

The Mathematics and Statistics of Infectious Disease Outbreaks

Michael Höhle¹

¹Department of Mathematics
Stockholm University, Sweden

L10: Univariate outbreak detection¹

¹LaMo: 2022-05-02 @ 20:57:25

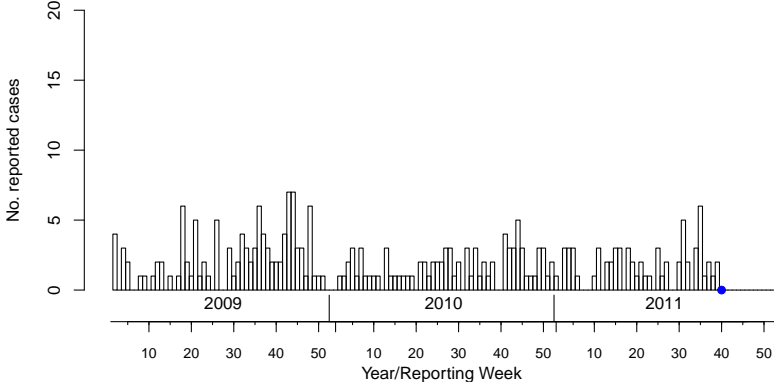
Overview

- 1 Monitoring of univariate count data time series
 - Statistical Framework for Aberration Detection
 - Simple Algorithm for Ad-Hoc Detection
 - Farrington algorithm and beyond
- 2 A System for Automated Outbreak Detection in Germany
- 3 Discussion

Outline

- 1 Monitoring of univariate count data time series
 - Statistical Framework for Aberration Detection
 - Simple Algorithm for Ad-Hoc Detection
 - Farrington algorithm and beyond
- 2 A System for Automated Outbreak Detection in Germany
- 3 Discussion

German Infection Protection Act (IfSG) data from the Robert Koch Institute (up to W40-2011):

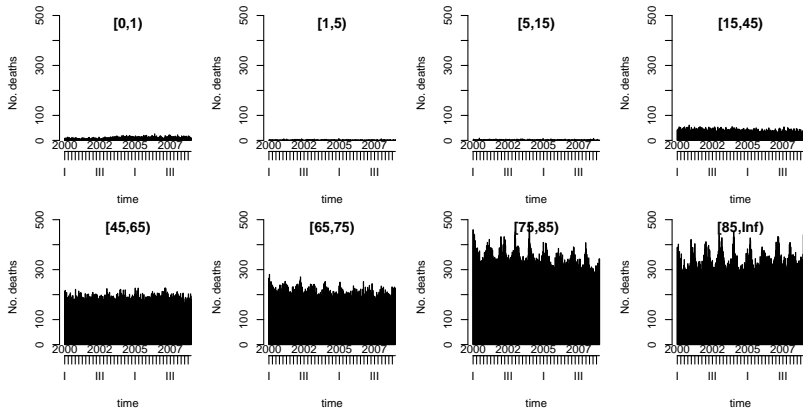


Example – The EuroMOMO project (1)

- European monitoring of excess mortality for public health action (EuroMOMO)
- Aim: develop and strengthen real-time monitoring of mortality across Europe in order to enhance the management of serious public health risks such as pandemic influenza, heat waves and cold snaps
- Main outcome of mortality monitoring: excess mortality
- In this course: Focus on monitoring aspect

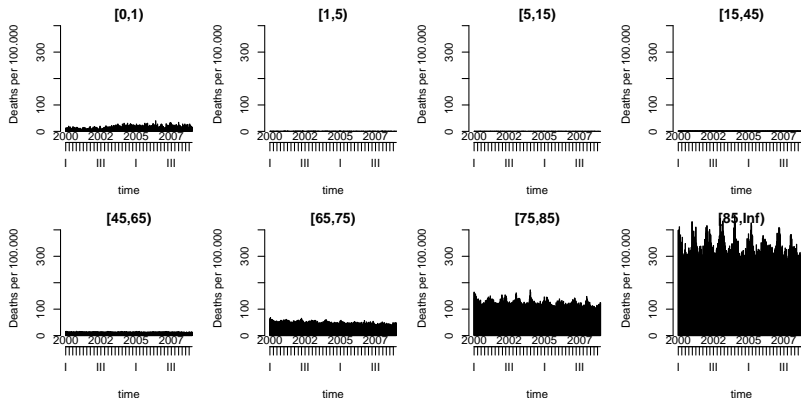
Example – The EuroMOMO project (2)

Weekly danish mortalities 2000-2008 in 8 age-groups as provided by Statens Serum Institute (Höhle and Mazick, 2010).



Example – The EuroMOMO project (2)

Weekly danish mortalities 2000-2008 in 8 age-groups as provided by Statens Serum Institute (Höhle and Mazick, 2010).



Outline

- 1 Monitoring of univariate count data time series
 - Statistical Framework for Aberration Detection
 - Simple Algorithm for Ad-Hoc Detection
 - Farrington algorithm and beyond

Statistical Framework for Aberration Detection (1)

- Univariate time series $\{y_t, t = 1, 2, \dots\}$ to monitor
- For each time t we differentiate between two underlying states: in-control (everything is fine) or out-of-control (something is wrong).
- At time $s \geq 1$, the available information is $\mathbf{y}_s = \{y_t; t \leq s\}$.
- Based on \mathbf{y}_s an automatic detection procedure has to decide if there is unusual activity at time s (or not).

Statistical Framework for Aberration Detection (2)

- The detectors are initially only based on the one-step-ahead predictive distribution at each time point (Shewhart-like control chart):
 - Let $G(y_s | y_1, \dots, y_{s-1}; \theta)$ be the distribution of Y_s in case everything is in-control.
 - If the actual observed value $Y_s = y_s$ is extreme in G , this is evidence against things being in-control.
 - The alarm threshold $a_{1-\alpha, s}$ at each time point is calculated as the $(1 - \alpha)$ 'th quantile of the predictive distribution. If $y_s > a_{1-\alpha, s}$ then we have an alarm.
- This can be generalized to more sequential control charts accumulating information, e.g. cumulative sum (CUSUM) methods.

Intermezzo: Estimation, prediction and uncertainty

- Data \mathbf{y} are the observed value of a random variable \mathbf{Y} characterized by a parametric model with density $f(\mathbf{y}; \boldsymbol{\theta})$.
- Aim: predict the value of a random variable \mathbf{Z} , which, conditionally on $\mathbf{Y} = \mathbf{y}$ has distribution function $G(\mathbf{z}|\mathbf{y}; \boldsymbol{\theta})$, *depending on $\boldsymbol{\theta}$* .
- Simplest form of the prediction problem:

$$Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} f(y; \boldsymbol{\theta}),$$

and the task is to predict $Z = Y_{n+1}$.

- In *time series 1-step-ahead prediction* the observations are correlated and the aim is to predict $\mathbf{Z} = Y_{n+1}$.

Outline

- 1 Monitoring of univariate count data time series
 - Statistical Framework for Aberration Detection
 - Simple Algorithm for Ad-Hoc Detection
 - Farrington algorithm and beyond

Example: Predicting a new $N(\mu, \sigma^2)$ observation (1)

- Let $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ with unknown μ and σ^2 . Then

$$\frac{Y_{n+1} - \bar{Y}}{s \sqrt{1 + \frac{1}{n}}} \sim t(n-1),$$

where $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ and $s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ are the sample mean and sample variance of \mathbf{Y} , respectively.

- A $(1 - 2\alpha) \cdot 100\%$ two-sided **prediction interval** (PI) is thus given by

$$\bar{Y} \pm t_{1-\alpha}(n-1) \cdot s \cdot \sqrt{1 + \frac{1}{n}}.$$

Example: Predicting a new $N(\mu, \sigma^2)$ observation (2)

- A *plug-in* $(1 - 2\alpha) \cdot 100\%$ two-sided **prediction interval** for Y_{n+1} is:

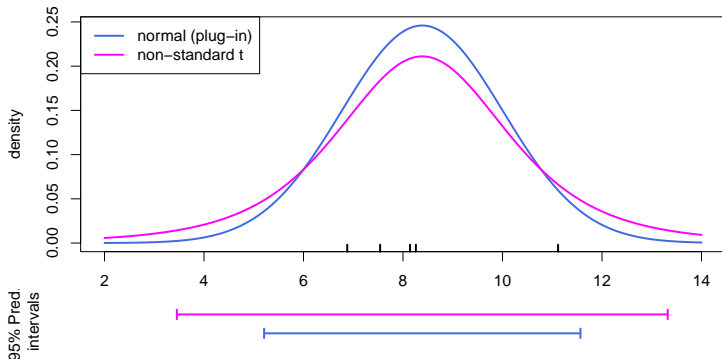
$$\bar{Y} \pm z_{1-\alpha} \cdot s.$$

- Both of these are not to be confused with a $(1 - 2\alpha) \cdot 100\%$ two-sided **confidence interval** for μ :

$$\bar{Y} \pm z_{1-\alpha} \cdot \frac{s}{\sqrt{n}}.$$

Example: Predicting a new $N(\mu, \sigma^2)$ observation (3)

- Illustration: PIs based on $n = 5$ observations from $N(\mu, \sigma^2)$.



- For $n = 5$ the 95% plug-in PI corresponds to a 85% PI. The 95% CI for μ is 7.2–9.6, which only corresponds to a 46% PI.

Summary: Ad-Hoc Outbreak Detection Algorithm

- Predict value y_s at time $s = (s^w, s^y)$ using a set of reference values from window of size $2w + 1$ up to b years back.
- Let $n = b(2w + 1)$ and compute threshold as the upper 97.5% quantile of the predictive distribution for y_s , i.e.

$$a_{0.975,s} = \bar{y} + t_{0.975}(n-1) \cdot s \cdot \sqrt{1 + \frac{1}{n}}.$$

- Sound alarm, if $y_s > a_{0.975,s}$.

Outline

- 1 Monitoring of univariate count data time series
 - Statistical Framework for Aberration Detection
 - Simple Algorithm for Ad-Hoc Detection
 - Farrington algorithm and beyond

Challenges of surveillance data

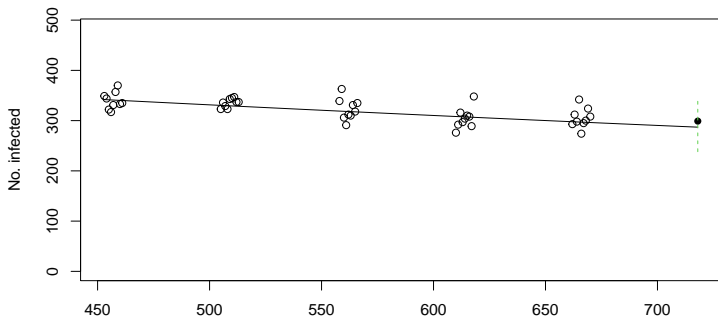
Issues making the statistical modelling and monitoring of surveillance time series a challenge:

- Lack of clear case definitions
- Under-reporting and reporting delays
- Often no denominator data
- Seasonality
- Low number of reported cases
- Presence of past outbreaks
- Existence of concurrent “explanatory” processes

Farrington algorithm (1) – basic model

- Predict value y_s at time $s = (s^w, s^y)$ using a set of reference values from window of size $2w + 1$ up to b years back.

Prediction at time $t=718$ with $b=5, w=4$



- Fit overdispersed Poisson generalized linear model (GLM) to the $b(2w + 1)$ reference values where $E(y_t) = \mu_t$, $\text{Var}(y_t) = \phi \cdot \mu_t$ with $\log \mu_t = \alpha + \beta t$ and $\phi > 0$.

Farrington algorithm (2) – outbreak detection

Predict and compare:

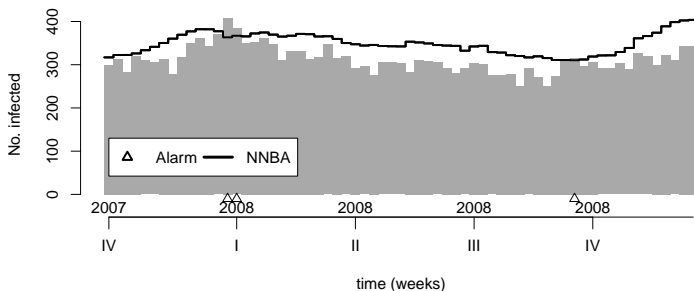
- An approximate $(1 - \alpha)$ one-sided prediction interval for y_s based on the GLM has upper limit
$$a_{1-\alpha,s} = \hat{\mu}_s + z_{1-\alpha} \cdot \sqrt{\text{Var}(y_s - \hat{\mu}_s)}$$
- If the observed y_s is greater than $a_{1-\alpha,s}$, then flag s as outbreak

Refinements of the algorithm include:

- Computation of the prediction interval on a transformed scale
- Use a re-weighted fit with weights based on Anscombe residuals in order to correct for outliers
- Low count protection

Application: Danish mortality data (age group 75-84 years)

- Results of the old and improved Farrington algorithm, respectively, with $w = 4$, $b = 5$ and $\alpha = 0.005$ starting at W40-2007:



Outline

- 1 Monitoring of univariate count data time series
- 2 A System for Automated Outbreak Detection in Germany
- 3 Discussion

System Design

- Salmon, Schumacher, and Höhle (2016) describes a system integrating outbreak detection algorithms into the epidemiological workflow

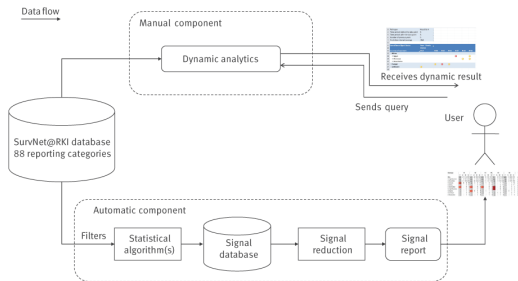


Figure source: Salmon, Schumacher, and Höhle (2016)

- Example of using machine learning methods for the more than 30,000 time series

Application on Salmonella Montevideo 2009-2010

Results from the extended Farrington procedure using last five years as reference values:

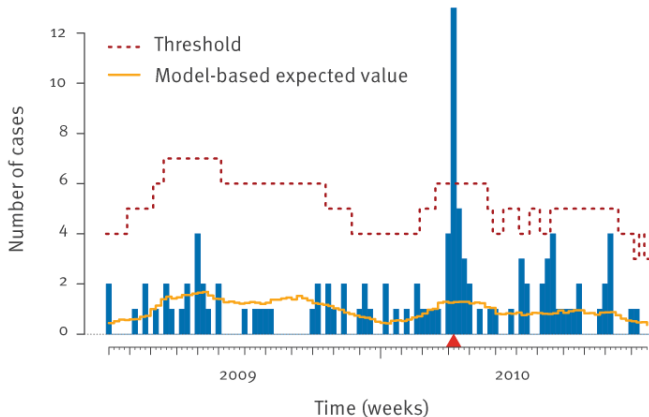


Figure source: Salmon, Schumacher, and Höhle (2016)

Salmonella Report for W41–46 of 2013

Weekly Report at National Level:

| Serotype | Week 41 | | | | Week 42 | | | | Week 43 | | | | Week 44 | | | | Week 45 | | | | Week 46 | | | |
|----------------------------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|
| | y_t | o_t | μ_t | U_t | y_t | o_t | μ_t | U_t | y_t | o_t | μ_t | U_t | y_t | o_t | μ_t | U_t | y_t | o_t | μ_t | U_t | y_t | o_t | μ_t | U_t |
| Salmonella, all serotypes | 466 | 27 | 512 | 691 | 373 | 23 | 485 | 650 | 370 | 16 | 461 | 620 | 356 | 15 | 439 | 601 | 411 | 8 | 417 | 580 | 290 | 14 | 390 | 540 |
| S. Typhimurium | 107 | 2 | 151 | 221 | 103 | 1 | 145 | 214 | 108 | 2 | 140 | 208 | 106 | 5 | 134 | 202 | 142 | 4 | 127 | 191 | 90 | 4 | 120 | 181 |
| S. Enteritidis | 158 | 11 | 154 | 230 | 123 | 12 | 142 | 212 | 115 | 11 | 131 | 194 | 84 | 4 | 124 | 189 | 80 | 1 | 116 | 182 | 62 | 2 | 107 | 168 |
| S. Infantis | 25 | 6 | 9 | 18 | 16 | 3 | 8 | 17 | 8 | 1 | 8 | 18 | 10 | - | 8 | 17 | 2 | - | 7 | 17 | 5 | - | 7 | 16 |
| S. Derby | 4 | NA | 5 | 11 | 2 | NA | 5 | 11 | 7 | NA | 5 | 11 | 3 | NA | 5 | 11 | 4 | NA | 5 | 11 | 1 | - | 5 | 11 |
| S. Manhattan | 7 | NA | 0 | 2 | 4 | NA | 0 | 2 | 4 | NA | 0 | 2 | 3 | NA | 0 | 2 | 3 | NA | 0 | 2 | NA | NA | 0 | 2 |
| S. Typhimurium, monophasic | 2 | NA | 0 | 2 | 2 | NA | 0 | 2 | 2 | NA | 0 | 2 | 6 | NA | 0 | 2 | 5 | NA | 0 | 3 | 3 | NA | 0 | 3 |
| S. Agona | 2 | NA | 1 | 4 | 7 | 4 | 1 | 4 | 2 | 1 | 1 | 4 | 3 | 2 | 1 | 4 | 1 | NA | 1 | 4 | 3 | 2 | 1 | 4 |
| S. Virchow | 4 | NA | 3 | 8 | 1 | NA | 3 | 8 | 3 | NA | 3 | 7 | 1 | NA | 3 | 7 | 5 | 1 | 3 | 7 | 1 | NA | 3 | 7 |
| S. Muenchen | 3 | NA | 1 | 4 | 3 | NA | 1 | 4 | NA | NA | 1 | 4 | 3 | NA | 1 | 4 | 2 | NA | 1 | 4 | NA | NA | 1 | 4 |

Table source: Salmon, Schumacher, and Höhle (2016)

Outline

- 1 Monitoring of univariate count data time series
- 2 A System for Automated Outbreak Detection in Germany
- 3 Discussion

Discussion

- The presented methods are implemented in the R package *surveillance* (Salmon, Schumacher, and Höhle, 2016)
- Developing, maintaining and improving automatic outbreak detection systems is an interdisciplinary activity!
 - Even more work could be put into user adaptation.
 - Delay adjusted monitoring (Salmon, Schumacher, Stark, et al., 2015)
- The system proved to be a good insurance against missing anything important – see e.g. Gertler et al. (2015)

Literature I



Gertler, Maximilian et al. (2015). “Outbreak of cryptosporidium hominis following river flooding in the city of Halle (Saale), Germany, August 2013”. In: *BMC Infectious Diseases* 15.1, p. 88. ISSN: 1471-2334. DOI: 10.1186/s12879-015-0807-1. URL: <http://www.biomedcentral.com/1471-2334/15/88>.



Höhle, M. and A. Mazick (2010). “Aberration detection in R illustrated by Danish mortality monitoring”. In: *Biosurveillance: A Health Protection Priority*. Ed. by T. Kass-Hout and X. Zhang. CRC Press, pp. 215–238. URL: https://staff.math.su.se/hoehle/pubs/hoehle_mazick2009-preprint.pdf.



RKI (2012). “Salmonella Newport-Ausbruch in Deutschland und den Niederlanden, 2011”. In: *Epidemiologisches Bulletin* 20. Available as http://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2012/Ausgaben/20_12.pdf, pp. 177–184.

Literature II



Salmon, M., D. Schumacher, and M. Höhle (2016). “Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance”. In: *Journal of Statistical Software* 70.10. Also available as vignette of the R package *surveillance*. DOI: [10.18637/jss.v070.i10](https://doi.org/10.18637/jss.v070.i10).



Salmon, M., D. Schumacher, K. Stark, and M. Höhle (2015). “Bayesian outbreak detection in the presence of reporting delays”. In: *Biometrical Journal* 57.6. <http://dx.doi.org/10.1002/bimj.201400159>, pp. 1051–1067.