Bayesian Hierarchical Model Algorithm for Automated Detection of Infectious Disease Outbreaks

Juliane Manitz[1] and Michael Höhle[2]

InFER 2011, Warwick

30th Mar 2011

[1] Institute for Mathematical Stochastics
Georg-August-University, Göttingen, Germany

[2] Department for Infectious Disease Epidemiology
Robert Koch Institute, Berlin, Germany

Motivation: German routine surveillance of Campylobacter

▶ Leading cause of enteritis illnesses in industrial countries (Robert Koch-Institut, 2005)
▶ Absolute humidity has been proven to influence the incidence
▶ Typical characteristics of surveillance time series

Process of Campylobacter infections in Germany

J. Manitz and M. Höhle   Bayesian model algorithm for automated outbreak detection
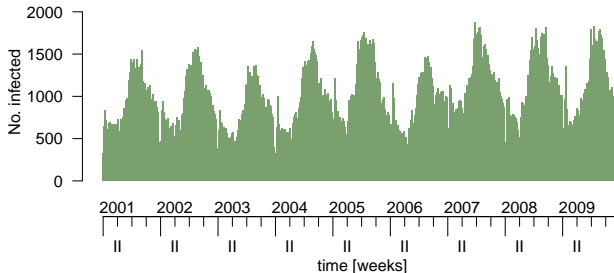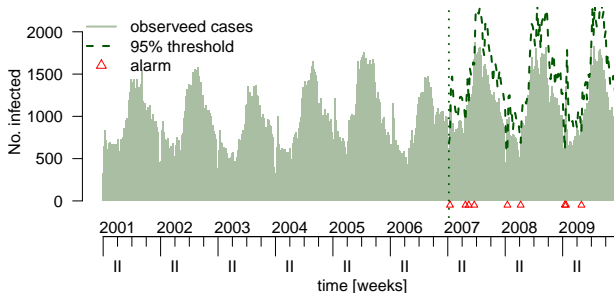
Aim of Surveillance

Public health view: Early-warning system to detect health risks, such as outbreaks of infectious diseases

Statistical view: Real-time online monitoring in routine surveillance data to detect aberrations for public health events

Process of Campylobacter infections in Germany

J. Manitz and M. Höhle   Bayesian model algorithm for automated outbreak detection

Outline

J. Manitz and M. Höhle   Bayesian model algorithm for automated outbreak detection

Theoretical Framework
Implementation
Simulation & Application
Conclusion
References

Statistical Foundation
Definition of Prior
Predictive Posterior

Statistical Foundation: Heisterkamp et al. (2006)

▶ Generalised additive quasi-Possion model (GAM) of the number of cases

$$y_t|\mu_t \overset{\text{iid}}{\sim} \text{Po}(\mu_t), \ t = 1, \dots, T$$
$$\log(\mu_t) = \eta_t = \beta_{0t} + \mathbf{x}_t'\boldsymbol{\beta}_x, \ t = 1, \dots, T$$

with overdispersion $\text{Var}(y_t|\mu_t) = \phi \cdot \text{E}(y_t|\mu_t) = \phi \cdot \mu_t$

▶ Time varying intercept as smoothed time effect

$$\beta_{0t} = \beta_0(t) = g_m(t) + u_t, \quad u_t \overset{\text{iid}}{\sim} \text{N}(0, \lambda^{-1})$$

stationary model (IID): $\quad g_{iid}(t) = \beta_0$
neighbour model (RW1): $\quad g_{rw1}(t) = \beta_{0,t-1}$
linear model (RW2): $\quad g_{rw2}(t) = 2\beta_{0,t-1} - \beta_{0,t-2}$

Theoretical Framework
Implementation
Simulation & Application
Conclusion
References

Statistical Foundation
Definition of Prior
Predictive Posterior

Heisterkamp et al. (2006): Advantages & Disadvantages

- ▶ Opposite to Farrington et al. (1996) all available data is used for modelling
- ▶ Likelihood inference for the GAM
- ▶ Threshold calculation using a plug-in prediction interval
    - ⇒ Uncertainty by parameter estimation is ignored
    - ⇒ Normal distribution assumption for MLE is questionable
- ▶ Alarm is triggered if observed y_t is exceeding treshold
- ▶ Enhanced computational efficiency by sequential model update procedure

Theoretical Framework
Implementation
Simulation & Application
Conclusion
References

Statistical Foundation
Definition of Prior
Predictive Posterior

Our Work: Bayesian Hierachical Model Algorithm

- ▶ Improvement of Heisterkamp et al. (2006) approach by using Bayesian inference for GAM: Deduce results only from posterior distribution

$$p(\boldsymbol{\beta}|\mathbf{y}_t) \propto p(\mathbf{y}_t|\boldsymbol{\beta})p(\boldsymbol{\beta}).$$

- ▶ Time varying intercept as random walk prior $p(\boldsymbol{\beta})$
- ▶ Regression modelling, i.e. covariates such as absolute humidity can be taken into account
- ▶ Threshold calculation using predictive posterior $p(y_T|y_1, \ldots, y_{T-1})$ as in Höhle (2007) which
  - ▶ includes directly uncertainty due to prediction and estimation
  - ▶ does not assume any further distribution properties

Theoretical Framework
Implementation
Simulation & Application
Conclusion
References

Statistical Foundation
Definition of Prior
Predictive Posterior

Definition of Prior

Time varying intercept: $\quad \beta_{0t} \quad \propto \quad \mathrm{N}(g_m(t), \lambda^{-1}) \quad$ random walk

Covariate paramters: $\quad \boldsymbol{\beta}'_x \quad \propto \quad \mathrm{N}(0, \mathbf{B}^{-1}) \qquad$ non-informative

I.e. the prior for $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0T}, \boldsymbol{\beta}'_x)'$ given λ is

$$p(\boldsymbol{\beta}) \propto \lambda^{\frac{\mathrm{rk}(\mathbf{A}_m)}{2}} \exp\left(-\frac{\lambda}{2}\boldsymbol{\beta}' \mathbf{A}_m \boldsymbol{\beta}\right),$$

where $\mathbf{A}_m = \begin{pmatrix} \boxed{\mathbf{D}'_m \mathbf{D}_m} & 0 \\ 0 & \mathbf{B} \end{pmatrix}$, with difference matrix $\mathbf{D}_m$.

E.g. for 2nd order random walk is $\mathbf{D}_{rw2} = \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \end{pmatrix}$.

Theoretical Framework
Implementation
Simulation & Application
Conclusion
References

Statistical Foundation
Definition of Prior
Predictive Posterior
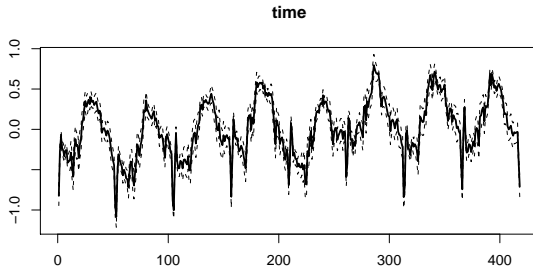
Predictive Posterior

The predictive posterior is the pdf for a new observation y_T at current time point T given the past

$$
\begin{aligned}
f(y_T|y_1,\ldots,y_{T-1}) &= \int f(y_T,\beta_{0T}|y_1,\ldots,y_{T-1}) \, d\beta_{0T} \\
&= \int f(y_T|\beta_{0T}) \, p(\beta_{0T}|y_1,\ldots,y_{T-1}) \, d\beta_{0T}.
\end{aligned}
$$

Alarm threshold $\xi_{1-\alpha}$ is defined as $100\% \cdot (1-\alpha)$-quantile of the predictive posterior.

Implementation using INLA

▶ Bayesian inference is usually very computer-intensive using MCMC methods

▶ Modern, efficient, and accurate GAM fitting by using Integrated Nested Laplace Approximation (INLA, Rue et al., 2009)

PostMean 0.025% 0.5% 0.975%

J. Manitz and M. Höhle   Bayesian model algorithm for automated outbreak detection

Implementation in R

- The realization with R inla is simple, but partially unstable
- Calculation of predictive posterior is done by Monte Carlo integration
- Implementation available as algo.hts() into R package surveillance (Höhle, 2007)

```
> algo.hts(disProgObj,
+          control=list(range=NULL, co.arg=NULL,
+                  prior='iid', family='poisson',
+                  alpha=0.05, mc.betaT1=100, mc.yT1=10))
```

Theoretical Framework
Implementation
Simulation & Application
Conclusion
References

Simulation Studies
Campylobacter data

Simulation Studies

Bayesian hierarchical model algorithm was compared with RKI method (inspired by Stroup et al., 1993), simple Bayes algorithm (Höhle, 2007), Farrington algorithm (Farrington et al., 1996) regarding

- ▶ Percentage of false alarm
- ▶ Realistic simulation data of different scenarios regarding trend, seasonality and distribution parameters for outbreak (Hutwagner et al., 2005)
- ▶ Computing time

Theoretical Framework
Implementation
Simulation & Application
Conclusion
References

Simulation Studies
Campylobacter data

Results for Bayesian hierarchical model algorithm

- ▶ 3.5-3.9% false alarms ⇒ not overly conservative
- ▶ 1-2 sec. per time point ⇒ slow, but still reasonable computing time
- ▶ approach is stable and well working in all scenarios, e.g. overall outbreak detection rates with trend and medium seasonality

	Bayesian Hierarchical Model						Farrington		Bayes	RKI
	Poisson			neg. Binomial			(4,0,2)		(4,0,2)	(4,0,2)
	iid	rw1	rw2	iid	rw1	rw2		rew		
spec	0.27	0.12	0.02	0.60	0.10	0.10	0.64	0.89	1.00	0.71
sens	0.76	0.94	0.99	0.49	0.95	0.95	0.83	0.56	0.01	0.75
mlag	5.86	12.26	18.46	2.80	12.50	12.50	4.14	0.88	0.00	2.80

⇒ our method could not be shown to be superior

Theoretical Framework
Implementation
Simulation & Application
Conclusion
References

Simulation Studies
Campylobacter data

Application to Campylobacter data including covariates

```
> control <- list(range=which(date>=as.Date('2007-01-01')),
+          co.arg=cbind(l1.hum,l2.hum,l3.hum,l4.hum,l5.hum),
+          prior='rw1', family='poisson')
> hts.hum <- algo.hts(cam.disProg, control=control))
```


Analysis of Campylobacter using hts(prior=rw1)
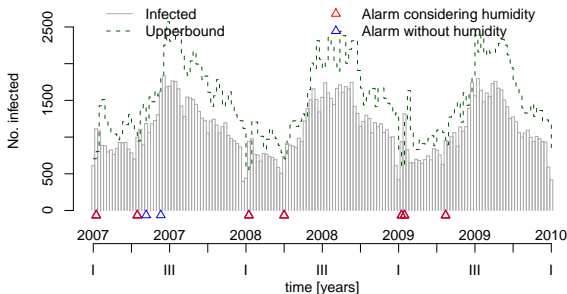
Conclusion and Discussion

- ▶ Consideration of all available knowledge
  - ▶ All past observations
  - ▶ Model includes covariates
  - ▶ Time series structure as prior knowledge
- ▶ Seamless inference
  - ▶ Simultaneous inclusion of estimate and prediction uncertainty
  - ▶ No further assumptions for calculation of threshold needed
  - ▶ Probabilistic decision making possible
- ▶ Stage of developement
  - ▶ Instabilities in numerical execution by R inla
  - ▶ Efficient update procedure needed
  - ▶ Superior performance compared to established methods could not be shown
- ▶ Potential for improvements and further developements

Farrington, C. P., N. J. Andrews, A. D. Beale, and M. A. Catchpole (1996). A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. *Journal of the Royal Statistical Society Series A*(159), 547–563.

Heisterkamp, S. H., A. L. M. Dekkers, and J. C. M. Heijne (2006). Automated detection of infectious disease outbreaks: hierachical time series models. *Statistics in Medicine* (25), 4179–4196.

Höhle, M. (2007). `surveillance`: An R package for the monitoring of infectious diseases. *Computational Statistics 22*(4), 571–582.

Hutwagner, L., T. Browne, G. M. Seeman, and A. T. Fleischauer (2005, February). Comparing Aberration Detection Methods with Simulated Data. *Emerging Infectious Diseases 11*(2), 314–316.

Robert Koch-Institut (2005). RKI-Ratgeber Infektionskrankheiten: Campylobacter-Infektionen. Online: http://www.rki.de. Last access: 5 Febuary 2010.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society 71*(2), 1–35.

Stroup, D., M. Wharton, K. Kafadar, and A. G. Dean (1993). Evaluation of a Method for Detecting Aberrations in Public Health Surveillance Data. *American Journal of Epidemiology 137*(3), 373–380.