Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics and Data Analysis 52 (2008) 4357–4368

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Count data regression charts for the monitoring of surveillance time series

Michael Höhle[a,b,*], Michaela Paul[c]

[a] *Department of Statistics, University of Munich, Munich, Germany*
[b] *Munich Center of Health Sciences, Germany*
[c] *Institute of Social and Preventive Medicine, University of Zurich, Switzerland*

Received 28 July 2007; received in revised form 14 February 2008; accepted 14 February 2008
Available online 10 March 2008

Abstract

Control charts based on the Poisson and negative binomial distribution for monitoring time series of counts typically arising in the surveillance of infectious diseases are presented. 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. Using the generalized likelihood ratio (GLR) statistic a monitoring scheme is formulated to detect on-line whether a shift in the intercept occurred. In the 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 auto-regressive epidemic component are discussed. Using Monte Carlo simulations run-length properties of the proposed schemes are investigated and the Poisson scheme is compared to existing methods. 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) and Farrington et al. (1996) were mainly based on the repeated use of confidence intervals – a method which did not take the inherent 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) and Woodall (2006). However, in our 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. 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

[*] Corresponding author at: Department of Statistics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany. Tel.: +49 (0) 89 2180 6404; fax: +49 (0) 89 2180 5040.
E-mail address: hoehle@stat.uni-muenchen.de (M. Höhle).

0167-9473/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.02.015

(Brown et al., 1975; Kim and Siegmund, 1989; Basseville and Nikiforov, 1998; Lai, 1995; Lai and Shan, 1999). However, when monitoring infrequent infectious diseases or considering a certain partition of age, sex and location, the resulting time series will contain a low number of counts. This makes Gaussian time series models such as AR or ARIMA models inappropriate and discrete distribution models should be used. Consequently, the detection of outliers should not be based on residuals from such Gaussian models. 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 to provide count data regression charts, which take the seasonal variation in the mean into account. Distributionally, we will consider both Poisson and negative binomial charts, because surveillance time series can display considerable overdispersion. Furthermore and contrary to the traditional CUSUM surveillance techniques, only the parametric form of the 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 count data regression model and discusses SPC techniques for detecting changes in the intercept parameter. The crux of the section is an efficient updating procedure for the so-called generalized likelihood ratio scheme in the Poisson case, which is compared with existing Poisson detectors. In Section 3 the negative binomial chart is tested on German salmonella data. Section 4 discusses extensions to more flexible alternatives, e.g. when the alternative consists in the addition of an epidemic component as in Held et al. (2005). Finally, Section 5 provides a discussion.

2. Detecting changes in a seasonal count data 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 we will in this paper assume that f is the negative binomial probability mass function with a fixed and known dispersion parameter α. Note that for $\alpha \to 0$ the Poisson distribution is obtained. Furthermore, we let f_{θ_0} and f_{θ_1} have respective means $\mu_{0,t}$ and $\mu_{1,t}$, which are functions of θ_0 and θ_1, respectively. Our 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 $H_0 : \mu_t = \mu_{0,t}, t = 1, \ldots$ has been collected to stop the sampling.

Mathematically speaking a seasonal log-linear model for the in-control mean is specified — a convenient form would e.g. be

$$\log \mu_{0,t} = \beta_0 + \beta_1 t + \sum_{s=1}^{S} \left(\beta_{2s} \cos(\omega s t) + \beta_{2s+1} \sin(\omega s t) \right). \tag{1}$$

In the above S is the number of harmonic waves to use, $\omega = \frac{2\pi}{T}$ and T is the period, e.g. for weekly data $T = 52$. However, other forms such as replacing the sum of harmonics by a T-periodic spline function as in Harvey et al. (1997) are imaginable.

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 will be shown in Section 2.1.

Example 1. Let $S = 1$, $\boldsymbol{\beta} = (1.5, 0, 0.6, 0.6)$, $\tau = 100$ and $\kappa = 0.4$, which roughly corresponds to a 50% increase in the number of cases. Fig. 1 shows a realization of $m = 120$ observations from the Poisson model.