# Monitoring of Surveillance Time Series by Count Data Regression Charts

Michaela Paul[†] and Michael Höhle[*‡]

[†]ISPM, University of Zurich, Switzerland, [*]Department of Statistics, University of Munich, Germany, [‡] Zentrum Mathematik, Technische Universität München, Germany

email: `michaela.paul@ifspm.uzh.ch, hoehle@stat.uni-muenchen.de`

## Abstract

We perform on-line change point detection in time series of counts arising from infectious disease surveillance using newly developed regression charts for the Poisson and negative binomial distribution. The in-control mean is assumed to be time-varying and linear on the log-scale with intercept and seasonal components. Using the generalized likelihood ratio (GLR) statistic a monitoring scheme is formulated to detect on-line whether a shift in the intercept occurred. In case of Poisson the necessary quantities of the GLR detector can be efficiently computed by recursive formulas. Extensions to more general alternatives e.g. containing an autoregressive epidemic component are discussed. A comparison of the Poisson scheme with existing methods is done using Monte Carlo simulations. The practicability of the methods is demonstrated by applying them to the observed number of salmonella hadar cases in Germany.

## 1 Introduction

Assume that the observations $x_1, x_2, \dots$ originate from some parametric distribution with density $f_\theta$ such that given the change-point $\tau$ the observations are independent realizations

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

- Our interest is to determine $\tau$ *on-line* – i.e. new observations are collected until one is convinced that a change has occurred.
- We will assume that $f$ is the negative binomial probability mass function with mean $\mu_t$ and a known dispersion parameter.
- A *stopping rule* is used to determine when enough evidence against $H_0 : \mu_t = \mu_{0,t}, t = 1, \dots$ has been collected to stop the sampling.

The out-of-control situation is characterized by a multiplicative shift

$$\mu_{1,t} = \mu_{0,t} \cdot \exp(\kappa) \qquad (1)$$

with $\log(\mu_{0,t})$ specified by cyclic regression.

More involved models for the shift are imaginable, one suggestion is

$$\mu^e_{1,t} = \mu_{0,t} + \lambda x_{t-1}, \quad t > 1, \qquad (2)$$

where $\lambda > 0$ and $\mu^e_{1,1} = \mu_{0,1}$. This auto-regressive model is discussed in Paul et al. (2008) as a model for time series of counts from infectious diseases.

## 2 The Generalized Likelihood Ratio (GLR) detector

The generalized likelihood ratio (GLR) statistic is defined as

$$GLR(n) = \max_{1 \le k \le n} \sup_{\theta \in \Theta} \left[ \sum_{t=k}^{n} \log \left\{ \frac{f_\theta(x_t|z_t)}{f_{\theta_0}(x_t|z_t)} \right\} \right]$$

with stopping rule $N_G = \inf \{ n \ge 1 : GLR(n) \ge c_\gamma \}$.

- Maximization of the log-likelihood has to be carried out over $\theta \in \Theta$ for each possible change time $k$ between 1 and $n$.
- Höhle and Paul (2008) show that this computational complexity can be reduced for the multiplicative shift (1) Poisson chart by clever recursive computations of the sums and sups.
- In the more general negative binomial setup or in setups with more flexible out-of-control models such as (2) this reduction is not possible.
- In case of large $n$ an alternative is the so-called *window-limited* GLR scheme where maximization is only performed for a moving window of $k$ values.
- Assuming a fixed out-of-control mean instead of computing the sup over all $\theta \in \Theta$ results in the LR detector.
- To investigate run-length properties of the GLR or LR scheme having time-changing means, Monte-Carlo sampling is necessary.

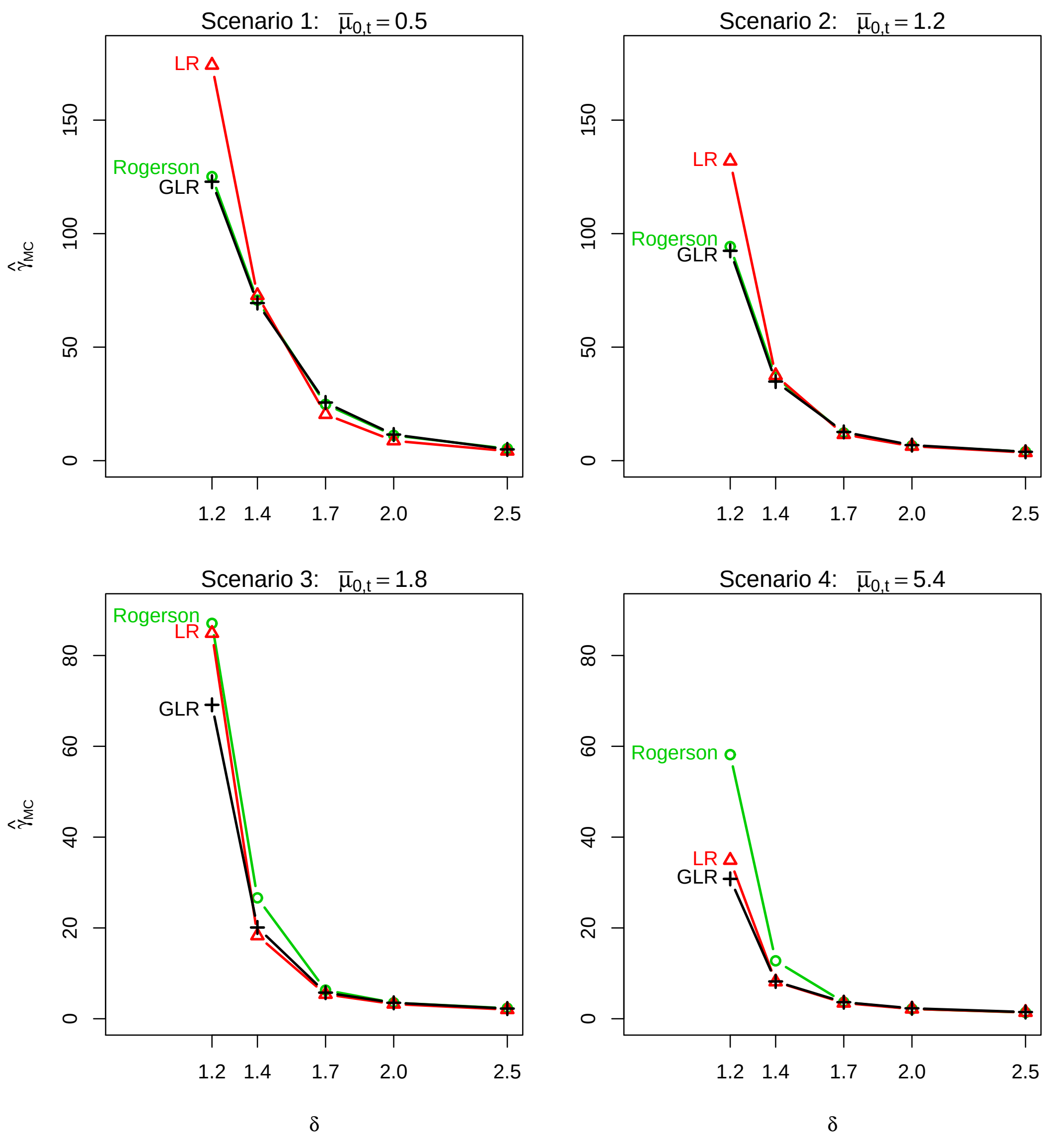
## 3 Comparison of time-varying mean Poisson detectors

The Poisson GLR-detector with multiplicative shift (1) is compared to

- the LR-detector,
- the modified Poisson CUSUM of Rogerson and Yamada (2004),
- the approximate Gaussian CUSUM of Rossi et al. (1999).

All three methods need an advance specification of which change to detect optimally.

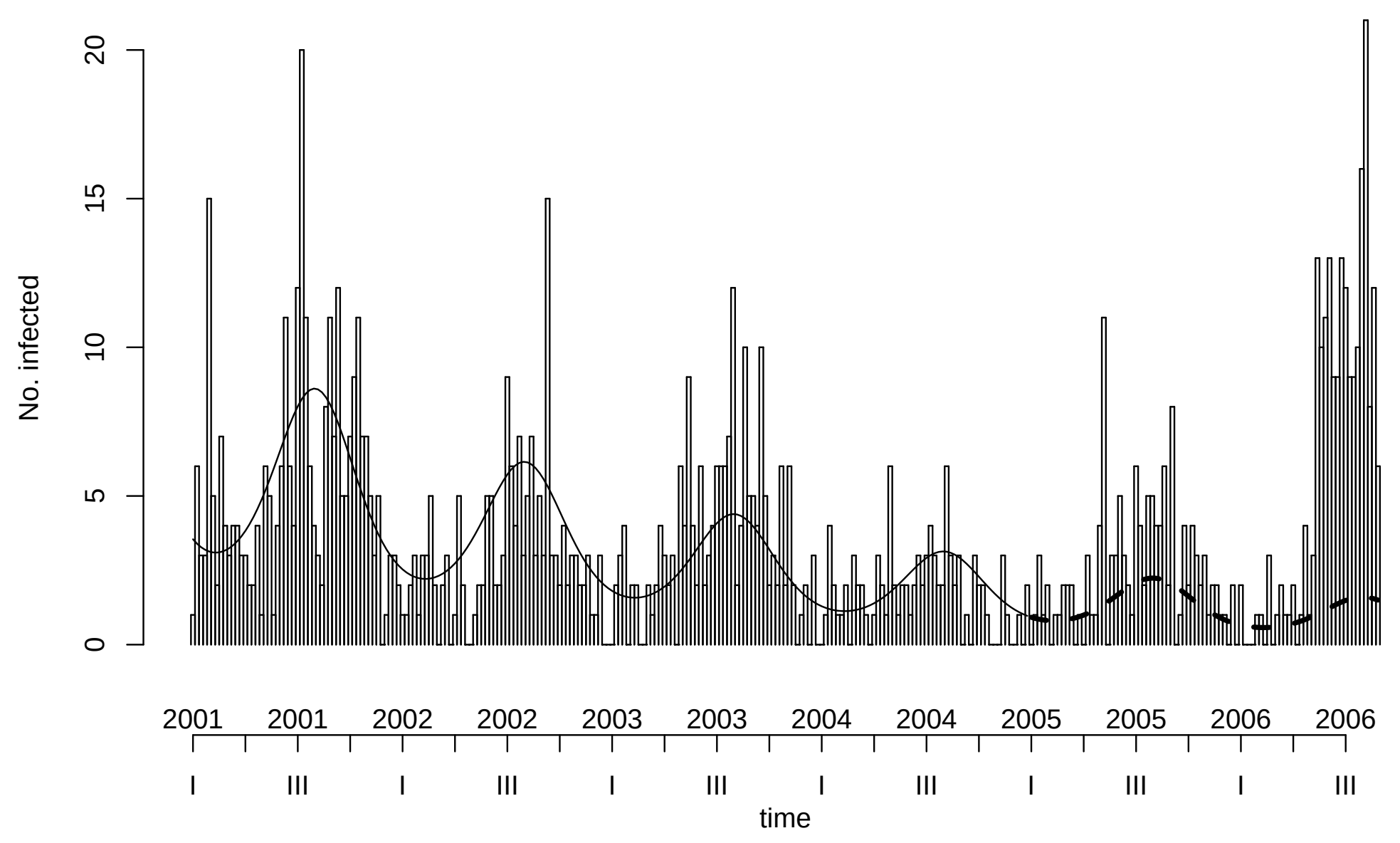
- We considered four scenarios with increasing mean and increasing seasonal amplitude of the monitored disease.
- The respective thresholds for all methods were selected to achieve an in-control average run length ($ARL_0$) of 500.
- The obtained $ARL_0$ using the method of Rossi et al. (1999) are much lower than the anticipated values whereas the other methods are doing a good job in obtaining the desired $ARL_0$.
- We estimated out-of-control ARLs for selected shift sizes $\delta$. The $ARL_1$ for the methods with fixed alternative hypothesis are sensitive to the choice of the shift size $\Delta$. The GLR detector performs best when the actual shift size $\delta$ is substantially lower.



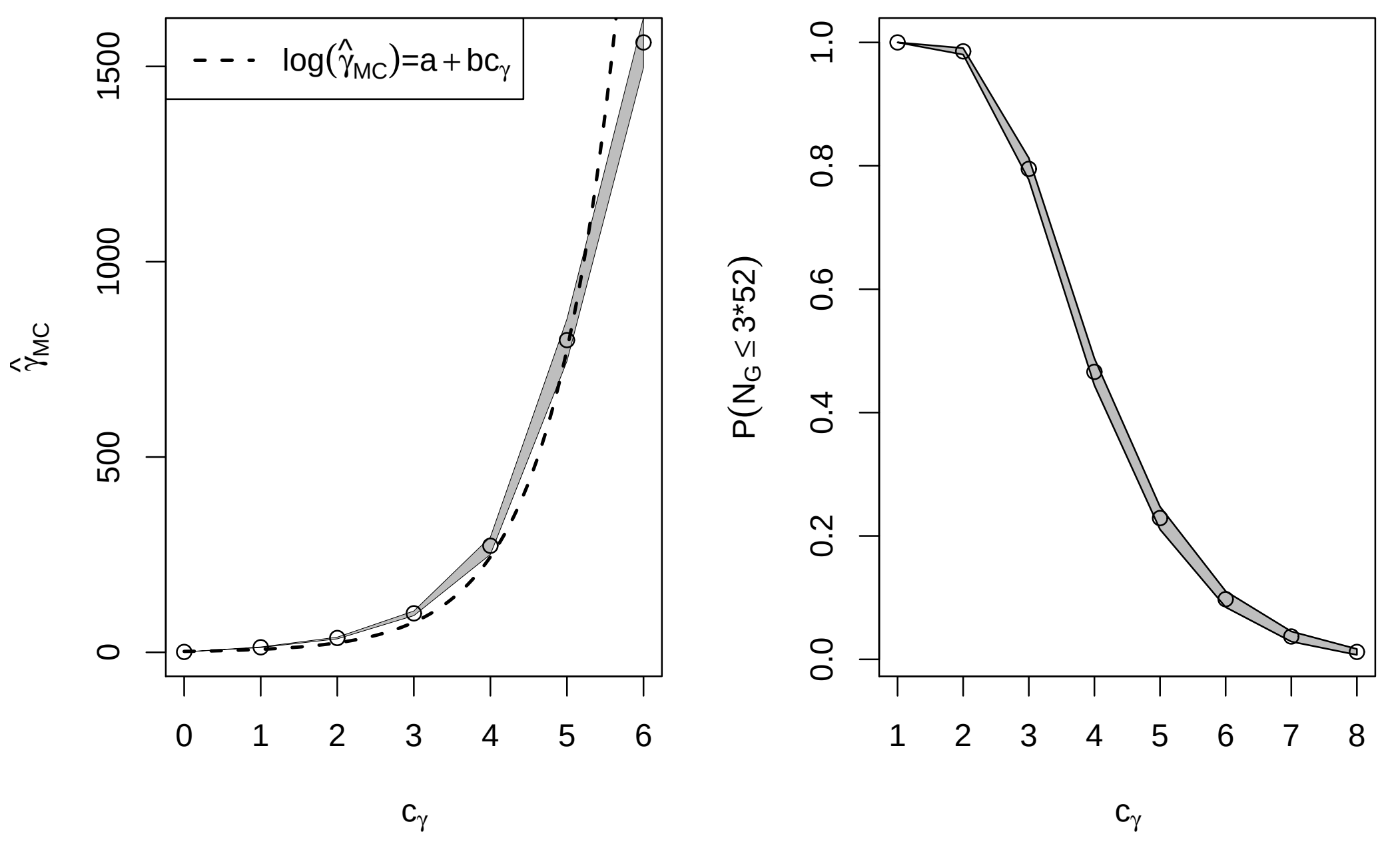
Monte Carlo estimated out-of-control ARLs for selected values of $\delta$ when $X_t \sim \text{Po}(\delta\mu_{0,t})$ and with the charts designed to detect a multiplicative shift of size $\Delta = 2$ .

## 4 Application to German salmonella data

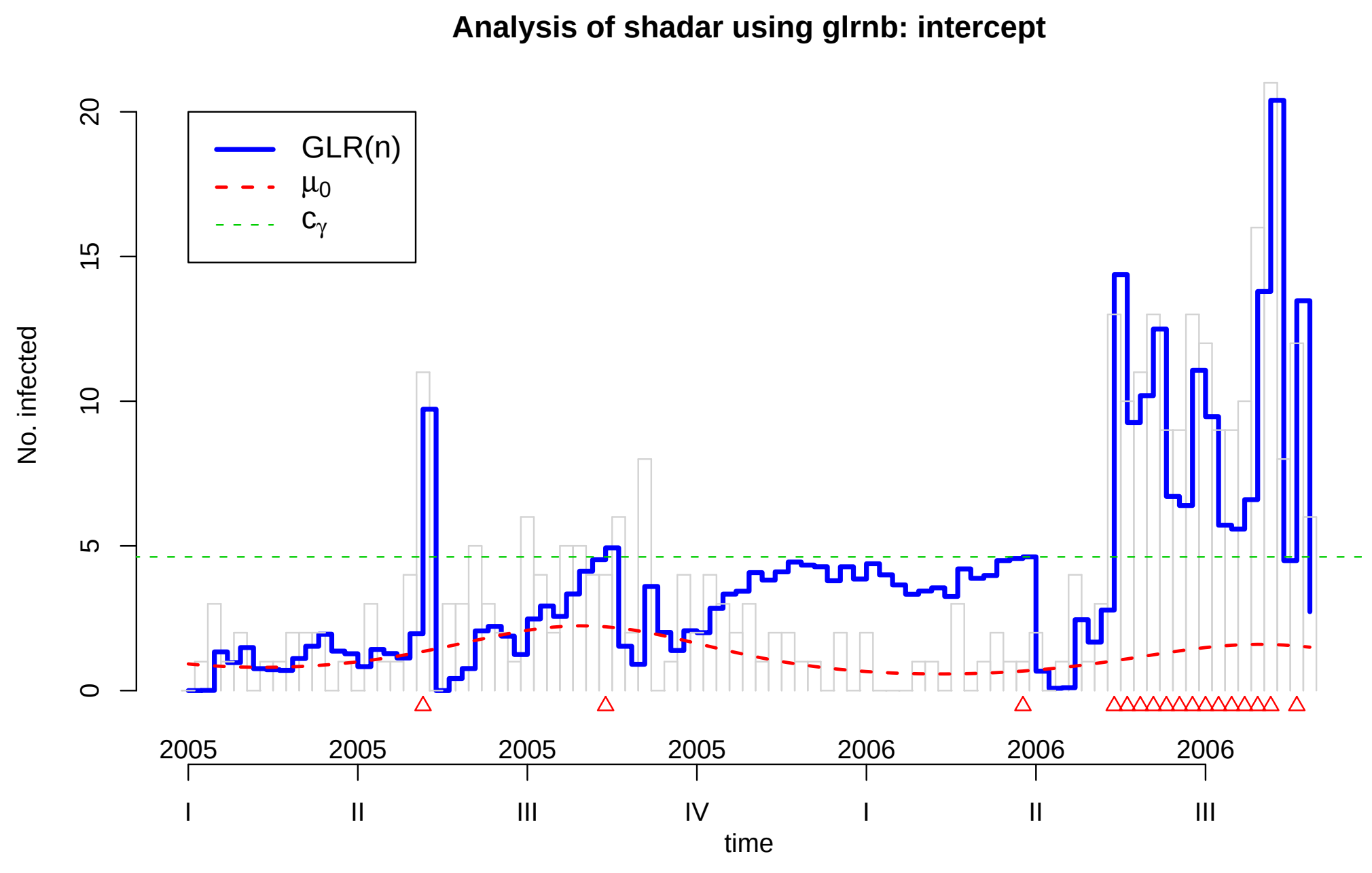
- During the year 2006 the German health authorities noted an increased number of cases due to *salmonella hadar* compared to the previous years.
- We fitted a log-linear negative binomial GLM to the data of $2001 - 2004$ and used this model to predict the expected number of cases $\hat{\mu}_{0,t}$ for the remaining 2 years.
- A window-limited negative binomial intercept chart with $\hat{\mu}_{0,t}$ as in-control mean is applied to the data of $2005 - 2006$.
- Direct Monte Carlo simulation was used to obtain a threshold $c_\gamma$ which results in a target ARL of $\gamma = 500$.



Weekly number of *salmonella hadar* cases in Germany taken from the SurvStat@RKI database. The dotted line shows the predicted number of cases $\hat{\mu}_{0,t}$ based on the 2001-2004 data.



Monte-Carlo estimated ARLs (left) and corresponding false alarm probability (right) as a function of $c_\gamma$ for the negative binomial intercept chart with $\hat{\mu}_{0,t}$ estimated from the hadar data. Also shown are point-wise 95% confidence regions.



$GLR(n)$ statistic superimposed on the observed data of 2005-2006. The triangles show all $n$, where $GLR(n) \ge c_\gamma$.

## 5 Software

The presented charts are implemented in the R package `surveillance` available from CRAN. Development versions are available from

`http://surveillance.r-forge.r-project.org/`

Furthermore, the package provides visualization of routinely collected surveillance data and a test-bench for new surveillance algorithms. Among others `surveillance` contains an implementation of

- the procedure of Farrington et al. (1996),
- the GLR detectors proposed by Höhle and Paul (2008),
- estimation routines for the model described in Paul et al. (2008).

Further information about the package can be obtained from Höhle (2007) or in the accompanying vignette.

## 6 Acknowledgements

Financial support was given by the Swiss National Science Foundation (SNF) and the Munich Center of Health Sciences.

## References

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

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

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

Paul, M., Held, L. and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data, *Technical report*, University of Zurich. `http://www.biostat.uzh.ch/research/manuscripts.html`.

Rogerson, P. A. and Yamada, I. (2004). Approaches to syndromic surveillance when data consist of small regional counts, *MMWR* **53(Suppl.)**: 79–85.

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