Hofmann, Mathias; Höhle, Michael; Held, Leonhard

Working Paper

A stochastic model for multivariate surveillance of infectious diseases

Discussion Paper, No. 394

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Hofmann, Mathias; Höhle, Michael; Held, Leonhard (2004) : A stochastic model for multivariate surveillance of infectious diseases, Discussion Paper, No. 394, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München,
https://doi.org/10.5282/ubm/epub.1764

This Version is available at:
https://hdl.handle.net/10419/31038

A stochastic model for multivariate surveillance of infectious diseases

Mathias Hofmann, Michael Höhle and Leonhard Held*

Department of Statistics, Ludwig-Maximilians-Universität München,

Ludwigstr. 33, 80539 München, Germany

23rd July 2004

Abstract

We describe a stochastic model based on a branching process for analyzing surveillance data of infectious diseases that allows to make forecasts of the future development of the epidemic. The model is based on a Poisson branching process with immigration with additional adjustment for possible overdispersion. An extension to a longitudinal model for the multivariate case is described. The model is estimated in a Bayesian context using Markov Chain Monte Carlo (MCMC) techniques. We illustrate the applicability of the model through analyses of simulated and real data.

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

1 Introduction

Surveillance of notifiable infectious diseases is a common task in most countries. One major goal is to detect outbreaks in disease incidence, in order to employ appropriate public health interventions. A similar task has gained increasing interest in the context of combatting bioterrorism attacks. For a recent review of the methodology used see Farrington and Andrews (2003).

How does such outbreak detection work? Most methods currently in use consider only a single time series of cases. Basically, suspiciously high number of cases are flagged as outbreaks. However, typically longitudinal space-time data is available, i.e. the number of cases in each district, say, and at each time point, e.g. day or week. Clearly, a lot of statistical efficiency will be lost if such longitudinal data is analysed separately, series by series.

A second issue arises regarding the mechanism how to detect outbreaks. In nearly all methods proposed, a simple statistical model is used to roughly describe the incidence in the past. For example, this could be a Poisson time series model with constant or linear time trend as in Farrington et al. (1996), or even assuming approximate normality of the observed counts (Stroup et al., 1989). For the current time point, observed cases are compared to the expected cases, under the fitted model to the past data. If the discrepancy is too large, i.e. the observed cases exceed some upper confidence limit for the predicted cases, an alarm will be flagged.

Farrington et al. (1996) have realized a problem with such a procedure: implicitly the method assumes that no outbreak has happened in the past. They suggest to downweight past observations with suspiciously large residuals under the simple model, and re-analyse the data with the new weights.

In this paper we follow a different strategy. First, we try to use a model that, at least

qualitatively, allows for outbreaks, in fact it can be justified as an approximation to the so-called SIR-model, often used to analyse person-to-person infections (see Anderson and Britton, 2000). More specifically, we will assume that the observed number of cases follows a so-called branching process with immigration (Guttorp, 1995). Branching processes are often used as approximations to SIR-model (Farrington et al., 2003, and the references therein), especially if there is no data on the number of susceptibles. This is nearly always the case for surveillance data, hence this model seems appropriate to describe the time course of surveillance data on infectious diseases. We will use a branching process model with Poisson distributed offsprings, i.e. the number of cases at time t is Poisson distributed with mean proportional to the number of infected case at time $t-1$.

However, a problem with branching processes is that, with probability one, the epidemic will either explode or die out in finite time, again clearly inappropriate for most infectious disease surveillance data, where endemic (i.e. quasi-stationary) incidence typically plays a large role. Therefore, we add an endemic part with constant rate to the branching process in order to describe the total observed counts. The resulting model is known as a *branching process model with immigration* (Guttorp, 1995). Under certain restrictions on the parameters, the model is now stationary, but is able to capture an epidemic behaviour in the observed counts.

To flag outbreaks, we also follow a different strategy and use the *predictive* distribution for the number of cases at the next time point. So, essentially, we assume that our model is "correct", in sharp contrast to the assumption of a "wrong" model in the methods currently in use for outbreak detection. One clear benefit of such a model-based approach is that the chosen model can be validated through residual and predictive checks. However adjustments for additional overdispersion in the observed counts are typically required,

because otherwise the residual variance will exceed the predicted variance for most data we have analysed. Technically this is no problem, essentially we replace the Poisson distribution for the observed counts with a negative binomial distribution.

Further motivation to adjust for overdispersion comes from the fact that the generation time of the branching process represents the spread of the epidemic. However, the (mean) length of the generations normally does not fit the length of the observation intervals, typically days or weeks. Besides, the generation time may be random. This can introduce an extra amount of variation in the data. Simulation studies showed, that discrepant generation and observation times result in overdispersion. However, the interpretation of the process as a classical branching process is now lost.

Finally, our model can easily be extended to the multivariate case, and we describe such an extension in this paper. This has the distinct advantage that model parameters can be estimated much more precisely based on all parallel time series and the statistical predictions will improve.

A few comments should be made on alternative autoregressive specifications for time series of counts. Diggle et al. (2002), Section 10.4 describe so-called transition models for Poisson counts within a generalized linear model framework. Unfortunately, in the log-linear Poisson case one cannot simply include past outcomes as explanatory variables, since the conditional mean will either grow exponentially in time, or can only describe negative association between outcomes. A modified model, proposed by Zeger and Qaqish (1988) addresses this problem and proposes a model somewhat similar to ours, since it can be interpreted as a *size-dependent branching process*. However, here the offspring rate of each case is inversely related to the number of cases at time $t-1$, to ensure stationarity, whereas we employ additional "immigration" in the branching process, enabling us to decompose the total incidence into endemic and epidemic cases.

Another modification to the autoregressive Poisson model, proposed in Knorr-Held and Richardson (2003), is to include additional *epidemic indicators* in the model, which decide if the autoregressive component is switched on or off. The indicators are modelled with a two-state hidden Markov model, and ensure that the process will not explode. This model has been successfully applied in a space-time context to rare infectious diseases, such as meningitis (Knorr-Held and Richardson, 2003), see also Diggle et al. (2003). However, a downside of this model is that, while it can model epidemic increases, the return to the "endemic" level must be abrupt by switching off the relevant indicator.

This paper is organized as follows. We first describe our model, both in the time series and multivariate case. Then we outline how to estimate the model using a Bayesian approach and Markov chain Monte Carlo (MCMC) techniques. We illustrate the performance of the model through several analyses of simulated and real surveillance data. Finally, we discuss several ways how to improve the model formulation.

2 Model

First we describe our model for a (equally spaced) time series Z_t, $t = 1, \ldots, n$, where Z_t is the number of observed cases at time t. The model assumes that Z_t is the sum of an endemic part, X_t, and an epidemic part, Y_t: $Z_t = X_t + Y_t$, $t = 1, \ldots, n$. The endemic part X_t is assumed to follow a Poisson distribution with fixed parameter ν. The epidemic part Y_t is assumed to have an autoregressive structure and infections are assumed to occur with rate proportional to the observed number of cases Z_{t-1} at time $t - 1$. The

model for X_t and Y_t is

$$\begin{aligned} X_t &\sim \text{Po}(\nu), \quad t = 1, 2, \ldots, n, \\ Y_t &\sim \begin{cases} \text{Po}\left(\omega_s \frac{\lambda\nu}{1-\lambda}\right) & t = 1, \\ \text{Po}(\lambda(Y_{t-1} + X_{t-1})) & t = 2, 3, \ldots, n. \end{cases} \end{aligned} \tag{1}$$

Here $\lambda \in (0, 1)$ is unknown and ω_s has a Gamma distribution with expectation equal to 1 and variance equal to $[\nu(1+\lambda)]^{-1}$, i.e. $\omega_s \sim \text{Ga}(\nu(1+\lambda), \nu(1+\lambda))$. This particular choice will be motivated below. The parameters X_t and Y_t are latent variables that are not observed, but can be estimated using MCMC.

The model corresponds to a branching process with immigration, as defined in Guttorp (1995), p. 99. For $t = 2, \ldots, n$, Z_t is the sum of Z_{t-1} independent random variables $L_{t,j}$, $j = 1, \ldots, Z_{t-1}$, each following a Poisson distribution with mean λ (the so-called offspring distribution) and the immigration variable X_t:

$$Z_t = \sum_{j=1}^{Z_{t-1}} L_{t,j} + X_t.$$

It can be shown (Guttorp, 1995, p. 99) that Z_t has a stationary distribution with mean $\mu_Z = \nu/(1-\lambda)$ and variance $\sigma_Z^2 = \nu/\{(1-\lambda)(1-\lambda^2)\}$. The epidemic part $Y_t = Z_t - X_t$ thus has stationary mean $\mu_Y = \lambda\nu/(1-\lambda)$ and variance $\sigma_Z^2 = \nu/\{(1-\lambda)(1-\lambda^2)\} - \nu$. The choice $\omega_s \sim \text{Ga}(\nu(1+\lambda), \nu(1+\lambda))$ in (1) simply results in a negative binomial distribution for Y_1 (after integrating out ω_s), with mean and variance equal to the stationary mean and variance of Y_t. Note that the parameters ν and λ represent the rate of the endemic and the infection rate of the epidemic part and do not depend on t. Figure 1 shows a simulation from this model using the parameter values $\nu = 50$ and $\lambda = 0.7$.

2.1 Overdispersion for the response variable

Typically, the observed time is not the same as the generation time of the branching process that shall represent the spread of the epidemic. The mean length of the generations normally does not fit the (typically arbitrary) length of the observation intervals. Besides, the generation time may be random. This can introduce an extra amount of variation into the model. Simulation studies showed, that it can therefore be useful to include parameters $\omega_t, t = 1, \ldots, n$ into the model, to adjust for possible overdispersion:

$$\begin{aligned} \omega_t &\sim \text{Ga}(\psi, \psi), \quad \psi > 0, \\ Y_t|\omega_t &\sim \begin{cases} \text{Po}(\omega_t \frac{\omega_s \lambda \nu}{1-\lambda}) & \text{for} \quad t = 1, \\ \text{Po}(\omega_t \lambda (X_{t-1} + Y_{t-1})) & \text{for} \quad t = 2, \ldots, n. \end{cases} \end{aligned}$$

It can be shown (DeGroot, 1970, p. 119) that the marginal distribution of Y_t integrating out ω_t is a negative binomial distribution,

$$Y_t \sim \begin{cases} \text{NegBin}\left(\omega_s \frac{\lambda \nu}{1-\lambda}, \psi\right) & \text{for} \quad t = 1, \\ \text{NegBin}\left(\lambda (X_{t-1} + Y_{t-1}), \psi\right) & \text{for} \quad t = 2, \ldots, n, \end{cases}$$

where $\text{NegBin}(\mu, \psi)$ denotes the negative binomial distribution with expectation μ and dispersion parameter ψ. Thus the marginal mean of Y_t is the same as in (1), but the marginal variance is now

$$V[Y_t] \;=\; E[Y_t] \left(1 + \frac{E[Y_t]}{\psi}\right),$$

hence larger. For $\psi \to \infty$ it can be seen that $V[Y_t] \to E[Y_t]$. Figure 2 shows a

simulation from this model using the parameter values $\nu = 50$, $\lambda = 0.7$ and $\psi = 10$. Note that, by introducing overdispersion to the model, the branching process interpretation is no longer valid, since there is no offspring distribution for which the Y_t can be seen as the sum of $Y_{t-1} + X_{t-1}$ offspring distributed random variables.

3 Estimation

The model is estimated using MCMC methods. Prior distributions are assumed for all unknown model parameters, and full conditionals are derived wherever possible to employ the Gibbs sampler, otherwise a Metropolis-Hastings algorithm is used. The distribution of the data $\boldsymbol{Z}$ and the parameters $\boldsymbol{\theta}$, including the latent variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, is given as

$$\begin{aligned} p(\boldsymbol{Z}, \boldsymbol{\theta}) &= p(\lambda)p(\nu) \\ &\quad \cdot \prod_{t=1}^{n} \left(P(Z_t|X_t, Y_t)P(X_t|\nu)\right) \prod_{t=2}^{n} P(Y_t|\lambda, Y_{t-1}, X_{t-1}) \\ &\quad \cdot P(Y_1|\lambda, \nu, \omega_s)p(\omega_s|\nu, \lambda). \end{aligned}$$

The prior distributions for the parameters λ and ν are

$$\lambda \sim \text{Beta}(\alpha_\lambda, \beta_\lambda), \quad \nu \sim \text{Ga}(\alpha_\nu, \beta_\nu).$$

The MCMC algorithm uses X_t as unknown auxiliary variables, i.e. generates samples of X_t, conditional on Z_t and all model parameters. This conditional distribution is a simple binomial distribution, and the value of Y_t is then determined through $Y_t = Z_t - X_t$.
The mixing parameter ω_s has a gamma full conditional, and this is also the case for the rate parameter ν. Only λ has a non-standard full conditional and here we use a simple Gaussian Metropolis random walk proposal for updating. All full conditionals

and Metropolis-Hastings steps are described in detail in the Appendix.

3.1 Overdispersion for Y_t

In the case of the Poisson-Gamma model used to obtain overdispersion, a $\text{Ga}(\alpha_\psi, \beta_\psi)$ prior is assumed for ψ. This parameter is then updated using a Metropolis-Hastings algorithms with a Gaussian random walk proposal. The distribution of the data $\boldsymbol{Z}$ and the parameters $\boldsymbol{\theta}$, now including ω_t and ψ, is then

$$\begin{aligned} p(\boldsymbol{Z}, \boldsymbol{\theta}) &= p(\lambda)p(\nu)p(\psi) \\ &\quad \cdot \prod_{t=1}^{n} (P(Z_t|X_t, Y_t)P(X_t|\nu)p(\omega_t|\psi)) \prod_{t=2}^{n} P(Y_t|\lambda, \omega_t, Y_{t-1}, X_{t-1}) \\ &\quad \cdot P(Y_1|\omega_1, \lambda, \nu, \omega_s)p(\omega_s|\nu, \lambda). \end{aligned}$$

The mixing parameters ω_t, $t = 1, \ldots, n$ all have gamma full conditionals. For more details see the Appendix.

3.2 Model comparisons

For model comparison, the deviance information criterion (DIC) described in Spiegelhalter et al. (2002) is used. It allows to compare models where the number of parameters is not clearly defined by considering the effective number of parameters in the model, p_D. The (saturated) deviance is

$$\begin{aligned} D_S(\boldsymbol{\theta}) &= -2 \log P(\boldsymbol{Z}|\boldsymbol{\theta}) + 2 \log P(\boldsymbol{Z}|\mu(\boldsymbol{\theta}) = \boldsymbol{Z}) \\ &= 2 \sum_{t=1}^{n} (Z_t \log(Z_t/\eta_t) - Z_t + \eta_t), \end{aligned}$$

using the convention that $0 \log 0 = 0$. Here $\eta_t = \nu + \omega_1 \omega_s \frac{\lambda\nu}{1-\lambda}$ for $t = 1$ and $\eta_t = \nu + \omega_t \lambda (X_{t-1} + Y_{t-1})$ for $t = 2, \ldots, n$.

The effective number of parameters in the model is calculated as

$$p_D = \overline{D_S(\boldsymbol{\theta})} - D_S(\overline{\boldsymbol{\theta}}),$$

and the DIC is

$$\begin{aligned} \text{DIC} &= D_S(\overline{\boldsymbol{\theta}}) + 2p_D \\ &= 2\overline{D_S(\boldsymbol{\theta})} - D_S(\overline{\boldsymbol{\theta}}), \end{aligned}$$

where $\overline{D_S(\boldsymbol{\theta})}$ is the posterior mean of the saturated deviance and $D_S(\overline{\boldsymbol{\theta}})$ is the saturated deviance at the posterior means of the parameters. A smaller value of the DIC indicates a more appropriate model.

3.3 Predictive distribution

One of the main aims is to compute the predictive distribution for the number of cases at time $n+1$. Using MCMC the predictive distribution of Z_{n+1} can be easily estimated. In every iteration k, a sample of Z_{n+1} is generated using the values of the k-th iteration of ν, λ, Y_n, X_n, ψ:

$$\begin{aligned} X_{n+1} &\sim \text{Po}(\nu), \\ \omega_{n+1} &\sim \text{Ga}(\psi, \psi), \\ Y_{n+1} &\sim \text{Po}(\lambda \omega_{n+1}(Y_n + X_n)), \\ Z_{n+1} &= X_{n+1} + Y_{n+1}. \end{aligned}$$

3.4 Estimation results

The model is first estimated for the simulated data, without and with overdispersion. Later the model is applied to Leptospirosis data observed in Rio de Janeiro in the time from January 1995 to December 1999.

3.4.1 The estimation of simulated data

The model is now estimated for the simulated data $\{Z_t\}$ shown in Figure 1. A Beta(7,3) prior distribution is assumed for λ and a Ga(10,0.2) for ν so that the prior mean is equal to the true value. Results of the posterior distribution for the parameters of interest as well as the posterior deviance and the posterior distribution of the stationary mean μ_Z and variance σ_Z^2 together with a 95% credibility interval are shown in Table 1. The mean and variance of the data is 169.29 and 370.01, respectively. The DIC is 104.92 where p_D is 2.07. The estimates of ν and λ agree quite well with the values of the simulation. Also, the estimates of the stationary mean and variance agree with the mean and variance of the data.

The model is now estimated with a different prior distribution. A Beta(5,3) prior distribution is assumed for λ and a Ga(10,0.1) for ν. The results are shown in Table 2. The DIC is now 105.53 where p_D is 2.01. There is some sensitivity with respect to the choice of the prior distributions. A reason for this may be that the number of time points is just 100. The deviance for the first choice of the prior is slightly smaller then for the second but there is virtually no difference in the DIC values.

3.4.2 The estimation of simulated data with overdispersion

The model is now estimated for the simulated data with overdispersion $\{Z_t\}$ shown in Figure 2. A Beta(7,3) prior distribution is assumed for λ, a Ga(10,0.2) for ν and

a Ga(10,1) for ψ so that the prior mean is again equal to the value of the simulation. Results of the posterior distribution for the parameters of interest as well as the posterior deviance together with a 95% credibility interval are shown in Table 3. The model with overdispersion can not be seen as branching process, which makes it difficult to derive the stationary variance of the Z_t depending on ψ. The mean and variance of the data is 175.17 and 5368.32, respectively. The DIC is 189.14 where p_D is 88.32. The estimates of λ, ν and ψ agree quite well with the values of the simulation.

The model is now estimated with different prior distribution for ψ, a Ga(20,1) distribution. The results are shown in Table 4. The DIC is 188.98 where p_D is 86.33. There is also some sensitivity with respect to the choice of the prior distribution of ψ. The deviance for the second choice of the prior distribution is again slightly higher, but there is virtually no difference in DIC values.

3.4.3 The estimation of Leptospirosis data

After the estimation of simulated data the model is now applied to the Leptospirosis data, observed in Rio de Janairo in the time from January 1995 to December 1999. The number of cases are shown in Figure 3. Leptospirosis is a bacterial infection usually caused by contaminated water. The data show one major outbreak that was caused by an inundation in combination with bad hygienic conditions. Clearly, the time series is dominated by this large outbreak. The mean and variance of the data is 6.78 and 1247.69, respectively. In most weeks there are very few cases observed except the big outbreak in the beginning of 1996. Therefore the parameter ν has to be small. The mean of the estimated stationary mean should be around the data mean. Because of the relation $\mu_{Z_t} = \nu/(1-\lambda)$, λ can not be very close to 1. A Beta(2,1) prior distribution is assumed for λ and a Ga(1,1) for ν. The results are shown in Table 5. The DIC is

now 1794.14 where p_D is 2.46. The estimated stationary mean agrees well with the data mean. However, the estimated stationary variance is much smaller then the variance of the data. The reason for this is, that given ν, the stationary variance of Z_t is determined by the stationary mean. The mean of the data is small, because most of the weeks there are few cases observed. The variance, however, is big because of the big outbreak. Therefore the model is not able to explain the outbreak very well, what can also be seen by the big deviance. One possibility to explain bigger outbreaks like this is to allow a time varying ν or λ, which will be discussed at the end of the article. Another possibility is to explain the higher variance by overdispersion.

Finally, a credibility interval for the predictive distribution has been calculated. We will always use a 95% credibility level. The last two observed values $Z_{n-1} = 1$ and $Z_n = 2$ are very small. However the next value $Z_{n+1} = 9$, not included in the analysis, is surprisingly high and could be the beginning of another outbreak. The credibility interval of the predictive distribution [0,6] does not include this observed value.

3.4.4 The estimation of Leptospirosis data with overdispersion

The model is estimated for the time series Leptospirosis data with overdispersion. The variance of the data is much bigger than the mean, so ψ must be small. A Beta(2,1) prior distribution is assumed for λ, a Ga(1,1) for ν and a Ga(1,2) for ψ. The results are shown in Table 6. The DIC is now 440.25 where p_D is 85.68. The estimated parameter ψ is very small. The DIC of the estimation with overdispersion is much smaller then of the estimation without overdispersion, so the model with overdispersion seems to fit the data much better. The credibility interval of the predictive distribution [0,9] is now larger and includes the observed value 9. This is because overdispersion explains an extra amount of variation.

4 Extension to the multivariate case

The model is now extended to a space-time model in order to describe the incidence of a disease, that is observed in I regions. In this basic space-time model the progress of the epidemic in on region is assumed to be independent from the other regions, given the model parameters ν, λ and ψ. However, all regions are linked since they share the same model parameters. The endemic part is assumed to be stationary in every region with mean depending of the population size of the region. The data $Z_{i,t}, i = 1, \ldots, I, t = 1, \ldots, n$ are given as I time series of length n. Additionally the proportion of the population of the I regions from the total population $\xi_i, i = 1, \ldots, I$ is known. In the following we use Z_{it} instead of $Z_{i,t}$. The model is given as

$$\begin{aligned} X_{it} &\sim \text{Po}(\nu\xi_i), \quad t = 1, 2, \ldots, n, \\ Y_{it} &\sim \begin{cases} \text{Po}\left(\omega_{s,i}\frac{\lambda\nu\xi_i}{1-\lambda}\right) & t = 1, \\ \text{Po}(\lambda(Y_{i,t-1} + X_{i,t-1})) & t = 2, \ldots, n, \end{cases} \\ Z_{it} &= X_{it} + Y_{it}, \end{aligned}$$

where $i = 1, \ldots, I$, $\lambda \in (0, 1)$ and $\omega_{s,i} \sim \text{Ga}(\nu\xi_i(1 + \lambda), \nu\xi_i(1 + \lambda))$. Note that the parameters ν and λ do not depend on i or t. Since we use the population proportions it follows that $\sum_{i=1}^{I} \xi_i = 1$, and hence the parameter ν can be seen as a parameter for the total population, since $\sum_{i=1}^{I} X_{it} \sim \text{Po}(\nu)$.

Z_{it} is a branching process with immigration for every region i. The introduction of overdispersion, the estimation, the predictive distribution and the model comparison follow the time series case, where ν is replaced by $\nu\xi_i$.

4.1 Estimation Results

The model is first simulated and estimated for the simulated data, for the case without and with overdispersion. Later the model is applied to Campylobacter data observed in Germany in the time from January 2001 to July 2003.

4.1.1 The estimation of simulated data

A realization of the model is simulated with parameters $n = 100$, $I = 5$, $\lambda = 0.7$ and $\nu = 50$. The ξ_i are 0.2 for every region. The model is now estimated for the simulated data $\{Z_{i,t}\}$. A Beta(7,3) prior distribution is assumed for λ and a Ga(10,0.2) for ν so that the prior mean is equal to the value of the simulation. The results are shown in Table 7. The mean and variance of the data, cumulated over the regions, is 170.79 and 465.2, respectively. The DIC is 473.54 where p_D is 4.45. The mean of the estimated parameters ν and λ are close to the estimates of the time series case, but the credibility intervals are smaller in the multivariate case. Although there are not more cases, the observation of more than one time series seems to supply more information.
The model has also been estimated with a different prior distribution. Now a Beta(5,3) prior distribution is assumed for λ and a Ga(10,0.1) for ν. The results are shown in Table 8. The DIC is 473.55 where p_D is 4.25. The model seems to be less sensitive to the prior distribution in the multivariate case then in the time series case, which makes sense, because the amount of data is much larger. Similar results have been obtained for simulated data with overdispersion.

4.1.2 The estimation of Campylobacter data

After the estimation of simulated data the model is now applied to the Campylobacter data, observed in the 16 states of Germany in the time from January 2001 to July 2003

over 129 weeks. The total counts in Germany are shown in the Figure 4. Campylobacter is a bacterial infection of the intestine that is mostly caused by contaminated food or transmitted by the excrements of pets or sometimes also by infected persons. It has a endemically dominated incidence. From Figure 4 it can be seen that there is a seasonal structure in the data. During the summer there is an increase of the observed cases, while in the winter the number of cases is smaller.

The mean and variance of the data is 282.62 and 10298.24, respectively. There are two apparent "outbreaks" in the summer, we will comment on this later. Therefore, the estimated λ is likely to be close to unity. The relation of ν and λ for the stationary mean then forces ν to be small. A Beta(9,0.5) prior distribution is assumed for λ and a Ga(10,1) for ν. The results are shown in Table 9. The DIC is 8239.83 where p_D is 12.84. The estimated parameter λ is rather large. The reason is that the model can not explain the seasonal structure of the data by the endemic part that is assumed to have a constant parameter ν. However, endemically dominated incidence often has a seasonal structure. It will therefore be necessary to allow for a time varying ν that can explain the seasonal structure by the endemic part. Additionally a seasonal component of ν would make sense, in order to get a better prediction. Figure 5 shows the credibility intervals of the predictive distributions for the 16 German states and the actually observed value at this time. The predictive distribution seems not to predict the real values very well. Four of the 16 observed values are larger than the upper credibility interval limit. The reason is that the model underestimates the variance of the data, the estimated stationary variance is smaller then the empirical variance of the data. A consequence is that the variance of the predictive distribution is also smaller. To get a better fit the data should be estimated allowing for overdispersion.

4.1.3 The estimation of Campylobacter data with overdispersion

The variance of the data is not so large compared with the mean. Therefore a Beta(9,0.5) prior distribution is assumed for λ, a Ga(10,1) for ν and a Ga(10,1) for ψ. The results are shown in Table 10. The DIC is now 3850.26 where p_D is 1416.74. The parameters ν and λ are similar to the estimation without overdispersion, while the estimation of ψ shows that there is some overdispersion. The DIC is smaller for the estimation with overdispersion then without overdispersion, The model with overdispersion seems to fit the data better, what can also be seen in the predictive distribution. The observed values are all, except one, inside the credibility intervals of the predictive distribution shown in Figure 6.

We conclude that the model with overdispersion seems to predict the development of the disease quite well. However, the assumption of a constant parameter ν is not very realistic. A better model could be obtained by allowing ν to have a seasonal pattern.

5 Discussion

We have proposed a new model to describe the typical temporal behaviour of surveillance data on infectious diseases. Using a Bayesian approach and MCMC, a predictive distribution for the future number of cases can easily be calculated. We believe that this distribution could form the basis for a model-based outbreak detection system. Also, we have outlined a multivariate extension in order to analyse the routinely collected longitudinal data on infectious diseases. Of course, there is a lot of scope for improvement of the model. We now outline a few areas that we currently consider.

In our model the transmission of the disease in the epidemic part is assumed to be independent in the I regions. In real epidemics an infectious disease often spreads from

one region to others. It may therefore be necessary to allow an individual of a region to infect individuals in the regions of the neighborhood. One way to model this is based on the theory of a multi-type branching process.

Within a region the epidemic part is assumed to be stationary. In most diseases with predominantly endemic part a seasonal structure can be observed. Therefore a time depending ν_{it} can be introduced into the model including a seasonal component. For example, one could use a parametric model with a few Fourier frequencies.

The parameter λ is assumed to be constant over time, and therefore has to be smaller than 1, otherwise the process has a positive probability to explode, which normally can not happen in surveillance data. On the other hand it is known that some infectious diseases have a basic reproduction number larger than 1, which would correspond to $\lambda > 1$. Therefore a time varying λ_t (with stationary mean smaller then one) might be interesting to consider. This would also allow to estimate the effect of public health interventions, which should result in a smaller λ parameter.

Besides these extensions of the model area-level covariates could be introduced in ν or λ as commonly done in ecological regression analyses. Another interesting question is, if it is possible to integrate an unknown underreporting rate in the model. Morton and Finkenstädt (2004) have included such a parameter in a spatio-temporal model for measles, however, they also had information on the number of susceptibles, which is typically not the case for ordinary surveillance data.

Acknowledgments

This work is supported by the German Science Foundation (DFG), SFB 386. We thank Marilia Sá Carvalho and the Robert-Koch Institute in Berlin for providing the data on Leptospirosis and Campylobacter, respectively.

References

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

DeGroot, M. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford: Oxford University Press.

Diggle, P. J., L. Knorr-Held, B. Rowlingson, T.-L. Su, P. Hawtin, and T. Bryant (2003). On-line monitoring of public health surveillance data. In R. Brookmeyer and D. Stroup (Eds.), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. Oxford University Press.

Farrington, C., N. Andrews, A. Beale, and M. Catchpole (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *J. R. Statist. Soc. A 159*, 547–563.

Farrington, C., M. Kanaan, and N. Gay (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics 4*, 279–295.

Farrington, P. and N. Andrews (2003). Outbreak detection: Application to infectious disease surveillance. In R. Brookmeyer and D. Stroup (Eds.), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. Oxford: Oxford University Press.

Gelman, A., G. Roberts, and W. Gilks (1996). Efficient metropolis jumping rules. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 5*, pp. 599–607. Oxford: Oxford University Press.

Guttorp, P. (1995). *Stochastic Modeling of Scientific Data*. Chapman & Hall. London.

Knorr-Held, L. and S. Richardson (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society, Series C (Applied Statistics) 52*, 169–183.

Morton, A. and B. F. Finkenstädt (2004). A discrete-time spatio-temporal sir model for disease transmission. Technical report, Department of Statistics, Warwick University, UK.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit(with discussion). *J. R. Statist. Soc. 64*, 583–639.

Stroup, D., G. Williamson, J. Herndon, and J. Karon (1989). Detection of aberrations in the occurrence of notifiable diseases surveillance data. *American Journal of Epidemiology 137*, 373–380.

Zeger, S. and B. Qaqish (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics 44*, 1019–31.

A Estimation

The time series model is given as a special case of the space-time model with $I = 1$ and $\xi_i = 1$. The full conditional for the parameter ν can be derived via

$$\begin{aligned}
p(\nu|\ldots) &\propto p(\nu)\prod_{i=1}^{I}\prod_{t=1}^{n}(P(X_{it}|\nu,\xi_i))\prod_{i=1}^{I}P(Y_{i1}|\lambda,\nu,\xi_i,\omega_{s,i}) \\
&\propto \nu^{\alpha_\nu-1}\exp(-\beta_\nu\nu)\prod_{i=1}^{I}\prod_{t=1}^{n}(\nu^{X_{it}}\exp(-\nu\xi_i))\prod_{i=1}^{I}\left(\nu^{Y_{i1}}\exp\left(-\frac{\lambda\nu\xi_i\omega_{s,i}}{1-\lambda}\right)\right) \\
&\propto \nu^{\alpha_\nu-1}\exp(-\beta_\nu\nu)\nu^{\sum_{i=1}^{I}\sum_{t=1}^{n}(X_{it})}\exp\left(-n\nu\sum_{i=1}^{I}\xi_i\right)\nu^{\sum_{i=1}^{I}Y_{i1}}\exp\left(-\frac{\lambda\nu}{1-\lambda}\sum_{i=1}^{I}\xi_i\omega_{s,i}\right) \\
&\propto \nu^{\alpha_\nu+\sum_{i=1}^{I}\sum_{t=1}^{n}(X_{it})+\sum_{i=1}^{I}Y_{i1}-1}\exp\left(-\left(\beta_\nu+n+\frac{\lambda}{1-\lambda}\sum_{i=1}^{I}\xi_i\omega_{s,i}\right)\nu\right),
\end{aligned}$$

hence

$$\nu|\ldots \sim \text{Ga}\left(\alpha_\nu+\sum_{i=1}^{I}\sum_{t=1}^{n}X_{it}+\sum_{i=1}^{I}Y_{i1},\beta_\nu+n+\frac{\lambda}{1-\lambda}\sum_{i=1}^{I}\xi_i\omega_{s,i}\right).$$

Instead of λ, $\tilde{\lambda} = \text{logit}(\lambda)$ will be updated using a Metropolis-Hastings algorithm, to be able to get candidates form a not truncated distribution using a random walk proposal, which simplifies the acceptance probability and to introduce covariates into the model. For the full conditional of λ applies

$$p(\lambda|\ldots) \propto p(\lambda)\prod_{i=1}^{I}\prod_{t=2}^{n}P(Y_{it}|\lambda,X_{i,t-1},Y_{i,t-1})\prod_{i=1}^{I}P(Y_{i1}|\lambda,\nu,\xi_i,\omega_{s,i})$$

The acceptance probability of the Metropolis-Hastings algorithm can be derived by change of variables. For $\tilde{\lambda}$ a Gaussian random walk proposal distribution with vari-

ance $\sigma^2_{\tilde{\lambda}}$ is used, where $\sigma^2_{\tilde{\lambda}}$ is tuned in order to get acceptance rates between 30% and 40% (see Gelman et al., 1996).

The parameters $\omega_{s,i}$ are updated using the Gibbs sampler since

$$\begin{aligned}
p(\omega_{s,i}|\ldots) &\propto p(\omega_{s,i})P(Y_{i1}|\nu,\lambda,\omega_{s,i}) \\
&\propto \omega_{s,i}^{\nu\xi_i(1+\lambda)-1}\exp(-\omega_{s,i}\nu\xi_i(1+\lambda))\frac{\left(\omega_{s,i}\frac{\lambda\nu\xi_i}{1-\lambda}\right)^{Y_{i1}}}{Y_{it}!}\exp\left(-\omega_{s,i}\frac{\lambda\nu\xi_i}{1-\lambda}\right) \\
&\propto \omega_{s,i}^{\nu\xi_i(1+\lambda)-1}\exp(-\omega_{s,i}\nu\xi_i(1+\lambda))\omega_{s,i}^{Y_{i1}}\exp\left(-\omega_{s,i}\frac{\lambda\nu\xi_i}{1-\lambda}\right) \\
&\propto \omega_{s,i}^{\nu\xi_i(1+\lambda)+Y_{i1}-1}\exp\left(-\omega_{s,i}\left(\nu\xi_i\left((1+\lambda)+\frac{\lambda}{1-\lambda}\right)\right)\right)
\end{aligned}$$

and therefore

$$\omega_{s,i}|\ldots \sim \text{Ga}\left(\nu\xi_i(1+\lambda)+Y_{i1},\nu\xi_i\left((1+\lambda)+\frac{\lambda}{1-\lambda}\right)\right).$$

The parameters (X_{it},Y_{it}) are updated in a block because of the dependence that is given by the equation $Z_{it}=X_{it}+Y_{it}$. The full conditional of (X_{it},Y_{it}) can be written as

$$P(X_{it},Y_{it}|\ldots)=P(Y_{it}|X_{it},\ldots)P(X_{it}|\ldots),$$

where $P(Y_{it}|X_{it},\ldots)$ is deterministic: $Y_{it}=Z_{it}-X_{it}$. The full conditional of X_{i1} is binomial distributed,

$$X_{i1}|Z_{i1},\ldots \sim \text{Bin}\left(Z_{i1}, \frac{\nu\xi_i}{\nu\xi_i + \omega_{s,i}\frac{\lambda\nu\xi_i}{1-\lambda}}\right),$$

and the full conditional of $X_{it}, t = 2,\ldots,n$ is

$$X_{it}|Z_{it},\ldots \sim \text{Bin}\left(Z_{it}, \frac{\nu\xi_i}{\nu\xi_i + \lambda(Y_{i,t-1} + X_{i,t-1})}\right)$$

Overdispersion for Y_{it}

In case of the Poisson-Gamma construction used to obtain overdispersion a $\text{Ga}(\alpha_\psi, \beta_\psi)$ prior is assumed on ψ. The parameter $\tilde{\psi} = \log(\psi)$ is then updated using Metropolis-Hastings algorithms with a Gaussian random walk proposal with variance σ^2_ψ. The full conditional of ψ is

$$p(\psi|\ldots) \quad \propto \quad p(\psi)\prod_{i=1}^{I}\prod_{t=1}^{n} P(\omega_{it}|\psi).$$

the acceptance rate of the Metropolis-Hastings algorithm can be derived by change of variables. For $\tilde{\psi}$ a Gaussian random walk proposal distribution with variance $\sigma^2_{\tilde{\psi}}$ is used, where $\sigma^2_{\tilde{\psi}}$ is again tuned in order to get appropriate acceptance rates.

The full conditional of ω_{i1} is

$$
\begin{aligned}
\omega_{i1}|\ldots &\propto p(\omega_{i1})P(Y_{i1}|\omega_{i1},\lambda,\nu,\xi_i\omega_{s,i}) \\
&\propto \omega_{i1}^{\psi-1}\exp(-\psi\omega_{i1})\frac{\left(\omega_{i1}\frac{\lambda\nu\xi_i\omega_{s,i}}{1-\lambda}\right)^{Y_{i1}}}{Y_{i1}!}\exp\left(-\omega_{i1}\frac{\lambda\nu\xi_i\omega_{s,i}}{1-\lambda}\right) \\
&\propto \omega_{i1}^{\psi-1}\exp(-\psi\omega_{i1})\omega_{i1}^{Y_{i1}}\exp\left(-\omega_{i1}\frac{\lambda\nu\xi_i\omega_{s,i}}{1-\lambda}\right) \\
&\propto \omega_{i1}^{\psi+Y_{i1}-1}\exp\left(-\omega_{i1}\left(\psi+\frac{\lambda\nu\xi_i\omega_{s,i}}{1-\lambda}\right)\right) \\
&\sim \mathrm{Ga}\left(\psi+Y_{i1},\psi+\frac{\lambda\nu\xi_i\omega_{s,i}}{1-\lambda}\right),
\end{aligned}
$$

and for $t=2,\ldots,n$ the full conditional of ω_{it} is

$$
\begin{aligned}
\omega_{it}|\ldots &\propto p(\omega_{it})P(Y_{it}|\lambda,\omega_{it},X_{i,t-1},Y_{i,t-1}) \\
&\propto \omega_{it}^{\psi-1}\exp(-\psi\omega_{it})\frac{(\lambda\omega_{it}(X_{i,t-1}+Y_{i,t-1}))^{Y_{it}}}{Y_{it}!}\exp(-\lambda\omega_{it}(X_{i,t-1}+Y_{i,t-1})) \\
&\propto \omega_{it}^{\psi-1}\exp(-\psi\omega_{it})\omega_{it}^{Y_{it}}\exp(-\lambda\omega_{it}(X_{i,t-1}+Y_{i,t-1})) \\
&\propto \omega_{it}^{\psi+Y_{it}-1}\exp(-\omega_{it}(\psi+\lambda(X_{i,t-1}+Y_{i,t-1}))) \\
&\sim \mathrm{Ga}\left(\psi+Y_{it},\psi+\lambda(X_{i,t-1}+Y_{i,t-1})\right).
\end{aligned}
$$

The update of the other parameters changes as follows:

$$
\begin{aligned}
\lambda|\ldots &\propto p(\lambda)\prod_{i=1}^{I}\prod_{t=2}^{n}P(Y_{it}|\lambda,\omega_{it},X_{i,t-1},Y_{i,t-1})\prod_{i=1}^{I}P(Y_{i1}|\omega_{i1},\lambda,\nu,\xi_i,\omega_{s,i}),\\
\nu|\ldots &\sim \text{Ga}\left(\alpha_\nu+\sum_{i=1}^{I}\sum_{t=1}^{n}(X_{it})+\sum_{i=1}^{I}Y_{i1},\beta_\nu+n+\frac{\lambda}{1-\lambda}\sum_{i=1}^{I}\xi_i\omega_{s,i}\omega_{i1}\right),\\
\omega_{s,i} &\sim \text{Ga}\left(\nu\xi_i(1+\lambda)+Y_{i1},\nu\xi_i\left((1+\lambda)+\frac{\omega_{i1}\lambda}{1-\lambda}\right)\right),\\
X_{i1}|Z_{i1},\ldots &\sim \text{Bin}\left(Z_{i1},\frac{\nu\xi_i}{\nu\xi_i+\omega_{i1}\frac{\lambda\nu\xi_i\omega_{s,i}}{1-\lambda}}\right),\\
X_{it}|Z_{it},\ldots &\sim \text{Bin}\left(Z_{it},\frac{\nu\xi_i}{\nu\xi_i+\lambda\omega_{it}(Y_{i,t-1}+X_{i,t-1})}\right).
\end{aligned}
$$

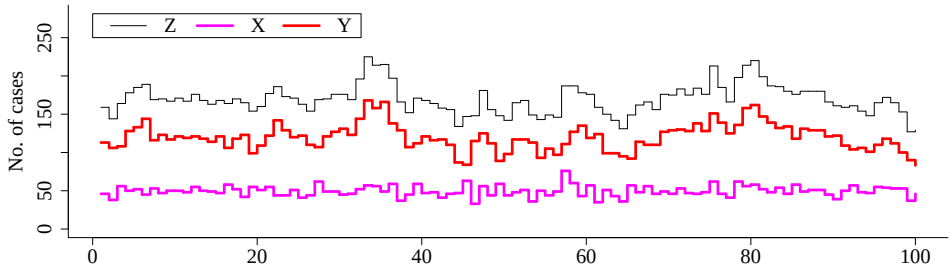

Figure 1: A Realization of X_t, Y_t and Z_t for model (1).

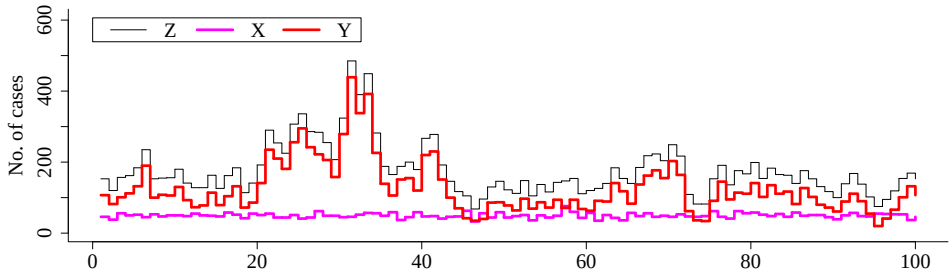

Figure 2: A Realization of X_t, Y_t and Z_t for model (1) with overdispersion.

	mean	std.dev	2.5%	97.5%
ν	44.89	9.03	28.62	63.64
λ	0.73	0.05	0.62	0.83
$D_S(\boldsymbol{\theta})$	102.85	1.87	100.80	107.77
μ	168.14	5.10	157.92	178.32
σ^2	377.42	74.71	273.51	546.41

Table 1: Estimation results for the simulated data

	mean	std.dev	2.5%	97.5%
ν	53.81	9.09	36.34	71.09
λ	0.68	0.05	0.58	0.79
$D_S(\boldsymbol{\theta})$	103.52	2.23	100.88	108.98
μ	168.49	4.15	160.39	176.84
σ^2	322.13	49.63	251.87	443.33

Table 2: Estimation results for the simulated data for a second prior distribution

	mean	std.dev	2.5%	97.5%
ν	42.26	7.34	27.67	56.83
λ	0.75	0.05	0.65	0.85
ψ	11.34	2.01	7.78	15.67
$D_S(\boldsymbol{\theta})$	100.82	14.31	74.85	130.37

Table 3: Estimation results for the simulated data with overdispersion

	mean	std.dev	2.5%	97.5%
ν	37.12	6.24	24.94	49.89
λ	0.78	0.04	0.70	0.86
ψ	14.73	2.25	10.65	19.44
$D_S(\boldsymbol{\theta})$	102.65	14.69	75.81	132.98

Table 4: Estimation results for the simulated data with overdispersion for a second prior distribution

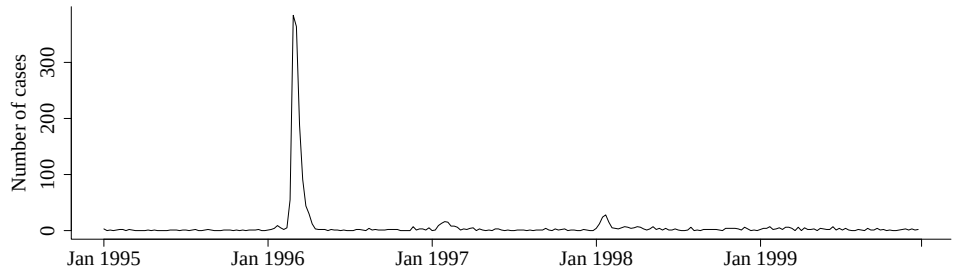

Figure 3: The Number of observed cases of Leptospirosis per week in Rio de Janeiro in the time from January 1995 to December 1999.

	mean	std.dev	2.5%	97.5%
ν	0.88	0.08	0.73	1.03
λ	0.87	0.02	0.82	0.91
$D_S(\boldsymbol{\theta})$	1791.68	2.24	1789.19	1797.25
μ	6.78	1.28	4.85	9.80
σ^2	28.96	11.18	15.42	56.40

Table 5: Estimation results for the Leptospirosis data without overdispersion

	mean	std.dev	2.5%	97.5%
ν	1.08	0.10	0.90	1.29
λ	0.63	0.09	0.47	0.82
ψ	0.51	0.12	0.31	0.76
$D_S(\boldsymbol{\theta})$	354.56	16.00	324.39	387.35

Table 6: Estimation results for the Leptospirosis data with overdispersion

	mean	std.dev	2.5%	97.5%
ν	45.30	5.05	35.36	54.91
λ	0.73	0.03	0.68	0.79
$D_S(\boldsymbol{\theta})$	469.09	2.81	465.15	475.91
μ	170.60	4.93	161.19	180.67
σ^2	374.69	41.05	310.46	470.42

Table 7: Estimation results for the simulated data

	mean	std.dev	2.5%	97.5%
ν	48.02	4.82	38.31	57.13
λ	0.72	0.03	0.66	0.78
$D_S(\boldsymbol{\theta})$	469.31	2.90	465.18	476.31
μ	170.75	4.56	161.99	179.82
σ^2	356.57	35.39	300.86	436.31

Table 8: Estimation results for the simulated data for a second prior distribution

Figure 4: The Number of observed cases of Campylobacter per week in Germany in the time from January 2001 to July 2003.

	mean	std.dev	2.5%	97.5%
ν	38.03	2.43	33.33	42.88
λ	0.95	0.00	0.94	0.95
$D_S(\boldsymbol{\theta})$	8226.99	11.02	8207.26	8250.48
μ	719.84	39.50	646.27	800.43
σ^2	7039.10	828.77	5635.48	8848.86

Table 9: Estimation results for the Campylobacter data without overdispersion

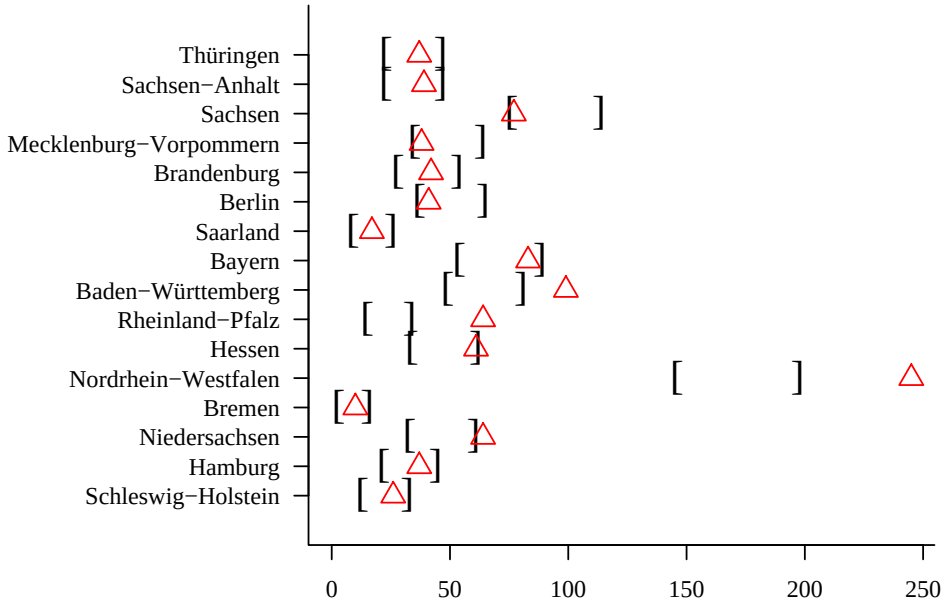

Figure 5: Posterior predictive credibility interval together with the observed value(indicated by a triangle)

	mean	std.dev	2.5%	97.5%
ν	36.60	4.00	28.97	44.36
λ	0.95	0.01	0.94	0.96
ψ	11.62	0.58	10.53	12.78
$D_S(\boldsymbol{\theta})$	2433.52	69.15	2298.82	2571.30

Table 10: Estimation results for the Campylobacter data with overdispersion

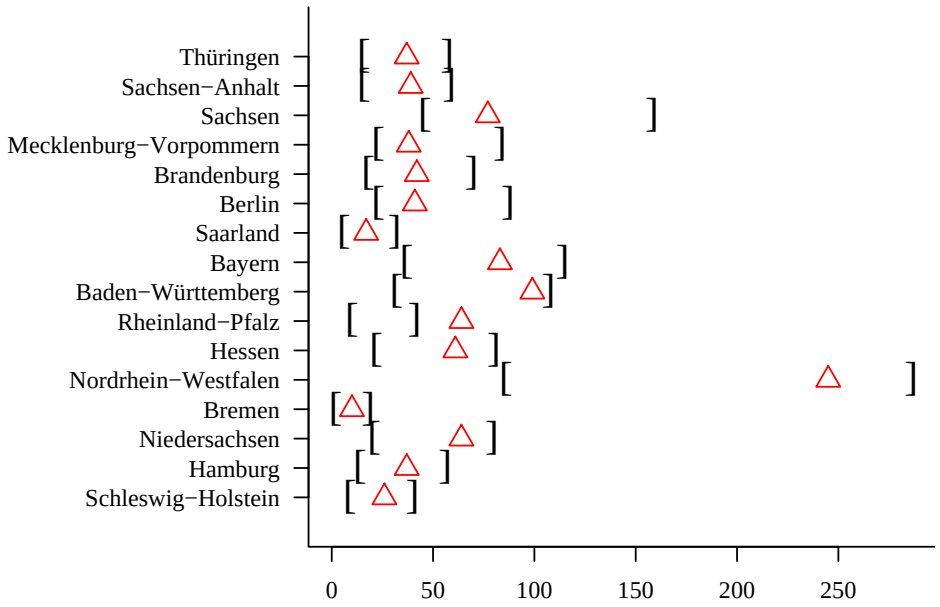

Figure 6: Posterior predictive credibility interval together with the observed value(indicated by a triangle)