The R-Package “surveillance”

Michael Höhle

Department of Statistics, University of Munich, Germany
email: `hoehle@stat.uni-muenchen.de`

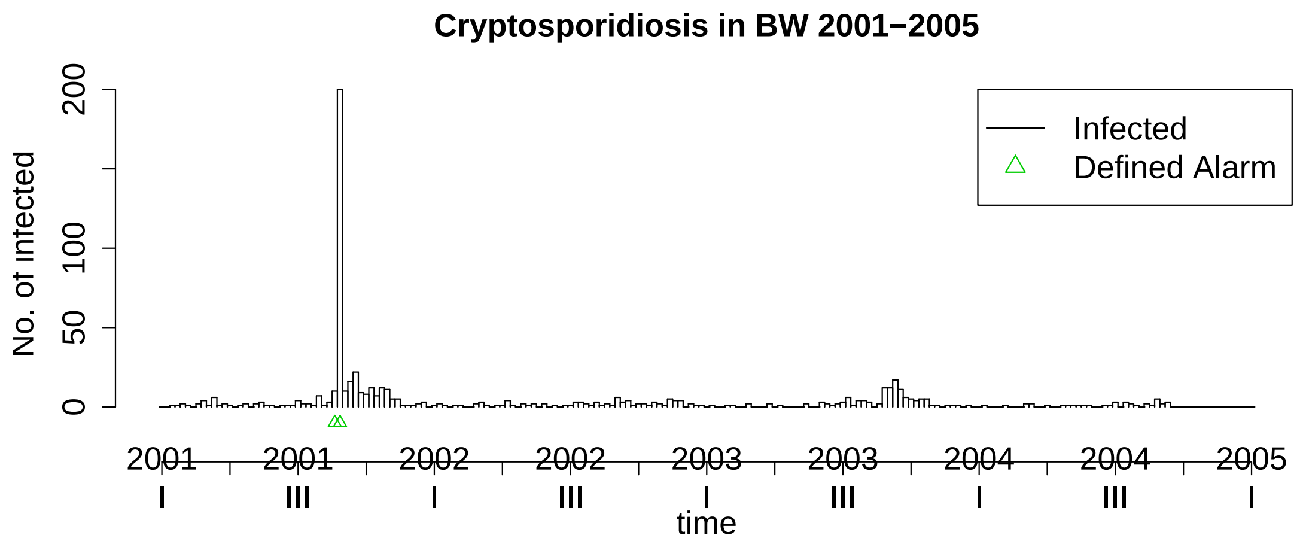
Abstract:

This poster introduces the R-package `surveillance` which provides a framework for the development of outbreak detection algorithms for surveillance data. The open-source package, available for download from the Internet, allows users to test new algorithms and compare their results with those of standard surveillance methods. Among others the package contains an implementation of the procedures described by Stroup et al. (1989), Farrington et al. (1996), and the system used at the Robert Koch Institute (RKI), Germany.
For evaluation purposes the package contains 14 example datasets drawn from the SurvStat@RKI Database maintained by the RKI, Germany. More comprehensive comparisons using simulation studies are possible by methods for simulating point source outbreak data using a hidden Markov model. To compare the algorithms, benchmark numbers like sensitivity, specificity and detection delay can be computed for entire sets of surveillance time series.

1 Surveillance Data

Surveillance data are typically collected on a weekly basis, we shall use the notation $\{y_{i:j}\}$ for the number of cases in week j of year i, where the years are indexed such that $i = 0$ is the most current year.

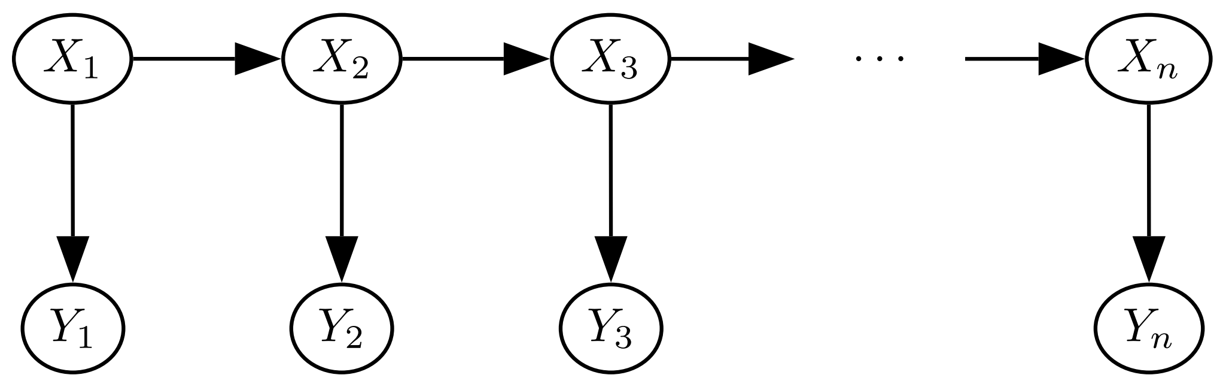
As an example the object `k1` describes cryptosporidiosis surveillance data for the German federal state Baden-Württemberg during 2001-2005. The peak in 2001 is due to an outbreak of cryptosporidiosis among a group of army-soldiers in boot-camp.



```
> k1 <- readData("k1")
> plot(k1, main = "Cryptosporidiosis in BW 2001-2005")
```

Time series from outbreaks with measles, Q-fever, salmonella, cryptosporidiosis, norovirus and hepatitis A are contained in the `data` directory. The data originate from the SurvStat@RKI database maintained by the Robert Koch Institute, Germany (Robert Koch-Institut, 2004).

The package also contains functionality to generate surveillance data containing point-source like outbreaks based on a Hidden Markov Model (HMM). A binary state $X_t, t = 1, \ldots, n$, denotes whether there was an outbreak and Y_t is the number of observed counts.



The observation Y_t is Poisson-distributed with log-link mean depending on a seasonal effect, a time trend and the outbreak indicator, i.e.

$$\log \mu_t = \alpha + \beta t + \gamma \sin(\omega \cdot (t + \varphi)) + \kappa X_t.$$

2 Surveillance Algorithms

A surveillance algorithm is a procedure using the reference values to create a prediction for the current week. This prediction is then compared with the observed $y_{0:t}$. If the observed number of cases is much higher than the predicted number, the current week is flagged for further investigations.

Currently, `surveillance` implements four different type of algorithms: The Centers for Disease Control and Prevention (CDC) method (Stroup et al., 1989), the Communicable Disease Surveillance Centre (CDSC) method (Farrington et al., 1996), the method used at the Robert Koch Institute (RKI), Germany, and a Bayesian approach documented in Höhle and Riebler (2005).

For the half-window width w (w_0 in year 0) and b years back let the reference values be

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

With $Y_1, \ldots, Y_n|\lambda \overset{\text{iid}}{\sim} \text{Po}(\lambda)$ and Jeffrey’s prior $\lambda \sim \text{Ga}(\frac{1}{2}, 0)$ the predictive posterior can be shown to be

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

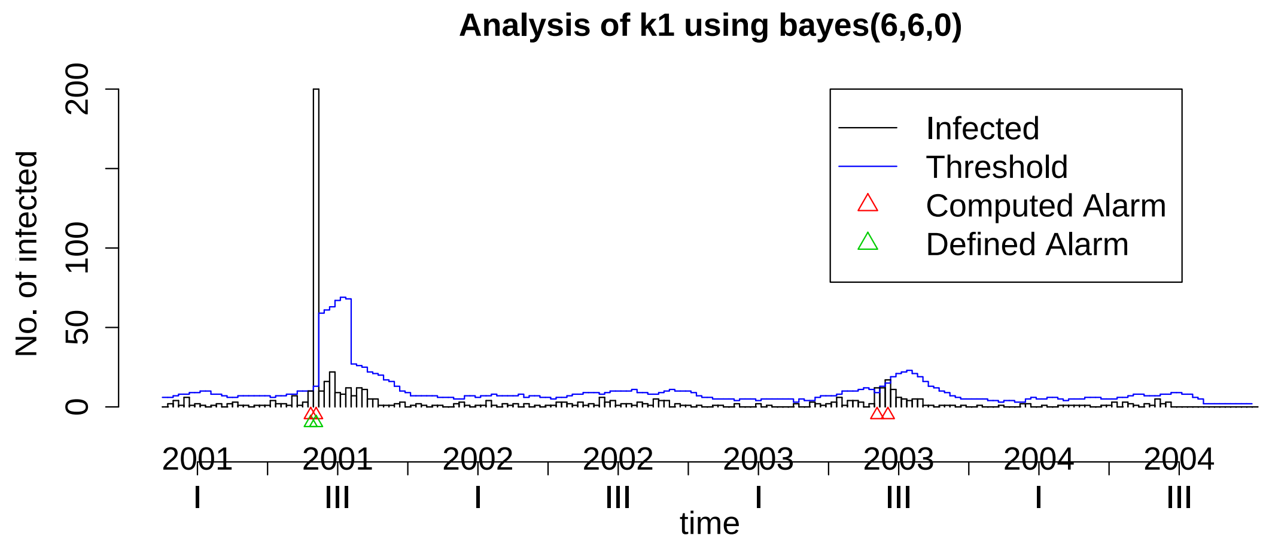
Given the quantile v of the predictive posterior, let y_B be the smallest value, such that

$$P(Y_{0:t} \leq y_B | R_{\text{Bayes}}) \geq v.$$

An alarm is raised if

$$y_{0:t} \geq y_B.$$

The following figure shows the `Bayes(0,6,0)` algorithm applied to the cryptosporidiosis data.



```
> ctrl <- list(range = 7:209,b=0, w=6, alpha=0.001)
> k1.b60 <- algo.bayes(k1, control = ctrl)
```

Several extensions of the simple Bayesian algorithm are imaginable: handling the over-dispersion of the data by using a negative-binomial distribution, integration of time trends and mechanisms correcting for past outbreaks. However, such methods would require simulation based methods like Markov Chain Monte Carlo or heuristic approximations in order to compute the required alarm thresholds from the predictive posterior.

3 Classification Power

If the true outbreak status is known, classical measures such as sensitivity (se) and specificity (sp) of the classification can be computed. To compute the various scores the function `algo.quality` can be used on a `SurvRes` object, e.g. the `k1.b60` object.

	TP	FP	TN	FN	sens	spec	dist	lag
bayes(0,6,0)	2	2	199	0	1.00	0.99	0.01	0.00

```
> algo.quality(k1.b60)
```

In the above, `dist` is the Euclidean distance from (se, sp) to the point $(1, 1)$ and `lag` is the mean delay until an outbreak is identified.

To compare the results of multiple algorithms on a single time series the function `algo.call` is used on a list of control objects – one for each algorithm. A test of multiple algorithms on a set of time series can be done as follows.

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

	TP	FP	TN	FN	sens	spec	dist	lag
rki(0,6,0)	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(0,6,0)	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)	25	26	2682	193	0.11	0.99	0.89	8.21

Here `outbrks` is a list containing the outbreaks as `SurvRes` objects. In both this study and in a simulation study the Bayesian approach seems to do quite well. Consult the work of Riebler (2004) for a more thorough comparison using simulation studies.

4 Discussion and Future work

The presented package provides a test-bench for integrating new surveillance algorithms. Extensions to the package could be more realistic mechanisms for the simulation of epidemics, e.g.

- multi-day outbreaks originating from single-source exposure with incubation time
- epidemics based on the Susceptible-Exposed-Infected-Recovered (SEIR) model

and the implementation of multivariate surveillance algorithms as in (Held et al., 2005).

The package Homepage:

`http://www.stat.uni-muenchen.de/~hoehle/software/surveillance/`

5 Acknowledgments

We are grateful to K. Stark and D. Altmann, RKI, Germany, for discussions and information on the surveillance methods used by the RKI. The research was conducted with financial support from the Collaborative Research Centre SFB 386 funded by the German research foundation (DFG).

References

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 data. *Statistical Modelling 5*, 187–199.

Höhle, M. and A. Riebler (2005). The R-Package 'surveillance'. SFB386 Discussion Paper 422, Department of Statistics, University of Munich.

Riebler, A. (2004). Empirischer Vergleich von statistischen Methoden zur Ausbruchserkennung bei Surveillance Daten. Bachelor’s thesis.

Robert Koch-Institut (2004). SurvStat@RKI. http://www3.rki.de/SurvStat. Date of query: September 2004.

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.