Biostatistics (2006), **7**, 3, *pp.* 422–437
doi:10.1093/biostatistics/kxj016
Advance Access publication on January 11, 2006

A two-component model for counts of infectious diseases

LEONHARD HELD*, MATHIAS HOFMANN, MICHAEL HÖHLE

Department of Statistics, Ludwig-Maximilians-Universität München,
Ludwigstrasse 33, 80539 München, Germany
leonhard.held@stat.uni-muenchen.de

VOLKER SCHMID

Institute of Biomedical Engineering, Imperial College London, London, UK

SUMMARY

We propose a stochastic model for the analysis of time series of disease counts as collected in typical surveillance systems on notifiable infectious diseases. The model is based on a Poisson or negative binomial observation model with two components: a parameter-driven component relates the disease incidence to latent parameters describing endemic seasonal patterns, which are typical for infectious disease surveillance data. An observation-driven or epidemic component is modeled with an autoregression on the number of cases at the previous time points. The autoregressive parameter is allowed to change over time according to a Bayesian changepoint model with unknown number of changepoints. Parameter estimates are obtained through the Bayesian model averaging using Markov chain Monte Carlo techniques. We illustrate our approach through analysis of simulated data and real notification data obtained from the German infectious disease surveillance system, administered by the Robert Koch Institute in Berlin. Software to fit the proposed model can be obtained from http://www.statistik.lmu.de/∼mhofmann/twins.

Keywords: Bayesian changepoint model; Epidemic modeling; Reversible jump Markov chain Monte Carlo; Surveillance data.

1. INTRODUCTION

This paper develops a stochastic model for the statistical analysis of surveillance data of infectious disease counts. This is a challenging task as such data have specific features, such as ‘seasonality’ and occasional ‘outbreaks’, which have to be taken into account. Also, as low counts are typical, normal approximations are often inadequate. Furthermore, the model should allow for overdispersion. Finally, a realistic model should be non-stationary, as many time series for surveillance data have a non-stationary pattern, for example, caused by an increasing vaccination coverage or other interventions.

Statistical methods in infectious disease epidemiology have been dominated by individual-based detailed modeling of the epidemic process (e.g. Becker, 1989, or Daley and Gani, 1999). In particular, chain-binomial and related continuous-time models, such as the susceptible-infected-removed model, have been used to estimate relevant parameters from detailed data on the infection process (Anderson and Britton,

*To whom correspondence should be addressed.

2000). However, such 'mechanistic' modeling is too ambitious for routinely collected surveillance data. For example, the non-availability of information on susceptibles makes detailed modeling of the infection process infeasible. Other common problems of surveillance data are underreporting or reporting delays (Diggle *et al.*, 2003).

On the other hand, despite their limitations, surveillance data have features that cannot be captured with standard 'empirical' models, say log-linear Poisson regression models. Too simple a model will not be able to capture the characteristics typical for surveillance data, so a compromise is needed between mechanistic and empirical modeling. The model we describe is such a compromise. For further discussion on the distinction between empirical and mechanistic models, see for example Pawitan (2001, pp. 4–6).

Although the benefits of 'model-based' inference and prediction seem to be generally well accepted in numerous scientific disciplines, this does not yet seem to have found the same resonance in the context of surveillance data. In particular, in the context of 'outbreak detection', a different strategy is the current standard (Stroup *et al.*, 1989; Farrington *et al.*, 1996; Kleinman *et al.*, 2004). For example, Farrington *et al.* fit a Poisson regression model to the time series at hand under the assumption that there is 'no' outbreak in the historic records. An upper threshold limit for the predictive distribution at the next time point is computed and compared with the actually observed counts, say Z_n. Note that Z_n has not been used in the fitting process of the regression model. If Z_n is larger than the threshold, an outbreak is flagged. However, there are problems associated with this algorithm as past surveillance data will typically contain outbreaks. Farrington *et al.* (1996) propose a reestimation of the model based on weighted observations, where observations with high residuals from the initial model are downweighted. However, this procedure is ad hoc, for example, it is not clear why the particular choice of weights is useful or why the reestimation procedure is not repeated further. In a similar spirit, Kleinman *et al.* (2004) use a generalized linear mixed model with additional spatial random effects to provide predictions of the expected number of cases in the absence of an outbreak and then compare observed case counts with those expected values.

While this and similar outbreak detection algorithms can be useful in practice (see Farrington and Andrews, 2003, for a review), a potentially promising alternative is to fit a fairly realistic model to the data at hand, in particular to allow for outbreaks in the model, and to base outbreak detection on the posterior distribution of suitable model parameters or on the predictive distribution of Z_{n+1}, the disease counts at time $n + 1$. We believe that our model is a significant step toward such a model-based outbreak detection system. Note that in this approach, Z_n is used to fit the model to the data at hand, in contrast to the algorithm described above. Furthermore, in our approach all available historic information on the disease enters. The method of Farrington *et al.* (1996), like many other outbreak detection procedures (e.g. Stroup *et al.*, 1989), ignores a large percentage of the data in order to avoid to deal with seasonal effects. More specifically, data are only considered at reference values from previous years close to the current week of interest: if we are currently in calendar week 8 in the year 2005, say, and use a 9-week window, only data from calendar weeks 4, 5, . . . , 12 from the previous years 2004, 2003, . . . will enter as reference values.

A further situation where realistic models for surveillance data are needed is in the field of ecological regression, where covariate information is related to the disease incidence. The covariates may be simply counts of other diseases as in Hubert *et al.* (1992) and Jensen *et al.* (2004), who relate past influenza counts to meningococcal disease incidence. In these interesting articles, purely empirical methods such as autoregressive moving average models assuming normality or log-linear Poisson models are used for the meningococcal disease counts to infer the effect of past influenza counts. However, no allowance is made for outbreaks in the model, neither for influenza nor for meningococcal disease, so the results are based on an assumption which is unlikely to hold. A suitable multivariate version of the model proposed in this paper will allow for outbreaks and will therefore be more appropriate for relating influenza counts to meningococcal disease incidence.

Our starting point is a simple branching process model with autoregressive parameter λ and Poisson offspring, which is essentially the epidemic component of our model that allows for outbreaks in the data.

It can be viewed as an approximation to the so-called chain-binomial model, which is perhaps the best-studied stochastic model for infectious disease data in small populations. Note that we do not attempt to provide a mechanistic model as this would assume that the time unit in which the data are collected equals the generation time, i.e. the time between 'generations' of infectives (e.g. Daley and Gani, 1999, Chapter 4). This is rarely the case in practice and the autoregressive parameter λ therefore cannot be interpreted as the basic reproduction number R_0, the mean number of offspring, for which nice mathematical threshold theorems exist. However, a similar qualitative threshold feature is still available in our model (see below).

The model is extended in order to (a) allow for an influx of endemic cases, so that realizations from the model will not either explode or die out with probability one (depending on the actual value of λ), (b) include seasonal terms in the endemic rate, and (c) replace the Poisson with a negative binomial observation model in order to adjust for overdispersion. The additional influx of endemic cases implies that whenever $\lambda \geqslant 1$ an outbreak will occur, while for $\lambda < 1$ the process will be stable. Inclusion of overdispersion through latent random effects can be seen as an attempt to adjust for unobserved covariates or mechanisms that affect the disease incidence. For example, overdispersion can be caused by the fact that the generation time of many infectious diseases does not equal the time unit in which the data are collected or simply by the influence of unobserved covariates that affect the disease incidence.

Likelihood inference in this model with time-constant parameter λ has been described in Held *et al.* (2005). A central feature of the current paper is to let the autoregressive threshold parameter λ of the branching process model vary over time. Reasons for doing this are manifold, for example, the infectiousness might change through public health measures, such as increasing vaccination coverage, or through external factors that influence the spread of the infectious agent. Another scenario where λ will effectively be decreasing is when the number of susceptibles decreases.

While we would like to allow for a smooth change of λ over time, we also want to capture sudden changes in infectiousness. A state-space or dynamic model (e.g. Jørgensen *et al.*, 1999; Fahrmeir and Knorr-Held, 2000) with an autoregressive or random walk prior on λ is therefore not appropriate as it does not allow for such sudden changes. A Bayesian changepoint model (e.g. Denison *et al.*, 2002; Fearnhead, 2006) with unknown number of changepoints is better suited to this setting. The locations of the changepoints are also treated as unknown and the threshold parameter λ is assumed to be constant within any subsequent changepoints. Through Bayesian model averaging, the estimated time-changing λ may still be smooth because it is obtained through averaging over different changepoint models of variable dimension with different locations of the changepoints (Green, 1995; Clyde, 1999).

For illustration, we consider two time series obtained from the German infectious disease surveillance system, administered by the Robert Koch Institute in Berlin (Robert Koch Institute, 2005): weekly surveillance counts on Hepatitis A and Hepatitis B in Germany, 2001–2004. Both are liver diseases caused by viral infections. For Hepatitis A, infections can occur in situations ranging from isolated cases of disease to widespread epidemics. Hepatitis A is particularly common in tropical regions. The Hepatitis B virus can cause lifelong infection, cirrhosis (scarring) of the liver, liver cancer, liver failure, and death.

In Figure 1(a), we see that the Hepatitis A counts have a yearly seasonal pattern with occasional outbreaks. Our interest is if our model is able to identify outbreaks and separate them from the seasonal pattern. The time series for Hepatitis B, Figure 1(b), has a clearly non-stationary decreasing incidence trend, but no immediate signs of seasonality or occasional outbreaks. Vaccination against Hepatitis B has been recommended in Germany since 1995 for all newborns, infants, and particular risk groups. Vaccination coverage has increased since then and this is apparently reflected in the weekly counts of new infections. Here, we are interested in whether our model is able to capture the non-stationarity and if the increasing vaccination coverage is reflected in the estimates of the time-changing autoregressive parameter.

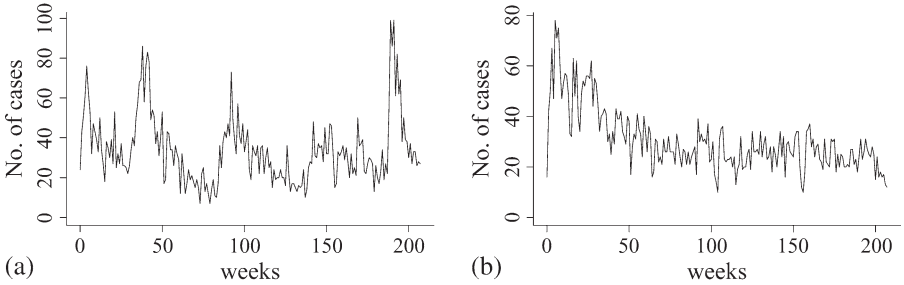

Fig. 1. Time series of weekly surveillance counts on Hepatitis A and Hepatitis B in Germany, 2001–2004.

2. Model

To begin, let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ denote the time series of, say, weekly counts of infectious disease cases. Our model is specified through the conditional distribution of $Z_t|Z_{t-1}$, so we condition on the observed count Z_0 at time $t = 0$. We now assume that Z_t follows a Poisson branching process with immigration and time-varying parameters λ_t and ν_t:

$$\begin{aligned} Z_t &= X_t + Y_t, \quad t = 1, \ldots, n \text{ with} \\ X_t &\sim \text{Po}(\nu_t), \text{ and} \\ Y_t|Z_{t-1} &\sim \text{Po}(\lambda_t Z_{t-1}). \end{aligned}$$

Here, the observed number of counts Z_t is decomposed into two (unknown) components: X_t and Y_t, which are assumed to be independent. Following Held *et al.* (2005), we call these two quantities the ‘endemic’ and ‘epidemic’ components, respectively. The distinction between endemic and epidemic incidence is quite common in dynamic models for infectious disease counts (e.g. Finkenstädt *et al.*, 2002; Knorr-Held and Richardson, 2003).

For further motivation, the introduction of the epidemic component can be seen as an attempt to allow for temporal dependence (beyond parametric seasonal patterns) and for occasional outbreaks in surveillance data. Indeed, the model for the endemic component alone is just a simple log-linear Poisson regression model, which could be fitted by the standard generalized linear model machinery. Farrington *et al.* (1996) use a similar endemic model to fit surveillance data under the assumption that no outbreak has occurred.

More technically, and in the spirit of Cox (1981), we could also call the two components the ‘parameter-driven’ and the ‘observation-driven’ model components. Note that we allow both model parameters ν_t and λ_t to vary over time. For constant ν and λ, Z_t is a simple branching process with immigration (e.g. Guttorp, 1995) with stationary mean $\nu/(1 - \lambda)$ for $\lambda < 1$. This result holds not only in the Poisson case but also for any other discrete distribution with non-negative support and finite expectation, for example, for the negative binomial distribution used in Section 2.5.

Knowledge of the stationary mean allows for a useful interpretation of λ. First note that the stationary endemic incidence is ν and the epidemic incidence therefore has stationary mean $\nu/(1 - \lambda) - \nu = (\lambda\nu)/(1 - \lambda)$. Hence, λ is simply the ratio of epidemic to total mean incidence. Pragmatically, we may use a similar interpretation for λ_t in the time-dependent case, as ν_t will cancel. Clearly, this interpretation holds only for $\lambda_t < 1$ since for $\lambda_t \geqslant 1$ the process is not stationary and will eventually explode. A useful quantitative measure for outbreak detection is therefore the posterior probability $P(\lambda_t \geqslant 1)$. For example, we may flag an alarm if this probability is above 1%.

We finally compare our model to the one proposed by Knorr-Held and Richardson (2003), who let $1 + Z_{t-1}$ enter multiplicatively as an explanatory variable in ν_t. The effect of the previous counts is modulated by latent 0–1 indicators, which are assumed to follow a two-stage hidden Markov model. One problem of this formulation is that there are essentially only two levels of incidence, an endemic and an epidemic one. In contrast, our model can have many levels of incidence as λ_t is time changing. Furthermore, the branching process model is a more natural approach for infectious disease data since the effect of previous counts enters additively on the Poisson intensity and not multiplicatively. (For further discussion, see Held *et al.*, 2005).

2.1 *The endemic component*

The endemic component of the process is driven by the parameter ν_t. Many data on infectious disease surveillance exhibit strong seasonality. We therefore model $\log \nu_t$ as the sum of L harmonic waves plus an intercept,

$$\log \nu_t = \gamma_0 + \sum_{l=1}^{L} (A_l \sin(\rho l t + \phi_l)), \tag{2.1}$$

where A_l is the amplitude of the corresponding sine curve, ϕ_l is the phase shift, and ρ is the base frequency. For weekly data, $\rho = 2\pi/52$ is the obvious choice. It is well known (e.g. Diggle, 1990) that (2.1) can be rewritten as

$$\log \nu_t = \gamma_0 + \sum_{l=1}^{L} (\gamma_{2l-1} \sin(\rho l t) + \gamma_{2l} \cos(\rho l t)), \tag{2.2}$$

so with $s_{t0} = 1$ and

$$s_{tj} = \begin{cases} \sin\left(\frac{\rho t (j+1)}{2}\right), & \text{for } j = 1, 3, \ldots, 2L-1, \\ \cos\left(\frac{\rho t j}{2}\right), & \text{for } j = 2, 4, \ldots, 2L, \end{cases}$$

(2.2) can be reduced to a simple linear regression of the form $\log \nu_t = \sum_{j=0}^{J} \gamma_j s_{tj}$, where $J = 2L$. For the two series considered in this paper, we only use $L = 1$ harmonic wave, since higher-order frequencies turned out to be insignificant in a likelihood analysis with constant λ (see Held *et al.*, 2005).

2.2 *The epidemic component*

The epidemic component of the process is driven by the parameter sequence $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, which is assumed to be a piecewise constant with unknown number of changepoints K and unknown location of the changepoints $\theta_1 < \cdots < \theta_K$. More specifically, we assume the following model:

$$\lambda_t = \begin{cases} \lambda^{(1)}, & \text{if } t = 1, 2, \ldots, \theta_1, \\ \lambda^{(k)}, & \text{if } t = \theta_{k-1} + 1, \ldots, \theta_k, \\ \lambda^{(K+1)}, & \text{if } t = \theta_K + 1, \ldots, n, \end{cases}$$

where $\theta_1 < \theta_2 < \cdots < \theta_K$ are the K unknown changepoints, so $\theta_k \in \{1, 2, \ldots, n-1\}$ for all $k \in \{1, 2, \ldots, K\}$. For $K = 0$, there is no changepoint and $\lambda_t = \lambda^{(1)}$ for all $t = 1, \ldots, n$.

2.3 *Prior assumptions*

The proposed model is particularly well suited for Bayesian inference via Markov chain Monte Carlo (MCMC). For this, we first need to specify prior distributions for the parameters in the endemic and epidemic components. For the regression coefficients, we set $\boldsymbol{\gamma} \sim N(0, \sigma_\gamma^2 \boldsymbol{I})$ with $\sigma_\gamma^2 = 10^6$, which corresponds to highly dispersed independent normal priors for each coefficient.

More interesting is the prior on the partition model. We have used the following settings: the number K of changepoints is assumed to be uniformly distributed among the possible values $\{0, 1, \ldots, n-1\}$, i.e. $P(K = k) = 1/n$, $k = 0, 1, \ldots, n-1$. For given $K > 0$, the location of the changepoints $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, where $\theta_1 < \theta_2 < \cdots < \theta_K$, is again uniformly distributed among all possible configurations,

$$P(\boldsymbol{\theta}|K = k) = \binom{n-1}{k}^{-1}.$$

Let A_i be the event that there is a changepoint at location i. For fixed $K = k$, we easily see that the prior probability that there is a changepoint at location i, conditional on $K = k$ changepoints, is $P(A_i|K = k) = k/(n-1)$. For the uniform prior on K above, the unconditional prior probability for a changepoint at any arbitrary location i is hence

$$P(A_i) = \sum_{k=0}^{n-1} \frac{k}{n-1} \frac{1}{n} = \frac{1}{2}.$$

Note that two events A_i and A_j, $i \neq j$, are dependent in this formulation as, for example, $P(A_i|A_j) = 2/3$ and $P(A_i|\bar{A}_j) = 1/3$, where $\bar{A}_j$ denotes the event that there is no changepoint at location $j \neq i$. Note also that this probability is not conditional on K; conditional on K, $P(A_i|A_j, K)$ will of course decrease compared to $P(A_i|K)$. More generally, it can be shown that the following result holds for the unconditional probability and all $i \neq j_1 \neq \cdots \neq j_m$:

$$P\left(A_i|A_{j_1}, \ldots, A_{j_l}, \bar{A}_{j_{l+1}}, \ldots, \bar{A}_{j_m}\right) = \frac{1+l}{2+m}. \tag{2.3}$$

Interestingly, this resembles 'Laplace's rule of succession' (see, for example Bernardo and Smith, 1994, p. 272). This result is important for computing the (posterior) predictive distribution for future counts Z_{n+1} (see Section 2.6).

Finally, for $\lambda^{(k)}$, $k = 1, \ldots, K+1$, we specify independent exponential distributions with mean $1/\xi$ and variance $1/\xi^2$. It is possible to perform robust analysis by placing a gamma hyperprior $\text{Ga}(\alpha_\xi, \beta_\xi)$ (with mean α_ξ/β_ξ) on the inverse mean ξ. This choice implies that the marginal prior distribution for $\lambda^{(k)}$ is gamma–gamma (see Bernardo and Smith, 1994, p. 120). In our applications, we use $\alpha_\xi = \beta_\xi = 10$ in which case the gamma–gamma marginal of $\lambda^{(k)}$ turns out to be simply an F-distribution with degrees of freedom equal to 2 and 20. This choice gives a marginal prior probability of 0.39 to the event $\lambda^{(k)} \geqslant 1$, while always favoring smaller values of $\lambda^{(k)}$, in the sense that the density function is monotonically decreasing. Of course, other choices could be made as well.

2.4 *Statistical analysis by MCMC*

The key to a successful application of MCMC methods to the specified model lies in the decomposition of Z_t into X_t and Y_t. While a likelihood-based approach (with time-constant λ) will use the probability mass function

$$P(\boldsymbol{Z}|Z_0) = \prod_{t=1}^{n} P(Z_t|Z_{t-1})$$

with

$$Z_t|Z_{t-1} \sim \text{Po}(\nu_t + \lambda Z_{t-1}),$$

and will hence ignore this decomposition (Held *et al.*, 2005), here we treat the variables X_t and Y_t as unknown auxiliary variables and update them explicitly in our MCMC algorithm. The benefit is that most parameter updates are now fairly simple. In particular, despite the apparent complexity of the changepoint model with unknown number of changepoints, if we condition on Y_t, updating the epidemic model parameters is straightforward, due to conjugacy and a specific marginalization trick. Conditional on X_t, updating the endemic model parameters is similar to MCMC algorithms in generalized linear models (Gamerman, 1997).

To be more specific, for fixed ν_t and λ_t, the auxiliary variables X_t and Y_t are updated in a block because of the linear dependence, given the observed data $Z_t = X_t + Y_t$. The conditional distribution of X_t and Y_t can be written as

$$P(X_t, Y_t|Z_t, \nu_t, \lambda_t) = P(Y_t|X_t, Z_t, \nu_t, \lambda_t)P(X_t|Z_t, \nu_t, \lambda_t),$$

where the first term $Y_t|X_t, Z_t, \nu_t, \lambda_t = Y_t|X_t, Z_t$ is deterministic: $Y_t = Z_t - X_t$. Due to the Poisson assumption for X_t and Y_t, the full conditional of $X_t, t = 1, \ldots, n$ is binomial:

$$X_t|Z_t, \nu_t, \lambda_t \sim \text{Bin}\left(Z_t, \frac{\nu_t}{\nu_t + \lambda_t Z_{t-1}}\right).$$

Update of the parameter vector $\boldsymbol{\gamma}$, which determines $\boldsymbol{\nu}$, is more involved. However, through conditioning on the auxiliary variables, the problem is equivalent to parameter estimation in a Bayesian log-linear Poisson regression model with response variable X_t. Here we use a Taylor approximation of second order to approximate the corresponding full conditional and to construct a suitable multivariate normal Metropolis–Hastings proposal (see, for example Rue and Held, 2005, Section 4.4). The algorithm works well across the wide range of datasets we studied, with acceptance rates typically between 80 and 85%.

Turning to the parameters in the endemic component, the key to a successful update lies again in conditioning on the auxiliary variables. Because the dimension of this model part is unknown, we employ reversible jump methodology (Green, 1995) for inference. In each step of our algorithm, we propose either to delete or to add a changepoint. It turns out to be advantageous to marginalize this step over $\boldsymbol{\lambda}$.

The exact algorithm we use proceeds in the following manner (Denison *et al.*, 2002). At each sweep of the algorithm, with probability $1/2$, we propose either to add or to delete a changepoint, with obvious modifications in the endpoint cases $K = 0$ and $K = n - 1$. If we add a new changepoint, the location of the changepoint is chosen uniformly among all possible locations, i.e. all locations where there is currently no changepoint. If we delete one, the proposed changepoint to be deleted is chosen uniformly among all current changepoints. At each step the acceptance probability is derived from the Metropolis–Hastings–Green algorithm (Green, 1995).

Let K^* be the proposed new number of changepoints, i.e. K^* is the current number of changepoints K plus or minus one. Consider first the case where a changepoint is proposed to be added, i.e. $K^* = K + 1$. Define $\boldsymbol{\theta}^*$ as the proposed new vector of ordered changepoints, with m the index of the proposed new changepoint θ_m and all other changepoints kept the same. The log-acceptance probability turns out to be

$$\begin{aligned}\log(a) = \min\Big(0, &\log(c) + \alpha_\lambda \log(\beta_\lambda) - \log\Gamma(\alpha_\lambda)\\ &+ \log\Gamma\left(\alpha_\lambda + Y_{[\theta_{m-1},\theta_m)}(t)\right) - \left(\alpha_\lambda + Y_{[\theta_{m-1},\theta_m)}(t)\right)\log\left(\beta_\lambda + Z_{[\theta_{m-1},\theta_m)}(t-1)\right)\\ &+ \log\Gamma\left(\alpha_\lambda + Y_{[\theta_m,\theta_{m+1})}(t)\right) - \left(\alpha_\lambda + Y_{[\theta_m,\theta_{m+1})}(t)\right)\log\left(\beta_\lambda + Z_{[\theta_m,\theta_{m+1})}(t-1)\right)\\ &- \log\Gamma\left(\alpha_\lambda + Y_{[\theta_{m-1},\theta_{m+1})}(t)\right) + \left(\alpha_\lambda + Y_{[\theta_{m-1},\theta_{m+1})}(t)\right)\log\left(\beta_\lambda + Z_{[\theta_{m-1},\theta_{m+1})}(t-1)\right)\Big),\end{aligned}$$

where $Y_{[a,b)}(t) = \sum_{a<t\leqslant b} Y_t$. In the case where m is the index of a changepoint θ_m proposed for removal, the log-acceptance rate is simply the negative of the above. Note that the acceptance probability

is essentially the ratio of the marginal likelihoods (see Denison *et al.*, 2002) of the proposed new changepoint model and the current one. The constant c is only relevant in the endpoint cases with

$$c = \begin{cases} 0.5, & \text{for } K = 0 \text{ or } K = n - 2, \\ 2, & \text{for } K^* = 0 \text{ or } K^* = n - 2, \\ 1, & \text{in all other cases.} \end{cases}$$

An alternative algorithm is the forward–backward method proposed in Fearnhead (2006) for direct simulation of the changepoints $\boldsymbol{\theta}$, given all other variables.

Given the changepoints $\boldsymbol{\theta}$, we can easily simulate from the full conditional distribution of $\boldsymbol{\lambda}$ via

$$\lambda^{(k)}|\cdots \sim \text{Ga}\left(1 + Y_{[\theta_{k-1},\theta_k)}(t), \xi + Z_{[\theta_{k-1},\theta_k)}(t-1)\right),$$

$k = 1, \ldots, K+1$. Note that since we have marginalized over $\boldsymbol{\lambda}$ in the update of $\boldsymbol{\theta}$, it is important to update first $\boldsymbol{\theta}$ and then $\boldsymbol{\lambda}$ because we perform essentially a joint update of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ based on the factorization $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\cdots) = p(\boldsymbol{\theta}|\cdots) \times p(\boldsymbol{\lambda}|\boldsymbol{\theta}, \cdots)$.

Finally, the full conditional of the inverse mean ξ of $\lambda^{(k)}$ is $\text{Ga}(\alpha_\xi + K + 1, \beta_\xi + \sum_{k=1}^{K+1} \lambda^{(k)})$, from which it is easy to sample.

2.5 *Adjustments for overdispersion*

The Poisson assumption is unlikely to hold in many circumstances, and some method of handling extra-Poisson variation is required. To adjust for overdispersion, we introduce a further set of independent auxiliary variables $\omega_t \sim \text{Ga}(\psi, \psi)$, $t = 1, \ldots, n$, in the model:

$$X_t|\omega_t \sim \text{Po}(\omega_t \nu_t),$$
$$Y_t|Z_{t-1}, \omega_t \sim \text{Po}(\omega_t \lambda_t Z_{t-1})$$

so $Z_t|\omega_t \sim \text{Po}(\omega_t(\nu_t + \lambda_t Z_{t-1}))$. It can easily be shown that, integrating out ω_t, the distribution is now negative binomial, $Z_t|Z_{t-1} \sim \text{NegBin}\,(\nu_t + \lambda_t Z_{t-1}, \psi)$ where $\text{NegBin}(\mu, \psi)$ denotes the negative binomial distribution with expectation μ and dispersion parameter ψ. Thus, the conditional mean $E[Z_t|Z_{t-1}]$ is the same as in the Poisson case, but the variance is now

$$V[Z_t|Z_{t-1}] = E[Z_t|Z_{t-1}]\left(1 + \frac{E[Z_t|Z_{t-1}]}{\psi}\right),$$

hence larger. For $\psi \to \infty$, it can be seen that $V[Z_t|Z_{t-1}] \to E[Z_t|Z_{t-1}]$ and we get back to the Poisson case.

Algorithmically, the introduction of the mixing variables ω_t is simple to handle in all updating steps described in Section 2.4. The full conditional of the mixing parameters ω_t is again gamma: $\omega_t|\ldots \sim \text{Ga}(\psi + Z_t, \psi + \nu_t + \lambda_t Z_{t-1})$. Finally, the prior distributions for the parameter ψ is chosen as $\psi \sim \text{Ga}(\alpha_\psi, \beta_\psi)$. We will use $\alpha_\psi = 1$ and $\beta_\psi = 0.1$ so that both the prior mean and the prior standard deviation equal 10. Of course, other choices could be made as well. Since $\psi > 0$, we prefer to update $\tilde{\psi} = \log(\psi)$ with a simple Metropolis–Hastings Gaussian random walk proposal. The full conditional of ψ is

$$p(\psi|\cdots) \propto p(\psi) \prod_{t=1}^{n} p(\omega_t|\psi)$$

and the corresponding full conditional of $\tilde{\psi}$ can be obtained through a change of variable. The variance of the random walk proposal is tuned automatically within the algorithm in order to obtain a suitable acceptance rate between 30 and 50% (Gelman *et al.*, 1996).

2.6 *One-step ahead prediction*

Of particular interest in infectious disease surveillance are 'short-term' predictions, in particular one-step-ahead predictions. Our model is well suited to this setting since it is based on the entire available time series and does not assume that there are no outbreaks in the past. While outbreak detection could be based on the posterior probability $P(\lambda_n \geqslant 1)$, the predictive distribution of the number of new cases Z_{n+1} is perhaps of more direct public health importance.

We omit the technical details here, but note only that with obvious modifications, the model can be written down for data $Z_1, \ldots, Z_{n+1}$ where the count Z_{n+1} is missing. This allows us to simulate from the posterior predictive distribution of ν_{n+1} and λ_{n+1} and subsequently of $Z_{n+1}|Z_n \sim \text{Po}(\nu_{n+1} + \lambda_{n+1} Z_n)$ in the Poisson case. If we include overdispersion, samples from $\omega_{n+1} \sim \text{Ga}(\psi, \psi)$ based on the posterior samples of ψ are generated and subsequently $Z_{n+1}|Z_n \sim \text{Po}(\omega_{n+1}(\nu_{n+1} + \lambda_{n+1} Z_n))$ is simulated.

However, there is a simpler way to obtain the posterior predictive distribution of λ_{n+1} and Z_{n+1} based on a model for $Z_1, \ldots, Z_n$ only. Note that the predictive distribution of λ_{n+1} is a mixture of two components. Because of the independence of $\lambda^{(k)}$, $k = 1, 2, \ldots, K+2$, the first component, corresponding to the case that there is a changepoint between Z_n and Z_{n+1}, is the conditional prior distribution $\lambda^{(K+2)}|\xi \sim \text{Ga}(1, \xi)$, where ξ will be sampled from the posterior. The other component, which corresponds to the case of no changepoint between Z_n and Z_{n+1}, is the posterior of $\lambda^{(K+1)}$. The mixing weights are determined by the probability p, say, for a changepoint between Z_n and Z_{n+1}.

For fixed number of changepoints $K = k$ among $n - 1$ possible locations, the probability p is just $(K+1)/(n+1)$ [compare (2.3)]. In each iteration of the algorithm, we therefore simulate the posterior predictive distribution of λ_{n+1} with probability $(K+1)/(n+1)$ from the conditional prior distribution $\lambda^{(K+2)}|\xi \sim \text{Ga}(1, \xi)$, otherwise we set $\lambda_{n+1} = \lambda^{(K+1)}$. Note how nicely the posterior distribution of K determines the probability for a changepoint in the future, in the sense that the more changepoints there are in the past, the more likely is it that there will be a changepoint in the future.

We finally note that m-step predictions, if required, may be obtained by sequentially repeating this process, given the current number of breakpoints up to time $n + m - 1$. At this point it is worth noting that for long-term predictions, eventually only the posterior of the endemic part $\boldsymbol{\nu}$ will enter, while the epidemic part will reduce to the conditional prior distribution with large probability.

3. APPLICATION TO DATA

3.1 *Analysis of simulated data*

To study the flexibility of the changepoint model, we first present an analysis of simulated data ($n = 199$, $\rho = 2\pi/52$). The true $\boldsymbol{\lambda}$ sequence is piecewise constant with two changepoints at $\theta_1 = 39$ and $\theta_2 = 49$. The parameter λ switches from $\lambda^{(1)} = 0.7$ to $\lambda^{(2)} = 1.2$ and then back to $\lambda^{(3)} = 0.7$. The other parameters are chosen to reflect the behavior of a more common infectious disease with approximately 33 weekly cases, on average, in the stationary phase (i.e. for $\lambda = 0.7$): $\gamma_0 = \log(10) \approx 2.30$, $\gamma_1 = 0.5$, and $\gamma_2 = 1.5$. We do not allow for overdispersion ($\omega_t = 1$) and thus generate the data from a Poisson observation model. The data are analyzed with the proposed model and the results are shown in Figures 2 and 3.

The model is able to detect the changepoint structure very well. The posterior distribution of K is at the true value $K = 2$ with posterior probability around 0.66 [see Figure 3(a)] and the true locations of the changepoints are also well estimated [see Figure 3(b)] with a more precise estimation of the second changepoint at $t = 49$. This can be explained by the low disease incidence at the first changepoint, so the model has more information to precisely determine the location of the second. Consequently, the estimated $\boldsymbol{\lambda}$ sequence is smooth around the first changepoint, but abrupt at the second. Note also that the seasonal structure in the data has been estimated quite well [see Figure 2(c) and (d)].

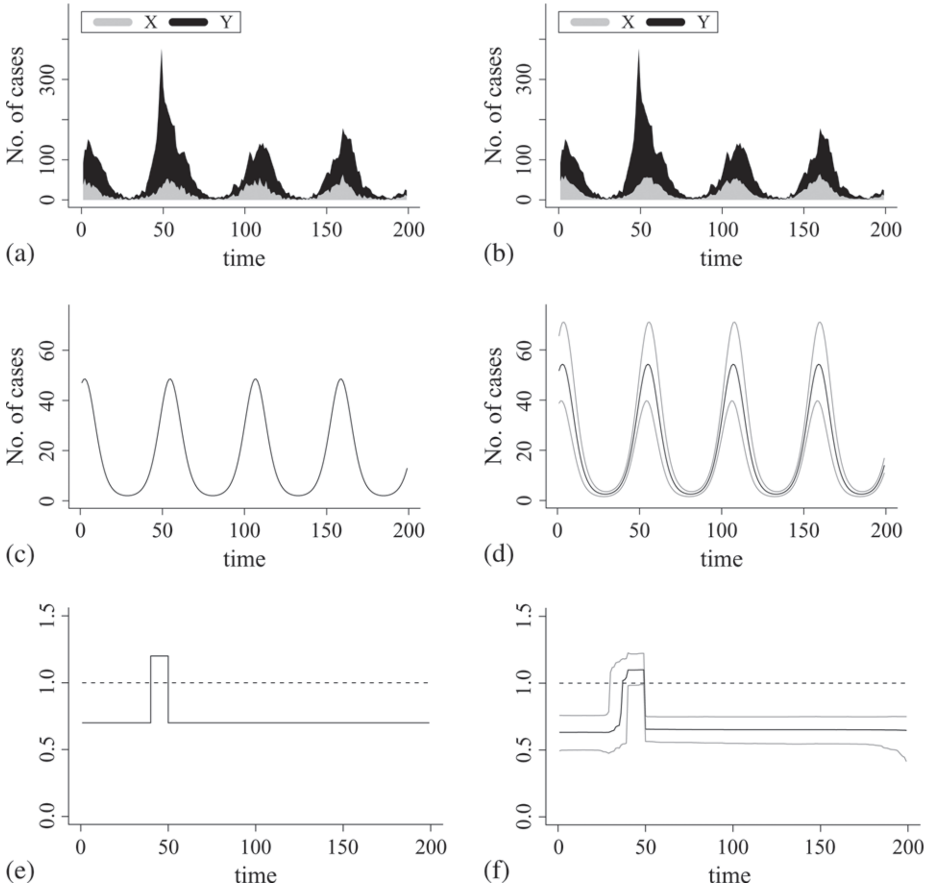

Fig. 2. Simulated data for known ν_t and λ_t (left panel) and posterior estimates (right panel).

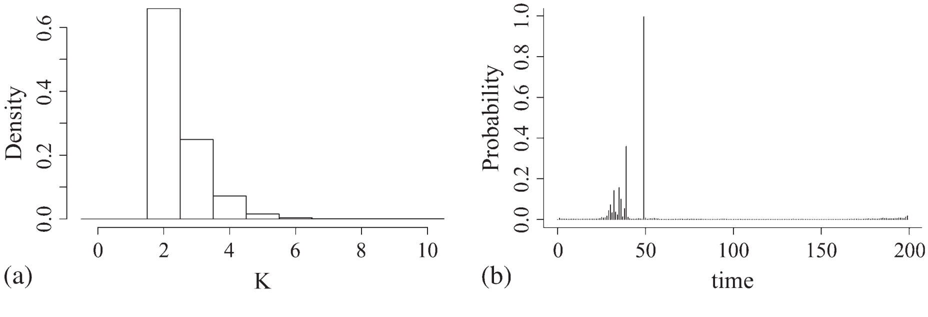

Fig. 3. Posterior distribution of K (left) and posterior probability for a changepoint at time t (right).

We have also simulated and analyzed data without any changepoint (i.e. $K = 0$) with $\lambda^{(1)} = 0.7$. Here the posterior probability of the true value $K = 0$ was 0.92. These simulation studies indicate that the model is able to detect a time-changing parameter λ_t very well and is able to separate it from the seasonal structure.

3.2 *Analysis of real data*

We analyze weekly surveillance data on Hepatitis A and B from Germany from the years 2001 to 2004 (208 weeks so $n = 207$) as introduced in Section 1.

3.2.1 *Hepatitis A.* Figure 4 displays the results from our model applied to the Hepatitis A time series. We find a strong seasonal pattern which peaks in December. Between 15 (in June) and 30 (in December) cases per week can be attributed to the regular endemic incidence pattern [see Figure 4(b)]. Retrospectively, there are two occasions where the model has detected unusual outbreaks, with $P(\lambda_t \geqslant 1|\mathbf{Z})$ clearly different from zero. A small outbreak has occurred in week $t = 169$ ($P(\lambda_t \geqslant 1|\mathbf{Z}) = 0.05$) and a second, more pronounced one in the two weeks $t = 188$ and $t = 189$ with $P(\lambda_t \geqslant 1|\mathbf{Z}) \approx 0.31$ in both weeks. This outbreak in the high holiday season (August) is discussed further in Anonymous (2004), and can be linked to holidaymakers in a certain hotel in Egypt. Outbreaks occurred also in other European countries.

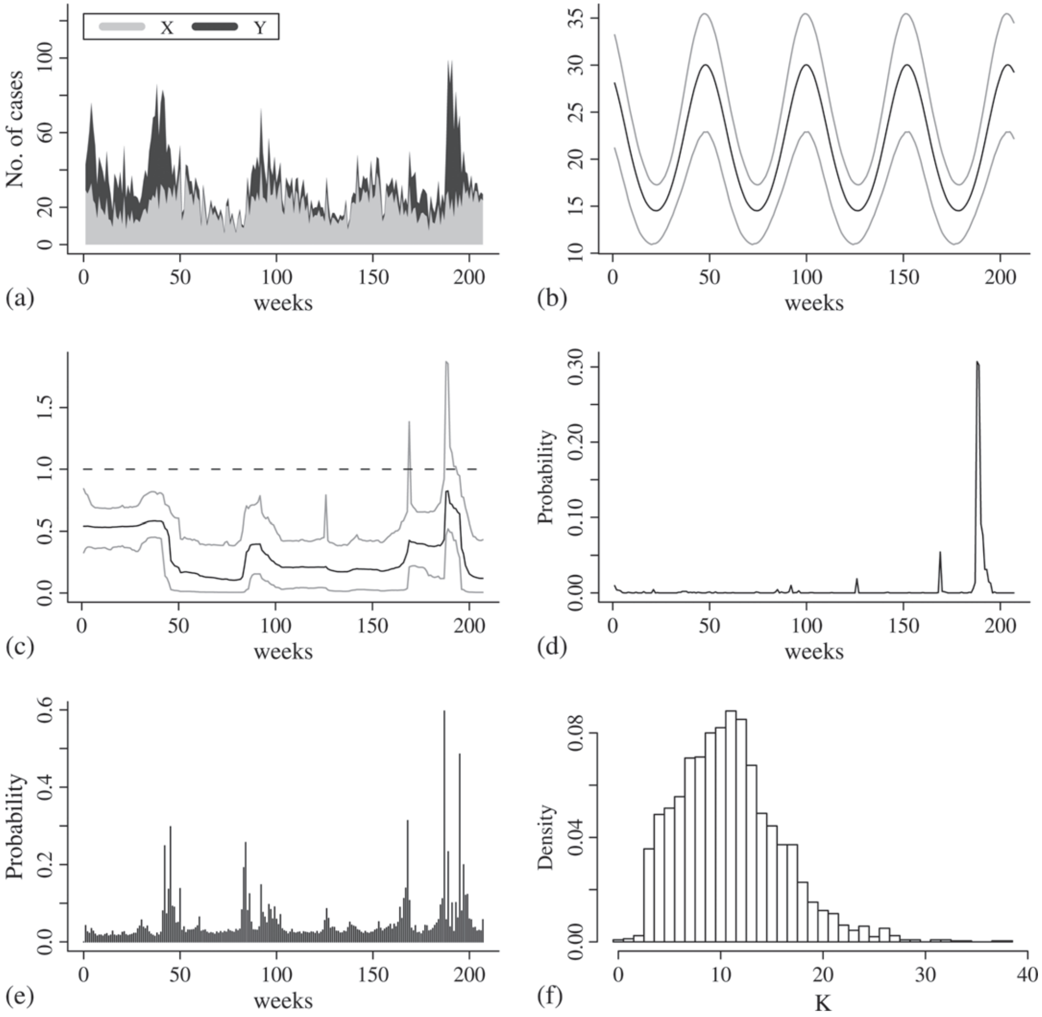

Fig. 4. Results for Hepatitis A.

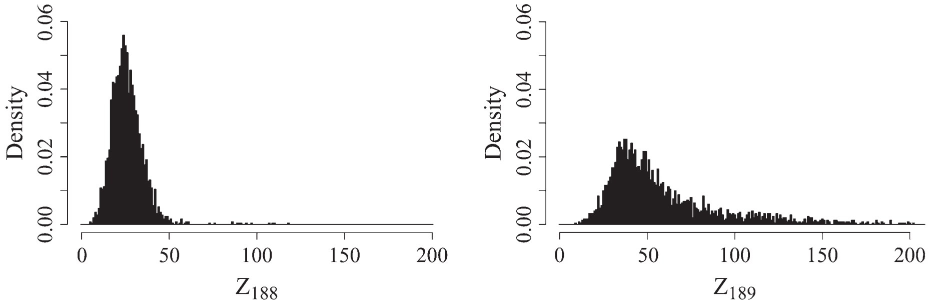

Fig. 5. Predictive distribution of $Z_{188}|Z_1, \ldots, Z_{187}$ (left) and $Z_{189}|Z_1, \ldots, Z_{188}$ (right).

To illustrate the predictive capabilities of the model, we show in Figure 5 the one-step-ahead predictive distribution of the number of cases for the weeks $t = 188$ and $t = 189$, based on the data up to week $t = 187$ and $t = 188$, respectively. These two weeks represent the beginning of the second outbreak: the observed counts are 22 in week 187, 54 in week 188, and 99 in week 189. It is interesting to see that the predictive distribution for week $t = 188$ is fairly symmetric, the observed value has some bline support with $P(Z_{188} \geqslant 54|Z_1, \ldots, Z_{187}) = 0.01$. However, the model immediately reacts to this unusually high value and the predictive distribution for $t = 189$ consequently has a long tail toward larger values. The observed number of cases is now well supported by the predictive distribution with $P(Z_{189} \geqslant 99|Z_1, \ldots, Z_{188}) = 0.13$.

For comparison, we have applied the Farrington *et al.* (1996) algorithm for outbreak detection in week $t = 188$. We used the implementation available in the R-package surveillance (Höhle and Riebler, 2005). An 11-week window has been chosen with 3 years of historical data (2001–2003). More data are not available as the German surveillance system had been set up in 2001. An outbreak is flagged if the actually observed number of counts is larger than an upper threshold, defined as the 99.9% quantile of the predictive distribution (based on a normal approximation of the transformed counts). The results depend on whether or not a linear time trend is included in the model. If included, the observed number of cases $Z_{188} = 54$ is flagged as an outbreak, as the upper threshold is 50.6. However, this is mainly due to the estimated (decreasing) time trend since the reference values in 2001 are unusually high, with up to 70 cases [see Figure 1(a)]. The algorithm does downweight these observations, but only to a certain degree, so the decreasing time trend remains significant. If we apply the algorithm without a time trend, the upper threshold is 77.2, so no alarm is flagged. Of course, these results also depend heavily on the nominal false-positive rate.

3.2.2 *Hepatitis B.* Figure 6 now displays the results from our model applied to the Hepatitis B time series. One can see that there is virtually no seasonality present, so the sinusoidal terms could as well have been omitted in the model. The autoregressive parameter λ_t decreases smoothly from values around 0.65 to values very close to 0.1, which can be interpreted as a consequence of the increasing vaccine coverage. The posterior mode of the number of changepoints is three; however, the possible locations of the changepoints are more dispersed than in the Hepatitis A example, so the estimated λ_t values are smoother than for Hepatitis A. This nicely illustrates the smoothing capabilities of the model through the Bayesian model averaging.

4. Discussion

In this paper, we have introduced a generic model for time series of infectious disease counts. The central assumption of the model is that the disease counts can be viewed as the sum of an endemic and an epidemic

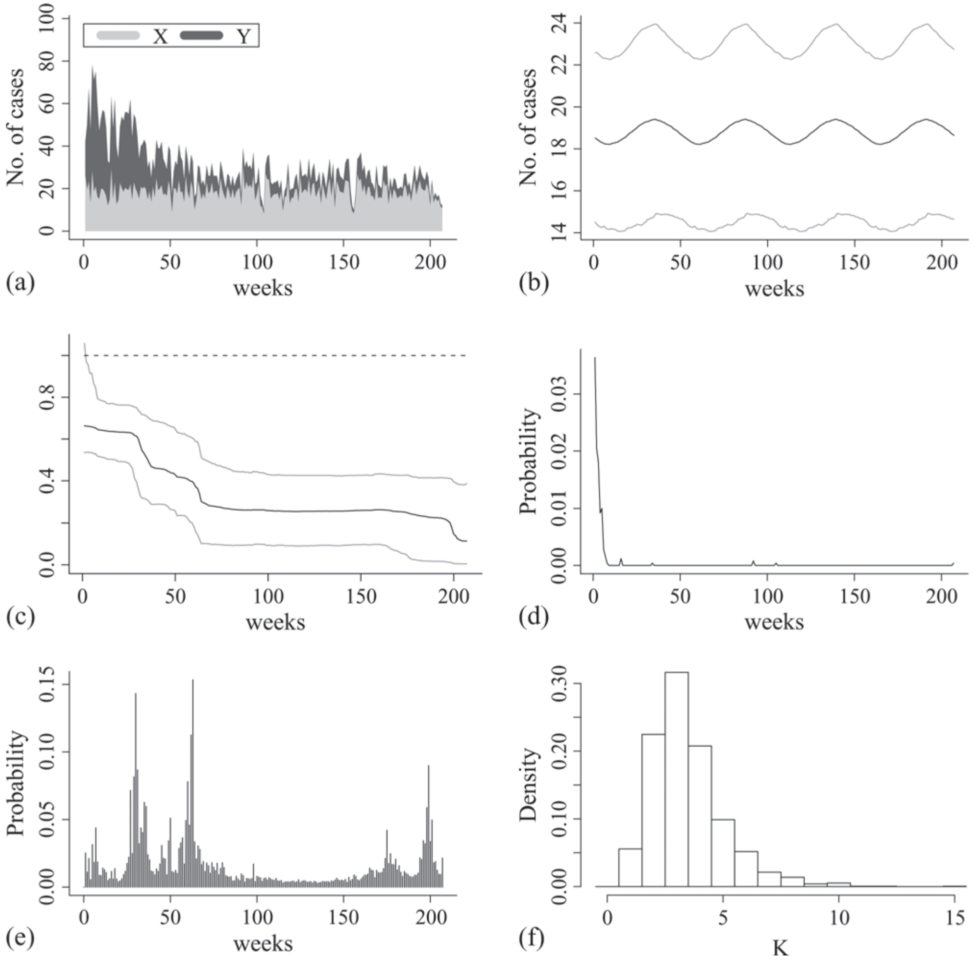

Fig. 6. Results for Hepatitis B.

component. The proportion of the epidemic component λ_t is allowed to vary over time according to a Bayesian multiple changepoint model. The inclusion of the epidemic component allows us to model time series of infectious disease counts with occasional outbreaks or other non-stationary features. Analyses of simulated and real data have illustrated how the model can be used for modeling diseases with increasing vaccination coverage (Hepatitis B) and for detecting and predicting outbreaks (Hepatitis A).

The motivation to include the autoregressive parameter in the model comes from a branching process formulation popular in infectious disease epidemiology. Nevertheless, the descriptive character of the model should be emphasized. In particular, the changepoint locations are purely random, which could be modified if a more detailed underlying structure is needed. For example, public health interventions to an outbreak of a certain disease might make it likely that a jump upward will be followed rapidly by an intervention, i.e. another changepoint but this time with a downward jump. This could be acknowledged in the model by a more sophisticated prior on the locations of the changepoints.

We now comment on some other extensions. For routine use in prospective disease surveillance, a sequential algorithm for inference will be helpful (Sonesson and Bock, 2003). Note in this context that the changepoint model used here has such a sequential representation based on (2.3), whereas our current implementation in twins is based on a retrospective analysis, given a fixed amount of data. Sequential updating of the parameter estimates could be based, for example, on particle filtering (Berzuini and Gilks, 2003) or the forward–backward algorithm (Fearnhead, 2006) and we currently consider such an algorithmic modification, which may require suitable approximations. For example, we might simply fix the estimates of the global model parameters $\boldsymbol{\nu}$ and update only $\boldsymbol{\lambda}$. A similar approach has been advocated in Brix and Diggle (2001) for spatiotemporal prediction (see also Diggle *et al.*, 2003). However, we note that all (retrospective) analyses in this paper take only little time compared to the weekly resolution on which surveillance data are typically collected. Nevertheless, a fast sequential algorithm will be useful for a detailed study of the predictive qualities of our model. For example, proper scoring rules, as discussed in Gneiting and Raftery (2004), or the probability integral transform (David, 1984; Gneiting *et al.*, 2005) could be used.

A multivariate or perhaps even spatial extension of our model is another area of potential value for applications. For example, in ecological regression, one might want to relate the endemic incidence ν or the epidemic parameter λ to area-level covariates. Also, the area of monitoring disease outcomes across multiple units is of great interest in practice (Marshall *et al.*, 2004; Kleinman *et al.*, 2004).

Acknowledgments

This work is supported by the German Science Foundation (DFG), Sonderforschungsbereich 386, Projekt B9: 'Statistical methodology for infectious disease surveillance'. We thank Klaus Stark from the Robert Koch Institute, Berlin, for epidemiological information on Hepatitis A and B and the editor, an anonymous referee, and Julian Besag, University of Washington, for helpful comments. *Conflict of Interest:* None declared.

References

Anderson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. Lecture Notes in Statistics, Volume 151. New York: Springer.

Anonymous (2004). Zu einer Häufung reiseassoziierter Hepatitis A unter Ägypten-Urlaubern. Epidemiologisches Bulletin Nr. 41, October 8, 2004, Robert Koch Institute, Berlin.

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

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Berzuini, C. and Gilks, W. R. (2003). Particle filtering methods for dynamic and static Bayesian problems. In Green, P. J., Hjort, N. L. and Richardson, S. (eds), *Highly Structured Stochastic Systems*. Oxford: Oxford University Press, pp. 207–227.

Brix, A. and Diggle, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society Series B* **63**, 823–841.

Clyde, M. (1999). Bayesian model averaging and model search strategies (with discussion). In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith A. F. M. (eds), *Bayesian Statistics 6*. Oxford: Clarendon Press.

Cox, D. (1981). Statistical analysis of time series. Some recent developments. *Scandinavian Journal of Statistics* **8**, 93–115.

Daley, D. J. and Gani, J. (1999). *Epidemic Modelling: An Introduction*. Cambridge: Cambridge University Press.

David, P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society Series A* **147**, 278–292.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Chichester: Wiley.

DIGGLE, P. J. (1990). *Time Series. A Biostatistical Introduction*. Oxford: Oxford University Press.

DIGGLE, P. J., KNORR-HELD, L., ROWLINGSON, B., SU, T.-L., HAWTIN, P. AND BRYANT, T. (2003). On-line monitoring of public health surveillance data. In Brookmeyer, R. and Stroup, D. F. (eds), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. Oxford: Oxford University Press, pp. 233–266.

FAHRMEIR, L. AND KNORR-HELD, L. (2000). Dynamic and semiparametric models. In Schimek, M. (ed.), *Smoothing and Regression: Approaches, Computation and Application*. New York: John Wiley & Sons.

FARRINGTON, C. P. AND ANDREWS, N. (2003). Outbreak detection: application to infectious disease surveillance. In Brookmeyer, R. and Stroup, D. F. (eds), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. Oxford: Oxford University Press, pp. 203–231.

FARRINGTON, C. P., ANDREWS, N., BEALE, A. D. AND CATCHPOLE, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society Series A* **159**, 547–563.

FEARNHEAD, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing* (in press).

FINKENSTÄDT, B. F., BJORNSTAD, O. N. AND GRENFELL, B. T. (2002). A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. *Biostatistics* **3**, 493–510.

GAMERMAN, D. (1997). Efficient sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* **7**, 57–68.

GELMAN, A., ROBERTS, G. O. AND GILKS, W. R. (1996). Efficient Metropolis jumping rules. In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. M. F. (eds), *Bayesian Statistics 5*. Oxford: Oxford University Press, pp. 599–607.

GNEITING, T., BALABDAOUI, F. AND RAFTERY, A. E. (2005). Probabilistic forecasts, calibration and sharpness. *Technical Report no. 483*. Department of Statistics, University of Washington.

GNEITING, T. AND RAFTERY, A. E. (2004). Strictly proper scoring rules, prediction, and estimation. *Technical Report no. 463*. Department of Statistics, University of Washington.

GREEN, P. J. (1995). Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

GUTTORP, P. (1995). *Stochastic Modelling of Scientific Data*. London: Chapman and Hall.

HELD, L., HÖHLE, M. AND HOFMANN, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance data. *Statistical Modelling* **5**, 187–199.

HÖHLE, M. AND RIEBLER, A. (2005). The R-Package ‘surveillance’. *Discussion Paper No. 422, SFB 386*. Department of Statistics, Ludwig-Maximilians-University Munich. Software available at http://www.statistik.lmu.de/∼hoehle/software/surveillance.

HUBERT, B., WATIER, L., GARNERIN, P. AND RICHARDSON, S. (1992). Meningococcal disease and influenza like syndrome: a new approach to an old question. *The Journal of Infectious Diseases* **166**, 542–545.

JENSEN, E. L., LUNDBYE-CHRISTENSEN, S., SAMUELSSON, S., SØRENSEN, H. T. AND SCHØNHEYDER, H. K. (2004). A 20-year ecological study of the temporal association between influenza and meningococcal disease. *European Journal of Epidemiology* **19**, 181–187.

JØRGENSEN, B., LUNDBYE-CHRISTENSEN, S., SONG, P. X.-K. AND SUN, L. (1999). A state space model for multivariate longitudinal count data. *Biometrika* **86**, 169–181.

KLEINMAN, K., LAZARUS, R. AND PLATT, R. (2004). A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology* **159**, 217–224.

KNORR-HELD, L. AND RICHARDSON, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Applied Statistics* **52**, 169–183.

MARSHALL, C., BEST, N., BOTTLE, A. AND AYLIN, P. (2004). Statistical issues in the prospective monitoring of health outcomes across multiple units. *Journal of the Royal Statistical Society Series A* **167**, 541–559.

PAWITAN, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

ROBERT KOCH INSTITUTE (2005). SurvStat@RKI. http://www3.rki.de/SurvStat. Accessed January 30, 2006.

RUE, H. AND HELD, L. (2005). *Gaussian Markov Random Fields. Theory and Applications*. Boca Raton, FL: CRC/Chapman and Hall.

SONESSON, C. AND BOCK, D. (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society Series A* **166**, 5–21.

STROUP, D. F., WILLIAMSON, G. D. AND HERNDON, J. L. (1989). Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in Medicine* **8**, 323–329.

[*Received March 29, 2005; first revision November 18, 2005; second revision December 12, 2005; accepted for publication January 10, 2006*]