Held, Hofmann, Höhle, Schmid:

A two-component model for counts of infectious diseases

Sonderforschungsbereich 386, Paper 424 (2005)

Online unter: http://epub.ub.uni-muenchen.de/

A two-component model for counts of infectious diseases

Leonhard Held*, Mathias Hofmann, Michael Höhle, Volker Schmid
Department of Statistics, Ludwig-Maximilians-Universität München,
Ludwigstr. 33, 80539 München, Germany

29th March 2005

Abstract

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. A 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 Bayesian model averaging using Markov chain Monte Carlo (MCMC) techniques. In analyses of simulated and real datasets we obtain promising results.

Keywords: Bayesian changepoint model; epidemic modelling; surveillance data; reversible jump Markov chain Monte Carlo

*Corresponding Author. E-mail: leonhard.held@stat.uni-muenchen.de

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. Furthermore, the model should appreciate that the responses are counts, as diseases with low counts are frequent and normal approximations are typically not useful. However, the model should also allow for overdispersion caused by e.g. unobserved covariates or mechanisms, that affect the disease incidence. 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 modelling of the epidemic process, e.g. Becker (1989). In particular, chain-binomial and related continuous-time models such as the SIR model have been used to estimate relevant parameters from detailed data on the infection process (Anderson and Britton, 2000). There are various reasons why such an approach is too ambitious and not applicable for routinely collected standard surveillance data: underreporting, reporting delay, non-availability of information on susceptibles, to mention just a few problems associated with *mechanistic* models for surveillance data.

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. In particular, a model too simple will not be able to capture the characteristics typical for surveillance data. A compromise is needed between *mechanistic* and *empirical* modelling. The model we describe is such a compromise. For further discussion on the distinction between *empirical* and *mechanistic* models see Pawitan (2001), Chapter 1.

Before proceeding, we give two longer examples where realistic statistical models of infectious disease surveillance data will be potentially useful. Although the benefits of *model-based* inference and prediction seems 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). For example, Farrington *et al.* fit a simple 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 y_n. Note that y_n has not been used in the fitting process of the regression model. If y_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 re-estimation of the model based on weighted observations, where observations with high residuals from the initial model are down-weighted. However, this procedure is ad-hoc, for example it is not clear why the particular choice of weights is useful or why the re-estimation procedure is not repeated further.

While this and similar outbreak detection algorithms can be useful in practice (see Farrington and Andrews, 2003, for a review) and the outbreak detection issue is not central to our paper, we want to emphasize that 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 y_{n+1}. We believe that our model is a significant step towards such a *model-based* outbreak detection system. Note that in this approach, y_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. Note that the Farrington *et al.* (1996), just 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 is 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 and use a nine week window, only data from calendar weeks $4, 5, \ldots, 12$ from the previous years 2004, 2003, and 2002, say, will enter as reference values.

A second 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 for example be simply counts of other diseases as in Hubert *et al.* (1992) and Jensen *et al.* (2004), who relate past influenza counts to meningococcal incidence. In these interesting articles descriptive methods such as ARMA models assuming normality or log-linear Poisson models are used for the meningococcal disease counts to infer the effect of the past influenza counts. However, no allowance is made for outbreaks in the model, neither for influenza nor for meningococcal disease. As a further example, we are currently studying the geographical variation of EHEC incidence in all districts of Germany in relationship to cow density. In general, such questions can be answered by suitable multivariate versions and modification of the model discussed in this paper, and we will give details on this in Section 4.

Our modelling strategy is as follows: We start with 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 seen as an approximation to the so-called chain-binomial model, which is perhaps the best studied stochastic models 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 index of the data denotes the *generation time* of the disease in question, not the *observation time*. This is rarely the case in practice and the autoregressive parameter λ can therefore not 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 in the sense that whenever $\lambda > 1$, an outbreak will occur while for $\lambda < 1$ the process will be stable.

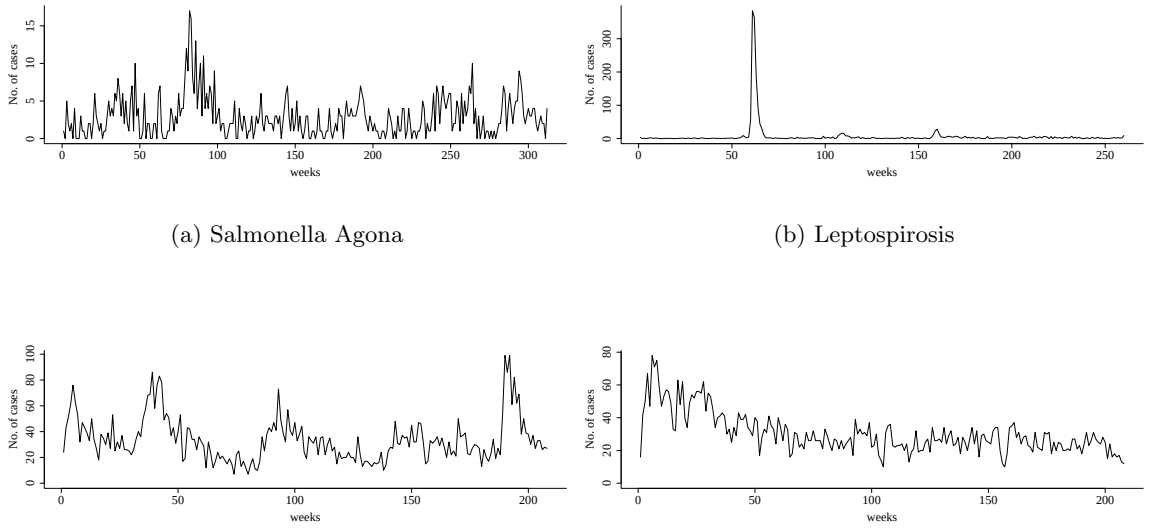
The model is extended in order to (a) allow for an influx of endemic cases, so that realisations from the model will not either explode or die out with probability one (depending on the actual value of λ), (b) include seasonal terms in the endemic rate and (c) switch from Poisson to a negative binomial observation model in order to adjust for overdispersion. Inclusion of overdispersion through latent random effects can be seen as an attempt to adjust for unobserved covariates or mechanisms, that do 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 *observation time* of surveillance data or simply by the influence of unobserved covariates that affect the disease incidence.

A central feature of our model is to let the autoregressive threshold parameter λ of the branching process model, to 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 due to 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 still want the model *also* to be able to capture sudden changes in infectiousness. A state-space or dynamic model (e.g. Jørgensen *et al.,* 1999, Fahrmeir & Knorr-Held, 2000, Rue and Held, 2005) 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, see also Fearnhead, 2004) 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 the following four time series, all collected on a weekly basis: (a) data on Salmonella Agona in the UK 1990-1995, (b) data on Leptospirosis counts in Rio de Janeiro, 1995 to 1999, and data on (c) Hepatitis A and (d) Hepatitis B in Germany, 2001-2004.

Figure 1 displays these four time series, which exhibit typical features of surveillance data: First, all of them, except perhaps Hepatitis B, show yearly seasonal patterns. Outbreaks can be seen for Salmonella Agona, Hepatitis A and, most obviously, for Leptospirosis. The time series for Hepatitis B shows a clearly non-stationary decreasing incidence, but no immediate signs of occasional outbreaks.



(a) Salmonella Agona

(b) Leptospirosis

(c) Hepatitis A

(d) Hepatitis B

Figure 1: The four time series on infectious disease counts that will be analysed in this paper.

2 Model

To begin, let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ denote the time series of, say, weekly counts of infectious diseases. Our model is specified through the conditional distribution of $Z_t|Z_{t-1}$, so we also need observed counts Z_0 at time $t = 0$ to condition on. We now assume that Z_t follows a generalized Poisson branching process with immigration,

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

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 those 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 GLM 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 simply ν and the epidemic incidence has therefore 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 $\lambda_t \geq 1$ the process is not stationary and will eventually explode. A useful quantitative measure for outbreak detection is therefore the posterior probability $P(\lambda_t \geq 1)$. For example, we may flag an alarm if this probability is above 1%, say.

We finally compare our model to the one proposed in Knorr-Held and Richardson (2003), who let the logarithm of $1 + y_{t-1}$ enter as an explanatory variable in the additive predictor. 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 *parameter-driven* or *endemic* component of the process is driven by the parameter ν_t. Most data on infectious disease surveillance data exhibit strong seasonality. We therefore model $\log \nu_t$ as the sum of L harmonic waves of different frequencies plus an intercept,

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

where A_l is the amplitude of the corresponding sine curve, ϕ_l 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 (1) can be rewritten as

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

so with $s_{t0} = 1$ and

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

equation (2) can be reduced to a simple linear regression form $\log \nu_t = \sum_{j=0}^{J} \gamma_j s_{tj}$, where $J = 2L$. For the four series considered in this paper we only use $L = 1$ harmonic waves, 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 *observation-driven* or *epidemic* component of the process is driven by the parameter sequence $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, which is assumed to be piecewise constant with unknown number of changepoints K and unknown location of the changepoints $\theta_1 < \ldots < \theta_K$. More specifically, we assume the following model:

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

where $\theta_k, k = 1, \ldots, K$ are the K unknown changepoints, so $\theta_k \in \{1, 2, \ldots, n-1\}$.

2.3 Prior assumptions

The proposed model is particularly well-suited for Bayesian inference. For this we first need to specify prior distributions for the parameters in the endemic and epidemic components. For the regression coefficients $\boldsymbol{\gamma}$ we set $\boldsymbol{\gamma} \sim \mathrm{N}(0, \sigma_\gamma^2 \boldsymbol{I})$ 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. $Pr(K = k) = 1/n$, $k = 0, 1, \ldots, n-1$. For given $K > 0$, the location of the changepoints $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, where $\theta_1 < \theta_2 < \ldots < \theta_K$, is again uniformly distributed among all possible configurations, i.e.

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

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 that there is a changepoint at location i, *conditional* on $K = k$ change points, 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

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

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* change point 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, the following result holds for the unconditional probability and all $i \neq j_1 \neq \ldots \neq j_m$:

$$P(A_i|A_{j_1}, \ldots, A_{j_l}, \bar{A}_{j_{l+1}}, \ldots, \bar{A}_{j_m}) = \frac{1+l}{2+m}, \tag{3}$$

see Held and Höhle (2005) for further details. 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 $1/\xi^2$, say. It is possible to perform robust analysis by placing a gamma hyperprior $G(\alpha_\xi, \beta_\xi)$ on the inverse mean ξ. This choice implies that the marginal prior distribution for $\lambda^{(k)}$ is gamma-gamma (see Bernardo and Smith, 1994, page 120). In our applications we use a standard exponential distribution, i.e. $\alpha_\xi = \beta_\xi = 1$ where the gamma-gamma

marginal of $\lambda^{(k)}$ turns out to be simply an F-distribution with both degrees of freedom equal to 2. This choice gives a marginal prior probability of exactly 0.5 to the event $\lambda^{(k)} \geq 1$, while always favouring smaller values of $\lambda^{(k)}$ (the density function has a unique mode at zero and is monotonously 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 (conditional) likelihood

$$p(Z_1, Z_2, \ldots, Z_n | Z_0) = \prod_{t=1}^{n} p(Z_t | Z_{t-1})$$

with

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

and will hence ignore this decomposition (see Held *et al.,* 2005, for details), 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

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

where the first term $Pr(Y_t | X_t, Z_t, \nu_t, \lambda_t) = Pr(Y_t | X_t, Z_t)$ 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. This algorithm has much in common with the "weighted least squares proposal" described in Gamerman (1997). The algorithm works fine across a 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, here we employ the reversible jump methodology (Green, 1995) for inference. In each step of our algorithm we propose to either delete or add a changepoint. It turns out to be advantageous to marginalize this step over $\boldsymbol{\lambda}$.

The exact algorithm we use proceeds as follows (Denison *et al.,* 2002). With probability 1/2 we either propose to add or 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. Each step is accepted with a certain probability, 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 where m, say, is the index of the proposed new changepoint θ_m with all other changepoints kept the same. The log-acceptance probability turns out to be

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

where $Y_{[a,b)}(t) = \sum_{a<t\leq b} Y_t$, for example. In the case where m is the index of a changepoint θ_m we want to remove, 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

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

An alternative to this algorithm is the *forward-backward* method proposed in Fearnhead (2004) for direct simulation of the changepoints $\boldsymbol{\theta}$ given the auxilliary variables.

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

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

$k = 1, \ldots, K+1$. 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 $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\ldots) = p(\boldsymbol{\theta}|\ldots) \times p(\boldsymbol{\lambda}|\boldsymbol{\theta}, \ldots)$.

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

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 will introduce a further set of independent auxiliary variables $\omega_t \sim \text{Ga}(\psi, \psi)$, $t = 1, \ldots, n$ in the model:

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

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, $Z_t|Z_{t-1} \sim \text{NegBin}\left(\nu_t + \lambda_t Z_{t-1}, \psi\right)$ 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

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

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: $\omega_t|\ldots \sim \text{Ga}\left(\psi + Z_t, \psi + \nu_t + \lambda_t Z_{t-1}\right)$. Finally, the prior distributions for the parameter ψ is chosen as $\psi \sim \text{Ga}(\alpha_\psi, \beta_\psi)$. Typically we will use $\alpha_\psi = 1$ and $\beta_\psi = 0.1$ so that the prior mean and 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

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

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 for this setting, since we fit our model to the entire available time series and do not attempt to fit a model, assuming there are no outbreaks. While outbreak detection could be based on the posterior probability $P(\lambda_n \geq 1)$, the predictive distribution of the number of new cases y_{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 counts Z_{n+1} are missing. This allows us to simulate from the posterior predictive distribution of ν_{n+1} and λ_{n+1} and subsequently of $Z_{n+1}|Z_n \sim Po(\nu_{n+1} + \lambda_{n+1}Z_n)$ 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 $Z_{n+1}|Z_n \sim Po(\omega_{n+1}(\nu_{n+1} + \lambda_{n+1}Z_n))$ 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. One component (which corresponds to the case that there is a changepoint between Z_n and Z_{n+1}) is, due to the independence of $\lambda^{(k)}$, $k = 1, 2, \ldots, K+2$, the conditional prior distribution $\lambda^{(K+2)}|\xi \sim G(1, \xi)$. Here the posterior samples of ξ enter. The other component, which corresponds to the case of no change point between Z_n and Z_{n+1} is the posterior of $\lambda^{(K+1)}$. The mixing weights are essentially determined by the probability p,

say, for a change point 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 equation (3)). In each iteration of the algorithm we hence simulate the posterior predictive distribution of λ_{n+1} with probability $(K+1)/(n+1)$ from the conditional prior distribution $\lambda^{(K+2)}|\xi \sim G(1,\xi)$, 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 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.

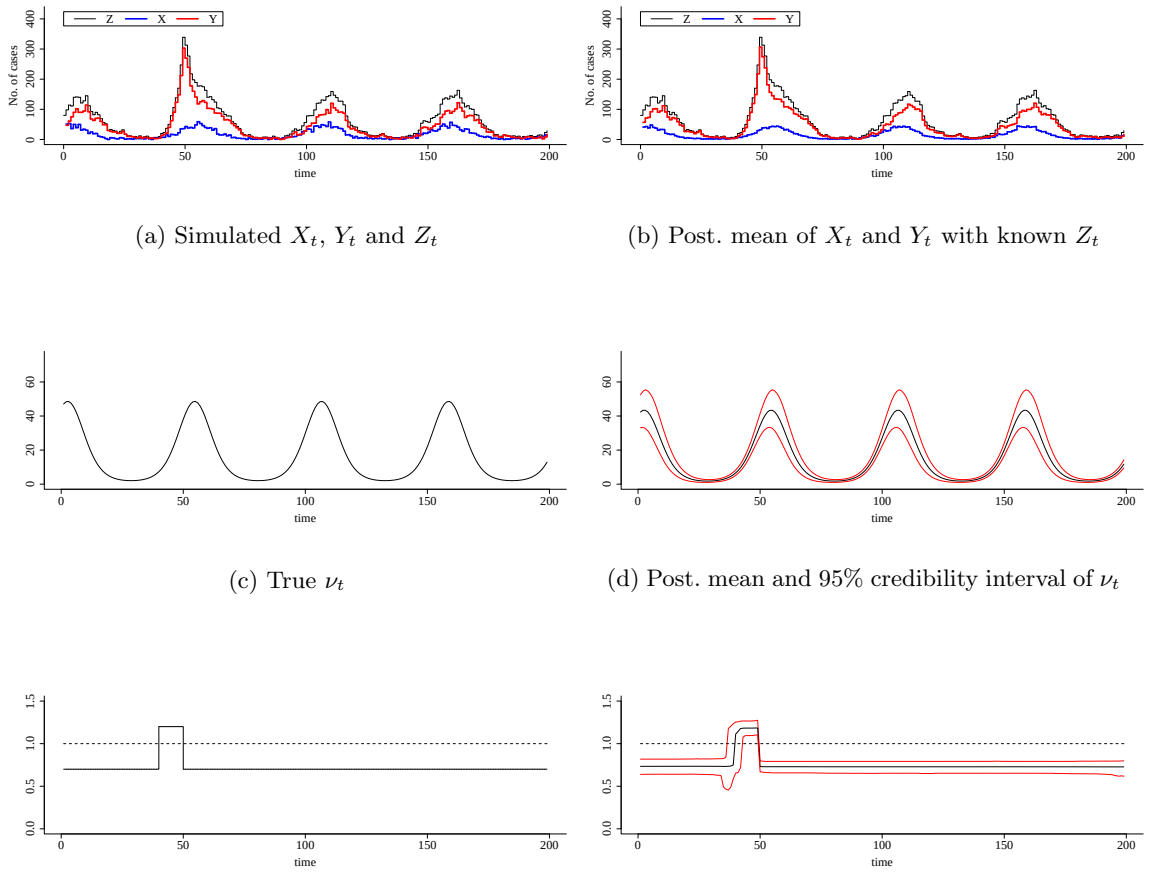
3 Application to data

In the following we present an analysis of a simple simulated series to investigate the performance of the model and analyse the four surveillance time series from Figure 1.

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 typical behaviour of a more common infectious disease: $\gamma_0 = \log(10) \approx 2.30$, $\gamma_1 = 0.5$ and $\gamma_2 = 1.5$. We do not allow for overdispersion ($\omega_t = 1$) so generate the data from a Poisson observation model. The data are analysed with the proposed model and the results are shown in Figure 2 and 3.

It can be seen that the model is able to detect the changepoint structure very well, the posterior model of K is at the true value $K = 2$ with posterior probability around 0.7 (Figure 3(a)), and the true locations of the changepoints are also well estimated (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 change point, so the model has more information to precisely determine the location of the second than of the first changepoint. 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 correctly (Figure 2(d)) .



(a) Simulated X_t, Y_t and Z_t

(b) Post. mean of X_t and Y_t with known Z_t

(c) True ν_t

(d) Post. mean and 95% credibility interval of ν_t

(e) True λ_t

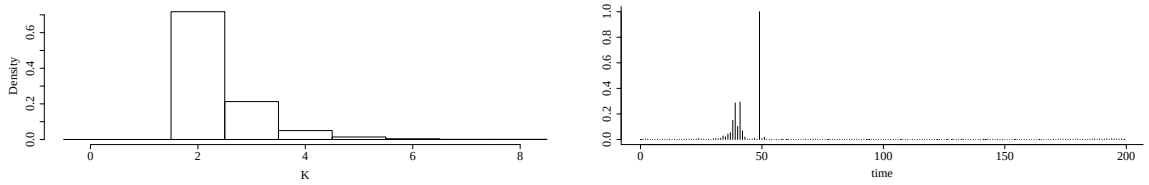
(f) Post. mean and 95% credibility interval of λ_t

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

We have also analysed similar data, but *without* any changepoint, i.e. $\lambda^{(1)} = 0.7$ and $K = 0$. Here the posterior probability of the true value $K = 0$ was larger than 0.9. 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.

3.1.1 Salmonella Agona

Salmonellosis is a bacterial gastrointestinal infection causing diarrhea, fever, or abdominal cramps. Figure 1(a) shows a time series of weekly counts for Salmonella Agona (a specific and uncommon serotype of Salmonella) in the UK, 1990-95, also reported also in Farrington



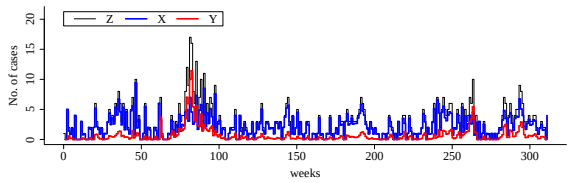
(a) Post. distribution of K

(b) Post. probability for a changepoint at time t

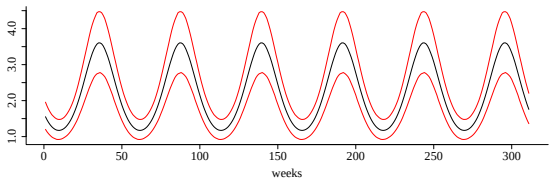
Figure 3: Post. distribution of K and posterior probability for a changepoint at time t

et al. (1996). Note that the time series shown in Farrington *et al.* (1996) is slightly different (and also slightly shorter), due to later modifications in the data-file.

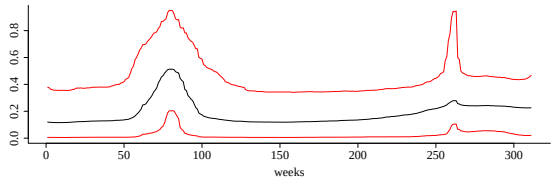
Figure 4 displays various features of the posterior distribution and associated parameters. Of particular interest is Figure 4(c), which displays the estimated time-changing sequence λ_t, $t = 1, \ldots, n$. An increase of λ_t can be seen around week 80, with large uncertainty when the increase actually started and when it ended. This can also be seen in Figure 4(e), which displays the unconditional probability for a changepoint at location t. Consequently, the estimated curve for $\boldsymbol{\lambda}$ is smooth. The posterior probabilities $\mathrm{P}(\lambda_t \geq 1|\boldsymbol{Z})$ increase here to values up to 0.02. A second shorter and less pronounced increase can be isolated around week 260 with the posterior probability $\mathrm{P}(\lambda_t \geq 1|\boldsymbol{Z})$ again increasing again to around 0.02. Interestingly, this second increase coincides with a reported outbreak of Salmonella Agona due to contaminated snacks (Anonymous, 1995). Apart from these two periods with increased infectiousness, λ is fairly constant with values between 0.1 and 0.3 and with $\mathrm{P}(\lambda_t \geq 1|\boldsymbol{Z})$ below 0.002.



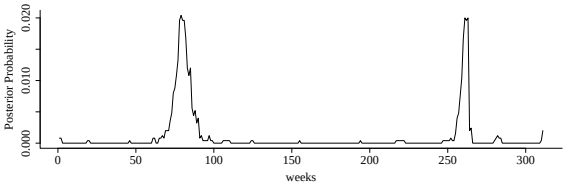
(a) Post. mean of X_t and Y_t with known Z_t



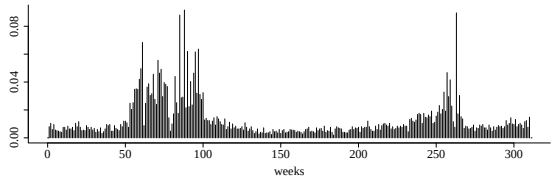
(b) Post. mean and 95% credibility interval of ν_t



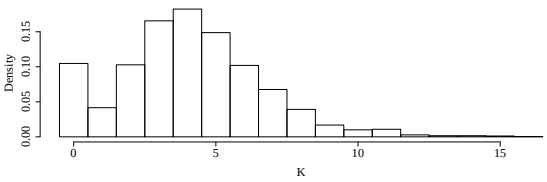
(c) Post. mean and 95% credibility interval of λ_t



(d) Post. probability $P(\lambda_t \geq 1|\boldsymbol{Z})$



(e) Posterior probability for a changepoint at location t



(f) Posterior distribution of K.

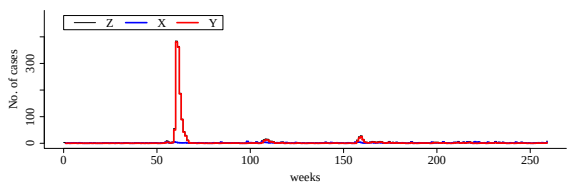
Figure 4: Results for Salmonella Agona

3.1.2 Leptospirosis

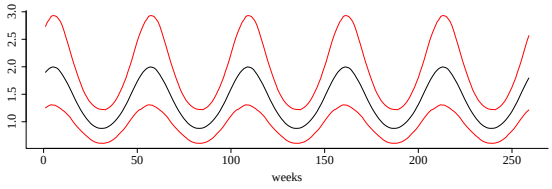
Leptospirosis is a bacterial disease that causes a wide range of symptoms such as high fever, severe headache, chills, muscle aches, and vomiting. Outbreaks of Leptospirosis are usually caused by exposure to water contaminated with the urine of infected animals. The data, shown in Figure 1(b), show one major outbreak that was caused by an inundation in combination with bad hygienic conditions.

Figure 5 now displays the results from our model. One can see a very volatile estimated $\boldsymbol{\lambda}$ sequence with a large number of changepoints. The obvious outbreak in the year 1996 is very well detected with λ_t well above 1. This outbreak is further discussed in Barcellos and

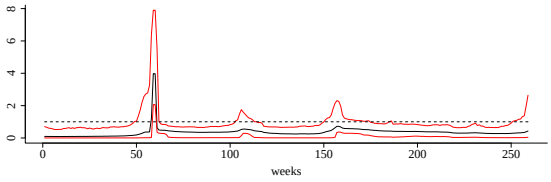
Sabraza (2000). At the end of the series, there is evidence that a new outbreak may occur with λ_t quickly rising. Indeed, the posterior mean of λ_t at the five last time points is 0.37 0.40, 0.42, 0.52 and 0.66 (median equal 0.32, 0.33, 0.34, 0.38 and 0.43) with corresponding posterior probabilities $P(\lambda_t \geq 1|\boldsymbol{Z})$ equal to 0.04, 0.06, 0.07, 0.12 and 0.19. The last five observations are 1, 3, 1, 2, and 9, so it is only the last observation that is suspiciously high. Of major interest is therefore the predictive distribution of Z_{n+1} which will be discussed separately in Section 4.



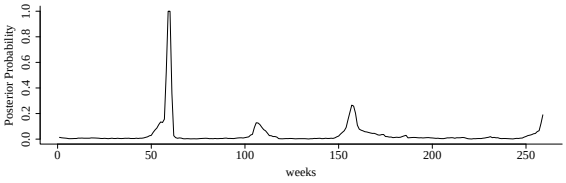
(a) Post. mean of X_t and Y_t with known Z_t



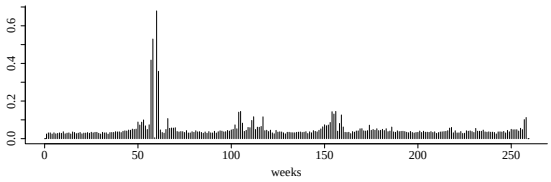
(b) Post. mean and 95% credibility interval of ν_t



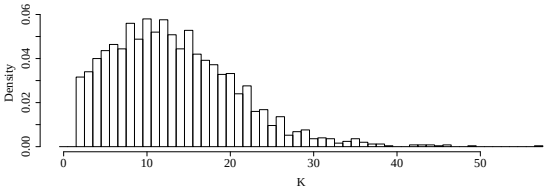
(c) Post. mean and 95% credibility interval of λ_t



(d) Post. probability $P(\lambda_t \geq 1|\boldsymbol{Z})$



(e) Posterior probability for a changepoint at location t



(f) Posterior distribution of K.

Figure 5: Results for Leptospirosis

3.1.3 Hepatitis A

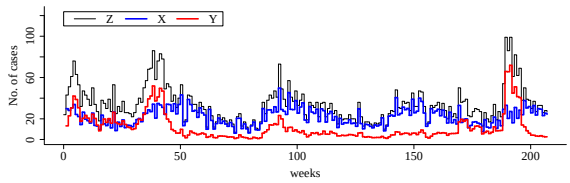
Hepatitis A is a liver disease caused by a virus. Hepatitis can occur in situations ranging from isolated cases of disease to widespread epidemics. Hepatitis A is particular common in tropical regions. We analyse weekly surveillance data on Hepatitis A from Germany from 2001 to 2004 (208 weeks).

Figure 6 displays the results from our model. One can see that there a strong seasonal pattern has been estimated 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 6(b). Retrospectively, there are two occasions where unusual outbreaks have been detected by our model, with $P(\lambda \geq 1|\boldsymbol{Z})$ clearly different from zero. A small one has occurred in week $t = 169$ ($P(\lambda_t \geq 1|\boldsymbol{Z}) = 0.02$) and a second more pronounced one in the two weeks $t = 188$ and $t = 189$ with $P(\lambda_t \geq 1|\boldsymbol{Z}) \approx 0.14$. This outbreak in the high holiday season (August) is discussed further in Anonymous (2004), and can be linked to holiday-makers in a certain hotel in Egypt. Outbreaks occurred also in other European countries.

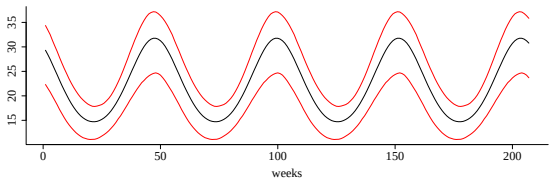
3.1.4 Hepatitis B

Hepatitis B is a serious disease caused by a virus that attacks the liver. The virus can cause life-long infection, cirrhosis (scarring) of the liver, liver cancer, liver failure, and death. Hepatitis B vaccine is available for all age groups to prevent hepatitis B virus infection. Vaccination against is Hepatitis B is recommended in Germany since 1995 for all newborns, infants and particular risk groups. Vaccination coverage has increased since then and it is therefore interesting to see if this is reflected in the weekly counts of new infections.

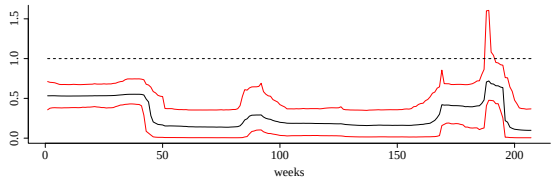
We analyse surveillance data on Hepatitis B from Germany from 2001 to 2004 (208 weeks). Figure 7 displays the results from our model. One can see that there is virtually no seasonality present, so the sinusoidal terms could have well been omitted in the model. The autoregressive parameter λ_t decreases from value around 0.65 to values well below 0.2 which can be interpreted as a consequence of an increasing vaccine coverage. There is no indication of an outbreak at the end of the series.



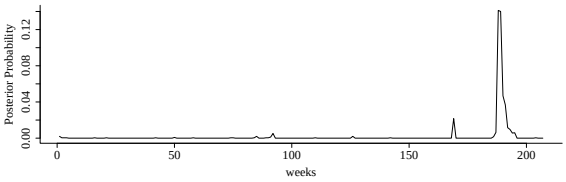
(a) Post. mean of X_t and Y_t with known Z_t



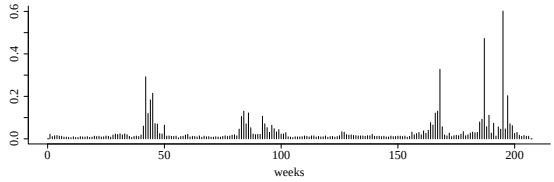
(b) Post. mean and 95% credibility interval of ν_t



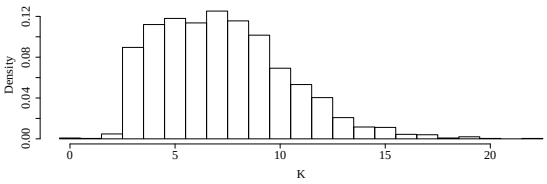
(c) Post. mean and 95% credibility interval of λ_t



(d) Post. probability $P(\lambda_t \geq 1|\boldsymbol{Z})$

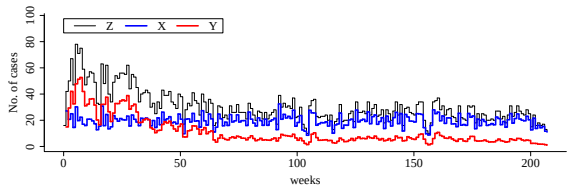


(e) Posterior probability for a changepoint at location t

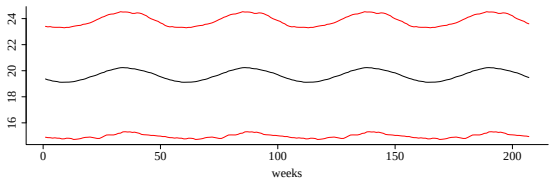


(f) Posterior distribution of K.

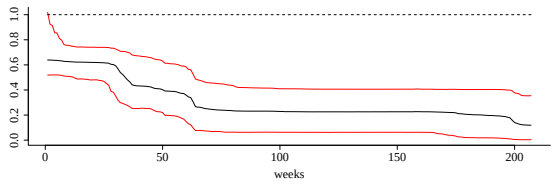
Figure 6: Results for Hepatitis A



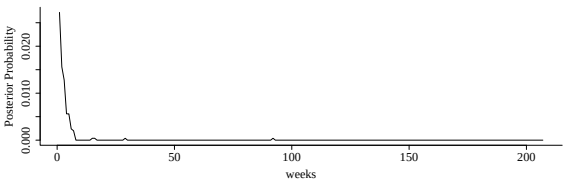
(a) Post. mean of X_t and Y_t with known Z_t



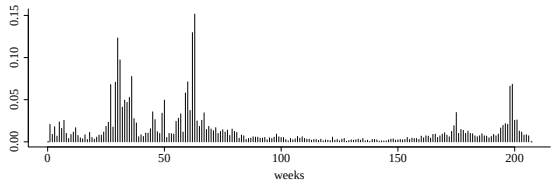
(b) Post. mean and 95% credibility interval of ν_t



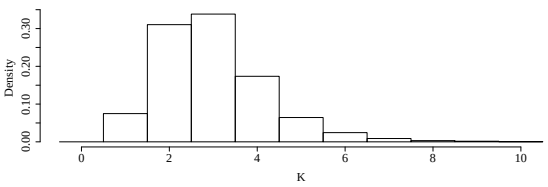
(c) Post. mean and 95% credibility interval of λ_t



(d) Post. probability $P(\lambda_t \geq 1|\boldsymbol{Z})$



(e) Posterior probability for a changepoint at location t



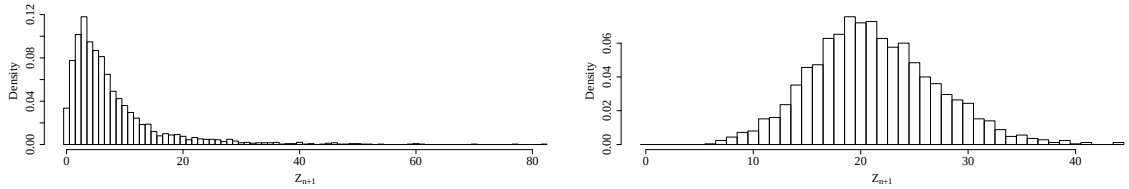
(f) Posterior distribution of K.

Figure 7: Results for Hepatitis B

4 Discussion

In this paper we have introduced a flexible model for time series of infectious disease counts. Analyses of simulated and real data have shown promising results. To illustrate the predictive facilities of the model we compare in Figure 8 the one-step-ahead predictive number of counts for Leptospirosis and Hepatitis B. For Leptospirosis, the predictive distribution for Z_{n+1} has a long tail with an upper 97.5% quantile of 30 cases, despite the fact that that the last observations are all below 10 $(\ldots, 1, 3, 1, 2, 9)$. The predictive distribution is extremely overdispersed, with an empirical variance nearly 10 times larger than the empirical mean. For comparison, Figure 8(b)

displays the corresponding predictive distribution for the Hepatitis B time series. Here there is no evidence for any local epidemic and the predictive distribution is only mildly overdispersed (variance to mean ratio of 1.6) with no tail at the upper end.



(a) Leptospirosis

(b) Hepatitis B

Figure 8: Posterior predictive distribution

We now comment on some further extensions. For routine use in prospective disease surveillance, a *sequential* algorithm for inference will be helpful (Sonesson and Bock, 2003). It is interesting to note in this context that the changepoint model used here has indeed such a sequential representation (Held and Höhle, 2005) whereas our current implementation 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, 2004) and we aim to develop such an algorithmic modification. However, this may require appropriate approximation techniques. 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 would like to note that all (retrospective) analyses in this paper take only little time compared to the weekly resolution in 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.

A multivariate or perhaps even spatial extension of our model is the other area with a lot of potential in applications. For example in ecological regression one might be interested to relate the disease incidence or infectiveness to area-level covariates. Also the area of monitoring disease outcomes across multiple units is of great interest in practice (Marshall *et al.,* 2004).

Acknowledgement

This work is supported by the German Science Foundation (DFG), SFB 386, Projekt B9: "Statistical methodology for infectious disease surveillance". We thank the Center for Disease Control, London, the National School of Public Health/FIOCRUZ, Rio de Janeiro, and the Robert-Koch Institute (RKI), Berlin for providing the data on Salmonella Agona, Leptospirosis, and Hepatitis A and B respectively.

References

Anderson, H. and Britton, T. (2000). *Stochastic Epidemic Models and their Statistical Analysis*, New York: Springer.

Anonymous (1995). An outbreak of Salmonella agona due to contaminated snacks. Commun Dis Rep CDR Wkly; 5: 29, 32

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

Barcellos, C. Sabraza, P.C. (2000). Socio-environmental determinants of the leptospirosis outbreak of 1996 in western Rio de Janeiro: a geographical approach. *International Journal of Environmental Health Research*, **10(4)**, 301-13.

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 *Highly Structured Stochastic Systems* (eds. P.J. Green, N.L. Hjort and S. Richardson), Oxford: Oxford University Press, 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.

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

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

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

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: *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance* (eds. R. Brookmeyer and D.F Stroup). Oxford University Press.

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

Farrington, C.P. and Andrews, N. (2003). Outbreak detection: Application to infectious disease surveillance. In: *Monitoring the Health of Populations* (eds. R. Brookmeyer and D.F. Stroup), Oxford: Oxford University Press, 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. (2004). Exact and efficient Bayesian inference for multiple changepoint problems. Technical Report, Department of Mathematics and Statistics, Lancaster University.

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 *Bayesian Statistics 5* (eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.M.F. Smith), pp. 599-607. Oxford: Oxford University Press.

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. Technical report, University of Munich.

Held, L. and Höhle, M. (2005). Properties of discrete-time changepoint models with unknown number of changepoints. In preperation.

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.

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.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields. Theory and Applications*. 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 notiable diseases surveillace data. *Statistics in Medicine*, **8**, 323-329.