

# A Cumulative Sum Approach to Syndromic Surveillance in Geographic Regions

Peter A. Rogerson

Department of Geography and National Center for Geographic Information and Analysis  
University at Buffalo, The State University of New York



## Introduction

Objective: to find as quickly as possible any significant increase in rates of reported syndromes in geographic areas.

Characteristics:

- (a) System should be capable of detecting increased rates quickly, while keeping the number of "false alarms" at an acceptable level
- (b) Observations may consist of small frequencies, necessitating the use of binomial or Poisson variables instead of normally distributed variables

Some previous approaches to surveillance in a health and disease context:

- (a) Roubertas (1989) was one of the first to suggest a cumulative sum approach to monitoring disease counts in small areas
- (b) Kleinman et al. (2002) have carried out surveillance by detecting outliers in a temporal sequence of observed binomial variables for multiple geographic regions
- (c) Farrington et al. (1997) discuss desirable features of a health surveillance system
- (d) Rogerson (1997, 2001) and Kulldorf (2001) take existing spatial statistical methods used for the detection of geographic clusters of disease, and modify them for use in surveillance, where new data become available, and repeated tests for emergent clusters are desired.

Here, I will use and develop further a cumulative sum approach for small counts assumed to follow a Poisson distribution.

Cumulative sum methods cumulate the deviations between observed and expected counts in a period; an alarm, or signal, is sent when the cumulative observed counts have exceeded the expected counts by some predetermined threshold.

I modify the usual cumulative sum approach to allow for the expected counts to vary from one time period to the next. The modified approach is applied to data on lower respiratory infection episodes reported by Boston area clinicians during the period January, 1996 to October, 1999.

## 2. Cumulative Sum (Cusum) Methods

- designed to detect sudden changes in the mean value of a quantity of interest
- widely used in industrial process control to monitor production quality
- rely upon assumptions that (a) the quantity being monitored is normally distributed, and (b) that the variable exhibits no serial autocorrelation.

Let the variable be converted to a z-score with mean 0 and variance 1. The cumulative sum, following observation  $t$  is defined as

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

where  $k$  is a parameter. A change in mean is signaled if  $S_t > h$ , where  $h$  is another parameter to be defined.

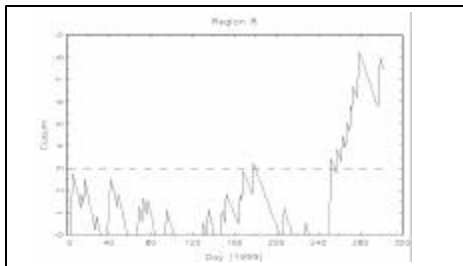
Thus values of  $z$  in excess of  $k$  are cumulated. The parameter  $k$  in this instance, where a standardized variable is being monitored, is often chosen to be equal to 1/2; in the more general case, it is often chosen to be equal to 1/2 the standard deviation associated with the variable being monitored.

The parameter  $h$  is chosen in conjunction with a predetermined acceptable rate "false alarms"; high values of  $h$  lead to a low probability of a false alarm, but also a lower probability of detecting a real change. Table 1 depicts the values of  $h$  associated with given average times until a false alarm. These times are called the "in-control" average run length, and are designated by the notation  $ARL_0$ . When  $k = 1/2$ , an approximation for  $ARL_0$  may be derived from:

$$ARL_0 \approx 2/(e^2 - \alpha - 1)$$

where  $\alpha = h - 1.166$ . One can make practical use of this approximation to choose the parameter  $h$  by first deciding upon a value of  $ARL_0$ , and then solving the approximation for the corresponding value of  $h$ . In the more general situation where a non-standardized variable is being monitored, the critical value of the cumsum is determined by multiplying the value of  $h$  by the standard deviation of the variable being monitored.

The choice of  $k=1/2$  minimizes the average out-of-control run length (that is, the time until a signal of change is sent when a real pattern change has occurred) for a given value of  $ARL_0$ .



## Poisson Cusum

When the variable being monitored has a Poisson distribution, other considerations are necessary for the determination of the parameters  $k$  and  $h$  (Lucas 1985).

Let  $I^{(0)}$  be the mean value of the in-control Poisson parameter. The signaling parameter  $h$  may be determined by first using  $I^{(0)}$  and the prespecified out-of-control parameter ( $I^{(1)}$ ) to find  $k$ . Following Lucas (1985), the corresponding  $k$ -value that minimizes the time to detect a change from  $I^{(0)}$  to  $I^{(1)}$  is

$$k = \frac{I^{(1)} - I^{(0)}}{\ln I^{(1)} - \ln I^{(0)}} \quad (1)$$

Then  $h$  can be found from the values of the parameter  $k$  and the desired ARL by using either a table (see, e.g., Lucas 1985), Monte Carlo simulation or an algorithm such as the one provide by White and Yeats (1996). For instance, if  $I^{(0)} = 4$ , one might desire to detect quickly a one standard deviation increase to  $I^{(1)} = 6$ . We would first find

$$k = \frac{6 - 4}{\ln 6 - \ln 4} = \frac{2}{\ln 1.5} \approx 4.93$$

Then, if we desired an ARL of approximately 420, we could use Table 2 of Lucas to find  $h = 10$ .

Suppose now that the expected in-control value associated with the Poisson or binomial random variable varies with time ( $I_t = 4, 2, \dots$ ). Simply implementing a cusum scheme with constant parameters would result in more false alarms if the actual values of  $I_t$  fluctuated from period to period.

Instead we will use values of the parameters  $k$  and  $h$  that are time-specific. The observed values,  $X_t$ , may then be used in the cumulative sum as follows (Rogerson 2002):

$$S_t = \max(0, S_{t-1} + (X_t - k_t)) \quad (2)$$

where the parameters  $c_t$  and  $k_t$  change from one period to the next, and their values are now discussed.

First  $h$  is chosen based upon the mean of the time-varying Poisson parameter, an associated value of  $k$ , and the desired ARL. Once  $h$  is chosen, next choose  $k_t$  based upon  $I_t^{(0)}$  and  $I_t^{(1)}$ :

$$k_t = \frac{I_t^{(1)} - I_t^{(0)}}{\ln I_t^{(1)} - \ln I_t^{(0)}} \quad (3)$$

Then  $c_t$  is chosen as the ratio to  $h$ , the value of the signaling parameter that would have been chosen in a usual Poisson scheme with desired ARL,  $k_t$  and constant values of  $I_t^{(0)}$  and  $I_t^{(1)}$  (designated  $h_0$ ). Thus  $c_t = h/h_0$ . The quantity  $c_t$  is therefore chosen so that observed counts,  $X_t$ , will make the proper relative contribution toward the signaling parameter  $h$  that is used in the actual cumsum. If for example  $h_0 = h$ , then the contribution  $X_t - k_t$  is scaled up by the factor  $h/h_0$ .

## 3. Data

Harvard Vanguard Medical Associates uses an automated record system for its 14 clinics in the Boston, Massachusetts area. Following each patient office visit, the clinician records diagnoses and International Classification of Disease (ICD-9) codes. Patient addresses are recorded; here they have been geocoded and assigned to census tracts.

Data on lower respiratory infection episodes were available for the period January, 1996 to October, 1999. During this time period, there were 80,683 episodes that could be assigned to one of the 566 census tracts in the study region.

## 4. Model for Expected Counts

A logit regression model was used for estimating expected counts in a region. The logistic transform of the probability of a visit is taken to be a linear function of the explanatory variables:

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + \dots + b_j x_j + e_i$$

where  $p_i$  is the probability of a visit in region  $i$ ,  $x_j$  is the value of explanatory variable  $j$  in region  $i$  and the  $b_j$ 's are the regression coefficients.

In comparison to the random effects model described by Kleinman et al. (2002), this has the advantage of having coefficients that are specific to individual regions. However, it is less convenient to construct a large number of models. More importantly, region-specific coefficients may not be reliable over time, especially when they are estimated from a small number of observations. An alternative would be to have region-specific dummy variables in a single equation, but this could potentially use a large number of degrees of freedom relative to the number of observations.

## 5. Poisson Cusum: Model and Results

To illustrate how the modified cumulative sum approach may be applied to monitor rare events, I arbitrarily chose a single census tract (the eighth tract in the data set). The first three years (1996-1998) were used as a base period, and monitoring began on January 1, 1999.

During the base period, there were 113 instances of lower respiratory events, an average of 0.1031 per day.

Logit model was used to regress the proportion of individuals in the census tract visiting a clinic on a given day against (a) weekend versus weekday (with weekday as the reference category), (b) month of year (with December as the reference category), (c) time, where the days of the base period were numbered from 1 to 1096, and (d) one- and two-day lagged values of the number of previous visits (to account for the possibility of temporal autocorrelation). Backward selection was used to eliminate insignificant variables. Table 2 displays the results. April, June, July, September, and October have significantly fewer visits, in comparison with December, during the base period. Not surprisingly, days that are weekends have significantly reduced odds of a visit. The time trend and the one- and two-day lagged observations were not significant.

Next, the estimated coefficients were used with information on the explanatory variables to generate an expected number of visits for each day for the period January-October, 1999.

An overall value of  $h$  was derived using  $ARL=300$ ,  $I_0 = 1031$ ,  $I_1 = 1031 + \sqrt{1031/2} = 26$ , and  $k$ , which from equation (1) is equal to 0.17. This leads to a value of  $h=3.0$ .

For each expected value ( $I_t$ ), a corresponding value of  $h_t$  was derived; these values were associated with an ARL of 300, and  $k_t$  values chosen using Equation (3) and an alternative expectation of  $I_t$ ,  $\sqrt{I_t/2}$ .

The Poisson cusum (Equation 2) was then started on January 1, 1999, using the observed number of visits, the expected number of visits, and values of  $h$  and  $k$  as described above. Figure 1 displays the cusum. A signal is sounded on day 178 (June 27) but this is not sustained for only one additional day; a more sustained signal is sounded on day 252 (September 9), and this change is sustained for the remainder of the period. One case on September 8 and two on September 9 were sufficient to sound the alarm. There were a total of 10 cases in the 28-day period between September 8 and October 6, 1999. This is an average of .36 cases per day, substantially higher than the mean of .10 observed during the base period.

## 6. Discussion

### Extensions to multiregional systems:

- one straightforward extension is to monitor  $p$  regions simultaneously; the only necessary modification is to multiply the desired ARL by  $p$  to implement a Bonferroni adjustment so that the average time to a false alarm in some region is equal to ARL. This results in a higher value of  $h$ .

- a more sophisticated extension would be to construct "local statistics" in association with each geographic unit. These would be defined as a weighted average of the region's observation and surrounding observations, where the weights would decline with distance. Cumulative sums associated with these local statistics would be monitored. Because the local statistics are spatially autocorrelated, a simple Bonferroni adjustment would result in too high a value of  $h$ , making it difficult to detect change. Work here should be devoted to developing more realistic adjustments.

### Extra-Poisson variation

- a common problem in assuming a Poisson distribution is that actual data exhibit overdispersion; the variance associated with observed counts is greater than the mean (while for a Poisson distribution it would be equal to the mean). Diggle et al. (1997) suggest allowing for the possibility that  $\text{Var}(X_i) = kI_i$ , where  $k > 1$ . An estimate of  $k$  is  $c^2/n$  where  $c^2$  is the usual chi-square statistic associated with the observed and fitted values, and  $n$  is the residual degrees of freedom associated with the regression model. The standard errors from the regression model should then be multiplied by  $\sqrt{k}$ . In the illustration here, we find  $k = 1.66$ , implying some overdispersion, and as a consequence, some of the parameters of the regression model in Table 1 are no longer significant.

More importantly, the overdispersion can cause premature signaling in the Poisson cusum. One approach would be the use of a cusum that assumed a negative binomial distribution; Muddapur (1974) indicates how the choice of the parameter  $k$  can be made in this instance, and the Markov chain approach outlined in White and Yeats (1996) could be used to determine the parameter  $h$ .

## References

Diggle, P., Elliott, P., Morris, S., and Shadick, G. 1997. Regression modeling of disease risk in relation to point sources. *Journal of the Royal Statistical Society Series A* 160: 491-505.

Kleinman, K., Lazarus, R., and Platt, R. 2002. A generalized linear mixed models approach for detecting incident clusters of disease: biological terrorism and other surveillance. Manuscript.

Kulldorf, M. 2001. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society Series A* 164: 61-72.

Lucas, J. M. 1985. Counted data cusums. *Techonometrics* 27: 129-144.

Muddapur, M.V. 1974. V-mask for the negative binomial process. *Journal of the Indian Statistical Association* 12: 31-38.

Raubertas, R. F. 1989. An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Statistics in Medicine* 8: 267-71.

Rogerson, P. 1997. Surveillance methods for monitoring the development of spatial patterns. *Statistics in Medicine* 16: 2081-2093.

Rogerson, P. 2001. Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society, Series A*, 164: 87-96.

Rogerson, P. 2002. Cumulative sum methods for Poisson and binomial variables with varying expectations. Submitted for publication.

White, C.H. and Yeats, J.B. 1996. ARLs and higher order run length moments for Poisson CUSUM. *Journal of Quality Technology* 28: 363-369.

## Acknowledgements

Ken Kleinman kindly provided the data on lower respiratory infection episodes. This work was supported in part by National Cancer Institute Grant R01 CA92693-01.

Table 1. In-Control ARLs (False Alarm Rates) for Various Values of  $h$

$h$	ARL <sub>0</sub>
2.5	69.9
2.6	76.9
2.7	85.8
2.8	95.6
2.9	106.5
3.0	118.6
3.1	131.9
3.2	146.7
3.3	163.1
3.4	181.2
3.5	201.2
3.6	223.4
3.7	247.9
3.8	275.0
3.9	304.9
4.0	338.1
4.1	374.7
4.2	415.3
4.3	460.1
4.4	509.6
4.5	564.4
4.6	625.0
4.7	691.9
4.8	765.0
4.9	847.8
5.0	938.2

Table 2. Coefficients in Logit Regression Model

Variable	Coefficient	Coefficient/Std. error
April	-0.952	-2.07
June	-1.435	-2.45
July	-1.222	-2.39
September	-0.950	-2.07
October	-0.993	-2.16
Weekend	-1.052	-3.69
Intercept	-7.935	