

Approximate Poisson CUSUM charts for the monitoring of time series with time-varying mean

Giuseppe Rossi, Simone Del Sarto and Marco Marchi

Abstract This paper presents an approximate Poisson control chart for monitoring time series of counts typically arising in the surveillance of health events. The CUSUM chart is formulated to detect on-line whether a shift in the control mean occurred. If a shift in mean occurs the system goes out-of-control. When the in-control mean is assumed to be time-varying with seasonal components and/or risk factors dependence, an approximate CUSUM based on a Poisson transformation to a Standard Normal distribution and with fixed reference and decision limits was proposed. Novel is that a varying reference limit can improve the performance of the approximate CUSUM, especially when the ARL (Average Run Length) in RQL (Rejectable Quality Level) and the ARL in AQL (Acceptable Quality Level) are equal to 3 and 500 respectively and the mean value of counts is very low (<2). For this specific Poisson chart the necessary quantities of the reference limit can be efficiently computed by a simple formula. Using Monte Carlo simulations run length properties of the proposed scheme are investigated. An application to a real dataset is also shown.

Keywords Cusum . monitoring . surveillance . time series

¹ Giuseppe Rossi, Unit of Epidemiology and Biostatistics, Institute of Clinical Physiology, CNR, Pisa; email: rossigi@ifc.cnr.it

Simone Del Sarto, Unit of Epidemiology and Biostatistics, Institute of Clinical Physiology, CNR, Pisa; email: simonedelsa@hotmail.com

Marco Marchi, Department of Statistics, University of Florence, Florence; email: marchi@ds.unifi.it

1 Introduction

Modern surveillance of health events has been the increased inspiration by statistical process control (SPC) techniques. However 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. Cumulative sum (CUSUM) method is widely used in a surveillance context to 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. In this context an approximate CUSUM method was proposed for count data (Rossi et al.,1999). Basically this time-varying in-control mean Poisson approach consists on transforming the observed counts $x_1, x_2, \dots \sim \text{Po}(\mu_0, t)$ to normality and on monitoring the resulting residuals by an ordinary iid Gaussian CUSUM. Simulation studies (Rossi et al.,1999; Rogerson and Yamada,2004; Höhle and Paul,2008) underlined that the performance of the proposed approach is problematic in case of very low count number mean (<2), particularly when the ARL (average run length) in RQL (rejectable quality level) is equal to 3 and the ARL in AQL (acceptable quality level) is 500. In these cases the discrepancy between the mean time (ARL, average run length) before a false alarm (i) in an in-control state (AQL, acceptable quality level) and (ii) in an out-of-control state (RQL, rejectable quality level) can thus be ample and instable. The aim of this paper is to provide a suitable solution to overcome this limitation, particularly when the ARL in RQL and the ARL in AQL are equal to 3 and 500 respectively. Section 2 presents the basic approximate CUSUM technique for detecting changes in time-varying in-control mean and the modified procedure for improving the performance of the approximate scheme. A formula is provided for approximating the reference limit k values as function of specified values of the expected count (λ_0) for the in- and out-of control values of ARL. In Sect. 3 the proposed scheme is tested by simulations. Sect. 4 shows an application to real data. Finally, Sect. 5 provides a discussion.

2 Detecting changes in a time-varying mean Poisson chart

When the variable being monitored has a Poisson distribution, CUSUM methods are designed to detect sudden changes in the mean value of data consisting of count with distribution: $X_t \sim \text{Po}(\lambda_0)$.

The CUSUM is

$$S_t = \max(0, S_{t-1} + X_t - k)$$

where X_t is the observed count, k is the reference limit. A change in mean is signaled if $S_t \geq h$, where h is the decision limit.

Values of X_t in excess of k are accumulated. Parameters k and h are chosen in conjunction with a predetermined acceptable rate of false alarms (ARL₀: Average Run Length in a Acceptable Quality Level), delay to detect a real increase (ARL₁: Average Run Length in a Rejectable Quality Level) and expected in-control count λ_0 .

When the expected in-control value associated with the Poisson variable changes over time ($\lambda_{0,t}$; $t = 1, 2, \dots$), e.g. as a result of seasonal effects or other time changing behaviour, a CUSUM scheme with constant parameters would have misleading results if the actual values of λ_0 fluctuated from period to period about the constant assumed parameter. Instead a transformation to normality is used. One such normalizing transformation for data consisting of small counts is (Rossi et al. 1999):

$$Z_t = \frac{x_t - 3I_{0,t} + 2\sqrt{I_{0,t}x_t}}{2\sqrt{I_{0,t}}}$$

where X_t is the observed count and $\lambda_{0,t}$ is the expected count at time t .

If the variable of interest is converted to a z-score with mean 0 and variance approximately 1, the CUSUM following the observation t is defined as follows:

$$S_t = \max(0, S_{t-1} + Z_t - k)$$

with a constant value of k and h . For ARL₀ and ARL₁ values commonly used in literature, k and h values are 1.04 and 2.26 respectively when ARL₀=500 and ARL₁=3, and 0.6 and 3.80 respectively when ARL₀=500 and ARL₁=7.

This transformation can be misleading for small values of λ_0 (<2), in particular when the desired values of ARL₀ =500 and ARL₁=3 (Rossi et al.,1999; Rogerson and Yamada,2004; Höhle and Paul,2008). To overcome this limitation a CUSUM scheme with varying k is proposed. Time-specific values of the k parameter are used while the h value remains constant for desired values of ARL₀ and ARL₁. For example, when ARL₀ =500 and ARL₁ = 7 or 3 the h value is equal to 2.26 and 3.80 respectively.

The specific k_t values are then used in the CUSUM as follows:

$$S_t = \max(0, S_{t-1} + Z_t - k_t),$$

where the parameter k_t changes from one period to the next.

For desired values of ARL₀ and ARL₁, the k_t reference value is carried out using the following formula (Rossi et al. 1999):

$$k_t = Z_{1,t}/2 \quad (1)$$

where

$$Z_{1,t} = \frac{\left(\sqrt{\frac{l_{1,t}}{l_{0,t}}} + 1 \right)^2 - 4}{2\sqrt{\frac{1}{l_{0,t}}}}$$

is the Z transformation in ARL₁, while $\lambda_{1,t}$ and $\lambda_{0,t}$ are the out and in-control expected counts respectively. Therefore the k_t value is function of the expected count both in AQL ($\lambda_{0,t}$) and in RQL ($\lambda_{1,t}$).

In a Poisson scheme, given λ_0 and a desired ARL₀ and ARL₁, we can get λ_1 , k and h by a recursive procedure. The stopping-rule is based on the minimization of

$$\left| \frac{\text{calculated } ARL_0 - \text{desired } ARL_0}{\text{desired } ARL_0} \right| + \left| \frac{\text{calculated } ARL_1 - \text{desired } ARL_1}{\text{desired } ARL_1} \right|$$

For example, if the desired ARL₀ and ARL₁ are 500 and 3 respectively, the selection criteria is:

$$|(\text{calculated } ARL_0 - 500)/500| + |(\text{calculated } ARL_1 - 3)/3|.$$

For a Poisson variate $X \sim \text{Po}(\lambda_0)$ the out-of-control parameter (λ_1) used to determine the reference value k is specified as

$$\lambda_1 = \lambda_0 + s \sqrt{\lambda_0}, \quad (2)$$

where s is the change to detect. Given s and hence λ_1 , we can compute the k value as $k = (\lambda_1 - \lambda_0) / (\ln \lambda_1 - \ln \lambda_0)$ and then compute the decision limit h for given in-control parameter (λ_0), reference value k and desired ARL₀ (first step).

In a second step, the in-control average run length (ARL₀) for an upward Cusum scheme for a Poisson distribution is calculated using the Markov chain approach (Hawkins D.M. 1992) given h, k and λ_0 . In the same manner the out-control average run length (ARL₁) is calculated given h, k and λ_1 . For a given values of λ_0 and desired ARL₀ this procedure is repeated for different s values, or equivalently for different λ_1 and k values. The values of ARL₀, ARL₁ and the selection criteria are computed for each s value.

We chose the s value where the selection criteria is minimum. Given s, we can obtain λ_1 from the formula (2) and hence k_t from the formula (1). For example, when $\lambda_{0,t} = 0.62$ the approximate $k_t = 1.22$.

For the first and second step we used two functions of the R-package "surveillance" version 1.1-0 October 15, 2009: findH and arlCusum (Höhle M. 2007).

To illustrate the procedure, values of ARL₀, ARL₁ and s computed by the procedure for specified expected count (λ_0) and desired values of ARL₀=500 and ARL₁=3 are reported In Table 1.

Table 1: Values of ARL_0 , ARL_1 , s , λ_1 , k computed by the procedure for desired values of λ_0 , $ARL_0=500$ and $ARL_1=3$

λ_0	s	ARL_0	ARL_1	λ_1	k_t
0.22	3.36	499	2.9	1.80	1.28
0.39	2.93	499	3.0	2.22	1.17
0.52	2.98	664	3.0	2.67	1.20
0.62	2.99	503	2.8	2.97	1.22
0.79	2.46	528	3.4	2.98	1.03
1.01	2.63	506	3.0	3.65	1.11
1.23	2.50	537	3.2	4.00	1.07
1.52	2.66	492	2.8	4.80	1.14
1.75	2.52	539	3.0	5.08	1.10
1.95	2.50	526	3.0	5.44	1.09
2.13	2.47	507	3.0	5.73	1.09
2.04	2.44	490	3.1	5.53	1.07
2.33	2.44	494	3.0	6.05	1.08
2.58	2.45	555	3.0	6.52	1.09
2.79	2.39	518	3.1	6.78	1.06
2.96	2.28	521	3.2	6.88	1.02
4.17	2.32	504	3.1	8.91	1.05
5.11	2.41	492	2.8	10.56	1.10
6.07	2.20	495	3.2	11.49	1.01
7.28	2.27	504	3.0	13.40	1.05
8.00	2.26	509	3.0	14.39	1.05

3 Simulation study

A simulation study was performed to compare the standard procedure (with fixed k and h) and the proposed procedure (with time-varying k and fixed h). To assess the performance of the proposed procedure and the standard procedure two indexes of performance were considered: the expected number of samples between two false alarms (ARL_0) and the expected number of samples before a true alarm (ARL_1). The value of ARL_0 for each procedure and for each of the tabulated λ_0 values was estimated by a sequence of at list 200,000 random numbers from a Poisson distribution, corresponding to 400 sequences of 500 random numbers. Each sequence was generated by using the algorithm given by R-package. The same data sets were analyzed by each of the two approximate procedures and by the standard Poisson procedure. The evaluation of the old and new approximated procedure was made on the basis of the deviations of the ARL_0 and ARL_1 actual values from the expected ones derived from the standard Poisson procedure. When $ARL_0 = 500$ and $ARL_1 = 3$, for the approximate standard procedure the constant k and h values used were 1.01 and 2.29 respectively, while for the proposed procedure a constant h value of 2.29 and a k value specific for each expected count (λ_0) were used. Each

k value was computed by formula (1). Figure 1 shows the estimated false alarms for various λ_0 values. Results in Figure 1 show that the proposed approximate procedure with time-varying k is preferred to the standard approximate procedure, since it leads to actual values of the expected false alarms (ARL_0) that are closer to the desired ARL_0 . Figure 1 also shows that the actual ARL_0 of the proposed procedure is very near to that of the Poisson procedure also for small values of λ_0 (<2). The mean actual delay value (ARL_1) for the three procedure was estimated by 400 sequence of 50 random numbers from a Poisson distribution for each λ_1 value obtained by the procedure for a Poisson variable when $ARL_0 = 500$ and $ARL_1 = 3$. The expected mean delay value is 3. The estimated mean actual delay values for the three procedure are reported in Figure 2. The new proposed procedure with time-varying k shows the mean delays are very close to the expected ones, even when the expected count (λ_0) is less than 2.

Figure 1: Estimated false alarms (ARL_0) related to λ_0 values

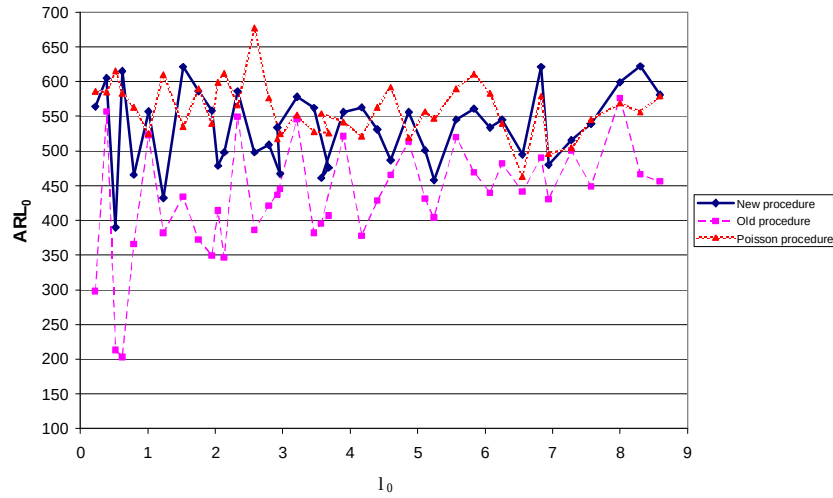
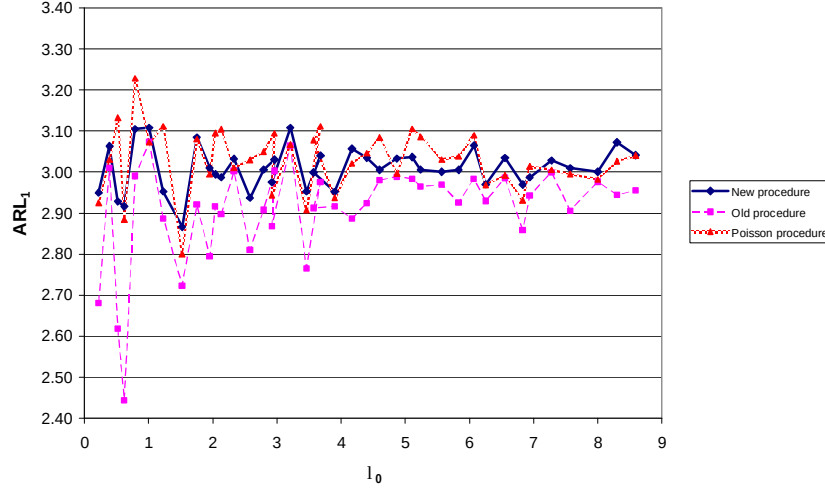
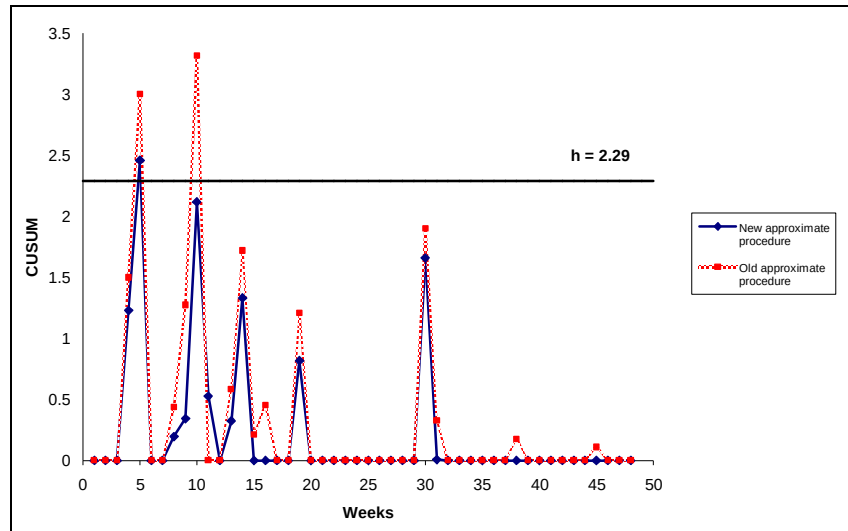


Figure 2: Estimated delay values (ARL_1) related to λ_0 values

4 Real data example

We used data from the Margherita Project of Regional Health Agency of Tuscany to illustrate monitoring of adverse events rates for Poisson data. Data on pneumonia during the stay in a intensive care unit of Tuscany Region over a period of one year were considered. Daily observations were aggregated over week. The expected weekly pneumonias in the considered intensive care unit were calculated on the base of the Regional weekly rates of the same year. The expected weekly pneumonias in the intensive care unit, as obtained on the base of the regional rates applied to the weekly at risk population, were 0.20. The old approximate CUSUM procedure was performed by using $h=2.29$ and $k=1.01$, corresponding to $ARL_0=500$ and $ARL_1=3$, while the new approximate procedure was performed by using $h=2.29$ and time-varying k . The old Cusum procedure produced two alarms (Figure 3); conversely, only one alarm was obtained by the new approximate Cusum procedure (Figure 3).

Figure 3: CUSUM carried out on pneumonia data

5 Discussion

The proposed Cusum procedure described in this paper seems to be more effective than the old approximate procedure when the expected count are very small. Simulations showed that the approximate procedure with time-varying k is sufficiently close to the standard Poisson Cusum.

References

1. Hawkins D.M.: Evaluation of Average Run Length of Cumulative Sum Charts for an Arbitrary Data Distribution. *Communications in Statistics – Simulation and Computation* **21**(4), 1001-1020 (1992).
2. Höhle, M.: Surveillance: An R package for the surveillance of infectious diseases. *Computational Statistics* **22**, 571-582 (2007).
3. Höhle, M., Paul, M.: Count data regression chart for the monitoring of surveillance time series. *Computational Statistics and Data Analysis* **52**, 4357-4368 (2008).
4. Rogerson, P., Yamada, I.: Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality. Weekly Report* **53**, 79-85 (2004).
5. Rossi, G., Lampugnani, L., and Marchi, M.: An approximate CUSUM procedure for surveillance of health events. *Statistics in Medicine* **18**, 2111-2122 (1999).