

# Surveillance of Infectious Disease Data using Cumulative Sum Methods

Michaela Paul<sup>1</sup>   Michael Höhle<sup>2</sup>   Leonhard Held<sup>1</sup>

<sup>1</sup>Institute of Social and Preventive Medicine  
University of Zurich

<sup>2</sup>Department of Statistics  
University of Munich

Swiss Statistics Meeting 2007  
Lucerne, 15 November

# Outline

- 1 Introduction
- 2 Cumulative Sum (CUSUM) schemes
  - Standard CUSUM schemes
  - Approximate Gaussian CUSUM
  - Modified Poisson CUSUM
- 3 Performance
- 4 Summary and Conclusions

# Introduction

## Aim

Detect a sudden increase in incidence as soon as possible

## Common method

Compare observed counts  $X_t$  with a predicted value  
If the deviation is too large, give an alarm

## Idea

Improve the detection ability by using e.g. Statistical Process Control (SPC) methods

## Standard CUSUM schemes

Assume that given change point  $\nu$

$$X_t \sim \begin{cases} F_{\theta_0}, & t = 1, \dots, \nu - 1 \quad (\text{in-control}) \\ F_{\theta_1}, & t = \nu, \nu + 1, \dots \quad (\text{out-of-control}) \end{cases}$$

with constant parameters  $\theta_0 < \theta_1$

### Aim

Detect this change point on-line as soon as possible after it has occurred

## Standard CUSUM schemes II

### How?

For each time point  $n$ : Test the hypotheses

$$H_0 : X_t \sim F_{\theta_0}, \quad t = 1, \dots, n$$

versus

$$H_1 : X_t \sim \begin{cases} F_{\theta_0}, & t = 1, \dots, \nu - 1 \\ F_{\theta_1}, & t = \nu, \nu + 1, \dots, n \end{cases}$$

with likelihood-ratio tests

Stop, i.e. give an alarm if  $H_0$  can be rejected

## Definition of the CUSUM

For members of a single-parameter exponential family:

### Decision Interval CUSUM

$$C_0 = 0, \quad C_n = \max(0, C_{n-1} + X_n - k), \quad n \geq 1$$

with stopping-rule  $N = \inf\{n : C_n \geq h\}$

$k$  is called **reference value** and  $h$  is the **decision interval**

The reference value  $k$  is completely determined by the parameter values in  $H_0$  and  $H_1$ :

- $X_t \sim \text{Poisson}$ :  $k = \frac{\lambda_1 - \lambda_0}{\ln(\lambda_1) - \ln(\lambda_0)}$
- $X_t \sim \text{Normal with fixed variance}$ :  $k = \frac{\mu_0 + \mu_1}{2}$

## Choice of CUSUM parameters

### Reference value $k$

Choose the size of shift for which you desire quickest detection

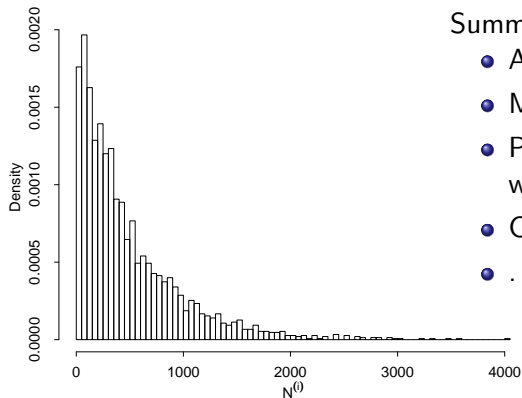
### Decision interval $h$

Choice should give good performance in terms of the run length

### Run length $N$

Random number of observations from the starting point up to the point at which the decision interval  $h$  is crossed

## Run length (RL) distribution



### Summaries of the RL distribution

- Average run length (ARL)
- Median run length
- Probability of a false alarm within the first  $m$  time points
- Conditional expected delay
- ...

**Example:**  $X_t \sim \text{Po}(5)$ ,  $t = 1, 2, \dots$   
 CUSUM with  $k = 6.1$ ,  $h = 11.2$

## Average Run Length

### In-control ARL

Mean time before a false alarm  $ARL_0 = E(N|\nu = \infty)$

### Out-of-control ARL

Mean time before the first true alarm  $ARL_1 = E(N|\nu = 1)$

ARLs for the CUSUM scheme can be computed by

- solving integral equations
- Markov chain approximation (Brook and Evans, 1972)
- Monte Carlo estimation

## Markov chain approach

Decision Interval CUSUM:

$$C_n = \max(0, C_{n-1} + X_n - k), \quad \text{with stopping rule } C_n \geq h$$

- $X_n$  discrete
- Rational reference value, say  $k = \frac{K}{M}$   
 $\Rightarrow$  range of all possible values of  $C_n$  is limited

$$\text{State } 0 \quad C_n = 0$$

$$\text{State } i \quad C_n = \frac{i}{M}, \quad i = 1, \dots, M \cdot h - 1$$

$$\text{State } M \cdot h \quad C_n \geq h$$

### ARL

Average time between visits to state  $M \cdot h$

## Time-varying parameters

- The incidence of infectious diseases is usually not constant but seasonally varying
- A CUSUM with constant in-control parameter is too restrictive

Now let  $X_t \sim \text{Poisson}$  with time-varying mean  $\lambda_t$

### Approaches

- Use a standard CUSUM and apply it to transformed observations
- Modify the CUSUM method itself and apply it to the original observations

## Approximate Gaussian CUSUM

Rossi et al. (1999)

Transform the counts  $X_t$  to normality using

$$Z_t = \frac{X_t - 3\lambda_t + 2\sqrt{X_t\lambda_t}}{2\sqrt{\lambda_t}} \stackrel{a}{\sim} N(0, 1)$$

and apply a Gaussian CUSUM with reference value  $k = \frac{\Delta}{2}$  to these standardized counts  $Z_t$

Other transformations, e.g.

- Pearson residual
- deviance residual

# Modified Poisson CUSUM

## Rogerson and Yamada (2004)

- 1 Compute time-varying reference values  $k_t$
- 2 Choose time-varying decision intervals  $h_t$  that give a target  $ARL_0$
- 3 Compute

$$S_t = \max\left\{0, S_{t-1} + \frac{h}{h_t}(x_t - k_t)\right\}, t \geq 1$$

- 4 Decide on alarm if  $S_t \geq h$

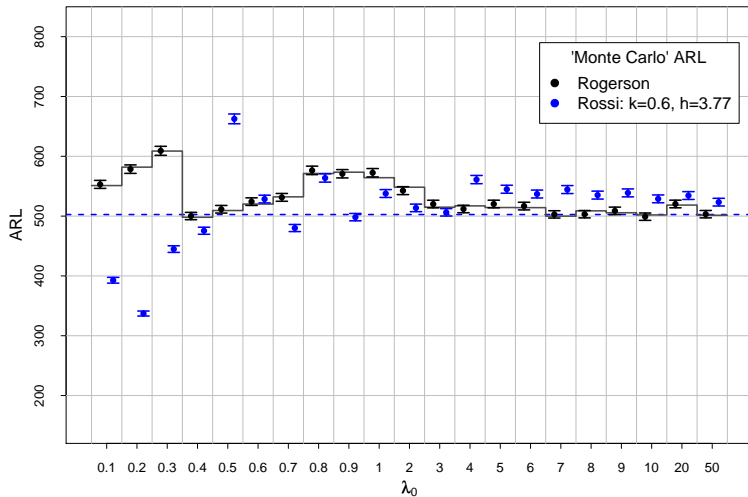
## ARL Performance

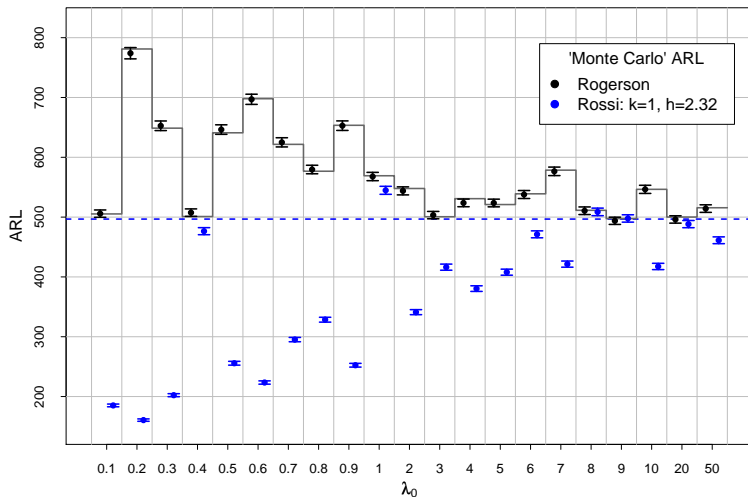
Simulated sequences

$$X_t \sim \begin{cases} \text{Po}(\lambda_{0,t}), & t = 1, \dots, \nu - 1 \\ \text{Po}(\lambda_{0,t} + \delta \sqrt{\lambda_{0,t}}), & t = \nu, \dots, L \end{cases}$$

CUSUM with

- $\Delta = 1.2$  and  $\Delta = 2$
- $ARL_0 = 500$

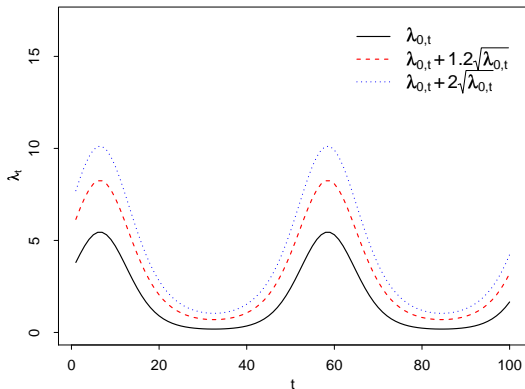
In-control ARL,  $\Delta = 1.2$ 

In-control ARL,  $\Delta = 2$ 

## Performance for time-varying in-control parameter

Assume  $X_t \sim \text{Po}(\lambda_t)$  with

$$\log(\lambda_t) = \alpha + \sum_{s=1}^S \left( \beta_s \sin\left(\frac{2\pi s}{52} t\right) + \gamma_s \cos\left(\frac{2\pi s}{52} t\right) \right)$$



## Average Run Lengths

	$\Delta = 1.2$		$\Delta = 2$	
	Rogerson	Rossi	Rogerson	Rossi
$\delta = 0$	548 (3.4)	481 (3.0)	572 (3.6)	290 (1.8)
$\delta = 1$	10.4 (0.04)	10.9 (0.05)	12.5 (0.07)	12.7 (0.07)
$\delta = 1.2$	7.8 (0.03)	8.0 (0.03)	8.8 (0.04)	9.0 (0.05)
$\delta = 2$	4.0 (0.01)	3.9 (0.01)	3.8 (0.01)	3.6 (0.01)
$\delta = 2.5$	3.1 (0.01)	3.0 (0.01)	2.8 (0.01)	2.7 (0.01)

CUSUMs are designed to detect a shift of size  $\Delta$

## Conditional Expected Delay

A shift hardly occurs immediately at the start of the surveillance  
⇒ use performance measures that also consider the change point

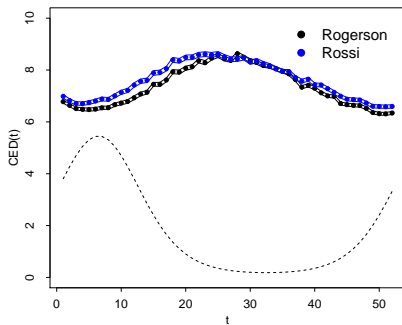
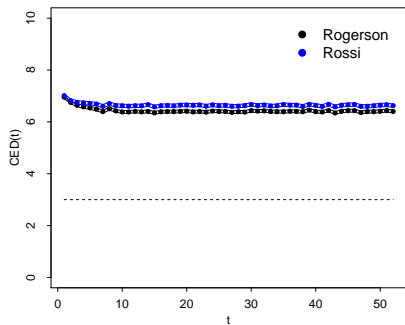
### Conditional Expected Delay

Average delay for an alarm when the change occurs at time  $t$

$$\text{CED}(t) = E(N - \nu \mid N \geq \nu, \nu = t)$$

Note that when the shift occurs at  $\nu = 1$ , the out-of-control *ARL* corresponds to  $\text{CED}(1) + 1$

# Conditional Expected Delay



## Summary and Conclusions

- Comparison of two approaches that allow for a time-varying in-control parameter
- Modified CUSUM of Rogerson and Yamada (2004) shows better ARL performance
- There are better performance measures than the ARL
- The size of shift is usually not known in advance  
⇒ extensions of the CUSUM to test for a shift to an unknown distribution

## References



Brook, D. and Evans, D. A. (1972).

An approach to the probability distribution of CUSUM run length.  
*Biometrika*, 59(3):539–549.



Hawkins, D. M. and Olwell, D. H. (1998).

*Cumulative Sum Charts and Charting for Quality Improvement*.  
Statistics for Engineering and Physical Science. New York: Springer.



Rogerson, P. A. and Yamada, I. (2004).

Approaches to syndromic surveillance when data consist of small regional counts.  
*Morbidity and Mortality Weekly Report*, 53/Supplement:79–85.



Rossi, G., Lampugnani, L., and Marchi, M. (1999).

An approximate CUSUM procedure for surveillance of health events.  
*Statistics in Medicine*, 18:2111–2122.