomial; Nonhomogeneous birth process; Nonhomogeneous death process; Reproductive power.

1. INTRODUCTION

Modeling 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. First, infected individuals arise in time which makes the data dependent: what happens on a certain time point depends on what happened before that. Second, these infecteds are, for various reasons, not immediately observed at the time of infection but after some detection time. The prevalence of the

*To whom correspondence should be addressed.

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 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 of right-censored infection times. These kinds of data are 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 are there 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 infected individuals as infectives (because there is no distinction); we will use infected in more general descriptions outside the context of the particular model (because in reality an infected individual could be in a latency phase and not yet infectious to others). In this paper, 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

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = -\beta(t)y(t)x(t). \tag{1.1}$$

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 (1.1) are described, the force of infection model and the reproductive power model, together with their nonhomogeneous 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 to 1998, the foot-and-mouth disease outbreak in Great Britain from 2001, and the Dutch avian influenza (H7N7) outbreak from 2003.

2. THE MODELS

2.1 *Force of infection*

Assume that the disease is irreversible and concentrate on monitoring the total number of susceptibles. Equation (1.1) 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

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = -\lambda(t)x(t). \tag{2.1}$$

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 remaining. The rate $\lambda(t) = \beta(t)y(t)$ itself depends on the fraction infectives. The force of infection model makes the following approximations:

1. The proportion of susceptibles at time zero, 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 (2.1) 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 (2.1) which is the differential equation of the nonhomogeneous death process with $X(t)$ the number of susceptibles at time t, $X(0) = x_0$ at time zero, and with $p_{x_t}(t)$ the probability that the number of susceptibles at time t is x_t:

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

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

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

and thus

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

the binomial distribution with success probability $\mathrm{e}^{-\rho(t)} = \exp\left[-\int_0^t \lambda(\tau)\mathrm{d}\tau\right] = S_\lambda(t)$ and $\lambda(t)$ the hazard rate. Note that x_t is a count, whereas $x(t)$ is a fraction. Also the expected value of $X(t)$ is the same as the solution of the deterministic equation (2.1) with $x(t)$ 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).

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 of 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 determined not only 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 proportion of susceptibles is large, $y(t)$, the proportion of infectives can approximately be taken as the proportion of infected

at time t, and hence $x(t) + y(t) \approx 1$. In this case, one can write (1.1) as a nonhomogeneous birth process:

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

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 (2.2) 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 the stochastic version of the nonhomogeneous birth process with $Y(t)$ the total number of infected at time t and y_0 the number of infected at time zero, we have the following differential equation:

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

with $p_{y_t}(t)$ the probability that the number of detected infected at time t is y_t. The solution has probability-generating function (Kendall, 1948)

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

and thus

$$P(Y(t) = y_t) = \binom{y_t - 1}{y_0 - 1}[\mathrm{e}^{\rho(t)}]^{y_0}[1 - \mathrm{e}^{\rho(t)}]^{y_t - y_0}, \quad y_t = y_0, y_0 + 1, \ldots$$

with $\mathrm{e}^{\rho(t)} = \exp\left[-\int_0^t \mu(\tau)\mathrm{d}\tau\right] = S_\mu(t)$ and with y_t now the number of infected, whereas $y(t)$ above is a fraction. 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. First, if for $S_\mu(t)$ the exponential distribution is taken, then this model is the same as that of 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.

Second, the expected value of $Y(t)$ is the same as the solution of the deterministic equation (2.2) 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.

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 (1.1) into account. We have assumed that this deterministic equation holds approximately in the first stage of the outbreak for the susceptible fraction and 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 (1.1) should also hold for the survival function and the distribution function because of this equivalence. In that case, the expected value of the nonhomogeneous 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

$$\frac{\mathrm{d}}{\mathrm{d}t}F(t) = \beta(t)F(t)(1 - F(t)) \quad \text{or} \quad \frac{\mathrm{d}}{\mathrm{d}t}S(t) = -\beta(t)S(t)(1 - S(t)) \tag{2.3}$$

for some nonnegative function $\beta(\cdot)$ (Kleibner and Kotz, 2003, p 52). This function $\beta(\cdot)$ can also depend on $F(t)$. When it does only depend on t, the solution of (2.3) reduces to the logistic form: $F(t) = [1 + \exp(-\int_0^t \beta(\tau)\mathrm{d}\tau)]^{-1}$ (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.

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, p 198).

For $a = 1$, the Burr distribution reduces to the Lomax (Pareto type II) distribution. The case $a = q$ is also known as the paralogistic 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 that 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

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

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).

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.

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 so much that it leads to 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 nonhomogeneous 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 nonhomogeneous 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 π as that fraction of the total number of infected that were generated by the y_0 infected that started the epidemic.

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, p 108) with, say, t_j, $j = 0, \ldots, n$, 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 nonhomogeneous 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 probability

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

where $h(t)$ is the discrete-time hazard rate. So at time t_j, the distribution of the number of infectives conditional on the number of infectives at t_{j-1} is

$$P\left(X(t_j) = x_{t_j} | X(t_{j-1}) = x_{t_{j-1}}\right) = \binom{x_{t_{j-1}}}{x_{t_j}} [1 - h(t_{j-1})]^{x_{t_j}} [h(t_{j-1})]^{x_{t_{j-1}} - x_{t_j}}. \tag{2.4}$$

The expected value $x_{t_j}(1 - h(t_j))$, $j = 0, \ldots, n$, might be called the individual or sample path profile (Lindsey, 2001). The difference between the individual and the underlying profile is that the former relates to the conditional model above, whereas the latter relates to the unconditional model. This model is very similar to the model of Becker (1989, Chapter 6, p 109).

If δ is the vector of parameters for the Burr distribution, then from (2.4) the log-likelihood $l(\delta)$ can be written as

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

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, 2005). The information matrix can be used to find the standard errors. With the nonhomogeneous 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 zero with y_0 infectives. So the number of infected at time t_1 has a shifted negative binomial with y_0 and probability $S(t_1)$. 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

$$P\left(Y(t_j) = y_{t_j} | Y(t_{j-1}) = y_{t_{j-1}}\right) = \binom{y_{t_j} - 1}{y_{t_{j-1}} - 1} [1 - h(t_{j-1})]^{y_{t_{j-1}}} [h(t_{j-1})]^{y_{t_j} - y_{t_{j-1}}},$$

$$y_{t_j} = y_{t_{j-1}}, y_{t_{j-1}} + 1, y_{t_{j-1}} + 2, \ldots.$$

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

$$P\left(Z(t_j) = z_{t_j} | Z(t_{j-1}) = z_{t_{j-1}}\right) = \binom{z_{t_j} + z_{t_{j-1}} - 1}{z_{t_{j-1}} - 1} [1 - h(t_{j-1})]^{z_{t_{j-1}}} [h(t_{j-1})]^{z_{t_j}},$$

$$z_{t_j} = 0, 1, \ldots. \tag{2.5}$$

So, conditioning on the previous observed number of cases makes the process start again in the same way as it was started at time zero. 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 (2.5) the log-likelihood $l(\delta)$ can be written as

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

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 nonhomogeneous death process giving the maximum likelihood estimates and, by using the information matrix, the standard errors can also be obtained.

3. THREE OUTBREAKS

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 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 nonhomogeneous birth process.

With the Burr distribution, the parameter q only affects the right tail. As can be seen from Figure 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 error of the logarithm of this parameter is very large, see Table 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 Akaike's Information Criterium (AIC) for the Burr distribution is 222.9. With the

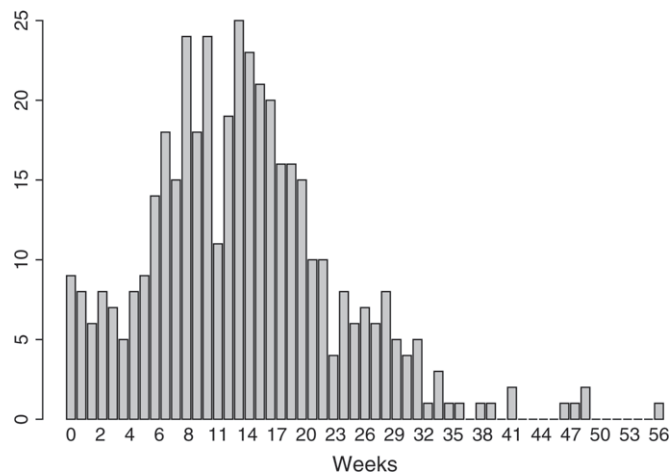

Fig. 1. New cases in the classical swine fever outbreak in The Netherlands, 1997–1998 by week.

Table 1. *Parameter estimates for the classical swine fever outbreak*

Burr distribution			Inverse Burr distribution		
Parameter	Estimate	Standard error	Parameter	Estimate	Standard 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

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 1. The individual profile and the underlying profile for the inverse Burr are in Figure 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 3. To get a sense of the statistical uncertainty at the estimated model, 100 sample paths from the estimated model were drawn and for each sample path, the reproductive power function was calculated. These reproductive power functions are shown in Figure 3 as gray lines.

Stegeman *and others* (1999) distinguish several phases of the outbreak according to the measures taken to control the spread of 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

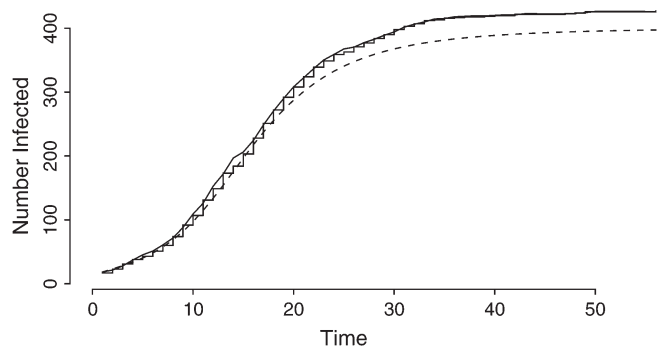

Fig. 2. The classical swine fever outbreak data (step line). The individual profile (solid line) and the underlying profile (dashed line) for the inverse Burr.

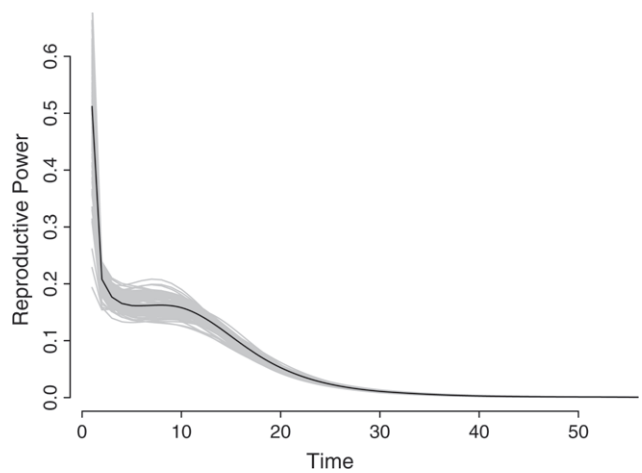

Fig. 3. The reproductive power for the inverse Burr for the classical swine fever outbreak data (black line) and the reproductive power calculated from sample paths drawn from the underlying estimated process (gray lines).

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 neighborhood of 0.3, then decreases quickly to a value just below 0.2 where it stays approximately until week 10–11 and then drops again until about week 20 after which it still decreases but less fast. The simulated reproductive power lines also follow this pattern. Most uncertainty seems to be at the start of the outbreak but the periods can also be seen in these simulated lines.

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 (Figure 4). The epidemic was controlled by movement restrictions, stamping out of infected flocks, and preemptive culling of flocks in the neighborhood of infected flocks. In total, 1255 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 *and others* (2003).

For the same reasons as for the classical swine fever outbreak, the nonhomogeneous birth model is more appropriate here. As compared to the classical swine fever outbreak, the data are more dispersed. 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 gives a better fit. The logarithms of the estimates are in Table 2 and Figure 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 *and others* (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

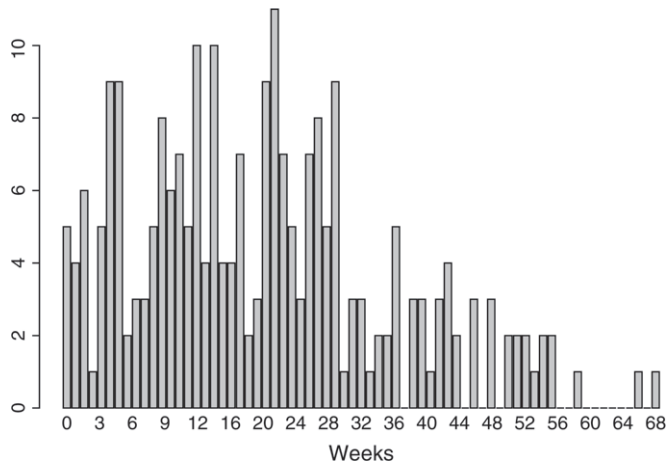

Fig. 4. New cases in the avian influenza (H7N7) outbreak by day.

Table 2. *Parameter estimates for the avian influenza (H7N7) outbreak*

Burr distribution			Inverse Burr distribution		
Parameter	Estimate	Standard error	Parameter	Estimate	Standard 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

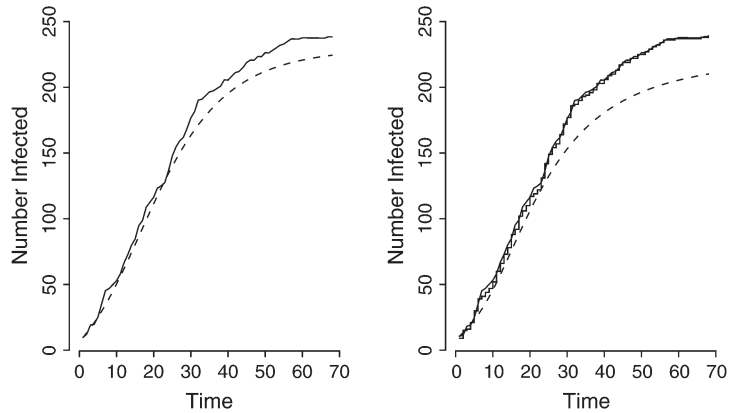

Fig. 5. The avian influenza (H7N7) outbreak data (step line). The individual profile (solid line) and the underlying profile (dashed line). Left for the Burr and right for the inverse Burr.

were only detected in the Gelderse Vallei. By the end of March, the virus had escaped to the Southern part of The Netherlands; between zero and five outbreaks per day were observed throughout most of April. The plot of the reproductive power shown in Figure 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 afterward 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.

To show the statistical uncertainty at the estimated model, 100 sample paths from the estimated model were drawn and for each sample path the reproductive power function was calculated, both for the Burr and for the inverse Burr case. These reproductive power functions are shown in Figure 6 as gray lines. These simulated lines also show the periods mentioned above. Most uncertainty is at the start of the outbreak, but whether a line starts low or high there is a clear change at approximately 5 days. Note that this is also the case with the Burr distribution although there the estimated reproductive power function is a rather smooth curve at 5 days.

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 *and others*, 2001). For more information on foot-and-mouth disease, see the web site

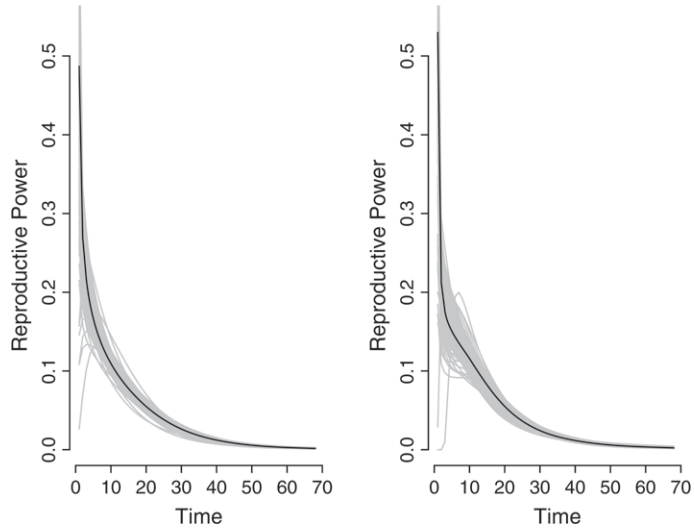

Fig. 6. The reproductive power for the Burr (left,black line) and the inverse Burr (right,black line) for the avian influenza (H7N7) outbreak data. The gray lines are the reproductive power functions calculated from sample paths drawn from the underlying estimated process.

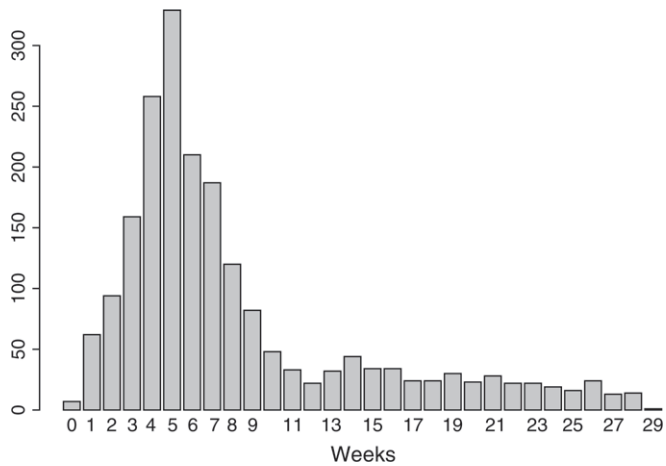

Fig. 7. New cases in the foot-and-mouth disease outbreak by week.

of the British Department of environment foot and rural affairs: http://www.defra.gov.uk/footandmouth. There were in total 2015 detected farms registered.

As can be seen from Figure 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 3; the underlying profile and the individual profile for both the distributions are in Figure 8. When comparing the parameter estimates of the inverse Burr with the outbreaks discussed above, one will note that they

Table 3. *Parameter estimates for the foot-and-mouth disease outbreak*

Burr distribution			Inverse Burr distribution		
Parameter	Estimate	Standard error	Parameter	Estimate	Standard 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

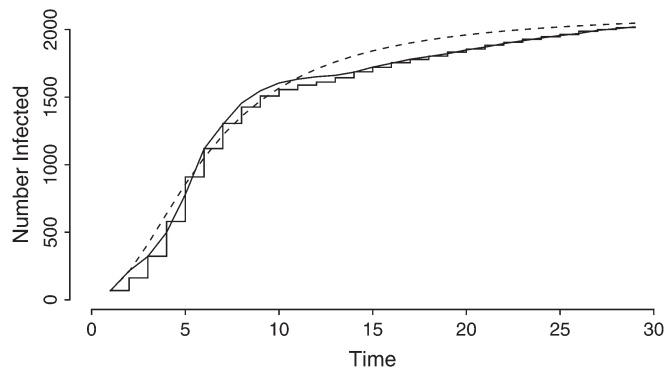

Fig. 8. The foot-and-mouth disease outbreak data (step line). The individual profile (solid line) and the underlying profile (dashed line). Left for the Burr and right for the inverse Burr.

are quite 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 where after it stabilizes on a value around 25. This can also be seen from the individual profiles in Figure 8 where both the Burr and the inverse Burr have a difficulty to follow the sharp bend in the curve. The reproductive power for the inverse Burr distributions is given in Figure 9. This function drops fast until about 5 weeks then decrease more slowly until about 10 weeks, after which the decrease levels off. Also here, 100 sample paths from the estimated model were drawn and for each sample path, the reproductive power function was calculated. These reproductive power functions are shown in Figure 9 as gray lines.

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. Discussion

The generalized nonlinear models described in this paper incorporate some issues met when an epidemic disease outbreak is studied. The dependence of the data is handled with a nonhomogeneous death or a nonhomogeneous birth process. The prevalence is related to the censored infection times in such a way that the distribution function and thus the survival function agree with the first equation of the SIR model for infection spread, incorporating the fact that cases are reported with some delay. The nonhomogeneous 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

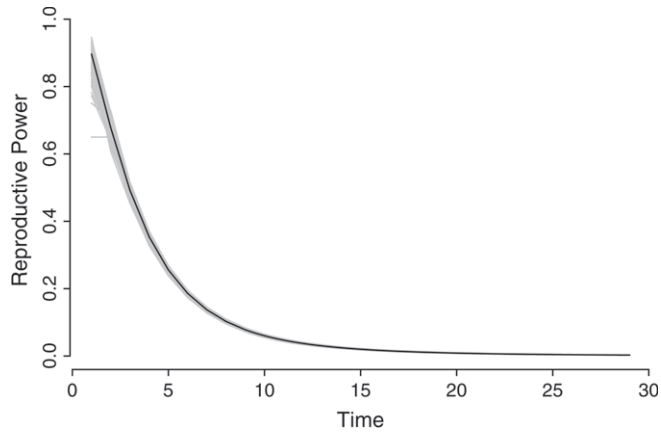

Fig. 9. The reproductive power for the inverse Burr (black line) for the foot-and-mouth disease outbreak data together with the reproductive power functions calculated from sample paths drawn from the underlying estimated process (gray lines).

by modifying the survival function with a final-size parameter in the same way as is done in long-term survival modeling. 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 large standard errors for some parameter estimates. 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 sheeps 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 article focuses on the application of the nonhomogeneous birth process, since with spontaneous outbreaks this seems the most appropriate. The nonhomogeneous 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 whether 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.

ACKNOWLEDGMENTS

We thank Jim Lindsey for helpful comments on an early draft of this paper. The authors thank the referees for their many detailed comments and suggestion which helped to improve this paper. *Conflict of Interest:* None declared.

REFERENCES

BECKER, N. G. (1989). *Analysis of Infectious Disease Data*. London: Chapman and Hall.

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

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

FERGUSON, N., DONNELLY, C. AND ANDERSON, R. (2001). The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science* **292**, 1155–1160.

KENDALL, D. G. (1948). On the generalized "birth-and-death" process. *Annals of Mathematical Statistics* **19**, 1–5.

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, NJ: John Wiley.

LANGBERG, N. A. (1980). The convergence in distribution of some simple epidemics. *Mathematical Biosciences* **50**, 273–284.

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

R DEVELOPMENT CORE TEAM (2005). *R: A language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

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

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

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.

[*Received February 1, 2006; first revision June 2, 2006; second revision August 30, 2006; accepted for publication August 30, 2006*]