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

Hepatitis A in Berlin 2001–2006

> 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)
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 $\lambda \sim \text{Ga}(\frac{1}{2}, 0)$:

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

Analysis of aggregate(ha) using 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

<table>
<thead>
<tr>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>sens</th>
<th>spec</th>
<th>dist</th>
<th>mlag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.00</td>
<td>0.00</td>
<td>78.00</td>
<td>2.00</td>
<td>0.50</td>
<td>1.00</td>
<td>0.50</td>
</tr>
</tbody>
</table>

> algo.quality(ha.b62)
### 14 selected time series

- measles
- Q fever
- salmonella
- cryptosporidiosis
- 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](http://www3.rki.de/SurvStat)
- Each surveillance algorithm is applied to all 14 time series
### Comparison of Algorithms (2)

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

---

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

> algo.summary(lapply(outbrks, all2one))
```
A control chart known from statistical process control

Cumulative Sum (CUSUM)

In control situation $X_1, \ldots, X_n \sim N(0, 1)$. 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 for count data $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Po}(m)$ 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)

Analysis of aggregate(ha) using cusum: rossi

> 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))
S4 class \texttt{sts} for surveillance data as \textit{multivariate} time series of counts

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

- Visualization of \texttt{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)
The volume of surveillance data requires automatic detection algorithms → 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


