# Controlling Correlated Processes of Poisson Counts.

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics

# Some introductory words . . .

This talk is based on the paper

**Weiß, C. H.**:

*Controlling Correlated Processes of Poisson Counts*.

August 23, 2006.

*Compare conference CD-ROM.*

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

# 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

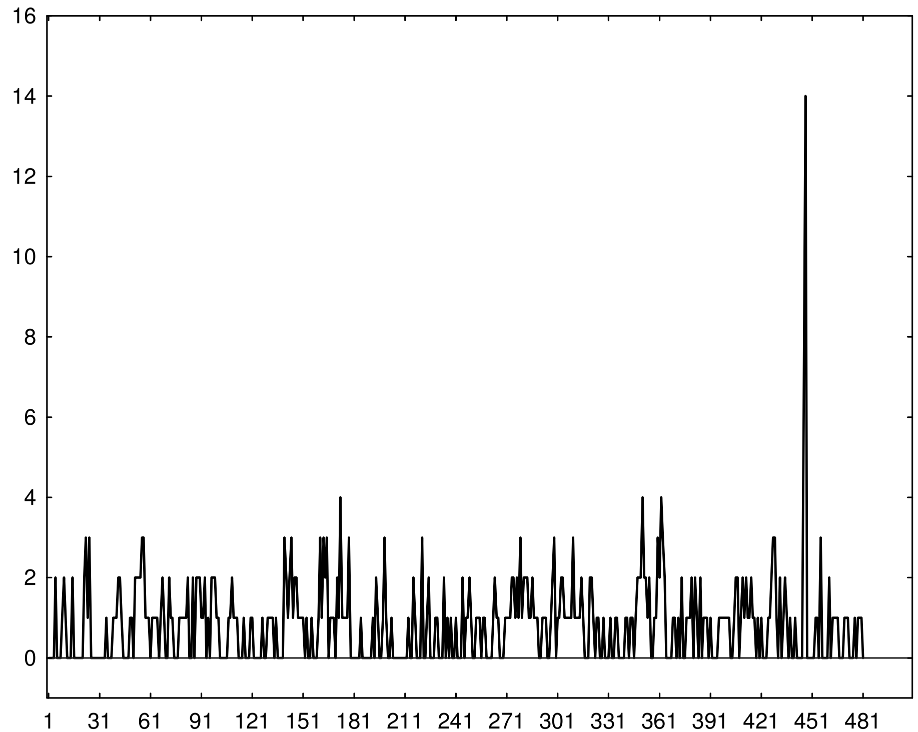
$\Rightarrow$ for each minute during the periods observed:

number of *different* IP addresses registered per minute, between 10 o'clock a.m. and 6 o'clock p.m.

$\Rightarrow$ time series of length 481 each.

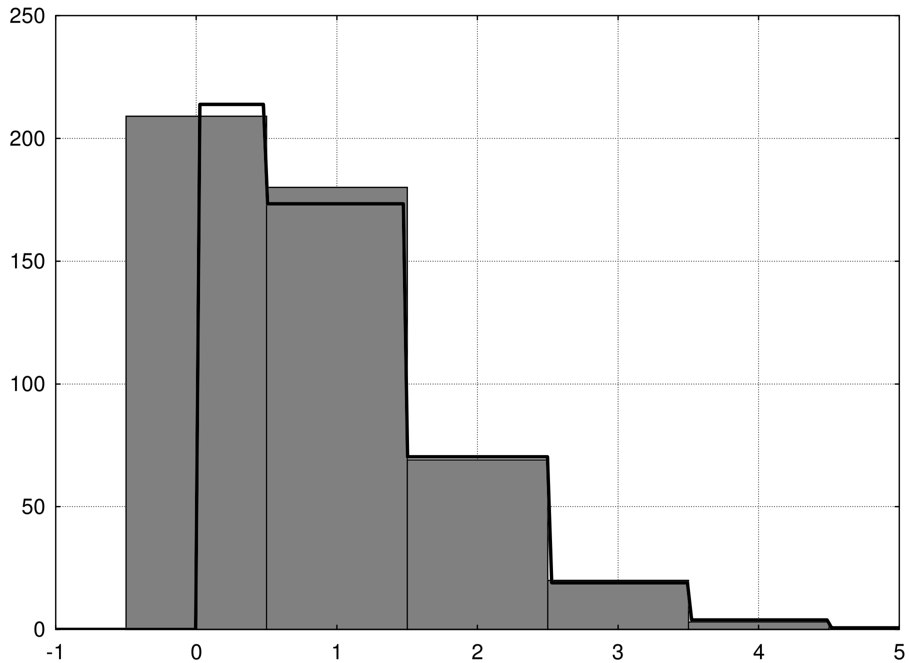
**Example:** (continued)

Data collected on November 29th, 2005:



**Example:** (continued)

Data collected on November 29th, 2005:



**Example:** (continued)

Data collected on November 29$^{\text{th}}$, 2005:

Outlier at time $t = 447$: 14 IP addresses of form 195.93.60.xxx.

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.

**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 no control schemes for autocorrelated Poisson data!

# Poisson
# INARMA Processes

## Definition & Properties

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

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

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

where $X_i$ are independent Bernoulli trials, $B(1, \alpha)$, also independent of $N$ → *counting series*.

We say: $\alpha \circ N$ arises from $N$ by *binomial thinning*
'$\circ$' is called *binomial thinning operator*.

**Interpretation** of $\alpha \circ N$:

- Population of size $N$ at a certain time $t$.
- Later at time $t+1$: population shrinked, because some individuals died.
- Assume that individuals die independently of each other with probability $1-\alpha$

$\Rightarrow$ *Number of survivors* is given by $\alpha \circ N$.

Probabilistic operation *binomial thinning* replaces scalar multiplication in definition of usual ARMA models.

Reason: $\alpha \cdot N \notin \mathbb{N}_0$ for any $\alpha \in (0;1)$.

Justification: $E[\alpha \circ N] \;=\; \alpha \cdot E[N]$, but $\alpha \circ N \in \mathbb{N}_0$.

$\Rightarrow$ **INARMA models** (**in**teger-valued **ARMA**)

**Definition of INAR(1) process:**

Let $(\epsilon_t)_{\mathbb{N}}$ be i.i.d. process with range $\mathbb{N}_0$, let $\alpha \in [0;1]$. A process $(N_t)_{\mathbb{N}_0}$, which follows the recursion

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

where the thinning operations at each time $t$ are performed independently of each other, and independently of $(\epsilon_t)_{\mathbb{Z}}$ and $(N_s)_{s<t}$, and where $\epsilon_t$ is independent of $(N_s)_{s<t}$, is called an *INAR(1) process*.

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

**Properties of stationary INAR(1) process:**

- expectation $E[N_t] \ = \ \dfrac{\mu_\epsilon}{1-\alpha}$,

- variance $V[N_t] \ = \ \dfrac{\alpha\mu_\epsilon + \sigma_\epsilon^2}{1-\alpha^2}$,

- autocorrelation $\rho(k) \ := \ Corr[N_t, N_{t-k}] \ = \ \alpha^k$,

- $(\epsilon_t)_{\mathbb{N}}$ i.i.d. $Po(\mu) \Rightarrow N_t \sim Po(\frac{\mu}{1-\alpha})$

**Interpretation & examples of INAR(1) process:**

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

- $N_t$: number of users accessing web server, $\epsilon_t$: number of new users, $\alpha \circ N_{t-1}$: number of previous users still active.
- $N_t$: number of faults in system or network, $\epsilon_t$: number of new faults, $\alpha \circ N_{t-1}$: number of previous faults not rectified yet.

**Examples of INAR(1) process:** (continued)

- $N_t$: number of unanswered complaints of customers, consisting of new and past complaints.

- $N_t$ products in circulation: products just been sold, and products sold in the past but still work.

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

The INAR(1) model applies well to typical tasks of statistical quality control!

$\Rightarrow$ From now on:

$(N_t)_{\mathbb{N}_0}$ 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 $\alpha = \alpha_0$.

# Controlling Poisson INAR(1) Processes

## Control Concepts

## $c$-Chart for Poisson INAR(1):

Realized values of $(N_t)_{\mathbb{N}}$ plotted on chart with

$$\begin{aligned} UCL &= \tfrac{\mu_0}{1-\alpha_0} + 3\sqrt{\tfrac{\mu_0}{1-\alpha_0}}, \\ \text{Center line} &= \tfrac{\mu_0}{1-\alpha_0}, \\ LCL &= \max\left\{0, \tfrac{\mu_0}{1-\alpha_0} - 3\sqrt{\tfrac{\mu_0}{1-\alpha_0}}\right\}. \end{aligned}$$

Standard AR(1) process: control of estimated residuals with Shewhart chart, compare Knoth & Schmid (2004).

Poisson INAR(1): $\widehat{\epsilon}_t \ := \ N_t \ - \ \alpha \cdot N_{t-1}$.

Properties: $E[\widehat{\epsilon}_t] = \mu$, $V[\widehat{\epsilon}_t] = (1+\alpha)\cdot\mu$,

$Corr[\widehat{\epsilon}_t, \widehat{\epsilon}_{t-k}] = 0, \quad k \geq 1.$

**Residual Control Chart:** Plot $(\widehat{\epsilon}_t)_{\mathbb{N}}$ with

$$UCL = \mu_0 + 3\sqrt{(1+\alpha_0)\cdot\mu_0},$$
$$\text{Center line} = \mu_0,$$
$$LCL = \mu_0 - 3\sqrt{(1+\alpha_0)\cdot\mu_0}.$$

Monitor conditional distribution of $N_t$ given $N_{t-1}$, with
$E[N_t|N_{t-1}] = \alpha \cdot N_{t-1} + \mu$, $V[N_t|N_{t-1}] = \alpha(1-\alpha) \cdot N_{t-1} + \mu$.
Conditional distribution changes for changing values of $N_{t-1}$
$\Rightarrow$ 'usual' control limits would change.

**Conditional Control Chart:** Plot statistic

$$T_t \; = \; \frac{N_t \; - \; \alpha_0 \cdot N_{t-1} \; - \; \mu_0}{3 \cdot \sqrt{\alpha_0(1-\alpha_0) \cdot N_{t-1} + \mu_0}}.$$

The process

$$N_t \text{ is assumed to be } \begin{cases} \text{in control} & \text{if } -1 \leq T_t \leq 1, \\ \text{out of control} & \text{if } T_t < -1 \text{ or } T_t > 1. \end{cases}$$

In fact, both residual chart and conditional chart monitor the conditional distribution of $N_t$, conditioned on $N_{t-1}$.

- $c$-**chart** concentrates on absolute value of $N_t$, hopefully sensitive to change in marginal distribution;

- **conditional control schemes** concentrate on observed transition, hopefully able to detect change in autocorrelation structure
(autocorrelation decreases $\Rightarrow$ larger jumps occur).

$4^{\text{th}}$ alternative: **Moving window** of length $w$.

Window sum $C_t^{(w)} := N_{t-w+1} + \ldots + N_t$, with

$$E[C_t^{(w)}] = \frac{w \cdot \mu}{1-\alpha}$$

and

$$V[C_t^{(w)}] = w \cdot \mu \cdot \frac{1+\alpha}{(1-\alpha)^2} \cdot \left(1 - \frac{2\alpha}{1-\alpha^2} \cdot \frac{1-\alpha^w}{w}\right).$$

**Controlling a Moving Average:** Window size $w$, step width $s$. Plot statistics $\ldots, T_t{}^{(w)}, T_{t+s}{}^{(w)}, \ldots$, defined by

$$T_t^{(w)} \;:=\; \frac{1}{w} \cdot C_t^{(w)},$$

with

$$UCL = \frac{\mu_0}{1-\alpha_0} + 3 \cdot \sqrt{\frac{\mu_0}{w} \cdot \frac{1+\alpha_0}{(1-\alpha_0)^2} \cdot \left(1 - \frac{2\alpha_0}{1-\alpha_0^2} \cdot \frac{1-\alpha_0^w}{w}\right)},$$

$$\text{Center line} = \frac{\mu_0}{1-\alpha_0},$$

$$LCL = \frac{\mu_0}{1-\alpha_0} - 3 \cdot \sqrt{\frac{\mu_0}{w} \cdot \frac{1+\alpha_0}{(1-\alpha_0)^2} \cdot \left(1 - \frac{2\alpha_0}{1-\alpha_0^2} \cdot \frac{1-\alpha_0^w}{w}\right)}.$$

# Controlling Poisson INAR(1) Processes

## Performance Study

**Situations considered:**

In control: $\mu = \mu_0$ and $\alpha = \alpha_0$

Out of control (→ marginal mean $\frac{\mu}{1-\alpha}$ affected)

- with $\mu = \mu_0$, but $\alpha = 1.2 \cdot \alpha_0$ or $\alpha = 0.8 \cdot \alpha_0$,

- with $\mu = 1.2 \cdot \mu_0$ or $\mu = 0.8 \cdot \mu_0$, but $\alpha = \alpha_0$.

Out of control (→ marginal mean $\frac{\mu}{1-\alpha}$ unaffected)

- with $\frac{\mu}{1-\alpha} = \frac{\mu_0}{1-\alpha_0}$, but $\alpha = 1.2 \cdot \alpha_0$ or $\alpha = 0.8 \cdot \alpha_0$.

**Situations considered:** (continued)

For each situation and combination of parameters, 10,000 INAR(1) processes were simulated with **mathematica 5** and controlled by one of the previous procedures. For each sequence, the time of the first alarm was measured ⇒ 10,000 $RL$ values ⇒ $ARL$.

All charts with 3-$\sigma$ limits, moving average charts with $w = 2, 5, 10$.

**Results:** ($\rightarrow$ detailed tables in article)

**In-Control State:** $\mu = \mu_0$ **and** $\alpha = \alpha_0$.

- $ARL$s of the $c$-chart vary heavily;

- $ARL$s of conditional charts not larger than 300;

- $ARL$s of moving average schemes increase for increasing values of $\mu_0$, $\alpha_0$ and/or $w$ $\Rightarrow$ quite robust against false alarms;

**Results:** (continued)

**Out-of-Control State:** $\mu = 1.2 \cdot \mu_0$ **and** $\alpha = \alpha_0$.

- $c$-chart gives good results for $\mu_0 \geq 3$, but moving average schemes clearly superior for $\mu_0 \geq 5$.
- Chart for $w = 10$ remarkable: out-of-control $ARL$s as low as for other schemes, but much better in-control $ARL$s.
- Conditional charts perform worst.

**Out-of-Control State:** $\mu = 0.8 \cdot \mu_0$ **and** $\alpha = \alpha_0$.

If $\mu_0 < 5$, none of the charts can be used. Moving average chart ($w = 10$) performs well for $\mu_0 \geq 5$.

**Results:** (continued)

**Out-of-Control State:** $\mu = \mu_0$ **and** $\alpha = 1.2 \cdot \alpha_0$.

All charts similar: If $\alpha_0$ is already large, then a further upward shift by 20 % is detected quickly.

(Note: $\frac{\mu}{1-\alpha}$ is increased)

**Out-of-Control State:** $\mu = \mu_0$ **and** $\alpha = 0.8 \cdot \alpha_0$.

None of the charts is able to detect such a shift, not even the conditional charts.

**Results:** (continued)

**Out-of-Control State:** $\frac{\mu}{1-\alpha} = \frac{\mu_0}{1-\alpha_0}$ **and** $\alpha = 1.2 \cdot \alpha_0$**.**
Conditional charts perform miserably ($\rightarrow$ lower jumps). Also performance of other charts inacceptable ($\rightarrow$ process mean unaffected).

**Out-of-Control State:** $\frac{\mu}{1-\alpha} = \frac{\mu_0}{1-\alpha_0}$ **and** $\alpha = 0.8 \cdot \alpha_0$**.**
$c$-chart and moving average charts cannot be used, their $ARL$s are even increased. Only conditional control charts can be applied, at least if $\alpha_0$ is large.

**Summary:**

| $\mu$ | $\alpha$ | $\frac{\mu}{1-\alpha}$ | Best chart |
|---|---|---|---|
| 1 | 1 | 1 | Moving average $w = 10$ |
| 1.2 | 1 | | Moving average $w = 10$ |
| 0.8 | 1 | | Moving average $w = 10$ if $\mu_0 \geq 5$ |
| 1 | 1.2 | | Moving average $w = 10$ |
| 1 | 0.8 | | None |
| | 1.2 | 1 | None |
| | 0.8 | 1 | Conditional charts |

# Thank You for Your Interest!

Christian H. Weiß
University of Würzburg
Institute of Mathematics
Department of Statistics