Höhle:

Poisson regression charts for the monitoring of surveillance time series

Sonderforschungsbereich 386, Paper 500 (2006)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Poisson regression charts for the monitoring of surveillance time series

Michael Höhle*

Abstract

This paper presents a Poisson control chart for monitoring time series of counts typically arising in the surveillance of infectious diseases. The in-control mean is assumed to be time-varying and linear on the log-scale with intercept and seasonal components. If a shift in the intercept occurs the system goes out-of-control. Novel is that the magnitude of the shift does not have to be specified in advance: using the generalized likelihood ratio (GLR) statistic a monitoring scheme is formulated to detect on-line whether a shift in the intercept occurred. For this specific Poisson chart the necessary quantities of the GLR detector can be efficiently computed by recursive formulas. Extensions to more general Poisson charts e.g. containing an autoregressive epidemic component are discussed. Using Monte Carlo simulations run length properties of the proposed schemes are investigated. The practicability of the charts is demonstrated by applying them to the observed number of salmonella hadar cases in Germany 2001-2006.

1 Introduction

A pleasant development in the design of algorithms for the surveillance of infectious diseases has been the increased inspiration by statistical process control (SPC) techniques. Early surveillance methods such as (Stroup et al., 1989; Farrington et al., 1996) were mainly based on repeated use of confidence intervals – a method which did not take the inane multiple testing structure of the problem into account. Modern surveillance as described in e.g. Lawson and Kleinman (2005) rely on knowledge gained in the SPC literature, e.g. Frisén (1992); Frisén and Wessman (1999); Woodall (2006). However, in my opinion, time series of counts from surveillance data exhibit special features, which the methods from SPC are not handling and for which special solutions have to be found.

*Department of Statistics, University of Munich, Ludwigstr. 33, 80539 München, Germany, Email: hoehle@stat.uni-muenchen.de

An example is the use of cumulative sum (CUSUM) methods, which in a surveillance context monitor counts or proportions. Here it is important to take covariate information into account, e.g. seasonal variations in the mean, adjustment for at-risk population or other explanatory variables. Basically what is needed are *regression charts* based on *generalized linear models* (GLMs). Regression charts with normal response are found in the statistics and engineering literature (Brown et al., 1975; Kim and Siegmund, 1989; Basseville and Nikiforov, 1998; Lai, 1995; Lai and Shan, 1999). Some attempts to regression charts based on GLMs are found in the SPC literature (Skinner et al., 2003) and in the surveillance literature (Rossi et al., 1999; Rogerson and Yamada, 2004a).

The aim of this paper is a pragmatic one: to provide a Poisson regression chart which takes the seasonal variation in the mean into account. Furthermore and contrary to the traditional surveillance techniques, only the parametric form of the Poisson-mean after the change-point should be specified in advance, the necessary parameters are then to be estimated from data at each instance.

This paper is organized as follows. Section 2 presents the basic seasonal Poisson regression model and discusses SPC techniques for detecting changes in the intercept parameter. Crux of the section is an efficient updating procedure for the so called generalized likelihood ratio scheme. In Sect. 3 the proposed scheme is tested on German salmonella data. Section 4 discusses an extensions of the seasonal Poisson model, where the alternative consists in the addition of an epidemic component as in (Held et al., 2005). Finally, Sect. 5 provides a discussion.

2 Detecting changes in a seasonal Poisson chart

Assume that the observations $x_1, x_2, \ldots$ originate from some parametric distribution with density f_θ such that given the change-point τ

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

Here, z_t denotes known covariates at time t. More specifically I will in this paper assume that f_{θ_0} and f_{θ_1} are the Poisson probability mass functions with respective means $\mu_{0,t}$ and $\mu_{1,t}$. The interest is to determine τ *on-line* – i.e. new observations are collected until one is convinced that a change has occurred. A *stopping-rule* thus determines when enough evidence against $H_0 : \mu_t = \mu_{0,t}$ has been collected to stop the sampling.

Mathematically speaking, the following seasonal Poisson-model for the in-

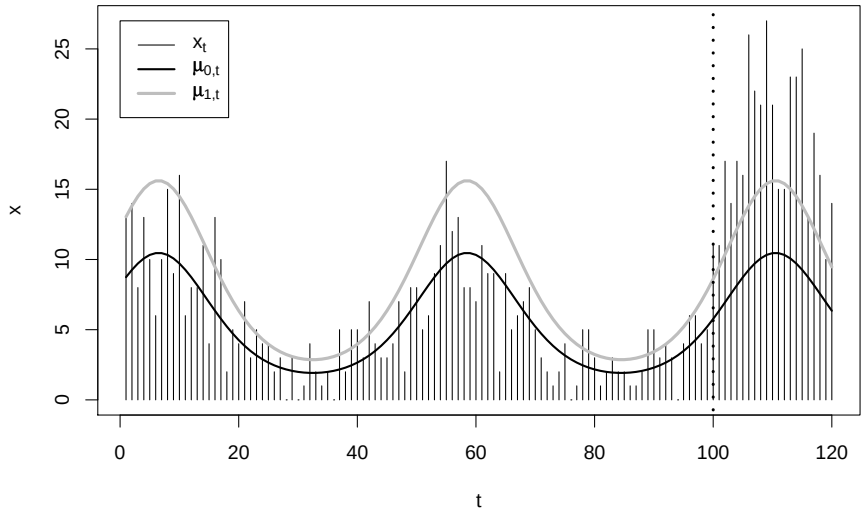


Figure 1: Realization from the model in Example 1 with a change-point at $\tau = 100$.

control situation is assumed:

$$\log \mu_{0,t} = \beta_0 + \sum_{s=1}^{S} \Big(\beta_{2s-1} \cos(\omega s t) + \beta_{2s} \sin(\omega s t) \Big) . \tag{1}$$

In the above $\omega = \frac{2\pi}{T}$ and T is the period, e.g. for weekly data $T = 52$. The out-of-control situation is characterized by a multiplicative shift

$$\mu_{1,t} = \mu_{0,t} \cdot \exp(\kappa), \tag{2}$$

which corresponds to an additive increase of the mean on the log-scale. In surveillance applications only increased rates are of interest, hence $\kappa \geq 0$ is assumed. A motivation of such an increase could be the introduction of a point-source causing an increased number of cases, e.g. contaminated food. Letting the increase be additive on the log-scale is – compared to the usual direct additive increase on mean – computationally advantageous as shown in Section 2.1.

Example 1*:* Let $S = 1$, $\boldsymbol{\beta} = (1.5, 0.6, 0.6)$, $\tau = 100$ and $\kappa = 0.4$, which roughly corresponds to a 50% increase in the number of cases. Figure 1 shows a realization of $m = 120$ observations from the model.

Several scenario with respect to the availability of the parameters are imaginable as discussed by Hawkins et al. (2003) in case of Gaussian observations.

1. All $2S + 2$ model parameters, i.e. $(\boldsymbol{\beta}, \kappa)$, are known.

2. The in-control parameters $\boldsymbol{\beta}$ are known, while $\boldsymbol{\theta} = \kappa$ is unknown and has to be estimated during the monitoring.

3. All model parameters are unknown, i.e. $\boldsymbol{\theta} = (\boldsymbol{\beta}, \kappa)$ has to be estimated.

A typical approach in the SPC and surveillance literature is to use part of the time series to estimate all parameters and then use a Scenario 1 monitoring scheme, for example the CUSUM stopping rule, which is based on the likelihood-ratio,

$$N = \inf\left\{ n \geq 1 : \max_{1 \leq k \leq n} \left[\sum_{t=k}^{n} \log\left\{ \frac{f_{\theta_1}(x_t|z_t)}{f_{\theta_0}(x_t|z_t)} \right\} \right] \geq c_\gamma \right\}. \tag{3}$$

In case of no-covariates the above can be shown to be optimal in the sense of minimizing the expected delay of detection among all monitoring schemes with an average run length (ARL) of $E(N) = \gamma$ when $\tau = \infty$ (Moustakides, 1986). With known θ_1 and θ_2, (3) can be given in recursive form as

$$l_0 = 0, \quad l_n = \max\left(0, l_{n-1} + \log\left\{ \frac{f_{\theta_1}(x_n)}{f_{\theta_0}(x_n)} \right\}\right), \; n \geq 1 \tag{4}$$

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

Scenario 1 schemes, however, ignore any uncertainty originating from the estimation of the in-control parameters $\boldsymbol{\beta}$. Shu et al. (2004) show how uncertainty from parameter estimation has an effect on e.g. the average run length in the case of normal distribution and Shewhart or exponentially weighted moving average (EMWA) control charts.

My interest is nonetheless the Poisson regression chart, where the theory is not as developed as in the Gaussian case. In a surveillance context, Scenario 3 would be most realistic, but with respect to speed and theoretical properties this situation is nevertheless the hardest to handle. Thus I settle on a pragmatic compromise – Scenario 2 based on the so called generalized likelihood ratio (GLR) scheme (Lai, 1995).

A generalization of the CUSUM scheme to this GLR form is to use the following stopping rule

$$N_G = \inf\left\{ n \geq 1 : \max_{1 \leq k \leq n} \sup_{\theta \in \Theta} \left[\sum_{t=k}^{n} \log\left\{ \frac{f_{\theta}(x_t|z_t)}{f_{\theta_0}(x_t|z_t)} \right\} \right] \geq c_\gamma \right\}. \tag{5}$$

In the above, a maximization of the log-likelihood has to be carried out over $\theta \in \Theta$ for each possible change time k between 1 and n. As n grows this task because infeasible: to determine whether $N_G \leq m$ for some specific data, the worst case number of operations is $O(m^3)$. Lai and Shan (1999) show for the Gaussian case how it is possible to reduce this complexity by

clever recursive computations of the sums and sups. I will show that this is also possible for the specific seasonal Poisson chart in (1) and (2). In more general Poisson setups such a computational reduction is although not always possible.

Contrary to (3) it is not possible to bring (5) into a recursive form. Prohibitive to the use of (5) is also that no exact or asymptotic methods exist (at least not to my knowledge) to compute the run length properties of such a Poisson scheme having time-changing means. Hence I will resort to Monte-Carlo sampling for the investigation of chart properties.

2.1 Recursive computations for the Poisson GLR detector

When using the seasonal Poisson chart and assuming fixed $\boldsymbol{\beta}$, the maximum likelihood estimator (MLE) of $\boldsymbol{\theta} = \kappa$ in (5) can be found analytically. Because

$$\log\left\{\frac{f_\theta(x_t|z_t)}{f_{\theta_0}(x_t|z_t)}\right\} = \kappa \cdot x_t + (1 - \exp(\kappa)) \cdot \mu_{0,t}, \;\; t = 1, 2, \ldots,$$

standard derivations show that the MLE for κ based on the observations $x_k, \ldots, x_n$ is

$$\hat{\kappa}_{n,k} = \log\left(\frac{\sum_{t=k}^{n} x_t}{\sum_{t=k}^{n} \mu_{0,t}}\right). \tag{6}$$

To enforce $\kappa \geq 0$ we will use $\hat{\kappa}^+_{n,k} = \max(0, \hat{\kappa}_{n,k})$. Furthermore,

$$l_{n,k} = \sup_{\theta\in\Theta} \sum_{t=k}^{n} \frac{f_\theta(x_t|z_t)}{f_{\theta_0}(x_t|z_t)} = \hat{\kappa}^+_{n,k} \sum_{t=k}^{n} x_t + \left(1 - \exp(\hat{\kappa}^+_{n,k})\right) \sum_{t=k}^{n} \mu_{0,t}.$$

Thus by recursively computing $s_x(k,n) = \sum_{t=k}^{n} x_t = s_x(k-1,n) + x_k$ and $s_\mu(k,n) = \sum_{t=k}^{n} \mu_{0,t} = s_\mu(k-1,n) + \mu_{0,k}$ we obtain an efficient computation of $\hat{\kappa}^+_{n,k}$ and $l_{n,k}$. This finding resembles the parallel recursive algorithms by Lai and Shan (1999) in case of Gaussian regression models, though the recursive computations only work for my one-parameter GLM model. Thus to determine whether $N_G \leq m$ for the above Poisson intercept chart, a worst case cost of $O(m^2)$ operations are needed. However, it is sufficient to terminate the computation once the first $l_n(k) > c$, i.e. the computation of the maximum in (5) is not really necessary. Even though $O(m^2)$ is an improvement, it is still more expensive than the $O(m)$ operations of the scheme in (4). The big advantage is however that we are not forced to specify κ in advance.

Example 2*:* Application of the GLR-CUSUM procedure with $c_\gamma = 5$ to the data from Example 1 is illustrated in Fig. 2. The stopping time occurs the first time $GLR(n) \geq c_\gamma$, where $GLR(n) = \max_{1\leq k\leq n} l_{n,k}$.

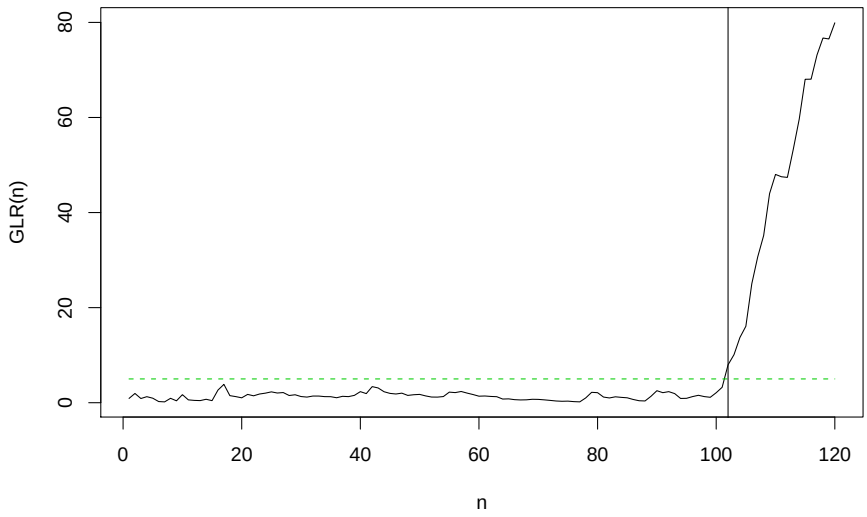


Figure 2: Plot of $GLR(n)$ as a function of n for the data in Fig.1. The stopping time with $c_\gamma = 5$ is $N_G = 102$.

The distribution of N_G (i.e. the run-length distribution) for a specific choice of c_γ can be obtained using direct Monte-Carlo simulation. Figure 3 shows a histogram of 2000 simulations each of length 4000 when $c_\gamma = 5$. Note that realizations with $N_G > 4000$ are thus truncated. A Monte-Carlo estimate of the ARL $\gamma = E(N_G)$ is easily calculated as the mean of these samples, i.e. $\hat{\gamma}_{MC} = 450.51$. The bounds of a $(1-\alpha)\cdot 100\%$ confidence interval for γ based on normal asymptotics are then given as $\hat{\gamma}_{MC} \pm z_{1-\alpha/2} \cdot \hat{\text{se}}(\hat{\gamma}_{MC})$, where $\hat{\text{se}}(\hat{\gamma}_{MC})$ is the Monte-Carlo estimate of $\text{Var}(\hat{\gamma}_{MC})^{\frac{1}{2}}$. Superimposed in the figure is also the density of the $\text{Exp}(\hat{\gamma}_{MC})$ distribution, as Lai (1995) shows $N_G \sim \text{Exp}(\gamma)$ as asymptotic result.

By wrapping a root search routine such as the secant-method around the estimation of $E(N_G)$ one can find a c_γ such that a desired ARL is approximately obtained. To compensate for the Monte-Carlo estimation a tolerant convergence criterion has to be used, e.g. to stop once $\hat{\gamma}_{MC} \pm 2 \cdot \hat{\text{se}}(\hat{\gamma}_{MC})$ is within $\gamma \pm 0.05\gamma$.

For comparison, Fig. 4 shows the ARL of the corresponding LR-Detector in (4) as a function of the pre-specified intercept change parameter κ. These values have to be compared to the 95% confidence intervals of $(431.60, 469.42)$ and $(5.16, 5.41)$ for the in-control and out-of-control ARL of the GLR scheme. This shows that the GLR-detector yields more false-alarms than the LR-detectors, but the out-of-control ARL of 5.28 is even faster than the LR-detector for the correct value of $\kappa^* = 0.4$. For the LR-detectors, the out-of-

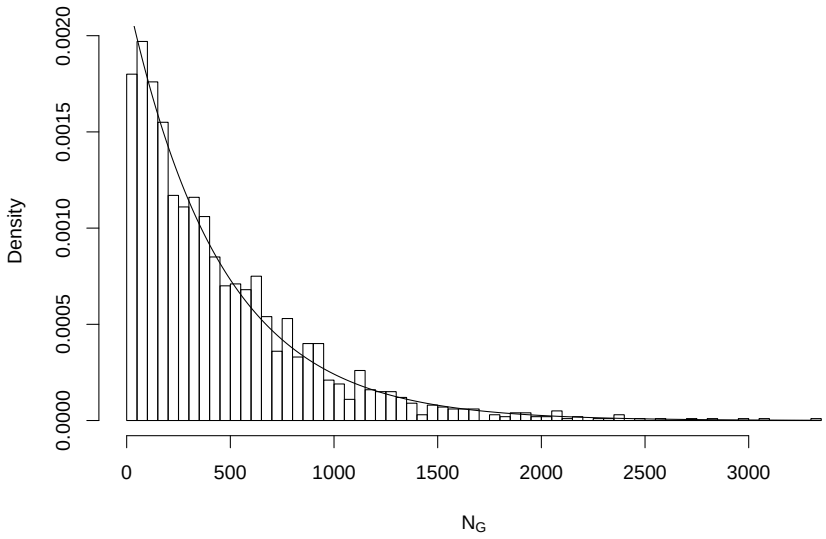


Figure 3: Histogram of the Monte-Carlo ARLs $\gamma^{(t)}, t = 1, \ldots, 2000$ when $c_\gamma = 5$.

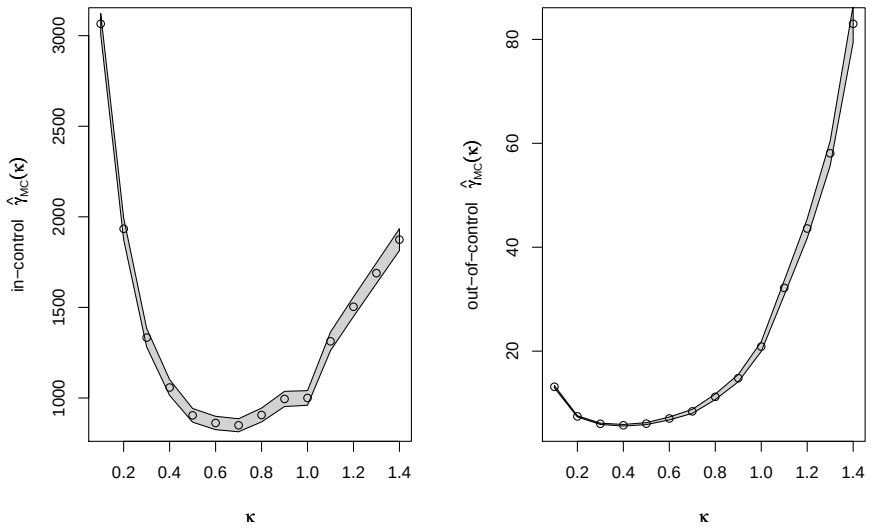


Figure 4: ARLs (circles) in the in-control and out-of-control situation for the LR-Detector in (4) with $c_\gamma = 5$ as a function of the pre-specified κ value. The shaded regions show 95% pointwise confidence intervals. A true value of $\kappa^* = 0.4$ was used in the out-of-control simulations.

control ARL is also sensitive to a correct specification of the true κ.

Instead of using the average run length as design constraint for the detector, Lai (1995) recommends the computationally less demanding false alarm probability $P_{\theta_0}(N_G \leq m)$ as design criterion. Another motivation against the mean is e.g. shown by Fig. 3: the mean is only a crude summary of the skew distribution. Especially there exists no guarantee that a large γ also implies a low probability of false alarm at the initial stages. Furthermore, to determine $P_{\theta_0}(N_G \leq m)$ by Monte-Carlo, a maximum of m operations are needed, which can be a substantial saving compared to direct simulation of $E(N_G)$. In applications of the seasonal Poisson chart it makes sense to select m such that m/T is integer. Thus one solution would be to use an iterative procedure to determine c_γ such that $P_{\theta_0}(N_G \leq m) = \phi$, where ϕ is some acceptable value.

3 Application

During 2006 the German health authorities noted an increased number of cases due to *salmonella hadar* compared to the previous years (Robert Koch Institute, 2006). Figure 5 shows the corresponding weekly number of cases in Germany for the years 2001-2006 taken from the SurvStat@RKI database (Robert Koch-Institut, 2006). To estimate $\boldsymbol{\beta}$ in (1) a GLM is fitted to the data from 2001-2004 resulting in $\boldsymbol{\beta} = (1.16, -0.45, -0.31)$. The line shows the expected number of cases $\hat{\mu}_{0,t}$ from this estimation. A likelihood ratio test confirms the necessity of the seasonal component associated with β_1 and β_2.

Figure 6(a) shows the connection between c_γ and the ARL γ when using the above $\boldsymbol{\beta}$ in the Poisson intercept chart. The superimposed line corresponds to the least-squares fit $\log(\hat{\gamma}_{MC}) = 1.17 + 1.00 c_\gamma$. Instead of a rather costly secant-search for a c_γ resulting in a target of $\gamma = 500$ the above interpolation is used to determine $c_\gamma \approx 5.09$. With this value one obtains a stopping time of $N = 227$ as shown in Fig. 7, which corresponds to week 19 in the year 2005. Looking at the counts this appears to be a sporadic outbreak. The traditional way to proceed the monitoring after the first alarm would be to restart monitoring beginning from time 228. A drawback of this approach is though, that it might take the chart a while to build up enough evidence against H_0 even if the alternative H_1 is also true at the subsequent time points. Instead the idea of Kenett and Pollak (1983) developed for the ordinary CUSUM is used: No resetting occurs and alarms are sounded until (if ever) $GLR(n) < c_\gamma$ again. With this modified procedure the explicit increase during 2006 is clearly detected.

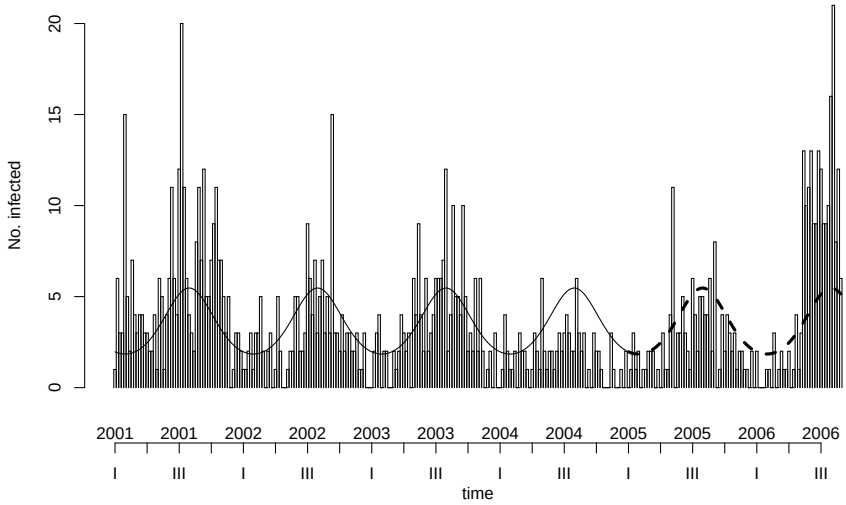
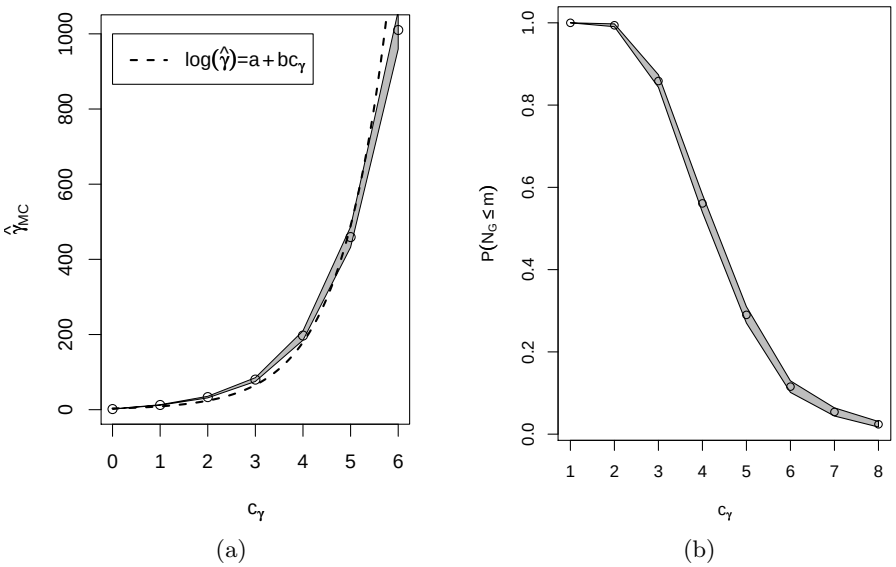


Figure 5: Weekly number of *salmonella hadar* cases in Germany. The line shows the expected number of cases $\hat{\mu}_{0,t}$ estimated from the first four years.



(a) (b)

Figure 6: (a) The Monte-Carlo estimated ARLs as a function of c_γ for model (1) with $\boldsymbol{\beta}$ estimated from the hadar data. Also shown are point-wise 95% confidence regions (b) Shows the corresponding $P_{\theta_0}(N_G \leq 3T)$ as a function of c_γ.

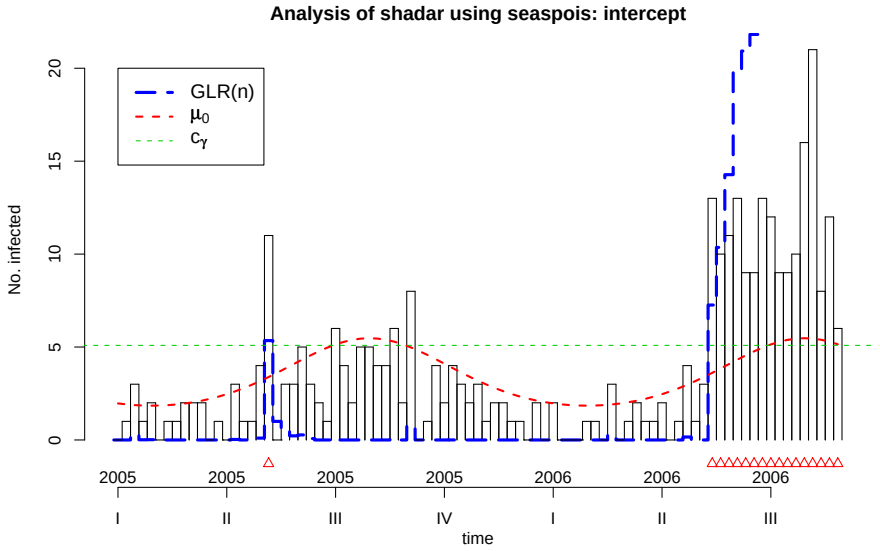


Figure 7: $GLR(n)$ for the salmonella hadar data superimposed on the observed data of 2005-2006. The triangles shows all n, where $GLR(n) \geq c_\gamma$.

4 Beyond the seasonal Poisson chart

The GLR detector for the Poisson means $\mu_{0,t}$ and $\mu_{1,t}$ given in (1) and (2) has the advantage of a fast implementation. However, more involved models for the shift are imaginable. As an example consider the alternative $\mu_{1,t}$ formulation

$$\mu^e_{1,t} = \mu_{0,t} + \lambda x_{t-1}, \quad t > 1, \tag{7}$$

where $\lambda > 0$ and $\mu^e_{1,1} = \mu_{0,1}$. This is an auto-regressive model in which past values enter as covariates. It corresponds to a branching process with immigration and is discussed by Held et al. (2005) as a model for time series of counts from infectious diseases. A motivation for a shift to $\lambda > 0$ could be due to the appearance of an epidemic component on top of the endemic component of an infectious disease, see Held et al. (2006) for a discussion.

Example 3*:* Let $S = 1$, $\boldsymbol{\beta} = (1.5, 0.6, 0.6)$, $\tau = 100$ as in Example 1 and let $\lambda = 0.4$. Figure 8 shows a realization of $m = 120$ observations from the epidemic model.

If the mean of the alternative is given by (7), no fast recursive updating is possible anymore. One reason is that for given n no explicit expression is available for the computation of the MLE $\hat{\lambda}_{n,k}, 2 \leq k \leq n$. Instead, one has to resort to iterative procedures to determine the MLE. For example

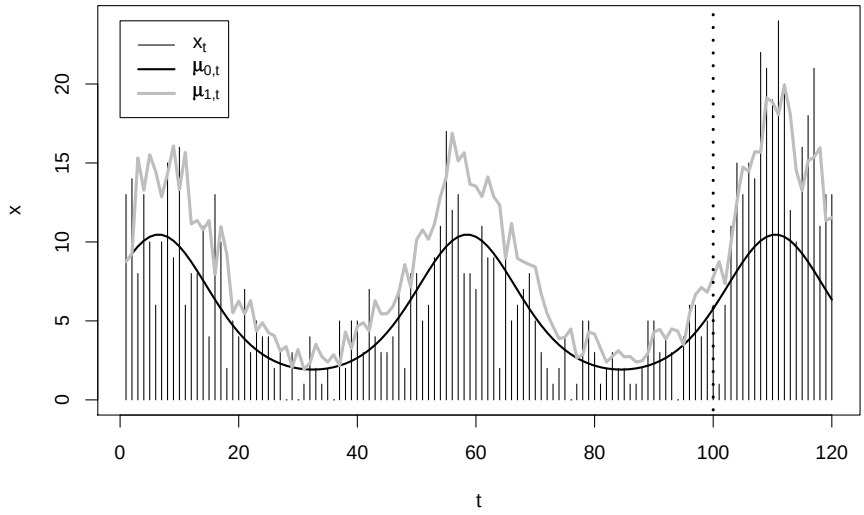


Figure 8: Realization from the model in Example 3 with a change-point at $\tau = 100$.

Newton-Raphson uses the update-rule

$$\lambda_{n,k}^{(s+1)} = \lambda_{n,k}^{(s)} + \frac{S(\lambda_{n,k}^{(s)})}{I(\lambda_{n,k}^{(s)})}, \quad \text{where}$$

$$S_{n,k}(\lambda) = \sum_{t=k}^{n} \frac{x_t x_{t-1}}{\lambda x_{t-1} + \mu_{0,t}} - \sum_{t=k}^{n} x_{t-1}, \quad \text{and}$$

$$I_{n,k}(\lambda) = \sum_{t=k}^{n} \frac{x_t x_{t-1}^2}{(\lambda x_{t-1} + \mu_{0,t})^2}$$

until convergence. As starting value for the computation of $\hat{\lambda}_{n,k}$ one can use $\lambda_{n,k}^{(0)} = \hat{\lambda}_{n,k+1}$. From this starting point convergence usually occurs in just a few update-steps. To enforce the constraint $\lambda_{n,k} > 0$ the optimization is done using $\phi_{n,k} = \log \lambda_{n,k}$ with corresponding changes in the Newton-Raphson update formula.

Another hindrance is that the likelihood ratios are not amenable to recursive updating anymore, because (7) inserted in (5) yields

$$l_{n,k} = \sum_{t=k}^{n} x_t \log\left(\hat{\lambda}_{n,k} \frac{x_{t-1}}{\mu_{0,t}} + 1\right) - \hat{\lambda}_{n,k} \sum_{t=k}^{n} x_{t-1}.$$

As the $\hat{\lambda}_{n,k}$ change for each k, the first sum now has to be computed for all terms $k \leq i \leq n$; the second term can still be computed recursively, though.

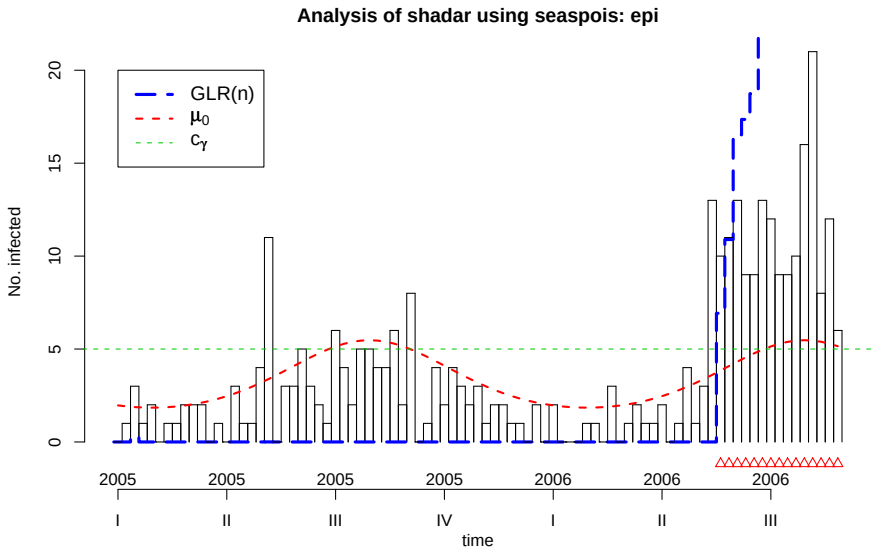


Figure 9: $GLR(n)$ for epidemic Poisson chart the salmonella hadar data superimposed on the observed data of 2005-2006. The triangle shows the first time $GLR(n) \geq c_\gamma$.

Without the possibility of recursive updating, one solution in case of large n is to use a so called window-limited GLR scheme as originally proposed by Willsky and Jones (1976). Here the maximization is not performed for all $1 \leq k \leq n$, but only for a moving window of $k \in \{n-M, \ldots, n-\tilde{M}+1\}$, where $0 \leq \tilde{M} < M$. The minimum delay $\tilde{M}$ is the minimal required sample size to obtain sufficiently reliable estimate of θ_1. In the univariate setup of the Poisson intercept chart with $\theta_1 = \kappa$, $\tilde{M} = 1$ is sufficient; similarly for the epidemic chart.

Example 4*:* Using $\tilde{M} = 1$, $M = 20$ and $c_\gamma = 6$ a probability of $P_{\theta_0}(\tilde{N}_G < 3 \cdot 52) = 0.0490$ is computed for the epidemic Poisson chart with fixed $\boldsymbol{\beta} = (1.16, -0.45, -0.31)$. Applying this detection scheme to the hadar data yields results as shown in Fig. 9 – specifically the first alarm is sounded at $N = 281$, which corresponds to week 21 in the year 2006. Again, no resetting occurs after the first alarm and all subsequent $GLR(n) \geq c_\gamma$ are flagged as alarms. Compared to the results of the intercept chart in Fig. 7, the epidemic model does not sound an alarm for the sporadic increase at $N = 227$. This is explained by its autoregressive nature focusing on person-to-person transmission rather than sudden spikes.

Instead of the rather specific mean structure in the seasonal Poisson chart from Sect. 2 one might consider the general form $\log \mu_{0,t} = \boldsymbol{\beta} \boldsymbol{z_t}$ and $\log \mu_{1,t} = \log \mu_{0,t} + \kappa z_{t,j}$, where $z_{t,j}$ is the j-th component of the covariate vector $\boldsymbol{z}_t$. Such a Poisson chart is a model, where the influence of the j-th component

is shifted from time τ. A closed-form expression for $\hat{\kappa}_{n,k}$ as in (6) is then only possible if the covariate $z_{t,j}$ is not time-varying, i.e. $z_{t,j} = z_j$. Otherwise an iterative scheme, similar to Newton-Raphson described above, has to be used to compute the MLE. Likewise, the $l_{n,k}$ can not be computed recursively anymore and window-limited GLR schemes appear as a way to obtain computationally feasible schemes.

5 Discussion

This paper introduced two Poisson regression charts to monitor time series of counts originating in the surveillance of infectious diseases. Based on a seasonal model for the mean of the Poisson distribution two types of changes were considered. Firstly, an additive shift in the mean on the log-scale motivated by point source outbreaks. Secondly, a sudden addition of an auto-regressive component motivated by e.g. person-to-person transmission of infectious diseases. A feature of the charts was a repugnance to specify more than the parametric form of the alternative. Using ideas from the engineering literature, efficient schemes were presented which makes the resulting parameter-estimation at each time instance feasible.

In applications, the Poisson distribution might not adequately address the vast over-dispersion found e.g. due to reporting bias. Here, the negative-binomial provides a more flexible framework (Held et al., 2005). Similarly, the cyclic regression with exponentiated harmonics used in (1) can be problematic and different models such as piecewise exponential curves might be better (Andersson et al., 2006). However, all these extensions are still possible within the GLM framework. Another aspect in the above developments is the potential to use the charts as basis for the surveillance of multivariate time series of counts as described e.g. by Rogerson and Yamada (2004b) or Sonesson and Frisén (2005). In particular it appears feasible to extend the proposed Poisson charts to the multivariate modelling described in (Held et al., 2005).

One important issue left is a more theoretical investigation of the optimality and characteristics of the proposed schemes. Literature exists on the asymptotic optimality of the GLR detector and its window-limited versions in the Gaussian case (Lai, 1995, 1998). Attempts have been made to generalize these results to the exponential family, however, additional complications arise when moving beyond the independent and identically distributed setting by the addition of covariates.

The direct Monte-Carlo estimation of ARLs utilized in this paper becomes problematic when γ is large. Techniques such as using control variates or importance sampling schemes known from ARL simulations for identical and

independent random variables (Jun and Choi, 1993; Lai and Shan, 1999) are not immediately applicable. A computationally easier quantity is therefore to use the failure rate $P_{\theta_0}(N_G \leq m)/m$. However, if this required probability is very small the precision of direct Monte-Carlo estimation can be very poor. Suggestions exists to use importance-sampling to substantially improve the precision – even in the non-Gaussian case (Lai and Shan, 1999; Chan and Lai, 2003). Work is although required to adapt this work to the presented charts.

At this point I would although like to emphasize the pragmatism of the presented approach: to combine ideas from engineering and SPC in order to develop a surveillance algorithm tailored for the specifics of infectious diseases. The data and surveillance methods presented in this paper have been implemented in the R-package `surveillance` available from the Comprehensive R Archive Network[1]. As the package also provides an implementation of classical surveillance methods, e.g. Farrington et al. (1996) or Rossi et al. (1999), it is straightforward to compare with the proposed Poisson charts using e.g. simulation studies.

6 Acknowledgments

The research was conducted with financial support from the Collaborative Research Centre SFB 386 funded by the German research foundation (DFG).

References

Andersson, E., Bock, D., and Frisén, M. (2006). Exploratory analysis of swedish influenza data. Technical Report 1:2006, Swedish Institute for Infectious Disease Control.

Basseville, M. and Nikiforov, I. (1998). *Detection of Abrupt Changes: Theory and Application*. Online version of the 1994 book published by Prentice-Hall, Inc. Available from http://www.irisa.fr/sisthem/kniga/.

Brown, R., Durbin, J., and Evans, J. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society, Series B*, 37(2):149–192.

Chan, H. and Lai, T. (2003). Importance sampling for generalized likelihood ratio procedures in sequential analysis. *Sequential Analysis*, 24:259–278.

[1] At the current time of writing the extensions concerning the GLR detector are implemented in the beta-version of `surveillance` available from `http://www.stat.uni-muenchen.de/~hoehle/software/surveillance`

Farrington, C., Andrews, N., Beale, A., and Catchpole, M. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society, Series A*, 159:547–563.

Frisén, M. (1992). Evaluations of methods for statistical surveillance. *Statistics in Medicine*, 11:1489–1502.

Frisén, M. and Wessman, P. (1999). Evaluations of likelihood ratio methods for surveillance. *Communications in Statistics: Simulation and Computation*, 28:597–622.

Hawkins, D., Qiu, P., and Kang, C. (2003). The changepoint model for statistical process control. *Journal of Quality Technology*, 35:355–366.

Held, L., Hofmann, M., Höhle, M., and Schmid, V. (2006). A two component model for counts of infectious diseases. *Biostatistics*, 7:422–437.

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.

Jun, C.-H. and Choi, M. (1993). Simulating the average run length for CUSUM schemes using variance reduction techniques. *Communications in Statistics - Simulation and Computation*, 22(3):877–887.

Kenett, R. and Pollak, M. (1983). On sequential detection of a shift in the probability of a rare event. *Journal of the American Statistical Association*, 78(382):389–395.

Kim, H.-J. and Siegmund, D. (1989). The likelihood ratio test for a changepoint in simple linear regression. *Biometrika*, 76(3):409–423.

Lai, T. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society, Series B*, 57:613–658.

Lai, T. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929.

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

Lawson, A. and Kleinman, K., editors (2005). *Spatial and Syndromic Surveillance for Public Health*. Wiley.

Moustakides, G. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387.

Robert Koch-Institut (2006). SurvStat@RKI. http://www3.rki.de/SurvStat. Date of query: September 2006.

Robert Koch Institute (2006). Epidemiologisches Bulletin 31. Available from http://www.rki.de.

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

Rogerson, P. and Yamada, I. (2004b). Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statistics in Medicine*, 23:2195–2214.

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

Shu, L., Tsung, F., and Tsui, K.-L. (2004). Run-length performance of regression control charts with estimated parameters. *Journal of Quality Technology*, 36:280–292.

Skinner, K., Montgommery, D., and Runger, G. (2003). Process monitoring for multiple count data using generalized linear model-based control charts. *International Journal of Production Research*, 41(6):1167–180.

Sonesson, C. and Frisén, M. (2005). Multivariate surveillance. In Lawson, A. and Kleinman, K., editors, *Spatial and syndromic surveillance for public health*, chapter 9, pages 153–166. Wiley.

Stroup, D., Williamson, G., Herndon, J., and Karon, J. (1989). Detection of aberrations in the occurence of notifiable diseases surveillance data. *Statistics in Medicine*, 8:323–329.

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.

Woodall, W. (2006). The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, 38(2):89–104.