Statistical Modelling 2015; **15(5)**: 457–475

Modelling volatility using a non-homogeneous martingale model for processes with constant mean on count data

Jan van den Broek[1]
[1]Faculty of Veterinary Medicine, Utrecht University, The Netherlands

Abstract: In this article a non-homogeneous martingale model is proposed to model volatility in a stochastic time series of count data with constant mean. The approach is derived from a general non-homogeneous birth-and-death process, in which the mean and the variance of population size can vary as a function of time. This model can be important in modelling early warning signals that there is going to be a change of state in a complex system. The net reproduction ratio obtained from fitting a non-homogeneous birth–death model can be used as an additional tool to compare this model with a model where there is no change in the mean over the observation period. These models and procedures are illustrated with quarterly Methicillin resistant staphylococcus aureus prevalence data registered since 2001 from three Acute Trusts of hospitals of the National Health Service in Great Britain.

Key words: count data; martingale; net reproduction ratio; non-homogeneous birth-death process; variance; volatility

Received March 2014; **revised** December 2014; **accepted** December 2014

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 behaviour. 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 where 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 erroneously an 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.

Address for correspondence: Jan van den Broek, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 7 3584 CL Utrecht The Netherlands. E-mail: j.vandenbroek@uu.nl

10.1177/1471082X14565358

Volatility can have different meanings in the literature. In the economic literature volatility is used for the constant diffusion coefficient in a geometric Brownian motion when modelling the (log) outcome of a financial instrument over time, such as with the famous Black-Scholes model.

Secondly, 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. Modelling 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 modelled 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 *et al.*, 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).

Volatility in this article is defined as the possible non-homogeneous variability in a measurement of interest over time.

Modelling volatility usually means modelling 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 $\sqrt{\mu_t}$, where 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, where the mean of the process is constant, but where 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.

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 becomes 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 foregoing modelling studies of data from 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 assumptions 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 article proposes a new non-homogeneous martingale model for (changing) volatility in count data with constant mean. As said, volatility is used here for the (homogeneous) variability of the outcomes of a stochastic process. In Section 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. In Section 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 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.

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 '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 non-linear—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 at 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 modelling 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 article 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 (Van den Broek and Nishiura, 2009). 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 following differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}p_y(t) = \lambda(t)(y-1)p_{y-1}(t) + \mu(t)(y+1)p_{y+1}(t) - (\lambda(t)+\mu(t))\,yp_y(t),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}p_0(t) = \mu(t)p_1(t) \tag{2.1}$$

where $p_y(t)$ represents the probability that the observed population size at time t is y. The parameters $\lambda(t)$ and $\mu(t)$ are referred to as the birth and death rates, respectively. Non-homogeneity is reflected by allowing both $\lambda(t)$ and $\mu(t)$ depend on time t.

From the birth rate it can be seen that $S_\lambda(t) = e^{-\int_0^t \lambda(\tau)\mathrm{d}\tau}$ 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 solution of 2.1 is a probability distribution, the derivation of which can be found in Van den Broek and Nishiura (2009).

The non-homogeneous martingale process is a special case of the above mentioned non-homogeneous birth-and-death process. In a population with a constant mean the birth and death rates are equal (i.e., $\lambda(t) = \mu(t)$ (say $\eta(t)$)). The probability distribution can be derived along the lines of Van den Broek and Nishiura (2009) and is given by:

$$P(Y(t)=0) = \delta^{y_0}(t), \tag{2.2}$$
$$P(Y(t)=y) = \frac{y_0}{y}\delta^{y+y_0}(t)\sum_{k=0}^{\min(y-1,y_0-1)}\binom{y_0-1}{k}\binom{y}{y-k-1}\left[\frac{(1-\delta(t))}{\delta(t)}\right]^{2(k+1)} \quad y \geq 1.$$

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. In (2.2), $\delta(t)$ is, by using $\eta(t)$ (instead of $\lambda(t)$ or $\mu(t)$), equal to $\delta(t) = \frac{\log(S_\eta^{-1}(t))}{1+\log(S_\eta^{-1}(t))} = \frac{\gamma(t)}{1+\gamma(t)}$.

The expected value of the population size at time t derived from (2.2) is

$$E(Y(t)|Y(0) = y_0) = y_0.$$

That is, the expected value of the population size for any time point t, conditional on time zero, is equal to the population size at time zero and thus the process is a martingale process. (See for a definition of a martingale: Fleming and Harrington, 1991, p. 22–23.) The variance of the population size is:

$$\text{var}(Y(t)|Y(0) = y_0) = 2y_0 \log(S_\eta^{-1}(t)) = 2y_0 \int_0^t \eta(\tau)\mathrm{d}\tau \qquad (2.3)$$

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 pp. 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 are 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 Fitting the model

According to the model described in Section 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 (Van den Broek and Heesterbeek, 2007; Van den Broek and Nishiura, 2009).

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 (2.2) 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:

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

where $h(t_{j-1})$ is the birth rate or the death rate in discrete time, depending on the event considered.

Using $h(t_{j-1})$, and defining $c(t_{j-1})$ as the discrete version of $\delta(t)$ in the non-homogeneous martingale model: $c(t_{j-1}) = \frac{\log(\frac{1}{1-h(t_{j-1})})}{1-\log(\frac{1}{1-h(t_{j-1})})}$, the conditional version of (2.2) is

$$P(Y(t_j) = 0 | Y(t_{j-1}) = y_{t_{j-1}}) = c(t_{j-1})^{y_{t_{j-1}}}, \tag{3.1}$$
$$P(Y(t_j) = y_{t_j} | Y(t_{j-1}) = y_{t_{j-1}}) = \frac{y_{t_{j-1}}}{y_{t_j}} c(t_{j-1})^{y_{t_j}+y_{t_{j-1}}} \times$$
$$\sum_{k=0}^{\min(y_{t_j}-1, y_{t_{j-1}}-1)} \binom{y_{t_{j-1}}-1}{k} \binom{y_{t_j}}{y_{t_j}-k-1} \left[\frac{(1-c(t_{j-1}))}{c(t_{j-1})}\right]^{2(k+1)} \quad y_{t_j} \geq 1.$$

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

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

and this can be referred to as sample path profile (Lindsey, 2004). 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:

$$\text{var}(Y(t_j) | Y(t_{j-1}) = y_{t_{j-1}}) = 2y_{t_{j-1}} \log\left(\frac{1}{1-h(t_{j-1})}\right). \tag{3.2}$$

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:

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

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 computations (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.

4 Modelling issues

4.1 Using the net reproduction ratio

To compare the martingale model derived above 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 are given elsewhere (Van den Broek and Nishiura, 2009). 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:

$$E(Y(t)|Y(0) = y_0) = y_0 \frac{S_\mu(t)}{S_\lambda(t)} = y_0 R(t),$$

$$\text{var}(Y(t)|Y(0) = y_0) = y_0 R(t) \left[1 + (2\gamma - 1)R(t)\right],$$

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 (Van den Broek and Nishiura, 2009) since it is the expected number of secondary cases per primary case at time t (Diekmann and Heesterbeek, 2000; Nishiura and Chowell, 2009). 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 (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 (Van den Broek and Nishiura, 2009). In other words, if non-removed individuals reproduce faster than 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 compartment model with compartments susceptible, infectives and removed—interpretation of 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 also changing.

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., Bailey, 1990).

Since in the present work a model with changing volatility is compared to a model with a changing mean and thus, because the variance is a function of the mean, 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 behaviour 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.

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 (Kleiber and Kotz, 2003).

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 centres. 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. Yet other modelling approaches can be found in Cooper *et al.* (2004), Grundmann *et al.* (2002) and in Grundmann and Hellriegel (2006). In contrast the present article 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. The model is a special case of the non-homogeneous birth–death model to which it can be compared in order to determine whether or not there is a change in mean over time. This comparison can be done in two ways: by using Akaike's information criterium (or the bias corrected version if the sample size is small) and by use of the net reproduction ratio.

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 those with low-risk susceptibility will evade infection for a longer period. 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 of 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 Public Health Enland (https://www.gov.uk/government/statistics/mrsa-bacteraemia-annual-data).

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 article 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 the bias corrected Akaike's information criterium. There is a change in volatility.
- The data from the Guy's & St. Thomas' Trust do not show a clear choice between some birth–death models and a martingale model, based on the bias corrected 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 the bias corrected 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 a exponential distribution, constant birth rate, and the removal times had a 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 a generalized gamma, gamma or a Weibull distribution. Table 1 compares the bias corrected AIC values (AICc) between the different models for each of the three Trusts mentioned above. The bias corrected version of the AIC is used since the sample sizes are not that large.

Figure 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 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 1 the martingale model with the gamma distribution gives the best fit according to AICc, indicating that the volatility of the process is changing. This martingale model has an AICc 3.63 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 AICc. 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 AICc) and has the same number of parameters. The (log)parameter estimates and their standard errors for the martingale model with the gamma distribution

Table 1 Bias corrected AIC's (AICc's) for 3 Trusts using 14 different models

Model	Survival distribution	Leeds Teaching	Guy's & St. Thomas'	King's College
Martingale	Generalized Gamma	2812.54	25045.76	263.02
	Gamma	280.20	248.88	261.06
	Weibull	281.85	248.14	261.22
	Exponential	282.22	249.71	259.01
Birth–death	Generalized Gamma	288.51	253.19	267.39
	Gamma	284.40	251.74	262.28
	Weibull	286.21	252.03	263.78
	Exponential	283.83	250.68	261.93
Birth–death with birth rate exponential	Generalized Gamma	287.81	251.15	262.06
	Gamma	286.10	250.12	259.82
	Weibull	285.76	251.81	261.31
Birth–death with death rate exponential	Generalized Gamma	288.23	248.42	263.70
	Gamma	286.19	251.53	261.22
	Weibull	286.11	253.03	261.76

Source: Author's own.

are: $\log(p) = 1.533(0.2599); \log(b) = -0.455(0.2248)$. From these estimates the standard deviations of the conditional process can be estimated using (3.2). These estimated conditional standard deviations are shown in Figure 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 1, the martingale model with a Weibull and the birth–death model with an exponential distribution for the reproduction times (a constant birth rate) and a gamma distribution for the removal times are close. Figure 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—and also the bias corrected one—judges the fit over the whole observation range. There can be evidence in the data, however,

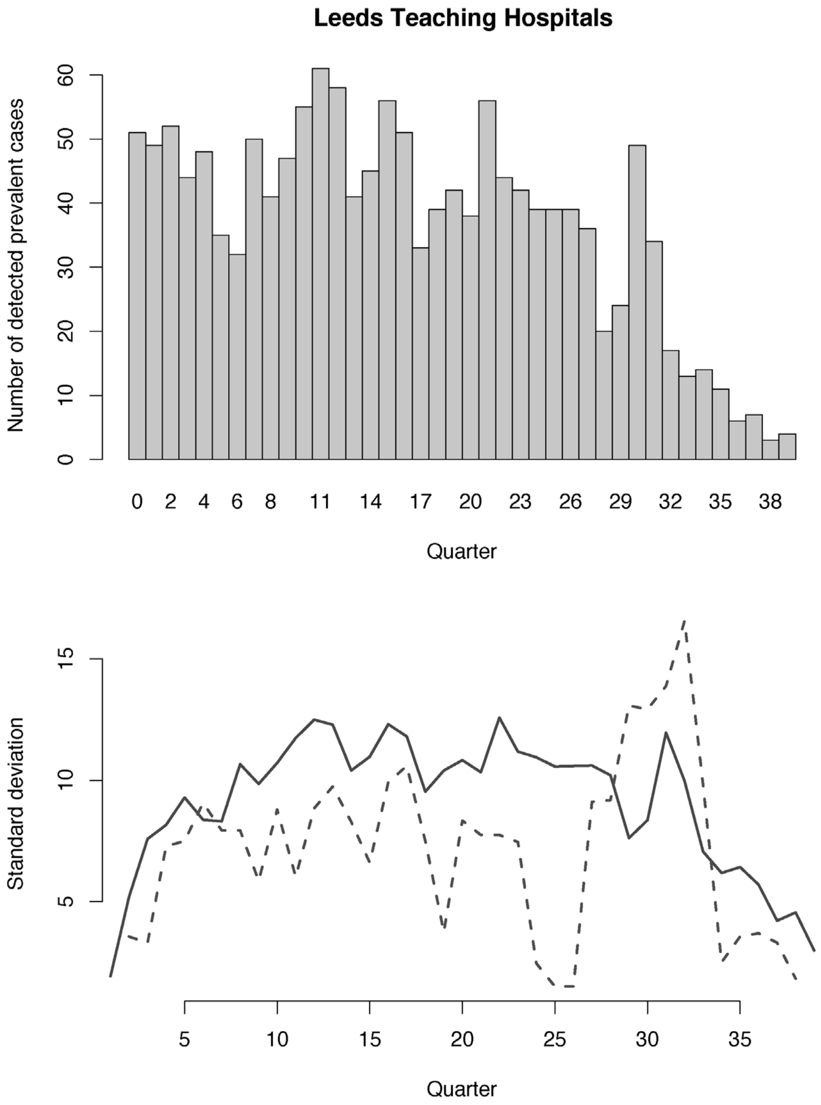

Figure 1 Leeds Teaching Hospital. Upper: the MRSA prevalence. Lower: the four-period moving standard deviation (dashed line) and the standard deviation from the conditional model (solid line).
Source: Author's own.

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 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

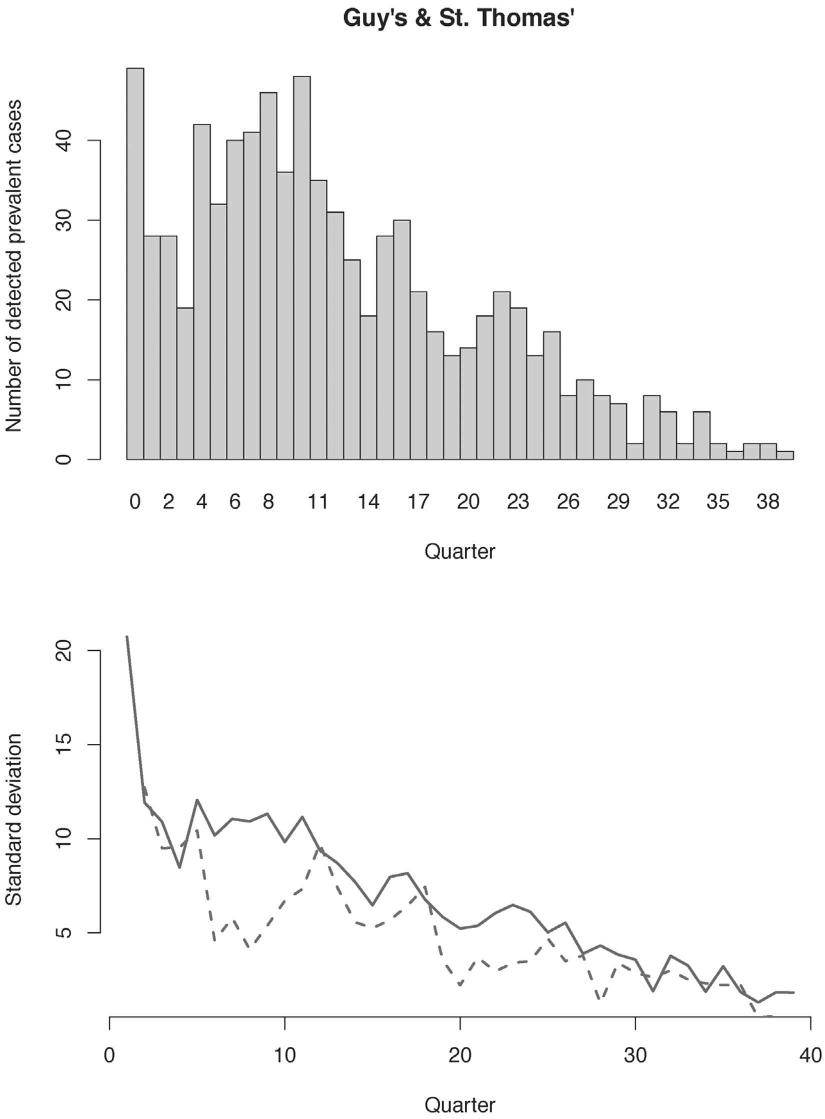

Figure 2 Guy's & St.Thomas'. Upper: the MRSA prevalence. Lower: the four-period moving standard deviation (dashed line) and the standard deviation from the conditional model (solid line).
Source: Author's own.

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

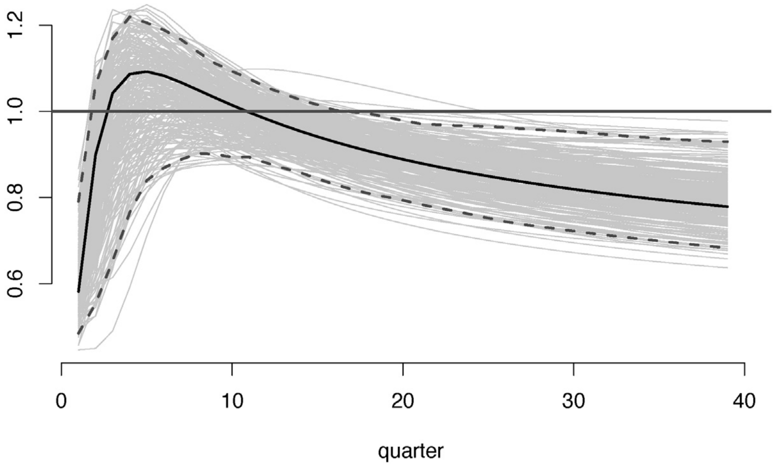

Figure 3 Guy's & St.Thomas'. Net reproduction ratio as a function of time (black solid line) and the net reproduction ratio calculated from sample paths drawn from the underlying estimated process (gray lines) with the 95% percentile lines (dashed lines).
Source: Author's own.

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 than 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. Five models are very close in fit as can be seen from Table 1—The martingale model with exponential distribution; the birth–death model with exponential survival functions; the birth–death model with exponential birth rate and generalized gamma or Weibull removal times and the birth–death model with exponential removal times and gamma reproduction times. Figure 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 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.

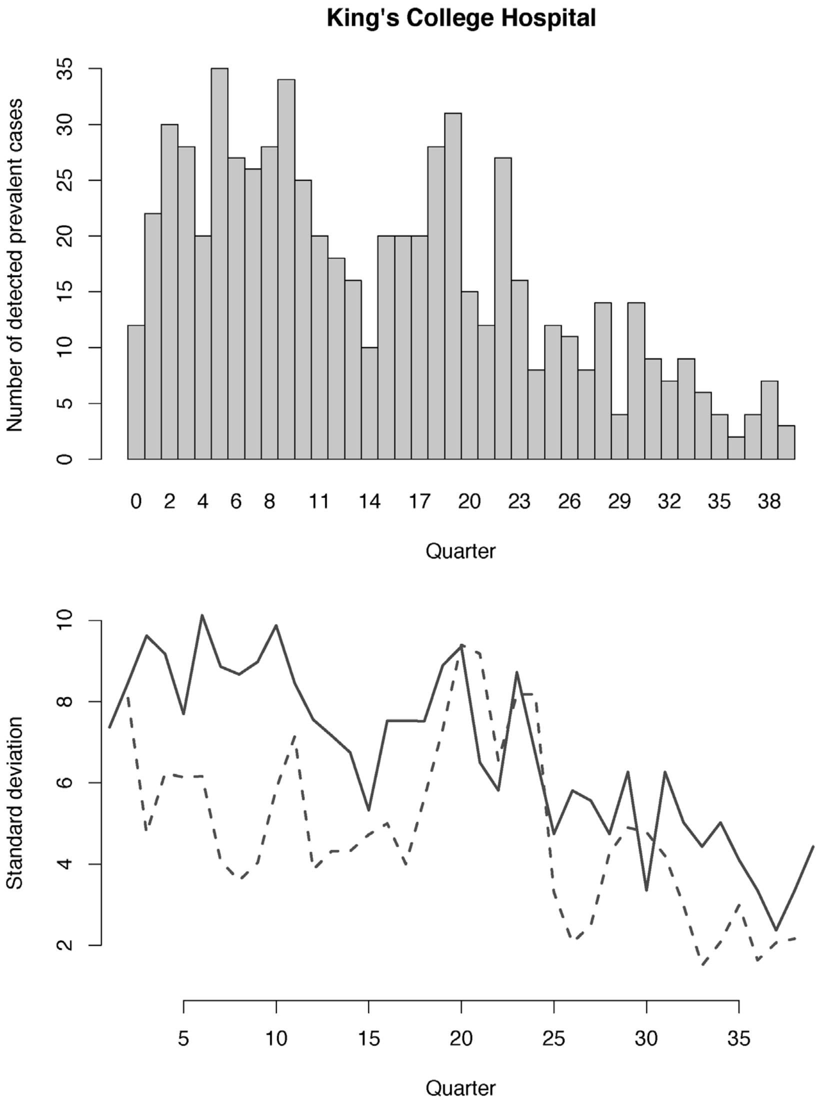

Figure 4 King's College Hospital. Upper: the MRSA prevalence. Lower: the four-period moving standard deviation (dashed line) and the standard deviation from the conditional model (solid line).
Source: Author's own.

6 Discussion

Modelling 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 where climate change is thought to cause an increase in volatility. Modelling volatility can be done separately

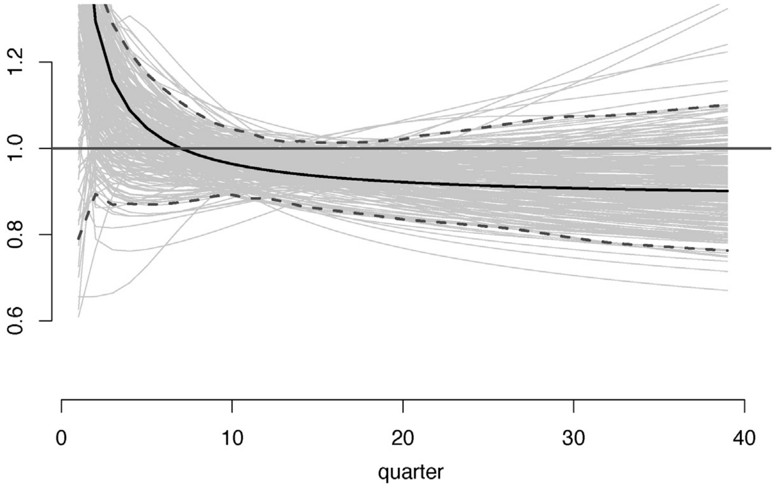

Figure 5 King's College Hospital. Net Reproduction ratio as a function of time (black solid line) and the Net reproduction ratio calculated from sample paths drawn from the underlying estimated process (gray lines) with the 95% percentile lines (dashed lines).
Source: Author's own.

from modelling 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 a 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 where 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 article 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 modelling a change in volatility while the mean stays the same, but can also 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 the bias corrected 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 reasonably 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.

Acknowledgements

I would like to thank Hiroshi Nishiura for his help on an early draft of the article.

References

Ahmed SA, Diffenbaugh NS and Hertel TW (2009) Climate volatility deepens poverty vulnerability in developing countries. *Environmental Research Letters*, **4**, 1–8.

Bailey NTJ (1990) *The elements of stochastic processes with applications to the natural sciences*, reprint edition. New York, USA: Wiley-interscience.

Becker NG 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**, 332–50.

Cooper BS, Medley GF, Stone SP, Kibbler CC, Cookson BD, Roberts JA, Duckworth G, Lai R and Ebrahim S (2004) Methicillin-resistant Staphylococcus aureus in hospitals and the community: stealth dynamics and control catastrophes. *Proceedings of the National Academy of Sciences*, **101**, 10223–28.

Diekmann O and Heesterbeek JAP (2000) *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. New York, USA: John Wiley & Sons.

Fleming TR and Harrington DP (1991) *Counting processes & survival analysis*. New York, USA: John Wiley & Sons.

Grundmann H, Hori S, Winter B, Tami A and Austin DJ (2002) Risk factors for the transmission of methicillin-resistant staphylococcus aureus in an adult intensive care unit: fitting a model to the data. *The Journal of Infectious Diseases*, **185**, 481–88.

Grundmann H and Hellriegel B (2006) Mathematical modelling: a tool for hospital infection control. *The Lancet Infectious Diseases*, **6**, 39–45.

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

Kleiber C and Kotz S (2003) *Statistical size distributions in economics and actuarial sciences.* Hoboken, New Jersey, USA: John Wiley.

Lewis AL (2000) *Option valuation under stochastic volatility: with matematica code.* Newport Beach, California, USA: Finance Press.

Lindsey JK (2004) *Statistical analysis of stochastic processes in time.* Cambridge, UK: Cambridge University Press.

Nishiura H and Chowell G (2009) The net reproduction ratio as a prelude to statistical estimation of time-dependent epidemic trends. In *Chowell G, Hyman JM, Bettencourt LMA and Castillo-Chavez C, eds. Mathematical and statistical estimation approaches in epidemiology.* New York: Springer, 103–21.

Pelupessy I, Bonten MJM and Diekman O (2009) How to assess the relative importance of different colonization routes of pathogens within hospital settings. *PNAS*, **99**, 5601–05.

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.

Scheffer M, Bascompte J, Brock WA, Brovkin V, Carpenter RC, Dakos V, Held H, Van Nes EH, Rietkerk M and Sugihara G (2009) Early-warning signals for critical transitions. *Nature*, **461**, 53–9.

Spiegelhalter DJ (2005) Problems in assessing the rates of infection with methicillin resistant Staphylococcus aureus. *BMJ*, **331**, 1013–15.

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

Van Den Broek J and Heesterbeek JAP (2007) Nonhomogeneous birth and death models for epidemic outbreak data. *Biostatistics*, **8**, 453–67.

Van Den Broek J and Nishiura H (2009) Using epidemic prevalence data to jointly estimate reproduction and removal. *Annals of Applied Statistics*, 3(4), 1505–20.