EWMA Monitoring of Correlated Processes of Poisson Counts.

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics

This talk is based on the articles

Weiß, C.H. (2008). *EWMA Monitoring of Correlated Processes of Poisson Counts*. Submitted to Quality Technology & Quantitative Management.

Weiß, C.H. (2007). Controlling correlated processes of Poisson counts. Quality & Reliability Engineering International 23(6), 741-754.

All references mentioned in this talk correspond to the references in this article.

Processes of Poisson Counts

Motivation

Count data arises in many different situations relevant for SQC. Often modelled by Poisson distribution.

Example: Application log data of Statistics web server:

```
84.170.62.177 - - [23/May/2005:11:16:36 +0200]
''GET /~weiss/kolloquium/ss_05_analysis_weiss.html HTTP/1.1'' 404 1361
''http://132.187.92.36/lehre/index.html''
''Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0; QXW0339m;
Q312461)''
```

containing information such as the host name of the user accessing a Web site, date and time of the request, etc.

Example: (continued)

Log data was transformed

$\Rightarrow$ for periods of two minutes length:

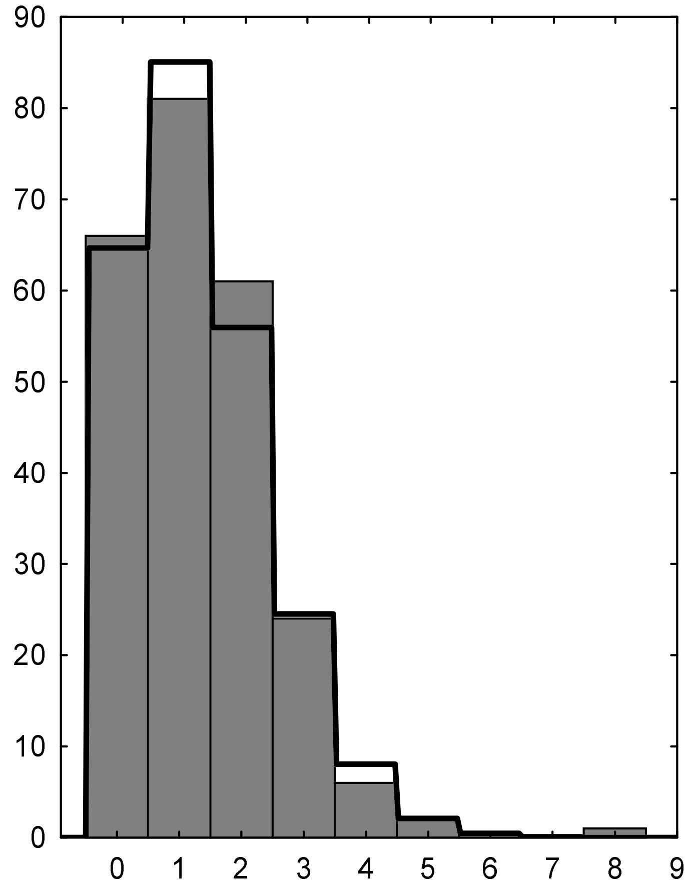
number of *different* IP addresses registered within period,

between 10 o'clock a.m. and 6 o'clock p.m.

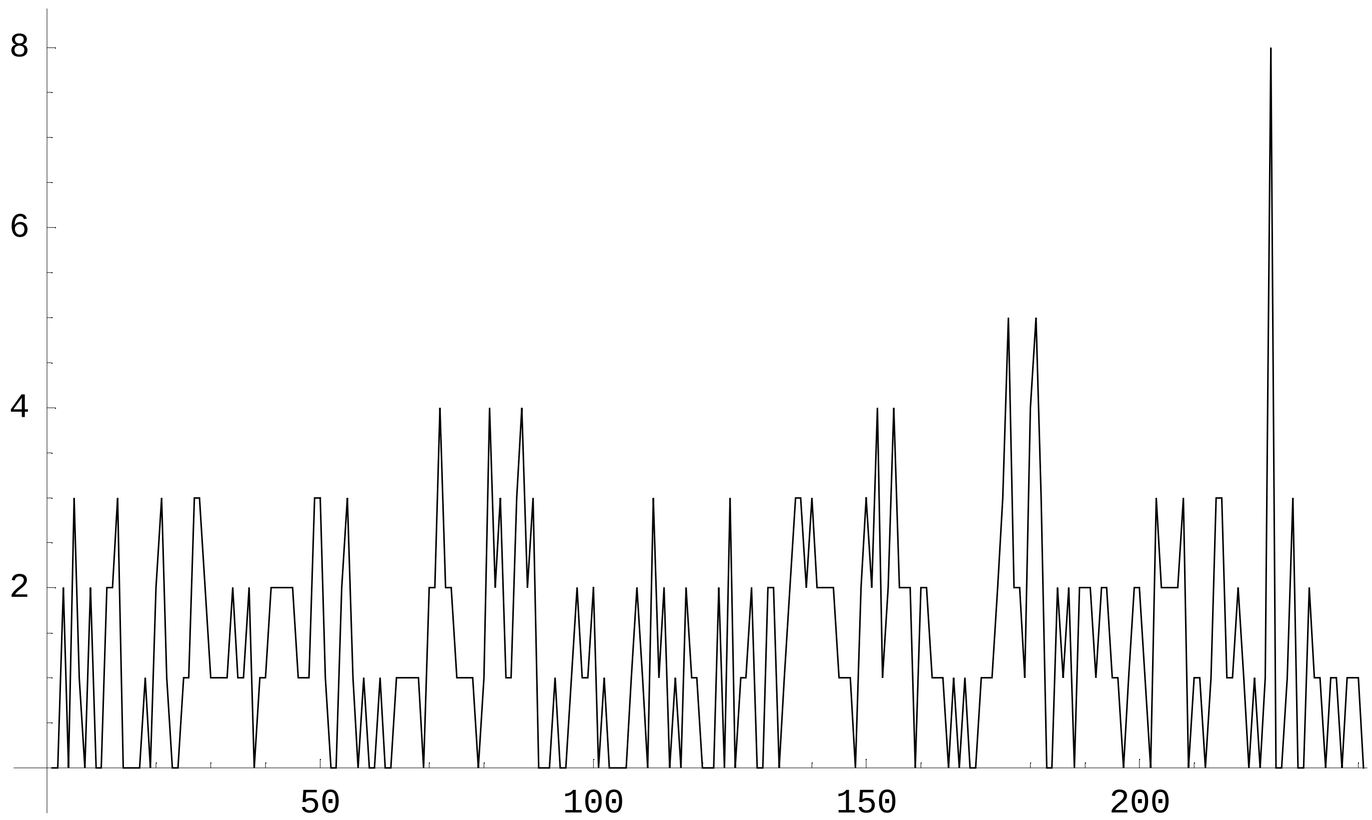
$\Rightarrow$ time series of length 241 each.

Example: (continued)

Data collected on November 29th, 2005: Histogram with fitted Poisson distribution $Po(1.32)$.



Example: (continued) Data from 29.11.2005:



Example: (continued)

Data collected on November 29th, 2005:

Outlier at time $t = 224$: 8 different IP addresses of form 195.93.60.xxx.

Altogether 191 accesses around 17:26 of this type.

Known phenomenon: Users of the AOL browser routed into internet through IP addresses of form 195.93.60.xxx. Any user gets permanently a new address of this area. Therefore: not possible to infer the user from the IP address.

After removing the outlier,

Weiß (2007) showed that the serial dependence structure of the time series can modelled very well by a stationary Poisson INAR(1) model.

Poisson INAR(1) Processes

Definition & Properties

Definition of INAR(1) process:

Let $(\epsilon_t)_{\mathbb{N}}$ be i.i.d. process with marginal distribution $Po(\mu)$, let $\alpha \in (0;1)$. Let $N_0 \sim Po(\frac{\mu}{1-\alpha})$. If the process $(N_t)_{\mathbb{N}_0}$ satisfies

$$N_t \;=\; \alpha \circ N_{t-1} \;+\; \epsilon_t, \qquad t \geq 1,$$

plus sufficient independence conditions, then it follows a stationary *Poisson INAR(1) model* with marginal distribution $Po(\frac{\mu}{1-\alpha})$.

McKenzie (1985), Al-Osh & Alzaid (1987, 1988)

Binomial thinning, due to Steutel & van Harn (1979):

N discrete random variable with range $\{0, \ldots, n\}$ or $\mathbb{N}_0$.

Binomial thinning

$$\alpha \circ N \;:=\; \sum_{i=1}^{N} X_i,$$

where X_i are independent Bernoulli trials $\sim B(1, \alpha)$.

Guarantees that right-hand side always integer-valued:

$$N_t \;=\; \alpha \circ N_{t-1} \;+\; \epsilon_t.$$

Interpretation of INAR(1) process:

$$\underbrace{N_t}_{\text{Population at time } t} = \underbrace{\alpha \circ N_{t-1}}_{\text{Survivors of time } t-1} + \underbrace{\epsilon_t}_{\text{Immigration}}.$$

Interpretation applies well to many real-world problems:

Examples of INAR(1) process: (continued)

- N_t: number of users accessing web server, ϵ_t: number of new users, $\alpha \circ N_{t-1}$: number of previous users still active.

- N_t: number of customers. ϵ_t: new customers, $N_{t-1} - \alpha \circ N_{t-1}$: customers lost at end of last period.
  $\rightarrow$ Brännäs et al. (2002): guest nights in hotels.

- . . . and many more.

The INAR(1) model . . .

- is of simple structure,
- essential properties known explicitly,
- is easy to fit to data,
- is easy to interpret,
- applies well to real-world problems, . . .

In a nutshell: A simple model for autocorrelated counts, which is well-suited for SPC!

Controlling Poisson INAR(1) Processes

Control Concepts

Poisson INAR(1) model:

$(N_t)_{\mathbb{N}_0}$ is stationary Poisson INAR(1) process with innovations $(\epsilon_t)_{\mathbb{N}} \sim Po(\mu)$. So $N_t \sim Po(\frac{\mu}{1-\alpha})$.

State of statistical control: $\mu = \mu_0$ and $\alpha = \alpha_0$.

Known control schemes for i.i.d. Poisson data:

- c- and u-chart, see Montgomery (2005)
- Q chart of Quesenberry (1991), considering skewness of Poisson distribution
- Poisson EWMA chart of Borror et al. (1998)
- CUSUM charts of Brook & Evans (1972), Lucas (1985)

But above count data obviously autocorrelated!

Weiß (2007) proposed the following control charts:

- c-Chart for Poisson INAR(1),
- Residual control chart,
- Conditional control chart,
- Moving average control chart.

Weiß (2007): Simulation study for ARL performance.

Result: None of charts can be applied universally.

The moving average charts are most suitable, but

- computation of plotted statistics relatively complex,
- in particular:
  Exact ARLs are extremely difficult to obtain.

Therefore, ...

The Combined EWMA Chart

Control Concept

The standard c-**chart**:

Observed counts N_t plotted on chart with control region

$$\mathcal{C}_c(l,u) := \{l,\ldots,u\}, \qquad l,u \in \mathbb{N}_0, \qquad 0 \leq l < \mu_{N,0} < u.$$

Process considered as being in control unless $N_t \notin \mathcal{C}_c(l,u)$.

The **Poisson EWMA chart** of Gan (1990):

Let $\lambda \in (0;1]$ and $q_0 \in \mathbb{N}_0$. Define

$$Q_t \;=\; \text{round}\big(\lambda \cdot N_t \;+\; (1-\lambda)\cdot Q_{t-1}\big), \quad t \in \mathbb{N}, \qquad Q_0 := q_0.$$

Statistics Q_t plotted on chart with control region

$$\mathcal{C}_e(l,u) := \{l,\ldots,u\}, \qquad l,u \in \mathbb{N}_0, \qquad 0 \leq l < u.$$

Process considered as being in control unless $Q_t \notin \mathcal{C}_e(l,u)$.

Theorem: (Weiß, 2008)

Let $(N_t)_{\mathbb{N}_0}$ be stationary Poisson INAR(1) process.
Then $(N_t, Q_t)_{\mathbb{N}}$ is bivariate Markov chain with range $\mathbb{N}_0^2$.
Transition probabilities

$$\begin{aligned} p(a,b|c,d) &:= P(N_t = a, Q_t = b \mid N_{t-1} = c, Q_{t-1} = d) \\ &= \mathbb{1}_{[b-\frac{1}{2};b+\frac{1}{2})}(\lambda a + (1-\lambda)d) \cdot p_{a|c}. \end{aligned}$$

Furthermore,

$$\begin{aligned} p_1(a,b|q_0) &:= P(N_1 = a, Q_1 = b \mid Q_0 = q_0) \\ &= \mathbb{1}_{[b-\frac{1}{2};b+\frac{1}{2})}(\lambda a + (1-\lambda)q_0) \cdot p_a. \end{aligned}$$

Combined EWMA Chart – Concept

Idea: Combine c- and EWMA chart, i. e., monitor both N_t and Q_t simultaneously.

Advantages:

- EWMA charts sensitive to small shifts, while large shifts often detected more quickly with Shewhart charts $\Rightarrow$ combine potential of both chart types.
- $(N_t, Q_t)_{\mathbb{N}}$ Markov chain $\Rightarrow$ exact ARL computation with approach of Brook & Evans (1972).

Combined EWMA Chart:

Let $l_c, u_c, l_e, u_e \in \mathbb{N}_0$ with $0 \leq l_c < u_c$ and $0 \leq l_e < u_e$.

Observed pairs (N_t, Q_t) plotted simultaneously on c-chart with control region $\mathcal{C}_c(l_c, u_c)$ and EWMA chart with control region $\mathcal{C}_e(l_e, u_e)$.

Process considered as being in control
unless $N_t \notin \mathcal{C}_c(l_c, u_c)$ or $Q_t \notin \mathcal{C}_e(l_e, u_e)$.

Combined EWMA Chart: ARL computation.

Exact ARL computation through solving appropriate system of linear equations

$$(\mathbf{I} - \mathbf{Q}) \cdot \boldsymbol{\mu} \; = \; \mathbf{1}.$$

Dimension of matrix determined by number of reachable in-control states.

For details and proofs, see Weiß (2008).

The Combined EWMA Chart

ARL Performance

Chart design $(\lambda, l_c, u_c, l_e, u_e)$ based on ARL_0:

1. Choose design (l_c, u_c) of c-chart such that ARL_0 larger than desired value.

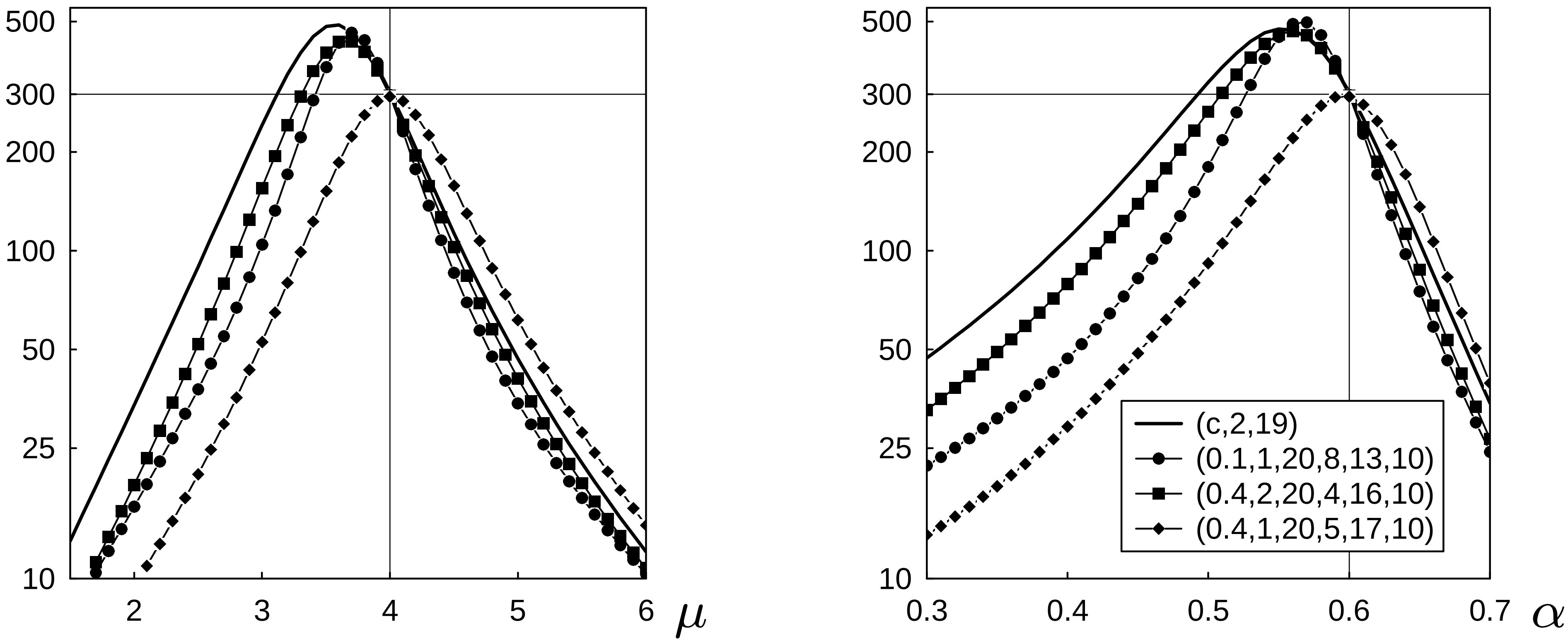
2. Choose sufficiently narrow design (l_e, u_e) of EWMA part such that total ARL_0 below desired value given $\lambda = 1$.

3. Decrease λ to adjust ARL_0 close to desired value.

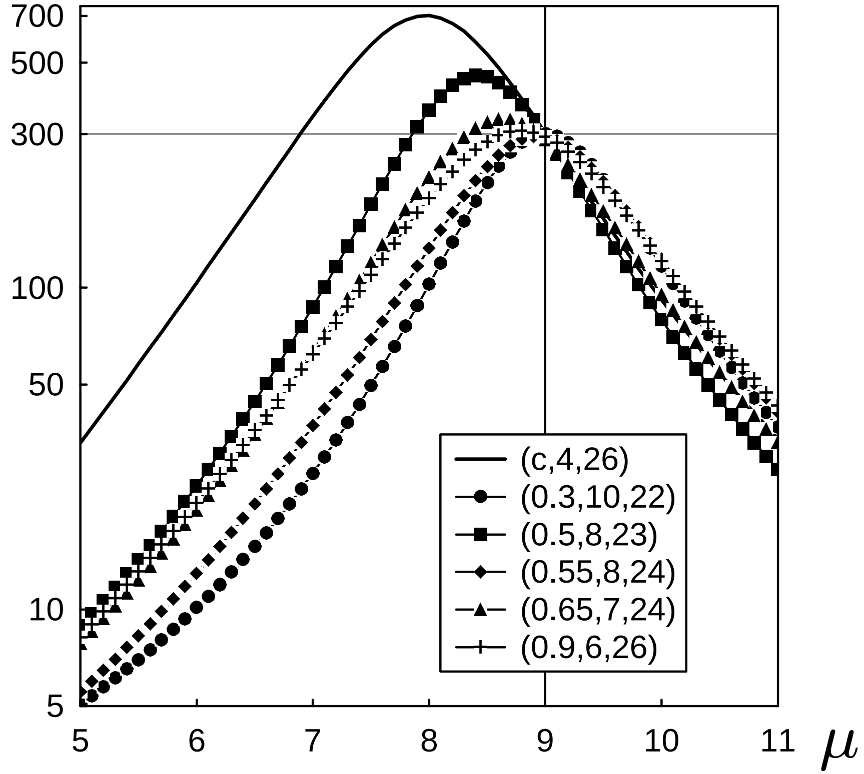
For tabulated design recommendations, see Weiß (2008).

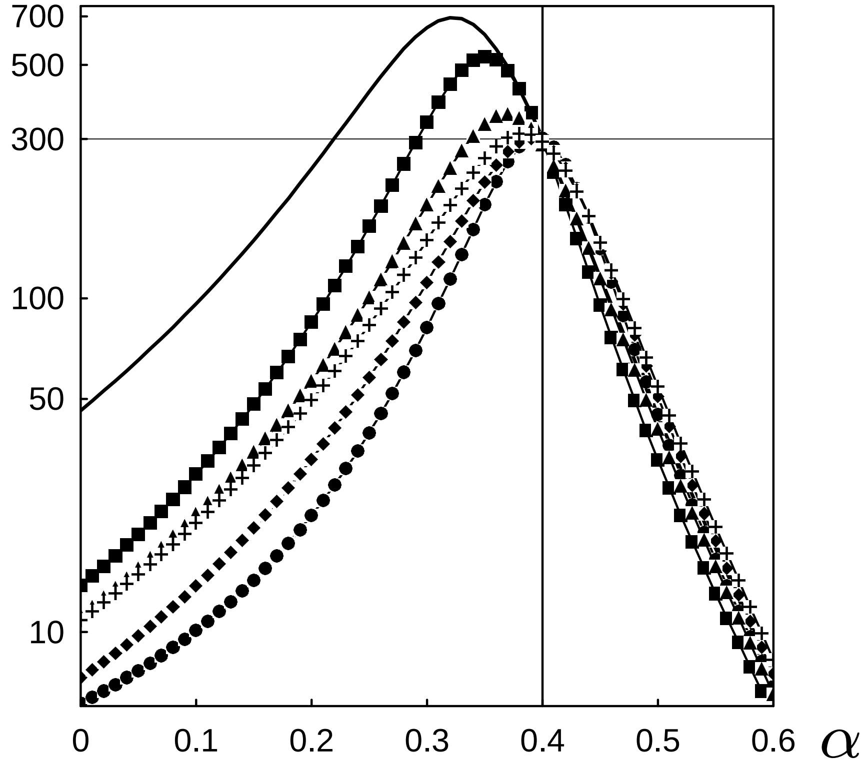
ARL of combined EWMA charts

with $(\mu_0, \alpha_0) = (4, 0.6)$ against μ and α:



ARL of combined EWMA charts with $(\mu_0, \alpha_0) = (9, 0.4)$ against μ and α:





Combined EWMA Chart – ARL Performance

Combined EWMA chart allows to significantly improve performance of single c-chart:

- ARL curves of respective c-chart always biased and perform worst among all charts.
- Additional EWMA part can either be used to improve sensitivity of chart to detect upward shift in μ and α.
- Or one can reach nearly unbiased ARL performance.

The Combined EWMA Chart

Real-data Example

Combined EWMA Chart – Real-data Example

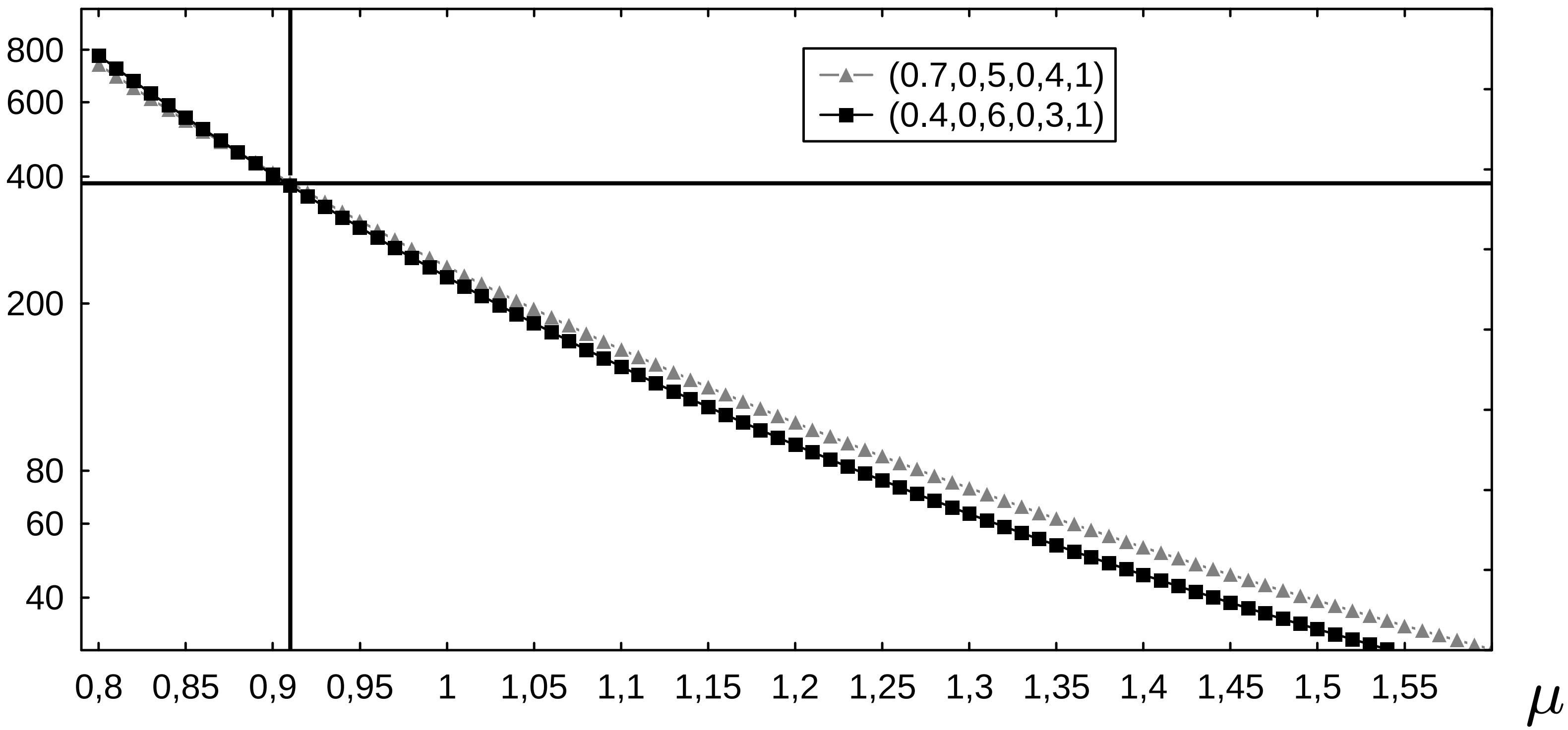
We continue the real-data example of IP counts.

In-control model, based on results from 29 November 2005:
INAR(1) model with parameters $\mu_0 = 0.91$ and $\alpha_0 = 0.29$

We consider time series $n_1, \ldots, n_{241}$ collected a week later on 6 December 2005, between 10 a.m. and 6 p.m.

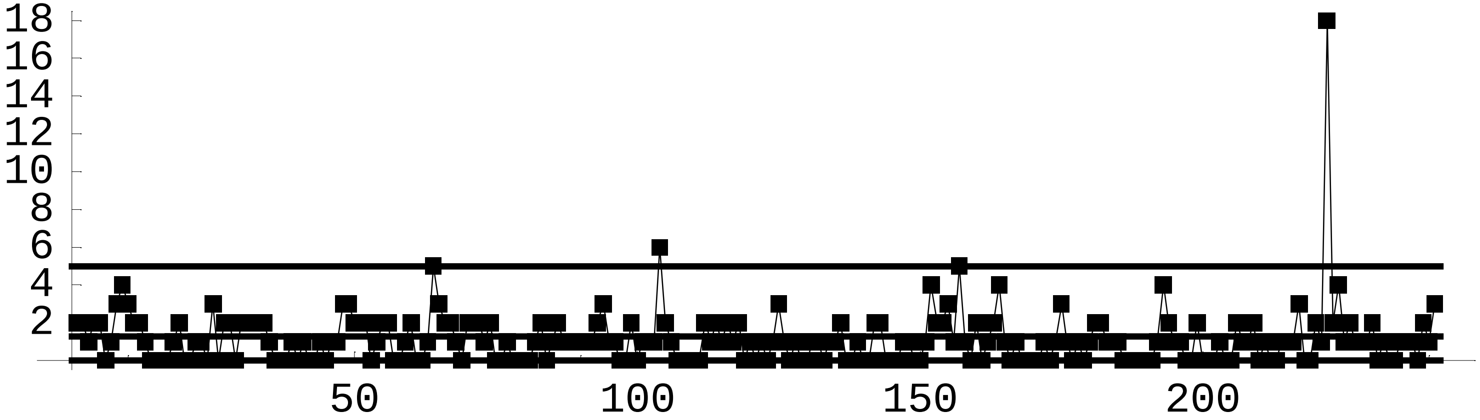
Design: $\lambda = 0.7,\ l_c = l_e = 0,\ u_c = 5,\ u_e = 4,\ z_0 = 1.$

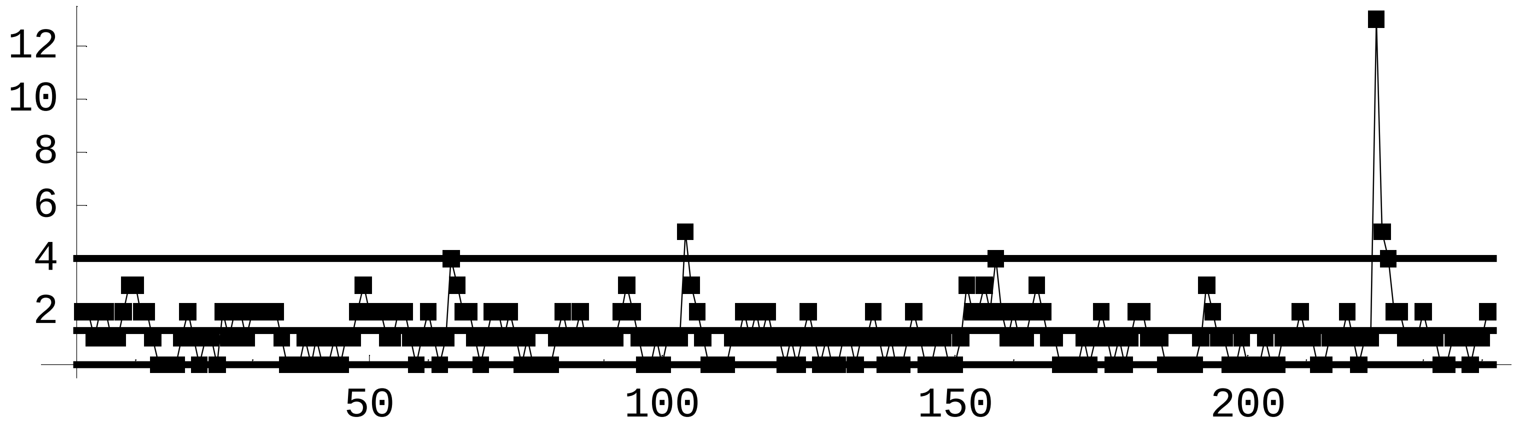
Design: $\lambda = 0.4,\ l_c = l_e = 0,\ u_c = 6,\ u_e = 3,\ z_0 = 1.$



Design: $\lambda = 0.7$, $l_c = l_e = 0$, $u_c = 5$, $u_e = 4$, $z_0 = 1$.

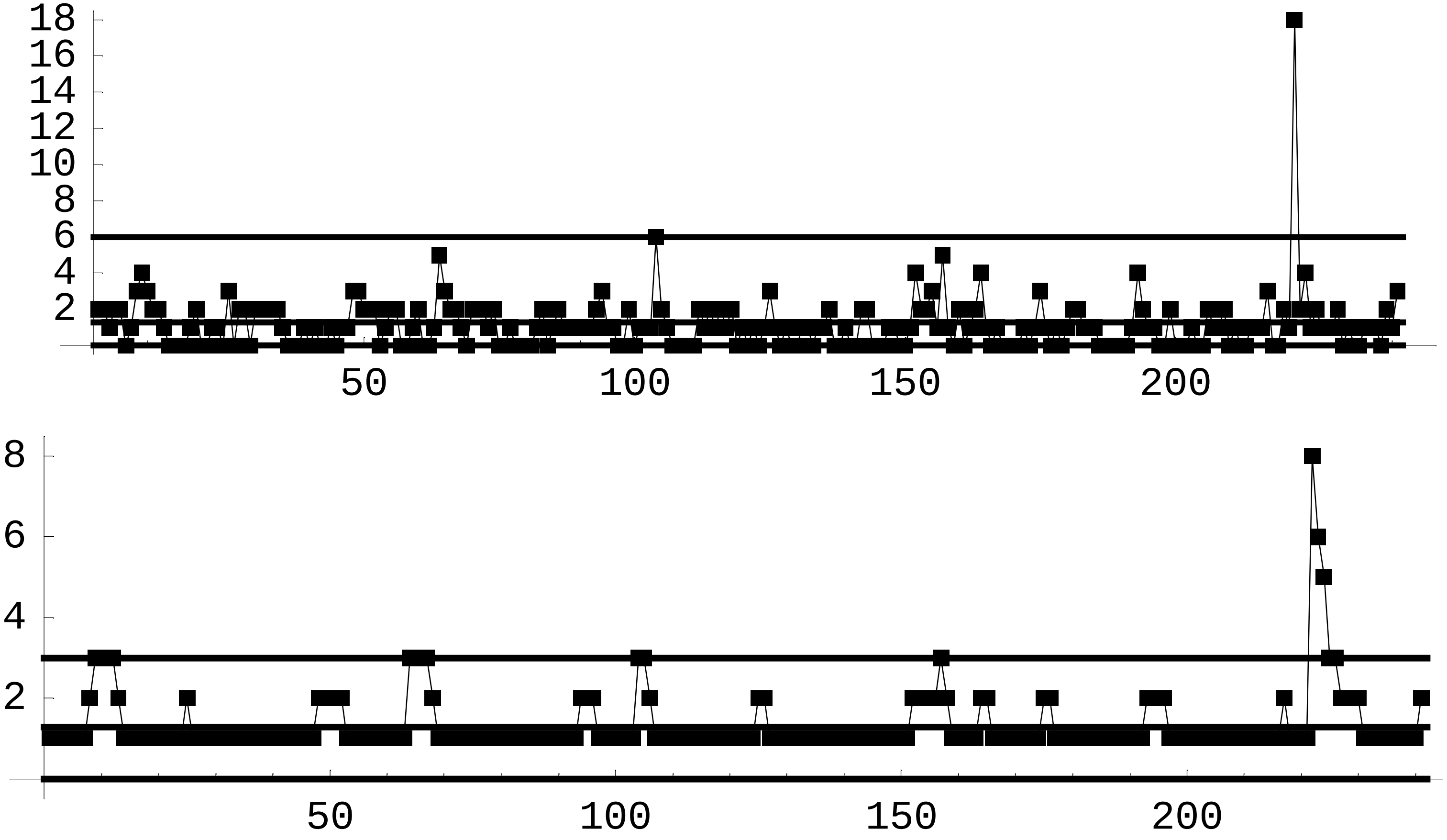
$ARL_0 = 386.991$.





Design: $\lambda = 0.4$, $l_c = l_e = 0$, $u_c = 6$, $u_e = 3$, $z_0 = 1$.

$ARL_0 = 381.05$.



Outlier at time $t = 222$ (period 17:22:00 to 17:23:59):

18 different IP addresses of the form 195.93.60.xxx.

$\Rightarrow$ Correct the data by setting $n_{222} := 1$.

Some of the chart still trigger an alarm ($n_{104} = 6$), perhaps false alarm.

Result: (Corrected) Data seems to follow the previously identified in-control model!

- INAR(1) model: simple, easily interpretable model, well-suited for real-world problems from SPC.

- Combined EWMA chart: Exact ARL computation with Markov chain approach.

- Concept of combining the two charts offers flexibility to design chart either in a universal manner (shift in any direction), or in a specific manner (certain type of shift).

Thank You
for Your Interest!

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics