# An R-package for the surveillance of infectious diseases

Michael Höhle

Department of Statistics
University of Munich

Compstat 2006
Rome, 28 August 2006

# Overview

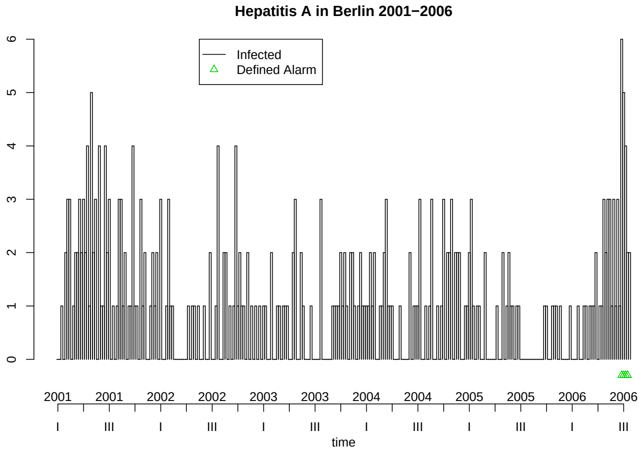
## Motivation

Software for the *use* and *development* of surveillance algorithms

## Features

- Visualisation of surveillance data and algorithm output
- Outbreak data from SurvStat@RKI and through simulation from a hidden Markov model
- Implementation of well-known surveillance algorithms
- Functionality to compare classification performance
- First steps towards multivariate surveillance

# Example of surveillance data




```
> data(ha)
> plot(aggregate(ha), main = "Hepatitis A in Berlin 2001-2006")
```

# Implemented Algorithms

- `cdc` – Centers for Disease Control and Prevention (Stroup et al., 1989)
- `rki` – Algorithm used by the Robert Koch Institute (RKI), Germany (Altmann, 2003)
- `bayes` – Simple Bayesian Approach (Höhle, 2006)
- `farrington` – Communicable Disease Surveillance Centre (Farrington et al., 1996)
- `cusum` – Cumulative Sum (CUSUM) for Poisson counts (Rossi et al., 1999)

# Surveillance Algorithms: Simple Bayes (1)

## Reference values for the current week (year):(week) $= 0 : t$

For half window-width $w$ ($w_0$ in year 0) and $b$ years back in time

$$R_{\text{Bayes}}(w, w_0, b) = \left( \bigcup_{i=1}^{b} \bigcup_{j=-w}^{w} y_{-i:t+j} \right) \cup \left( \bigcup_{k=-w_0}^{-1} y_{0:t+k} \right)$$

## Predictive posterior distribution

If $Y_1, \ldots, Y_n | \lambda \overset{\text{iid}}{\sim} \text{Po}(\lambda)$ and Jeffrey's priori $\lambda \sim \text{Ga}(\frac{1}{2}, 0)$:

$$Y_{0:t} | R_{\text{Bayes}} \sim \text{NegBin}\left( \frac{1}{2} + \sum_{y_{i:j} \in R_{\text{Bayes}}} y_{i:j}, \quad \frac{|R_{\text{Bayes}}|}{|R_{\text{Bayes}}| + 1} \right)$$

# Surveillance Algorithms: Simple Bayes (2)

## Threshold

Given quantile-parameter $\alpha$ compute smallest value $y_\alpha$, such that:

$$P(Y_{0:t} \leq y_\alpha | R_{\text{Bayes}}) \geq 1 - \alpha$$

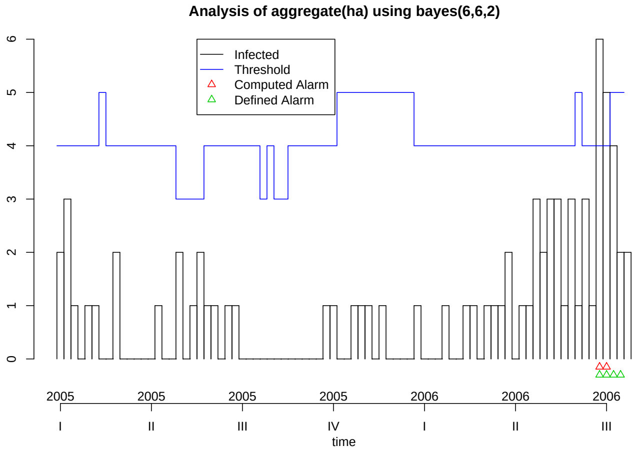
## Alarm

$$y_{0:t} \geq y_\alpha$$

## Problems

- Reference values belonging to an outbreak
- Over-dispersion

# Detection of Hepatitis A with Bayes(6,6,2)




```
> ctrl <- list(range = 209:290, b = 2, w = 6, alpha = 0.005)
> ha.b62 <- algo.bayes(aggregate(ha), control = ctrl)
```

# Classification Performance of Bayes(6,6,2) on `ha`

- Computation of sensitivity and specificity
- Euclidean distance between the points $(Se, Sp)$ and $(1, 1)$
- Expected delay before outbreak detection

| | TP | FP | TN | FN | sens | spec | dist | mlag |
|---|---|---|---|---|---|---|---|---|
| 1 | 2.00 | 0.00 | 78.00 | 2.00 | 0.50 | 1.00 | 0.50 | 0.00 |

```
> algo.quality(ha.b62)
```

# Comparison of Algorithms (1)

**14 selected time series**

measles, Q fever, salmonella, cryptosporidosis, Norwalk virus, hepatitis A

**Details**

- Each time series contains one outbreak as defined by the "Epidemiologisches Bulletin" published by the RKI.
- Data are collected from the SurvStat@RKI database

  `http://www3.rki.de/SurvStat`
- Each surveillance algorithm is applied to all 14 time series

# Comparison of Algorithms (2)

| | TP | FP | TN | FN | sens | spec | dist | mlag |
|---|---|---|---|---|---|---|---|---|
| rki(6,6,0) | 38 | 62 | 2646 | 180 | 0.17 | 0.98 | 0.83 | 5.43 |
| rki(6,6,1) | 65 | 83 | 2625 | 153 | 0.30 | 0.97 | 0.70 | 5.57 |
| rki(4,0,2) | 80 | 106 | 2602 | 138 | 0.37 | 0.96 | 0.63 | 5.43 |
| bayes(6,6,0) | 61 | 206 | 2502 | 157 | 0.28 | 0.92 | 0.72 | 1.71 |
| bayes(6,6,1) | 123 | 968 | 1740 | 95 | 0.56 | 0.64 | 0.56 | 1.36 |
| bayes(4,0,2) | 162 | 920 | 1788 | 56 | 0.74 | 0.66 | 0.43 | 1.36 |
| cdc(4*,0,5) | 65 | 94 | 2614 | 153 | 0.30 | 0.97 | 0.70 | 7.14 |
| farrington(3,0,5) | 37 | 53 | 2655 | 181 | 0.17 | 0.98 | 0.83 | 5.64 |

```
> all2one <- function(outbrk) {
+     survResList <- algo.call(outbrk, control = ctrl)
+     t(sapply(survResList, algo.quality))
+ }
> algo.summary(lapply(outbrks, all2one))
```

# CUSUM as Surveillance Algorithm (1)

- A control chart known from statistical process control

## Cumulative Sum (CUSUM)

In control situation $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0,1)$. Monitor shift to $N(\mu,1)$ by

$$S_t = \max(0, S_{t-1} + X_t - k), \quad t = 1, \ldots, n$$

where $S_0 = 0$ and $k$ is the *reference value*. Raise alarm if $S_t > h$, where $h$ is the *decision interval*.

- CUSUMs are better to detect sustained shifts
- Given $h$ and $k$ we can determine the *average run length* (ARL)

# CUSUM as Surveillance Algorithm (2)

- CUSUM for count data $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Po}(m)$ by transforming data to normality (Rossi et al., 1999)

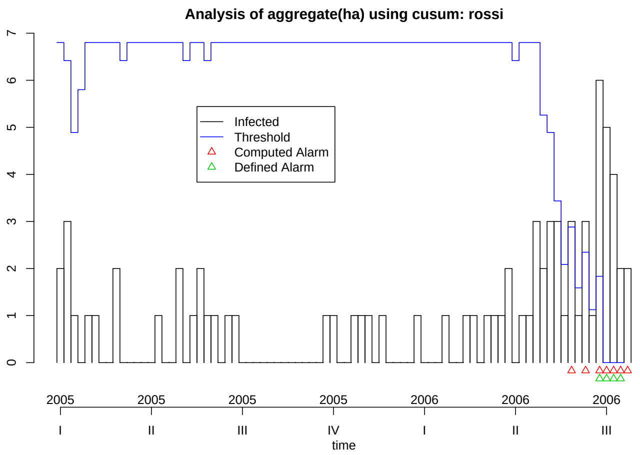
$$X_t = \frac{Y_t - 3m + 2\sqrt{m \cdot Y_t}}{2\sqrt{m}}$$

- Risk-adjust the chart by letting $m$ be time varying, e.g. as output of a GLM model

$$\log(m_t) = \alpha + \beta t + \sum_{s=1}^{S} (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)),$$

where $\omega_s = \frac{2\pi}{52} s$ are the Fourier frequencies.

# CUSUM as Surveillance Algorithm (3)




```
> kh <- find.kh(ARLa = 500, ARLr = 7)
> ha.cs <- algo.cusum(aggregate(ha), control = list(k = kh$k,
+     h = kh$h, trans = "rossi", range = 209:290))
```

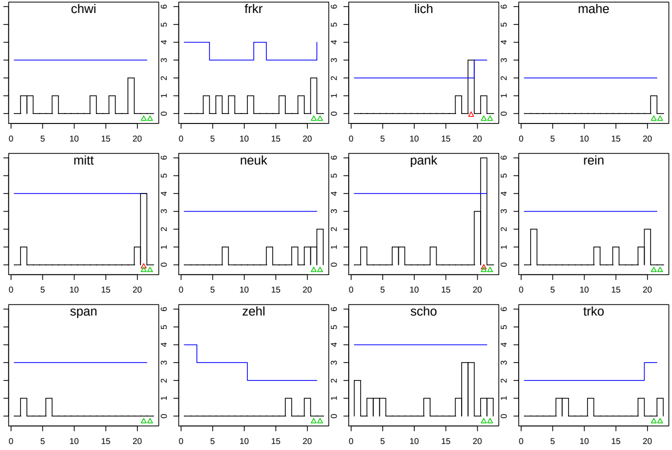
# Current developments (1)

- S4 class `sts` for surveillance data as *multivariate* time series of counts

```
> ha <- new("sts", ha, map = readShapePoly("berlin.shp",
+       IDvar = "SNAME"))
```

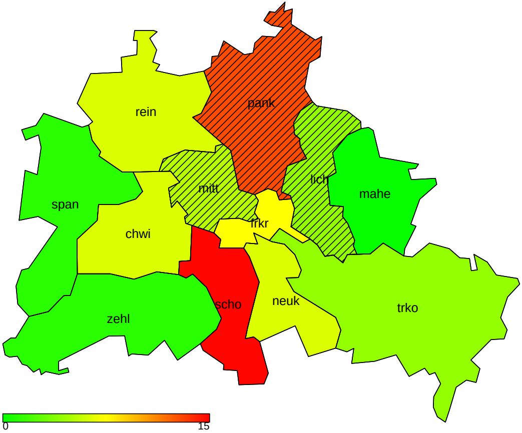
- Visualization of `sts` objects
- Surveillance for multivariate time series
  - Multivariate extensions of the univariate procedures
  - Multivariate GLM as in (Held et al., 2005) with CUSUM on the residuals
  - Adjusted CUSUM procedure as in (Rogerson and Yamada, 2004)

# Current developments (2) – Multivariate Bayes




```
> ha4 <- aggregate(ha, nfreq = 13)
> ha4.b62 <- algo.bayes(ha4, control = list(range = 52:73,
+       b = 2, w = 6, alpha = 0.001))
> plot(ha4.b62, type = observed ~ time | unit)
```

# Current developments (2) – GIS-Shapefiles




```
> plot(ha4.b62, type = observed ~ 1 | unit, axes = FALSE)
```

# Summing Up

- The volume of surveillance data requires automatic detection algorithms $\rightarrow$ data-mining
- `surveillance` offers an implementation for epidemiologist and a framework for developers
- The package is available from CRAN (current version is 0.9-1)
- Combining database, R, Sweave and LaTeX allows for easy generation of reports
- Multivariate surveillance is an active research area

# Literature I


Altmann, D. (2003). The Surveillance System of the Robert Koch Institute, Germany. Personal Communication.

Farrington, C., N. Andrews, A. Beale, and M. Catchpole (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society, Series A 159*, 547–563.

Held, L., M. Höhle, and M. Hofmann (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling 5*, 187–199.

Höhle, M. (2006). An R-package for the surveillance of infectious diseases. In *Proceedings of the CompStat 2006 conference, Rome, 28 Aug–1 Sep 2006*. To appear.

Rogerson, P. and I. Yamada (2004). Approaches to syndromic surveillance when data consists of small regional counts. *Morbidity and Mortality Weekly Report* (53), 79–85.

# Literature II


Rossi, G., L. Lampugnani, and M. Marchi (1999). An approximate CUSUM procedure for surveillance of health events. *Statistics in Medicine 18*, 2111–2122.

Stroup, D., G. Williamson, J. Herndon, and J. Karon (1989). Detection of aberrations in the occurence of notifiable diseases surveillance data. *Statistics in Medicine 8*, 323–329.