Control charts for the statistical surveillance of infectious diseases

Michael Höhle[1,2] Michaela Paul[3]

[1]Centre for Mathematical Sciences, Technische Universität München, Germany

[2]Department of Statistics, University of Munich, Germany

[3]Institute of Social and Preventive Medicine, University of Zurich, Switzerland

IBC 2008
Dublin, 15 July 2008

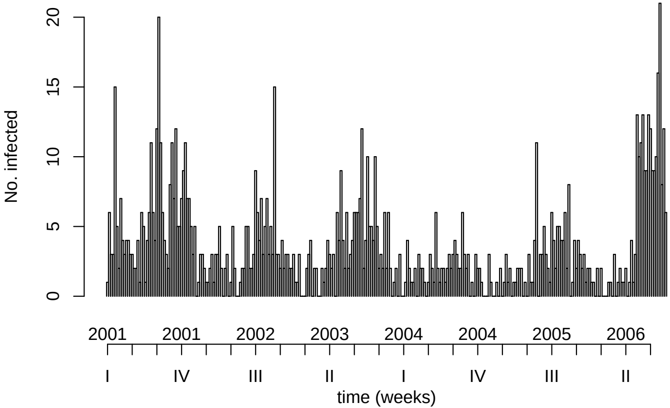
Outline

Motivation (1)

- *On-line* detection of changepoints in time series of counts originating from public health surveillance of infectious diseases
- Combine ideas from statistical process control (SPC) and generalized linear models (GLM) to develop a detector which takes the seasonal variation in surveillance data into account
- Encourage use by providing efficient implementation within the R-package `surveillance` (H., 2007) available from the Comprehensive R Archive Network (CRAN)

Motivation (2) - *Salmonella hadar* cases in Germany

- During 2006 the German health authorities noted an increased number of cases due to *salmonella hadar*



- Plot shows weekly number of cases in 2001-2006

CUSUM likelihood ratio detectors (1)

Assume that change-point τ is given as follows:

$$y_t|z_t, \tau \sim \begin{cases} f_{\theta_0}(\cdot|z_t) & \text{for } t = 1, \ldots, \tau - 1 \text{ (in-control)} \\ f_{\theta_1}(\cdot|z_t) & \text{for } t = \tau, \tau + 1, \ldots \text{ (out-of-control)} \end{cases}$$

where z_t denotes known covariates at time t and f_θ is e.g. the negative binomial or the Poisson probability function parametrized by θ.

Likelihood ratio (LR) based stopping time

$$N = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} \left[\sum_{t=k}^{n} \log \left\{ \frac{f_{\theta_1}(y_t|z_t)}{f_{\theta_0}(y_t|z_t)} \right\} \right] \geq c_\gamma \right\}.$$

Seasonal count data GLM (1)

Let the seasonal count data GLM be

$$y_t \sim \mathrm{NegBin}(\mu_{0,t}, \alpha),$$

where the limit $\alpha \to 0$ corresponds to $\mathrm{Po}(\mu_{0,t})$. Furthermore,

$$\begin{aligned} \log \mu_{0,t} &= \beta_0 + \beta_1 t + g(t), \\ \log \mu_{1,t} &= \log \mu_{0,t} + \kappa \end{aligned}$$

Here, $g(t)$ is a function with period T, e.g.

$$g(t) = \sum_{s=1}^{S} \left(\gamma_s \cos(\omega_s t) + \delta_s \sin(\omega_s t) \right),$$

with $\omega_s = \frac{2\pi}{T} s$. Alternatively, $g(t)$ could be a T-periodic spline.

Generalized likelihood ratio detector (1)

- A problem of the LR scheme is that detection is only optimal for pre-specified θ_1.
- Generalization:

Generalized likelihood ratio (GLR) based stopping rule

$$N_G = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} \sup_{\theta_1 \in \Theta_1} \left[\sum_{t=k}^{n} \log \left\{ \frac{f_{\theta_1}(y_t | z_t)}{f_{\theta_0}(y_t | z_t)} \right\} \right] \geq c_\gamma \right\}$$

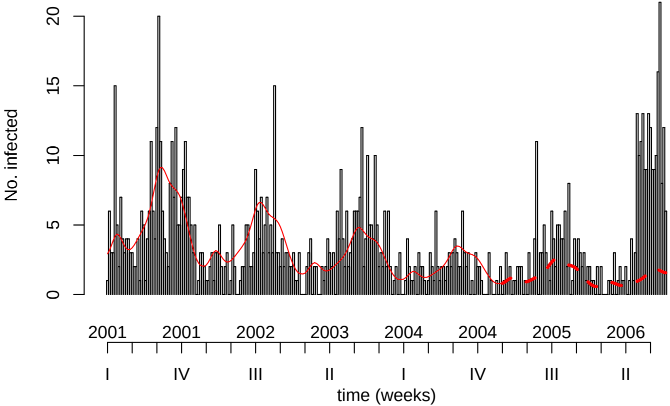
- No recursive updating as in LR-CUSUM possible: worst case number of operations to determine if $N_G \leq m$ is $O(m^3)$

Generalized likelihood ratio detector (2)

- For the Poisson case efficient computations are possible because necessary MLE is analytically available
- For the NegBin case the MLE $\hat{\kappa}_{n,k}$ has to be found by e.g. Newton-Raphson iterations
- Use $\hat{\kappa}_{n,k+1}$ as starting value for the computation of $\hat{\kappa}_{n,k}$
- Speedup the GLR detector by using a *window-limited* approach as proposed by Willsky and Jones (1976). Maximization only for a moving window of $k \in \{n-M, \ldots, n\}$, where $M \geq 1$

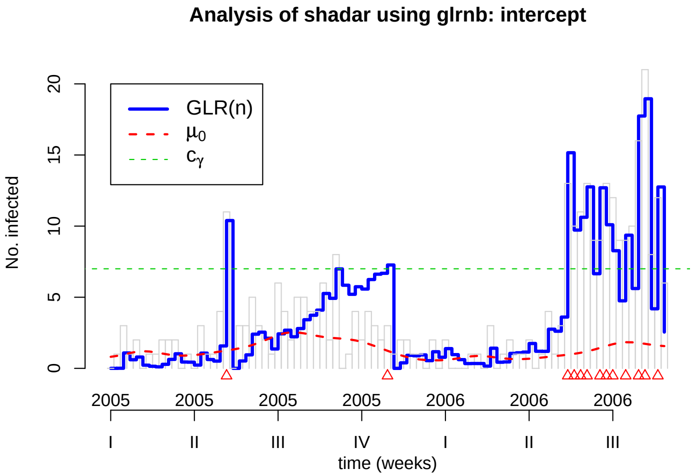
Applying the GLR detector to salmonella hadar (1)

- A seasonal negative binomial GLM is fitted to the training period



- The fitted model is used to predict $\mu_{0,t}$ of the test period

Applying the GLR detector to salmonella hadar (2)



Comparison with two existing methods

- We performed a comparison with two existing surveillance methods for count data time series having time varying mean (Rossi et al., 1999; Rogerson and Yamada, 2004)
- However, these methods require pre-specified κ, where detection should be optimal ($\kappa = 0.69$ was chosen)
- Simulation study with 4 different seasonal means: between 0.13-5.00 cases per week with different amplitudes of the harmonics
- Comparison of average run length (ARL), i.e. $ARL_0 = \mathrm{E}(N|\tau = \infty)$ and $ARL_1 = \mathrm{E}(N|\tau = 0)$, for various sizes of the true excess in out-of-control state

(a) Rossi: $ARL_0 = 500$ detector for $2\mu_{0,t}$

	$\delta = 1.0$	$\delta = 1.2$	$\delta = 1.4$	$\delta = 1.7$	$\delta = 2.0$	$\delta = 2.5$
μ_t^a	114 (1.4)	73.2 (0.76)	52.3 (0.62)	26.0 (0.37)	12.1 (0.19)	5.1 (0.06)
μ_t^b	307 (3.9)	95.0 (0.89)	47.3 (0.59)	16.3 (0.23)	7.4 (0.09)	3.7 (0.03)
μ_t^c	291 (3.7)	81.0 (0.84)	28.2 (0.41)	6.3 (0.09)	3.4 (0.02)	2.1 (0.01)
μ_t^d	418 (5.4)	49.7 (0.63)	10.5 (0.15)	3.5 (0.02)	2.2 (0.01)	1.5 (0.01)

(b) Rogerson: $ARL_0 = 500$ detector for $2\mu_{0,t}$

	$\delta = 1.0$	$\delta = 1.2$	$\delta = 1.4$	$\delta = 1.7$	$\delta = 2.0$	$\delta = 2.5$
μ_t^a	578 (7.2)	125.1 (0.91)	69.7 (0.76)	25.5 (0.36)	10.8 (0.16)	5.2 (0.04)
μ_t^b	505 (6.4)	94.0 (0.87)	36.3 (0.44)	12.2 (0.15)	6.6 (0.05)	3.9 (0.02)
μ_t^c	537 (6.9)	86.7 (0.87)	27.2 (0.38)	6.3 (0.08)	3.6 (0.02)	2.3 (0.01)
μ_t^d	496 (6.4)	57.8 (0.68)	13.0 (0.17)	3.6 (0.03)	2.2 (0.01)	1.4 (0.01)

(c) GLR: $\widehat{ARL}_0 = 500$ detector with $c_\gamma = (4.7, 4.95, 4.98, 5.08)$

	$\delta = 1.0$	$\delta = 1.2$	$\delta = 1.4$	$\delta = 1.7$	$\delta = 2.0$	$\delta = 2.5$
μ_t^a	533 (6.9)	124.0 (0.92)	69.3 (0.72)	26.3 (0.36)	11.2 (0.17)	5.0 (0.05)
μ_t^b	509 (6.4)	92.1 (0.82)	35.1 (0.38)	12.6 (0.14)	6.8 (0.06)	3.9 (0.03)
μ_t^c	506 (6.4)	70.3 (0.72)	20.2 (0.26)	5.7 (0.05)	3.5 (0.02)	2.2 (0.01)
μ_t^d	480 (6.2)	30.5 (0.32)	8.2 (0.07)	3.6 (0.02)	2.3 (0.01)	1.5 (0.01)

Beyond the seasonal count data detector (1)

- More involved models for the shift are possible, e.g. the auto-regressive epidemic model from Held et al. (2005)

$$\mu^e_{1,t} = \mu_{0,t} + \lambda y_{t-1}, \quad t > 1,$$

where $\lambda > 0$ and $\mu^e_{1,1} = \mu_{0,1}$

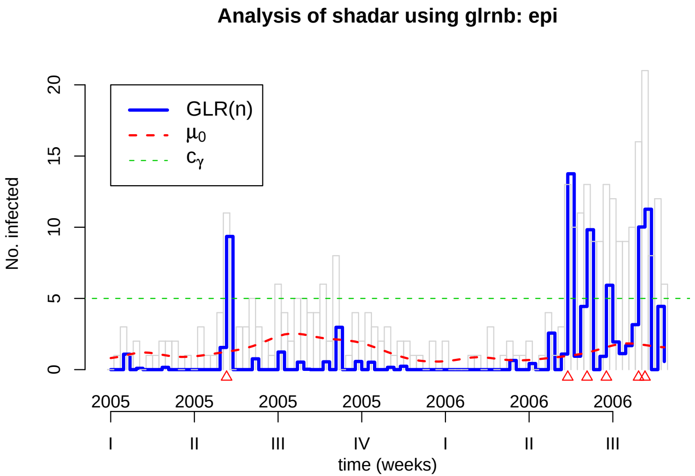
- Multivariate count data time series extension: $y_{it} \sim \text{NegBin}(\mu_{0,it}, \alpha)$, $i = 1, \ldots, m$ with

$$\log \mu_{0,it} = \beta_{0i} + \beta_{1i} t + g(t)$$

and $\log \mu_{1,it} = \log \mu_{0,it} + \kappa_i$

Beyond the seasonal count data detector (2)

Autoregressive model with $M = 26$ and $c_\gamma = 5 \Rightarrow$
$P(\tilde{N}_G < 3 \cdot 52 | \tau = \infty) = 0.049$



Discussion

- Combination of SPC and classical GLMs to obtain changepoint detector for count time series documented in H. and Paul (2008).
- Model for $\mu_{0,t}$ is crucial.
- We are working on adapting ideas to the binomial distribution, e.g. for the monitoring of varicella sentinel data

Literature I

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance data. *Statistical Modelling*, 5:187–199.

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

Höhle, M. and Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Computational Statistics and Data Analysis*, 52(9):4357–4368.

Rogerson, P. and Yamada, I. (2004). Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report*, 53:79–85.

Rossi, G., Lampugnani, L., and Marchi, M. (1999). An approximate CUSUM procedure for surveillance of health events. *Statistics in Medicine*, 18:2111–2122.

Willsky, A. and Jones, H. (1976). Generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic control*, 21:108–112.