Surveillance of infectious diseases using likelihood ratio detectors

Michael Höhle[1,2]

[1]Department of Statistics
University of Munich

[2]Munich Center of Health Sciences
University of Munich

Recent Advances in the Statistical Analysis of Count and Survival Data
Zurich, 25-26 September, 2007

Outline

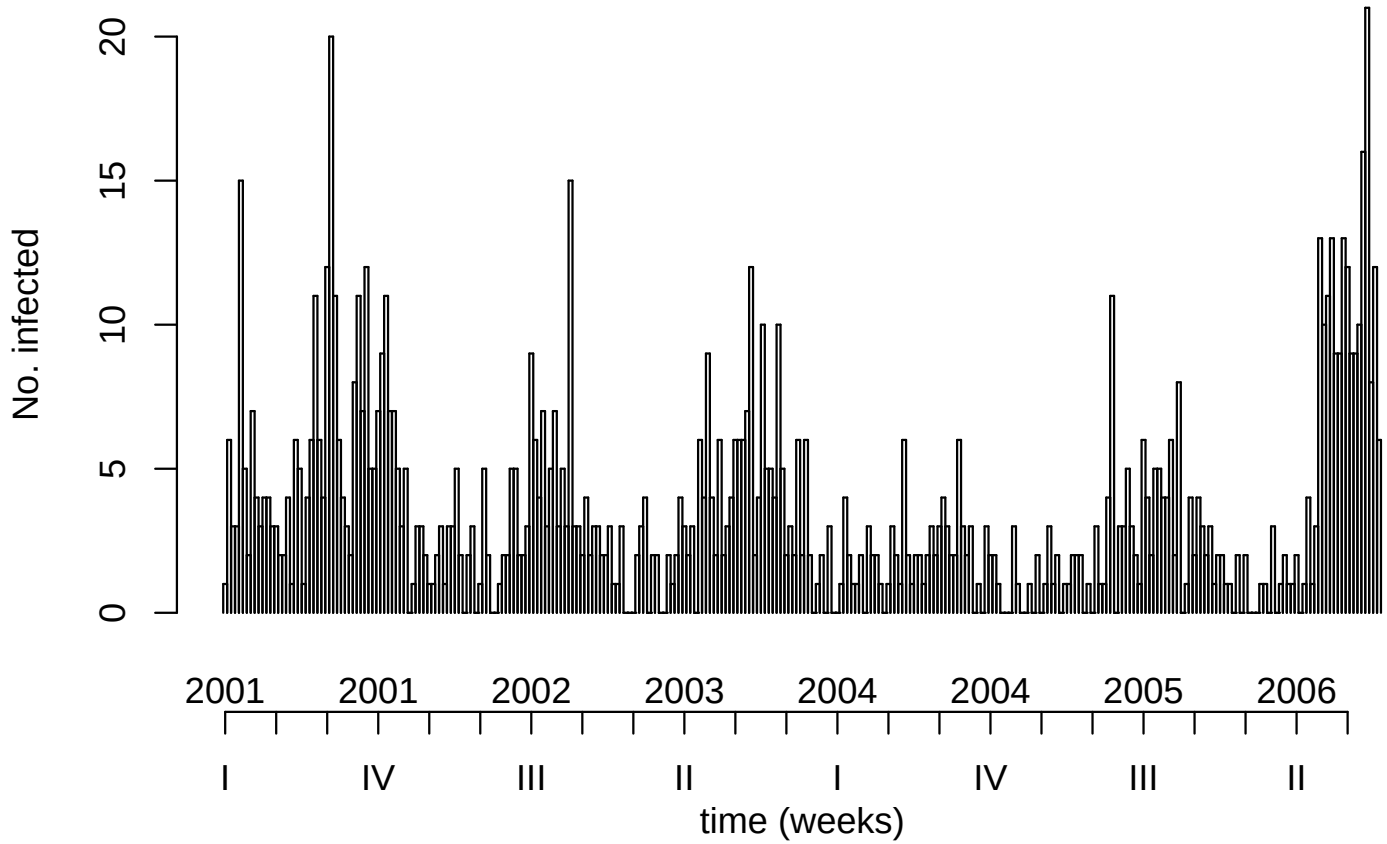
1. Motivation: statistical surveillance of infectious diseases
2. Outbreak detection based on likelihood ratio detectors
3. Evaluating performance
4. Possible extensions of the proposed Poisson detector
5. Discussion

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` (Höhle, 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 2001-2006

Motivation (3)

- Classical surveillance algorithms deal with seasonality by creating predictive intervals based on a subset of the historical values

  $\longrightarrow$ Using only a subset of the historical values is sub-optimal

- More orientation towards methods from SPC (CUSUM, EWMA, etc.)

  $\longrightarrow$ Process performance for *on-line* scheme more at focus

  $\longrightarrow$ Developed for the Gaussian case, but levels of aggregation and rare disease means low count

CUSUM likelihood ratio detectors (1)

Assume that given change-point τ

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

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

Likelihood ratio (LR) based stopping time

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

CUSUM likelihood ratio detectors (2)

With no covariates and pre-specified θ_0 and θ_1 the stopping rule can be written in recursive form:

Cumulative Sum (CUSUM)

$$l_0 = 0, \quad l_n = \max\left(0, l_{n-1} + \log\left\{\frac{f_{\theta_1}(y_n)}{f_{\theta_0}(y_n)}\right\}\right), \; n \geq 1$$

with stopping-rule $N = \inf\{n : l_n \geq c_\gamma\}$.

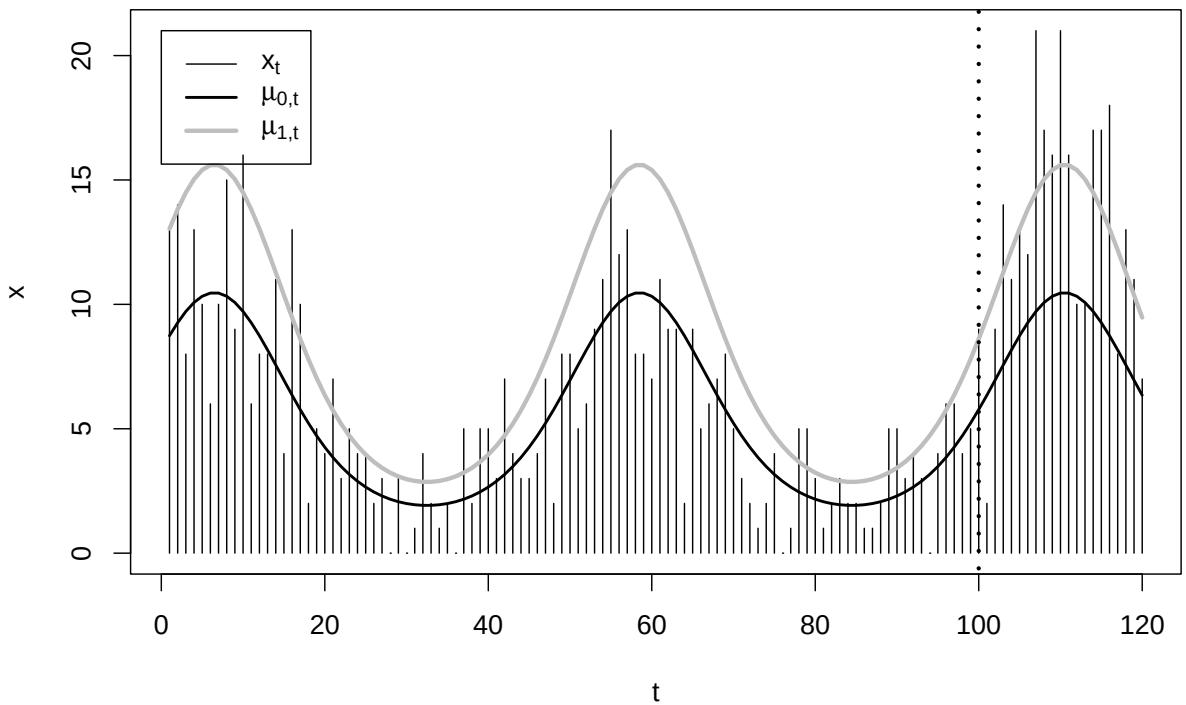
- To determine if $N \leq m$ takes at most $O(m)$ operations
- The CUSUM detector is optimal (in some technical sense) for the detection from θ_0 to θ_1.
- CUSUM for Poisson distribution described by Lucas (1985)

Seasonal Poisson GLM

What if in-control observations originate from seasonal Poisson GLM, i.e. $y_t \sim \text{Po}(\mu_{0,t})$ with

$$\log \mu_{0,t} = \alpha + \beta t + \sum_{s=1}^{S} \Big(\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t) \Big)$$

and $\omega_s = \frac{2\pi}{T} s$ with period T?



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 CUSUM possible: worst case number of operations to determine if $N_G \leq m$ is $O(m^3)$
- Lai and Shan (1999) show for the Gaussian case how it is possible to reduce this complexity by recursive least squares and clever treatment of the sums and sups

Generalized likelihood ratio detector (2)

The GLR detector rephrased:

$$
\begin{aligned}
l_{n,k} &= \sup_{\theta_1 \in \Theta_1} \left[\sum_{t=k}^{n} \log \left\{ \frac{f_{\theta_1}(y_t|z_t)}{f_{\theta_0}(y_t|z_t)} \right\} \right] \\
&= \left[\sup_{\theta_1 \in \Theta_1} \sum_{t=k}^{n} \log f_{\theta_1}(y_t|z_t) \right] - \left[\sum_{t=k}^{n} \log f_{\theta_0}(y_t|z_t) \right] \\
&= \sum_{t=k}^{n} \log \left\{ \frac{f_{\hat{\theta}_{n,k}}(y_t|z_t)}{f_{\theta_0}(y_t|z_t)} \right\},
\end{aligned}
$$

where $\hat{\theta}_{n,k} = \arg\sup_{\theta_1 \in \Theta_1} \sum_{t=k}^{n} \log f_{\theta_1}(y_t|z_t)$. Now $GLR(n) = \max_{1 \le k \le n} l_{n,k}$

GLR detector (3) – recursive computations

- For Poisson case with $\log \mu_{1,t} = \log \mu_{0,t} + \kappa$, efficient computations are possible as necessary MLE is analytically available:

$$\hat{\kappa}_{n,k} = \log\left(\frac{\sum_{t=k}^{n} y_t}{\sum_{t=k}^{n} \mu_{0,t}}\right).$$

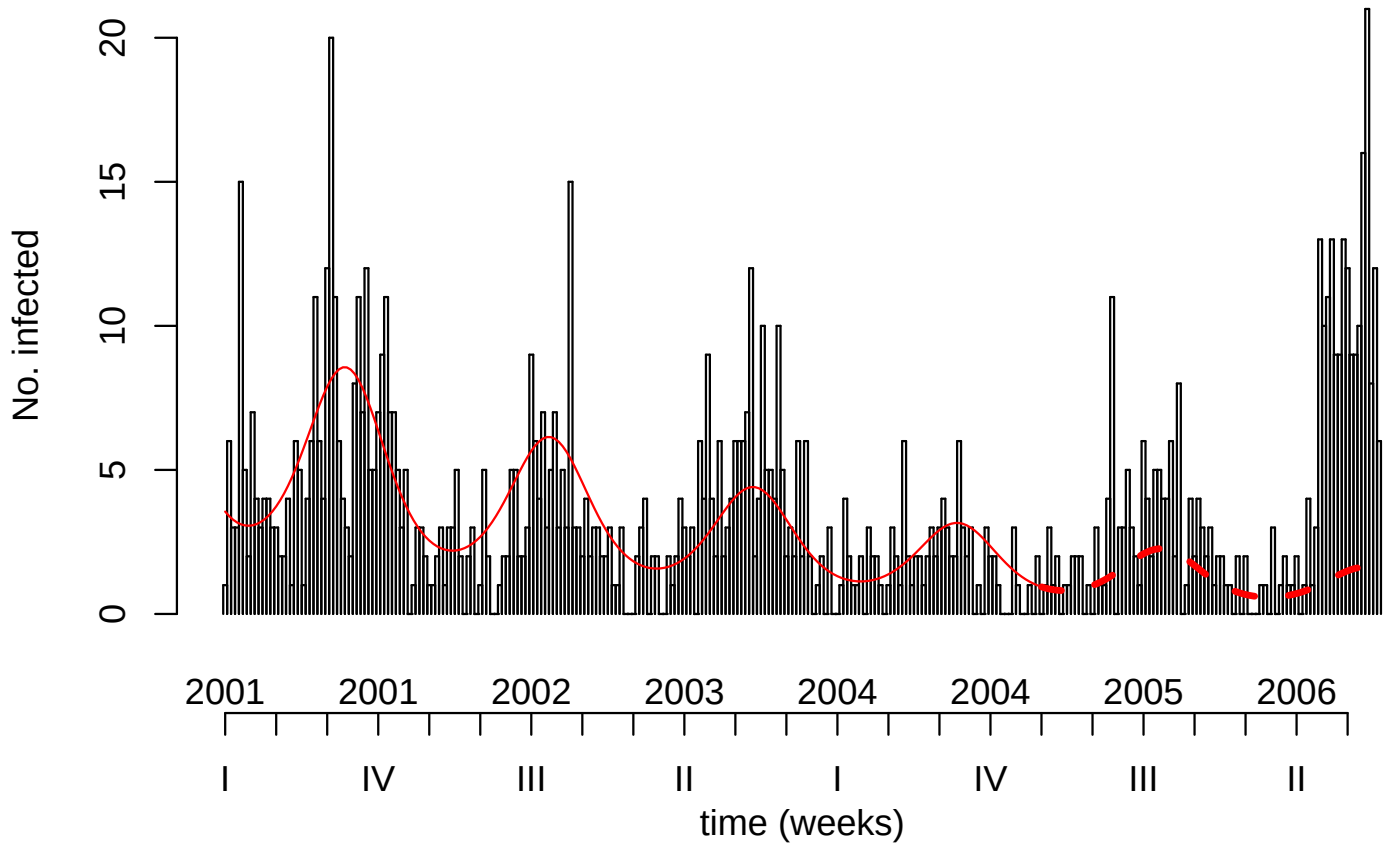
- Inserting $\theta = \kappa$ yields,

$$\begin{aligned} l_{n,k} &= \sup_{\theta \in \Theta} \sum_{t=k}^{n} \log \frac{f_\theta(y_t|z_t)}{f_{\theta_0}(y_t|z_t)} \\ &= \hat{\kappa}_{n,k} \sum_{t=k}^{n} y_t + (1 - \exp(\hat{\kappa}_{n,k})) \sum_{t=k}^{n} \mu_{0,t}. \end{aligned}$$

- By recursively computing $\sum_{t=k}^{n} y_t$ and $\sum_{t=k}^{n} \mu_{0,t}$, an efficient computation of $\hat{\kappa}_{n,k}$ and $l_{n,k}$ is available.

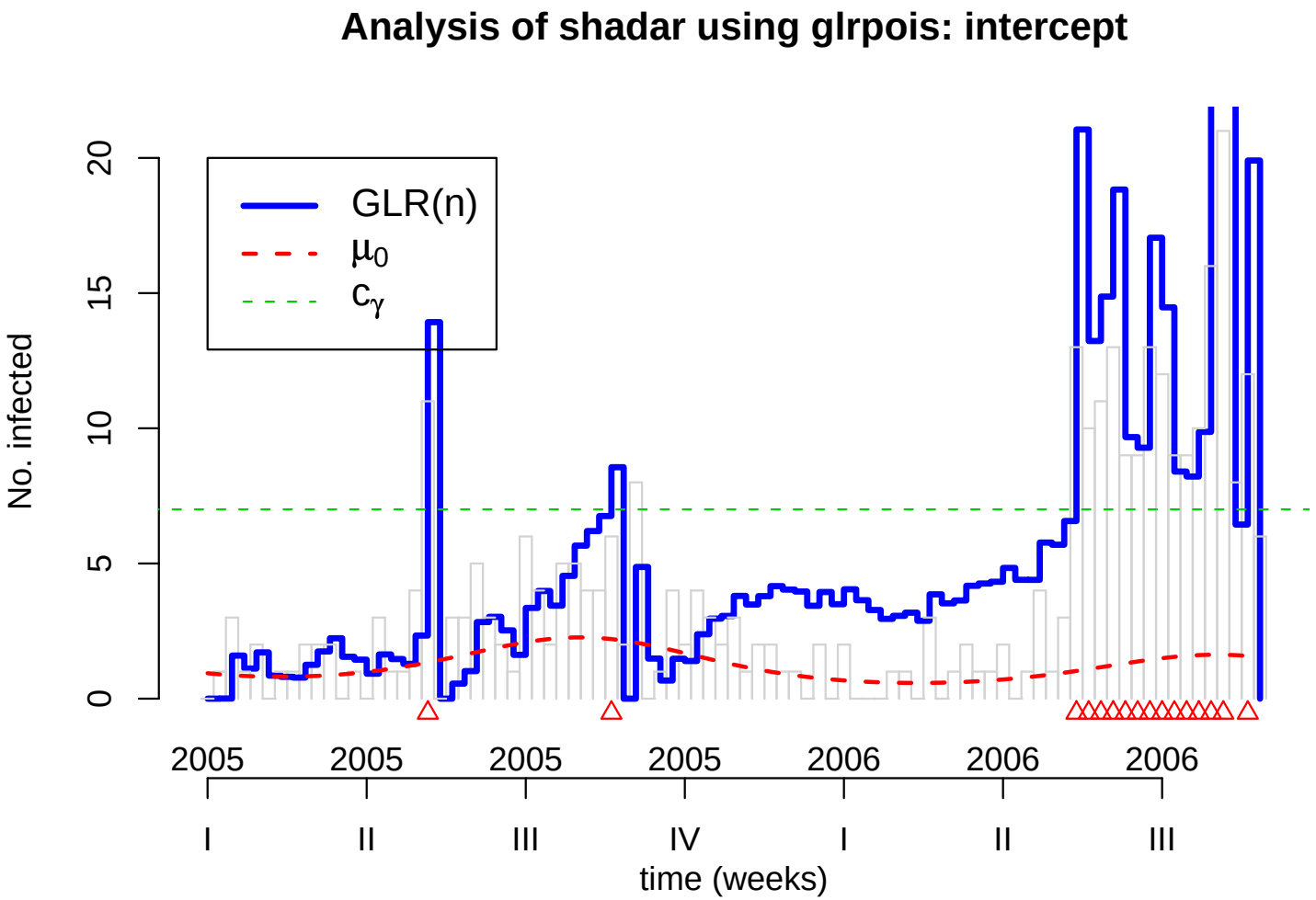
Applying the GLR detector to salmonella hadar (1)

- A $S = 1$ seasonal Poisson 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)



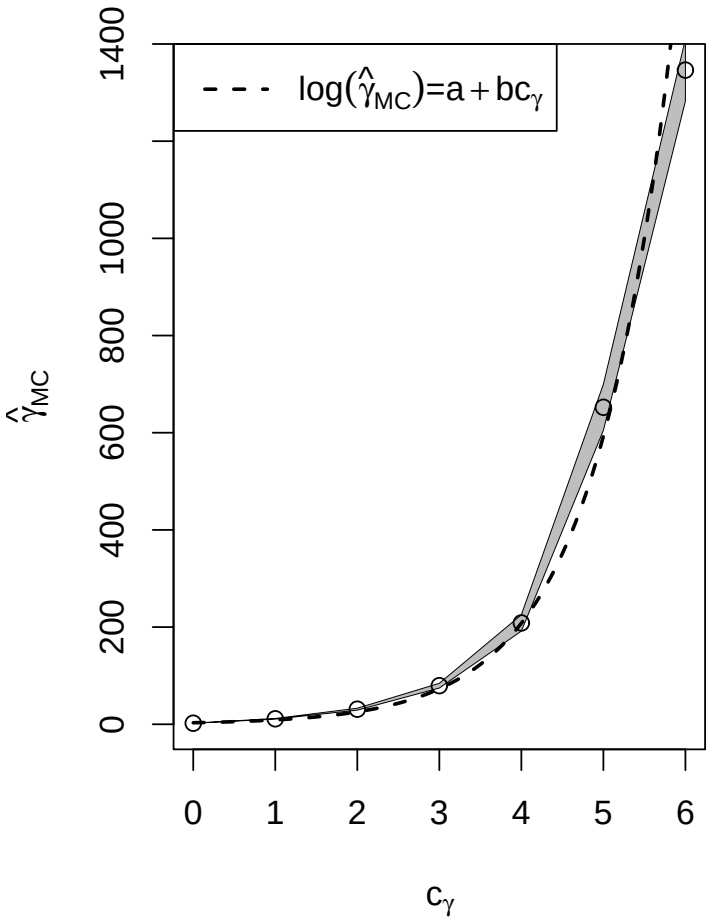
Evaluating performance of a surveillance algorithm

Choice of threshold in surveillance algorithms should be based on a performance measure:

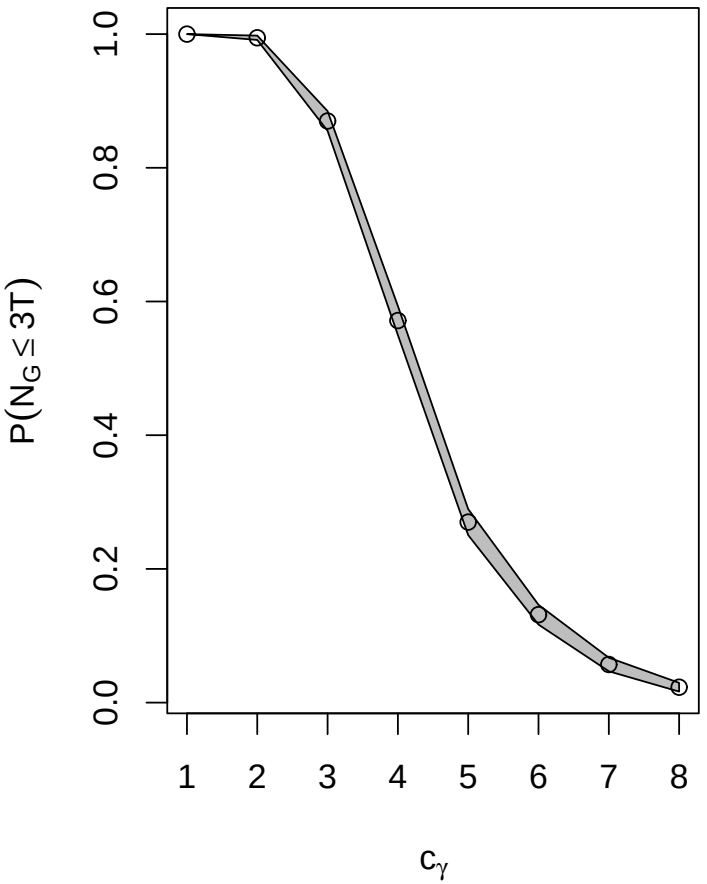
- Location parameters of the run length distribution, e.g. the average run lengths (ARLs) $ARL_0 = \mathrm{E}(N|\tau = \infty)$ or $ARL_1 = \mathrm{E}(N|\tau = 0)$.
- Conditional expected delay $E(N - \tau|\tau, N \geq \tau)$
- Probability of false alarm within first m time points, i.e. $P(N \leq m|\tau = \infty)$.
- Sensitivity, Specificity, ROC-Curves

Above criteria for a detector are rarely available as closed formulas $\rightarrow$ use Monte Carlo sampling instead

Average run length and probability of false alarm



(a) Monte-Carlo estimated $ARL(c_\gamma)$



(b) $P(N_G \leq 3T|\tau = \infty)$

Other approaches for seasonal count data (1)

- Rossi et al. (1999) suggest a Poisson CUSUM for time varying mean data by transformation to normality

$$x_t = \frac{y_t - 3\mu_{0,t} + 2\sqrt{\mu_{0,t} \cdot y_t}}{2\sqrt{\mu_{0,t}}}$$

  and applying a Gaussian CUSUM to the x_t's
- Chart performance is controlled by desired values for in-control mean ARL_0 and the out-of-control mean ARL_1
- Threshold c_γ and θ_1 selected by algorithms to determine ARL for iid. standard normal CUSUM

Other approaches for seasonal count data (2)

- Rogerson and Yamada (2004) compute time-varying parameters of the Poisson CUSUM to keep in-control ARLs fixed:

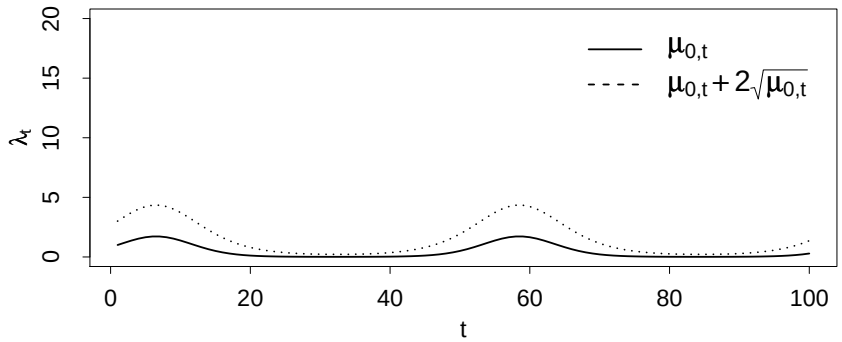
$$S_t = \max\{0, S_{t-1} + c_t(y_t - k_t)\}, \quad \text{with}$$

$$k_t = \frac{\mu_{1,t} - \mu_{0,t}}{\log(\mu_{1,t}) - \log(\mu_{0,t})},$$

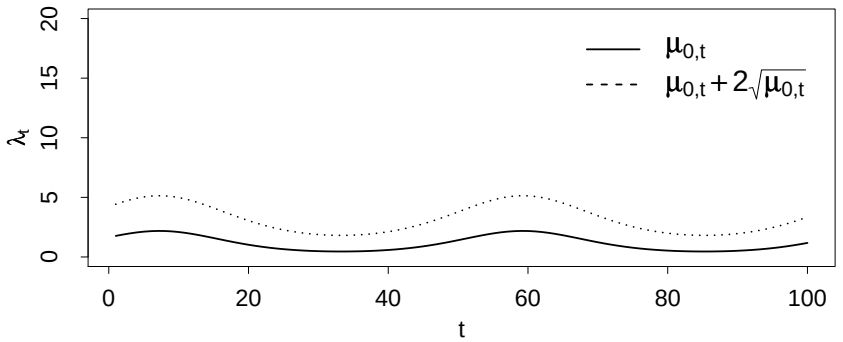
  with $\mu_{1,t} = \mu_{0,t} + s\sqrt{\mu_{0,t}}$ and $c_t = h/h_t$ scales the contribution of $(y_t - k_t)$.

- The decision interval h_t is determined at each time point as the decision interval of a Poisson CUSUM with reference value k_t having ARL_0.

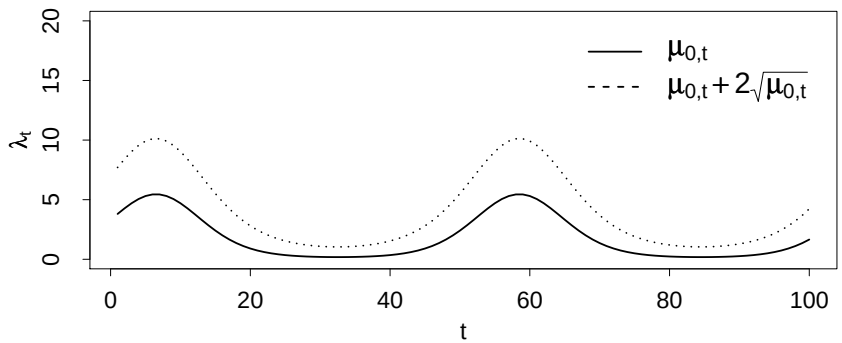
ARL Comparison on 4 simulated setups



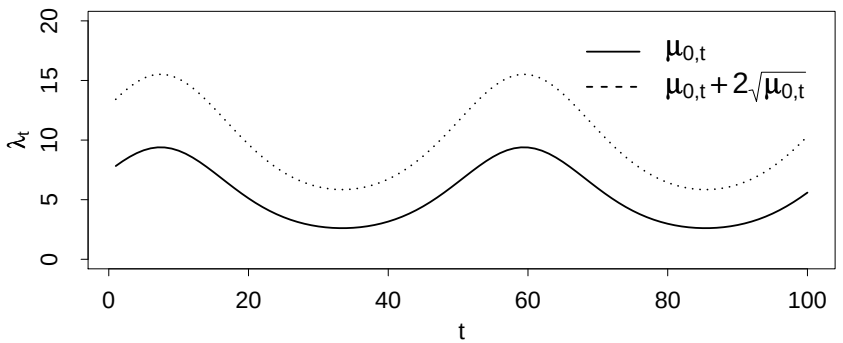
(a) μ_t^a: $\alpha = -2$, $\beta_1 = 1.8$, $\beta_2 = 1.8$



(b) μ_t^b: $\alpha = 0$, $\beta_1 = 0.6$, $\beta_2 = 0.5$



(c) μ_t^c: $\alpha = 0$, $\beta_1 = 1.2$, $\beta_2 = 1.2$



(d) μ_t^d: $\alpha = 1.6$, $\beta_1 = 0.5$, $\beta_2 = 0.4$

Data are $y_t \sim \text{Po}(\mu_t + \delta\sqrt{\mu_t})$, $\delta = \{0.0, 0.5, 1.0, 1.5, 2.0, 2.5\}$

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

	$\delta = 0.0$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$	$\delta = 2.5$
μ_t^a	113 (3.1)	25.7 (0.54)	11.9 (0.26)	6.4 (0.14)	4.0 (0.07)	3.1 (0.05)
μ_t^b	321 (9.1)	39.3 (0.97)	12.7 (0.33)	6.1 (0.13)	4.0 (0.07)	2.9 (0.05)
μ_t^c	284 (8.3)	41.3 (1.02)	12.7 (0.31)	5.9 (0.12)	3.6 (0.06)	2.7 (0.04)
μ_t^d	416 (11.6)	49.7 (1.24)	12.8 (0.30)	5.6 (0.11)	3.6 (0.06)	2.6 (0.04)

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

	$\delta = 0.0$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$	$\delta = 2.5$
μ_t^a	654 (17.4)	43.4 (1.12)	13.8 (0.37)	6.6 (0.15)	4.3 (0.07)	3.3 (0.05)
μ_t^b	576 (16.5)	45.1 (1.09)	12.6 (0.30)	6.2 (0.11)	4.3 (0.07)	3.2 (0.05)
μ_t^c	552 (15.6)	45.2 (1.14)	12.6 (0.31)	5.8 (0.11)	3.7 (0.06)	2.8 (0.04)
μ_t^d	526 (14.6)	48.3 (1.20)	12.0 (0.28)	5.5 (0.10)	3.6 (0.06)	2.6 (0.04)

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

	$\delta = 0.0$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$	$\delta = 2.5$
μ_t^a	519 (14.3)	34.1 (0.76)	12.0 (0.29)	6.1 (0.13)	4.1 (0.06)	3.2 (0.05)
μ_t^b	532 (15.2)	31.9 (0.64)	10.9 (0.21)	6.0 (0.10)	4.2 (0.07)	3.1 (0.05)
μ_t^c	492 (13.8)	33.2 (0.68)	11.1 (0.23)	5.7 (0.09)	3.7 (0.06)	2.8 (0.04)
μ_t^d	480 (13.0)	30.7 (0.60)	10.4 (0.18)	5.5 (0.09)	3.7 (0.05)	2.7 (0.04)

Beyond the seasonal Poisson detector (1)

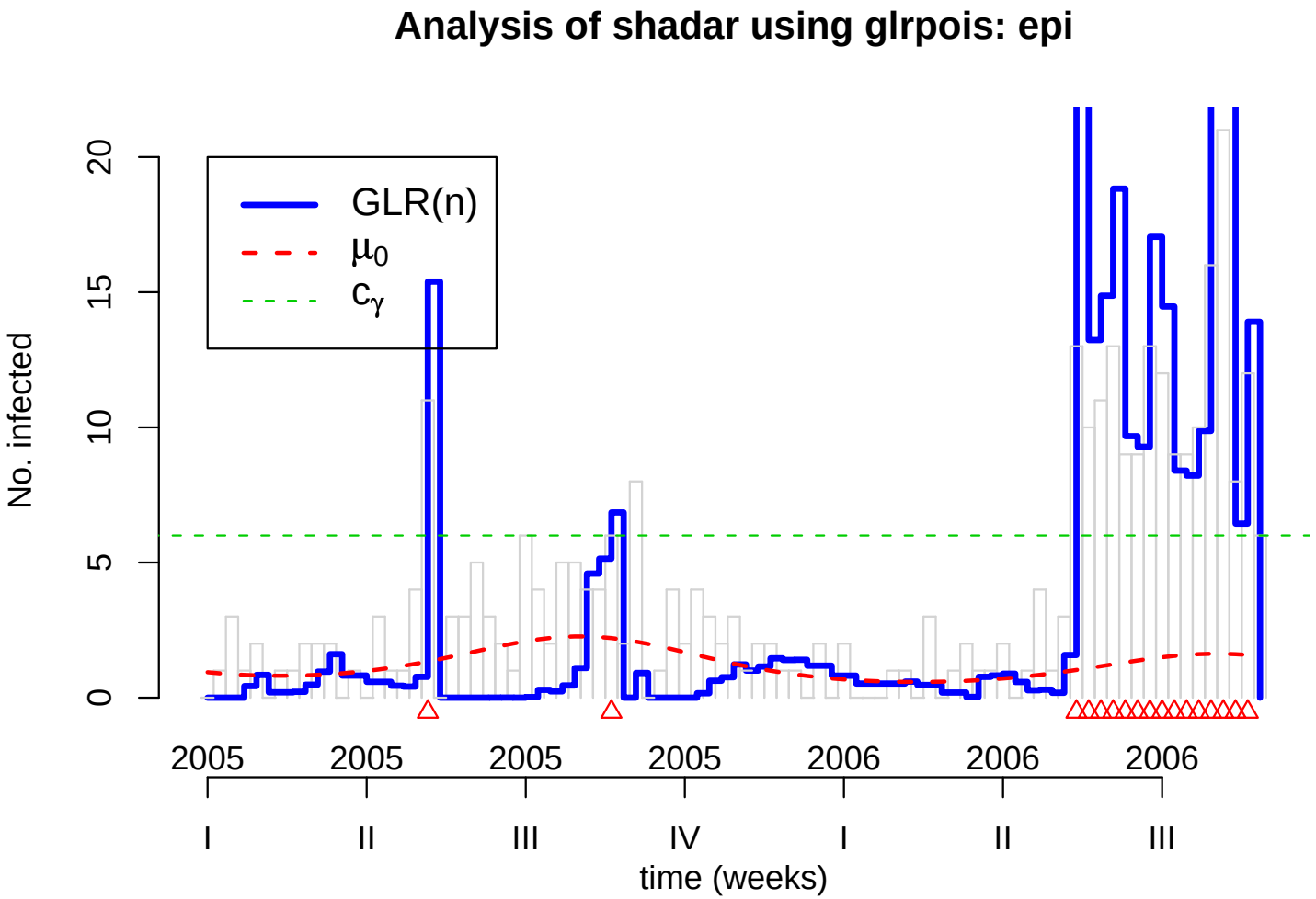
- More involved models for the shift are imaginable, e.g. the auto-regressive epidemic model from Held et al. (2005)

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

  where $\lambda > 0$ and $\mu^e_{1,1} = \mu_{0,1}$
- MLE only available by iterative procedure $\longrightarrow$ no fast recursive updating of GLR statistic $l_{n,k}$ possible
- Use $\lambda^{(0)}_{n,k} = \hat{\lambda}_{n,k+1}$ as starting value for $\hat{\lambda}_{n,k}$ computation
- Speedup by *window-limited GLR scheme* as proposed by Willsky and Jones (1976). Maximization only for a moving window of $k \in \{n - M, \ldots, n\}$, where $1 \leq M$

Beyond the seasonal Poisson detector (2)

With $M = 26$ and $c_\gamma = 6 \Rightarrow P(\tilde{N}_G < 3 \cdot 52 | \tau = \infty) = 0.042$



Beyond the seasonal Poisson detector (3)

- Extension to multivariate Poisson time series $y_{it} \sim \text{Po}(\mu_{0,it})$, $i = 1, \ldots, p$ with

$$\log \mu_{0,it} = \alpha_i + \beta_i t + \sum_{s=1}^{S} \Big(\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t) \Big)$$

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

- The components of the GLR statistic with $\boldsymbol{\theta} = (\kappa_1, \ldots, \kappa_p)$ are thus

$$l_{n,k} = \sup_{\boldsymbol{\theta} \in \Theta} \sum_{t=k}^{n} \log \left\{ \prod_{i=1}^{p} \frac{f_{\boldsymbol{\theta}}(y_{it}|z_{it})}{f_{\boldsymbol{\theta_0}}(y_{it}|z_{it})} \right\} = \sum_{t=k}^{n} \log \left\{ \prod_{i=1}^{p} \frac{f_{\hat{\boldsymbol{\kappa}}_{\mathbf{n,k,i}}}(y_{it}|z_{it})}{f_{\boldsymbol{\theta_0}}(y_{it}|z_{it})} \right\}$$

Discussion

- Nice combination of SPC and classical GLM to obtain changepoint detector for count time series
- Proposed Poisson GLR detector is rather specific
  - Dealing with overdispersion $\rightarrow$ negative binomial
  - For binomial distribution no fast recursive updating possible
  - Model for $\mu_{0,t}$ is crucial $\rightarrow$ self-starting CUSUMs
- Evaluation of run length, $P(N_G \leq m|\tau = \infty)$ and other criteria is only possible by Monte Carlo evaluation
  - tuning by control variates or importance sampling
- The paper accompanying this talk is (Höhle and Paul, 2007)

Acknowledgement

Persons:

- Michaela Paul, Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich
- Valentin Wimmer, Department of Statistics, University of Munich

Financial support:

- German Science Foundation (SFB386), 2003-2006

Literature I

Held, L., M. Höhle, and M. Hofmann (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*. Accepted for publication.

Höhle, M. and M. Paul (2007). Poisson regression charts for the monitoring of surveillance time series. Submitted.

Lai, T. and J. Shan (1999). Efficient recursive algorithms for detection of abrupt changes in signals and control systems. *IEEE Transactions on Automatic Control 44*, 952–966.

Lucas, J. M. (1985). Counted data CUSUMs. *Technometrics 27*(2), 129–144.

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

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

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