

# 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

ROeS 2007  
Bern, 11 September

# Outline

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

# Introduction

## Surveillance of infectious diseases

Aim is to detect a sudden increase in incidence as soon as possible

## Common method

Compare observed counts  $X_t$  with an expected value

If  $X_t$  is larger than some upper prediction interval, give an alarm

## Idea

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

## Standard CUSUM schemes

Assume that given an unknown 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 as soon as possible after it has occurred

### How?

Repeatedly test the hypotheses

$$H_0 : X_t \sim F_{\theta_0} \quad \text{versus} \quad H_1 : X_t \sim F_{\theta_1}$$

with likelihood-ratio tests

## 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**

$k$  is completely determined by the parameter values in  $H_0$  and  $H_1$

- $X_t \sim \text{Po}(\lambda)$ :  $k = \frac{\lambda_1 - \lambda_0}{\ln(\lambda_1) - \ln(\lambda_0)}$
- $X_t \sim \text{N}(\mu, \sigma^2)$ , with  $\sigma^2$  fixed:  $k = \frac{\mu_0 + \mu_1}{2}$

## Performance measures

### Run length $N$

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

Choice of  $h$  should be based on performance measures like

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

# 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
- Monte Carlo estimation

## Approximate Gaussian CUSUM

Now let  $X_t \sim \text{Po}(\lambda_t)$  with time-varying mean  $\lambda_t$

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 to these standardized counts  $Z_t$

## Modified Poisson CUSUM

### Rogerson and Yamada (2004)

Compute time-varying reference values  $k_t = \frac{\lambda_{1,t} - \lambda_{0,t}}{\ln(\lambda_{1,t}) - \ln(\lambda_{0,t})}$

and scale contributions  $(x_t - k_t)$  with  $c_t = h/h_t$  to keep the overall in-control ARL fixed.

$$S_0 = 0, \quad S_t = \max\{0, S_{t-1} + c_t(x_t - k_t)\}, \quad t \geq 1$$

with stopping-rule  $N = \inf\{t : S_t \geq h\}$

$h_t$  corresponds to the decision interval that gives the target  $ARL_0$  for a standard Poisson CUSUM with reference value  $k_t$

## ARL Performance

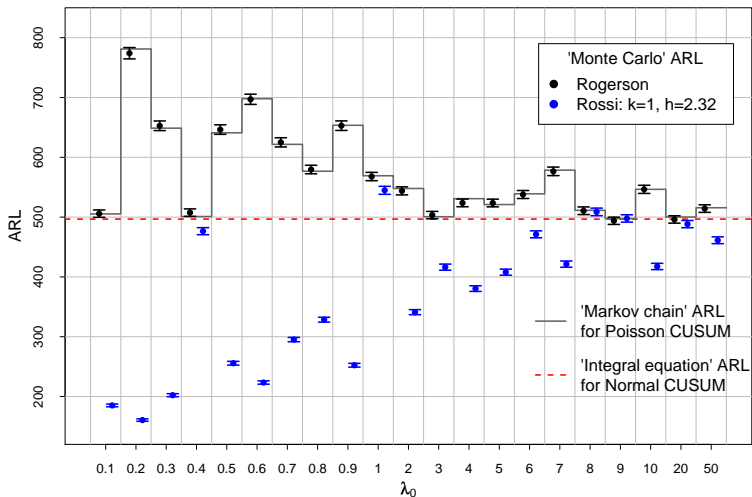
Simulate  $n = 25000$  sequences

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

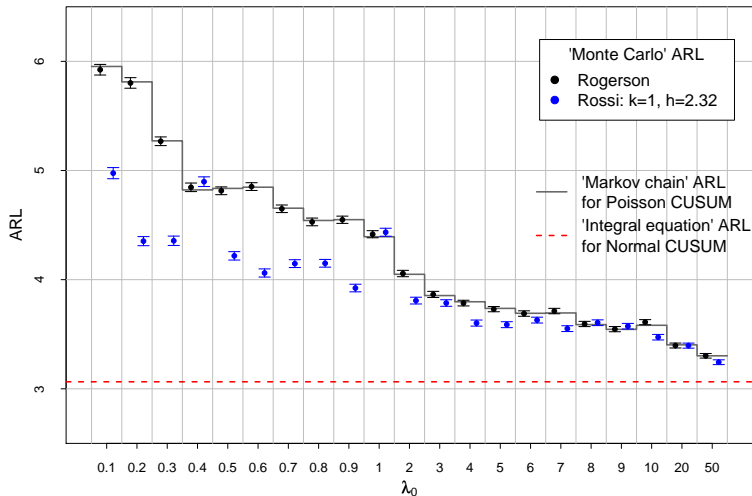
A CUSUM designed to detect a shift of size  $\Delta$  is applied

- $\Delta = 1.2$  ( $ARL_1=7$ ) and  $\Delta = 2$  ( $ARL_1=3$ )
- $ARL_0=500$
- $L$  sufficiently large

# In-control ARL



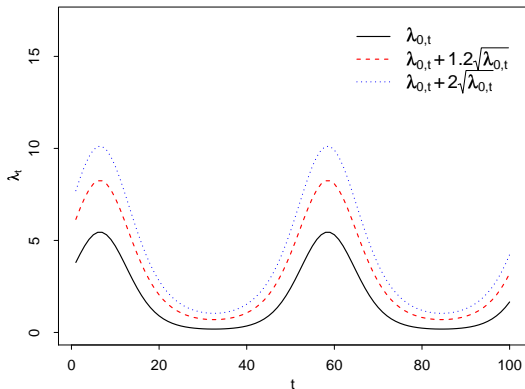
# Out-of-control ARL



# 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

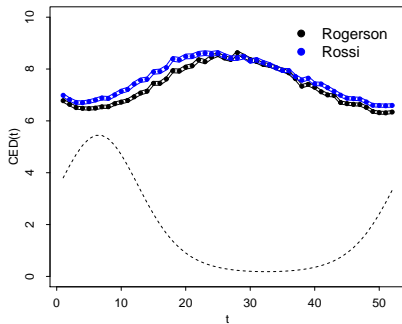
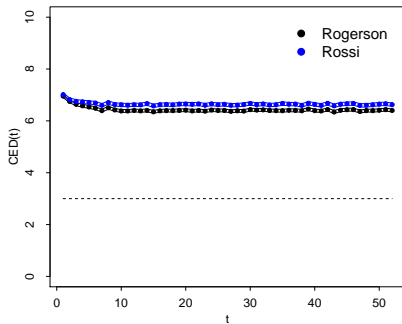
### Conditional Expected Delay

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

- CUSUM for (approximately) normal residuals leads to substantially more false alarms in case of small means
- There are better performance measures than the average run length