Modelling and surveillance of infectious diseases - or why there is an R in SARS

Michael Höhle[1,2]

[1]Department of Statistics, Ludwig-Maximilians-Universität München, Germany

[2]Centre for Mathematical Sciences, Technische Universität München München, Germany

useR!2008
Kaleidoscope II Session
Dortmund, 12 August 2008

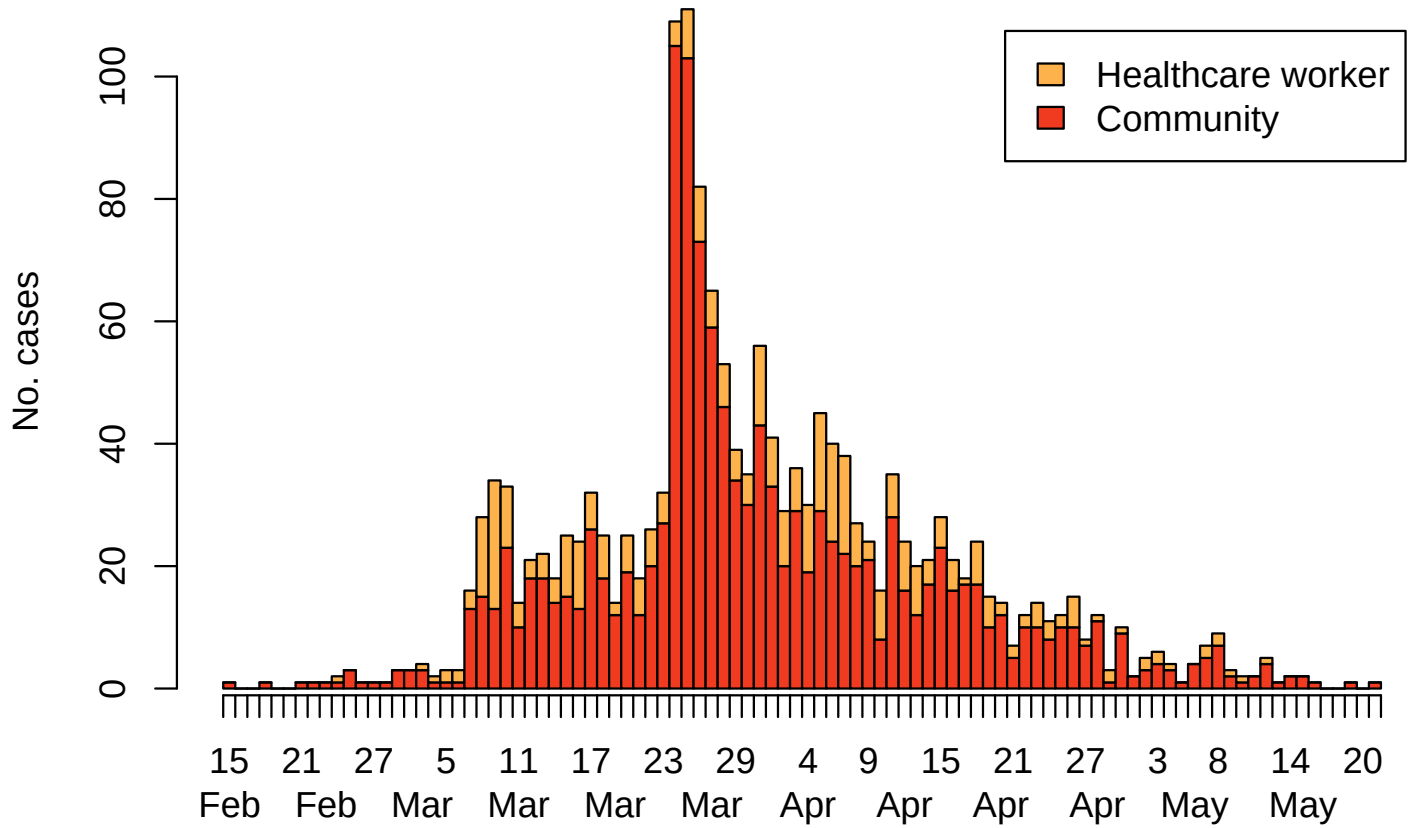
Motivation

- How can R assist in understanding and controlling infectious diseases – be it in human, plant or veterinary epidemiology.
- Two R packages exist:
  1. `RLadyBug` contains a set of functions for the simulation and parameter estimation in spatially heterogeneous SIR models.
  2. `surveillance` contains algorithms for the detection of aberrations in time series of counts arising from routine public health surveillance
- This talk intends to give an overview of using R for especially (1) – deeper mathematical details are suspended to the lunch break

```
> library("RLadyBug")
```

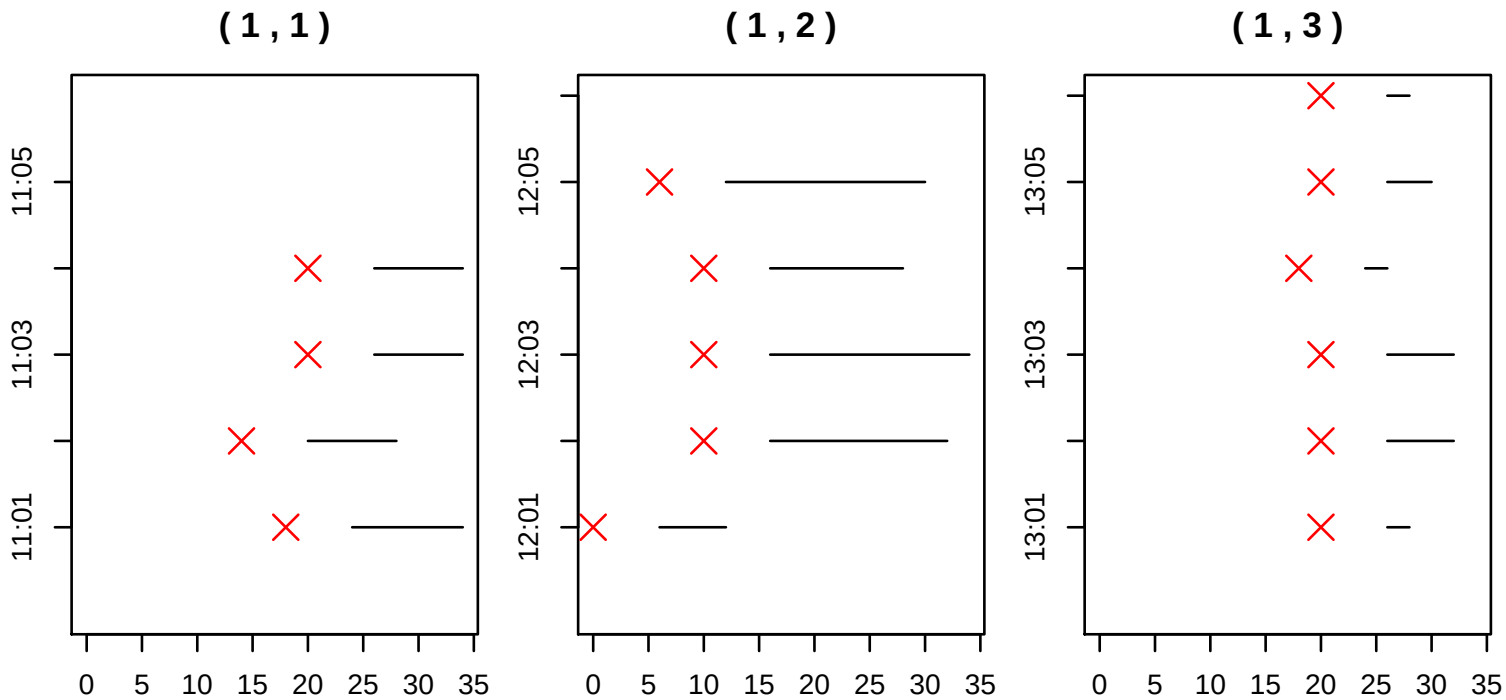
Applications (1): SARS in Hong Kong 2003

- Daily number of new cases of the severe acute respiratory syndrome (SARS) in Hong Kong (Anonymous, 2003)
- Epidemic curve created with package `epitools` (Aragon, 2007).



Applications (2): CSF Transmission Experiment

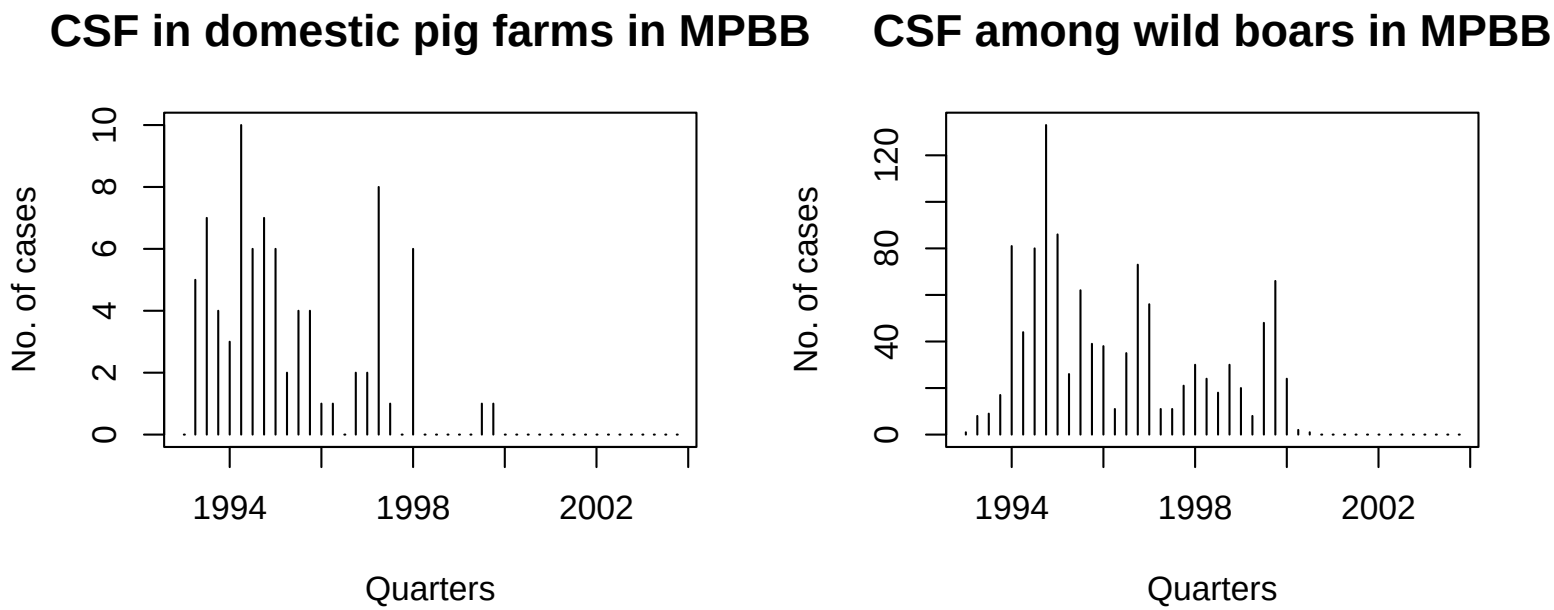
- Experiment by Laevens et al. (1999) with classical swine fever (CSF) using $\mathbf{S}(0) = (5, 5, 6)$ and $\mathbf{E} = (0, 1, 0)$.
- Event history of each pig with inoculation as origin




```
> data("laevens")
> plot(laevens, type = individual ~ time | position)
```

Applications (3) – CSF surveillance

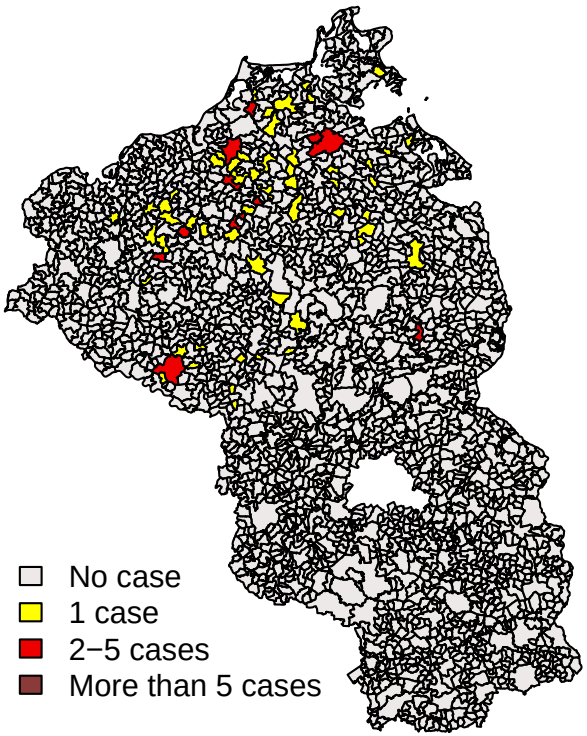
- Classical swine fever (CSF) in Brandenburg (BB) and Mecklenburg-Western Pomerania (MP), Germany
- Total of 81 infected farms out of 3290 during 1993-2004



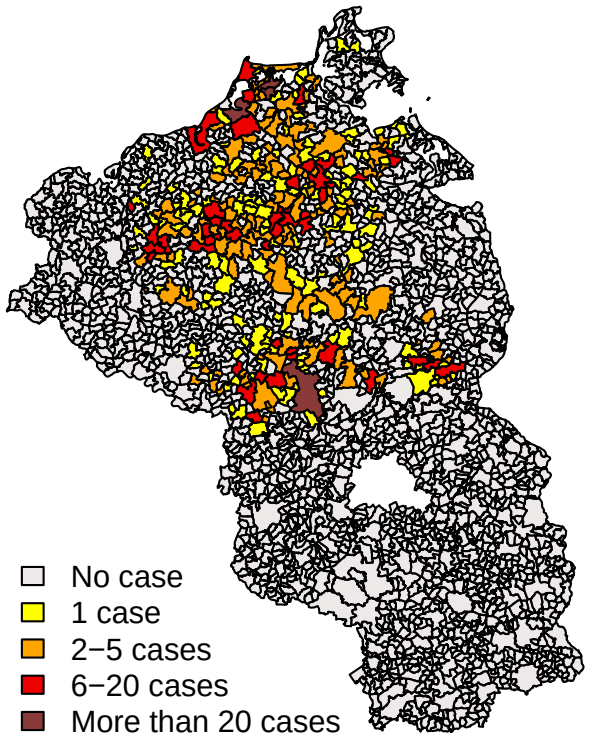
- Interest in investigating the connection between the CSF incidence among domestic pigs and wild boars

Applications (4) - Spatial incidence of CSF in MPBB

Domestic pigs in MPBB



Wild boars in MPBB



Stochastic epidemic models (1)

- SEIR model: A closed population of $n + m$ individuals divided into *s*usceptible, *e*xposed, *i*nfectious, and *r*ecovered
- $S(0) = n$, $E(0) = m$, $I(0) = 0$ and $R(0) = 0$
- At time t, an individual j meets infectious at rate

$$\lambda_j(t|\mathcal{H}_t) = \sum_{i=1}^{n+m} \mathbb{1}_{i \in \text{Infectious}(t)} \cdot f(i,j),$$

  where $f(\cdot) \geq 0$ is a function of the distance between i and j
- If a susceptible meets an infected, it becomes exposed

Stochastic epidemic models (2) – Distance kernels

1. Homogeneous model: $\forall i, j : f(i,j) = \beta > 0$ and hence

$$\lambda_j(t|\mathcal{H}_t) = \beta I(t)$$

2. Heterogeneous model: The population is made up of k units arranged on a grid in space. For j in unit u_j:

$$\lambda_j(t|\mathcal{H}_t) = \beta I_{u_j}(t) + \beta_\eta \sum_{u \in N(u_j)} I_u(t)$$

3. Heterogeneous model, where individuals have locations in $\mathbb{R}^2$ and $f(i,j)$ is a function on the Euclidean distance $\text{dist}(i,j)$

SARS in Hong Kong 2003

- Assuming a constant incubation time of 6.4 days and a constant recovery time of 34 days as suggested by the mean of the distributions in Donelly et al. (2003) we obtain

```
> data("hksars")
> print(m1 <- seir(hksars, hksars.opts.m1))
```

```
Calling LadyBug (monitor ladybug.system.out/err for progress)...
...
Parameter Estimations:
  Parameter:
          beta
    4.3984e-09
...
```

- Basic reproduction number $R_0 =$ `R0(m1, hksars)` $= 1.0012$.

CSF Transmission Experiment (1)

- Exposure times are not observed, instead of imposing we assume $T_E \sim \mathcal{G}a(\delta_E, \gamma_E)$ and $T_I \sim \mathcal{G}a(\delta_I, \gamma_I)$
- A Bayesian setting with MCMC is used for parameter inference

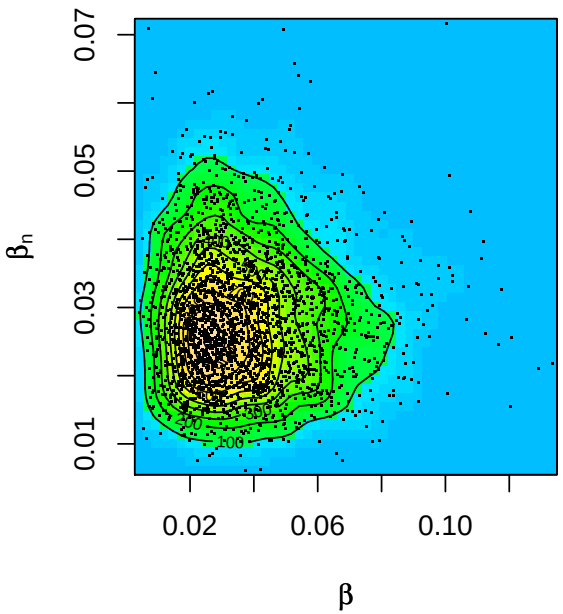
```
> print(m2 <- seir(laevens, laevens.opts.mcmc))
```

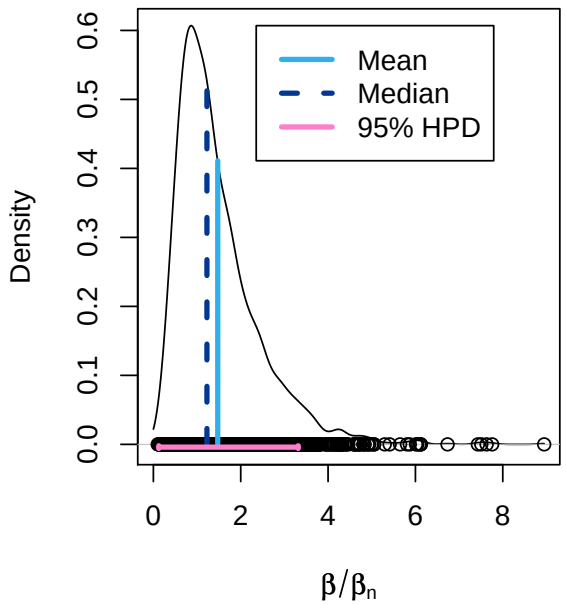
```
An object of class LBInferenceMCMC

Parameter Estimations (posterior mean from 2500 samples):
  Parameter:
    beta    betaN   gammaE   deltaE   gammaI   deltaI
 0.03706  0.02837 56.82000  9.37400  2.16200  0.25640

  StandardErrors (posterior std.dev. from 2500 samples):
     beta     betaN    gammaE    deltaE    gammaI    deltaI
 0.018500  0.009481 45.510000  7.761000  0.738100  0.097760
```

CSF Transmission Experiment (2)

- MCMC output can be further analysed by e.g. coda package
- Posterior density of β/β_η and R_0

```
> plot(m2, which = "betabetaN")
> quantile(R0(m2, laevens), c(0.025, 0.5, 0.975))
```






```
  2.5%    50%  97.5%
0.3245 0.6370 1.2426
```

CSF surveillance (1)

- CSF surveillance data consists of multiple outbreaks
- SIR Extension: Risk of infection consists of two components
  - endemic component: Time to infection from external sources modelled by a Cox model
  - epidemic component: Similar to heterogeneous SIR model with distance weighting of infectives
- Rate of infection has the following form

$$\lambda_j(t|\mathcal{H}_t) = \exp\left(h_0(t) + \mathbf{z}_j(t)'\boldsymbol{\alpha}\right) + \sum_{i=1}^{n+m} \mathbb{1}_{i\in\text{Infectious}(t)} \cdot f(i,j)$$

- When using a linear basis expansion of $f(i,j)$ this rate is similar to the conditional intensity of an additive-multiplicative hazard model from survival analysis

CSF surveillance (2)

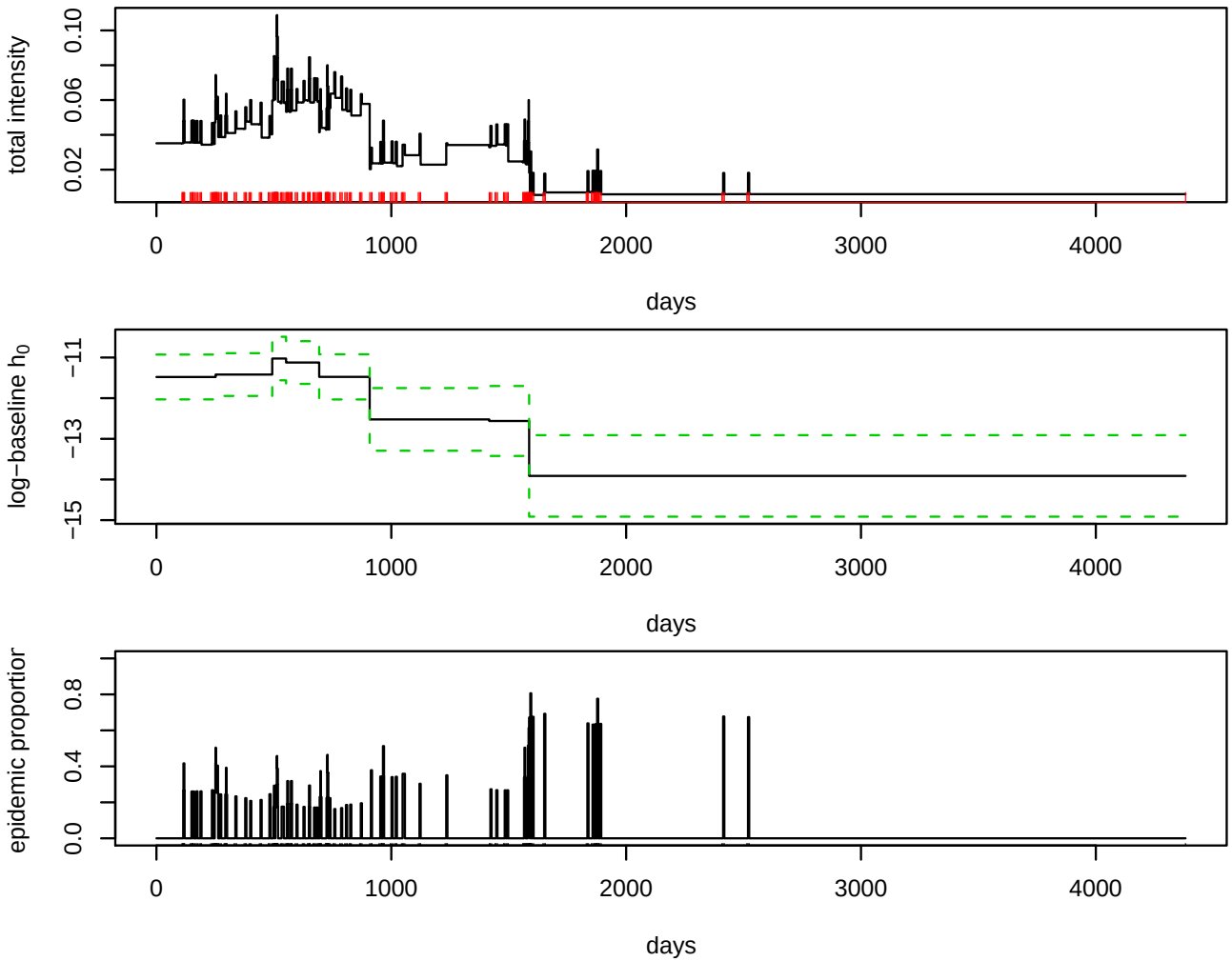
- Endemic component: piecewise exponential baseline and time varying covariates boars and vaccination area
- Epidemic component: $f(i,j) = \beta > 0$
- Inference using penalized loglikelihood with a model syntax similar to the `timereg` package (Scheike, 2006)

```
> m3 <- spatialSIR(Surv(start, stop, event) ~ fconst +
+     cox(boar) + cox(vacc), data = mpbb.evHist, ...)
> coef(m3)[c("fconst", "cox(boar)", "cox(vacc)")]
> diag(vcov(m3))[c("fconst", "cox(boar)", "cox(vacc)")]
```

```
   fconst cox(boar) cox(vacc)
3.814e-06 2.108e+00 1.261e+00

   fconst cox(boar) cox(vacc)
7.371e-12 9.263e-02 1.729e-01
```

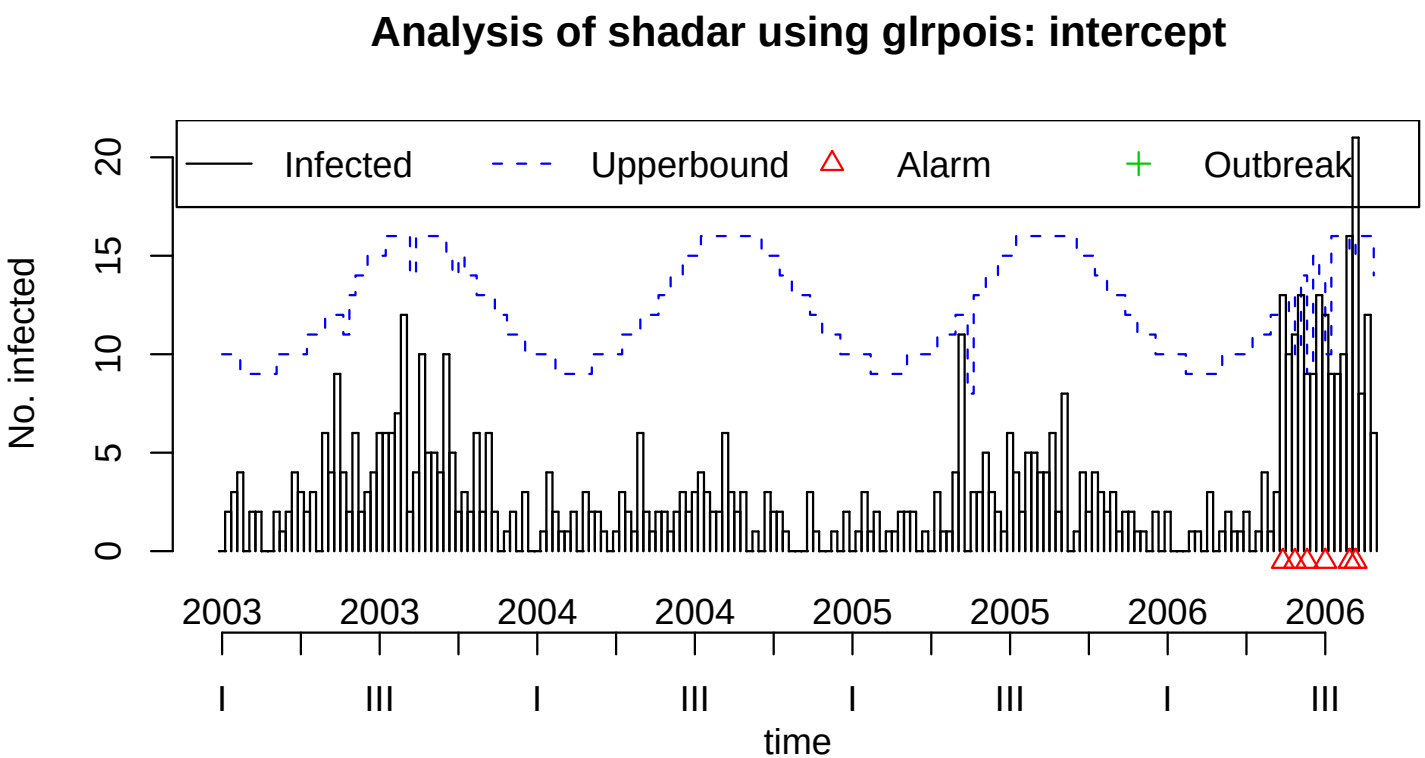
CSF surveillance (3)

- Plot of the total intensity $\sum_{j=1}^{n+m} \lambda_j(t|\mathcal{H}_t)$, the log baseline hazard $h_0(t)$ (with a 95% CI) and the epidemic proportion



The surveillance package (1)

```
> library("surveillance")
> data("shadar")
> control = list(range = 105:295, ret = "cases", alpha = 0,
+     c.ARL = 5)
> plot(algo.glrnb(shadar, control = control))
```



The surveillance package (2)

- Surveillance algorithms:
  - `cdc` – Stroup et al. (1989)
  - `farrington` – Farrington et al. (1996)
  - `cusum` – Rossi et al. (1999)
  - `roya` – Rogerson and Yamada (2004) (Experimental)
  - `lr` and `glr` – H. and Paul (2008)
- Comparsion of surveillance algorithms using sensitivity, specificity and its variants in simulations
- Surveillance time series models:
  - `hhh` - Held et al. (2005); Paul et al. (2008)
  - `twins` - Held et al. (2006) (Experimental)

Discussion

- First steps towards R functionality for infectious disease modelling. More complex and realistic models imaginable.
- Packages contain many additional visualization and simulation procedures (Sellke construction, Ogata's modified thinning)
- Combining database, R, Sweave/odfWeave and LaTeX/OpenOffice allows for easy generation of daily bulletins or reports

```
> motd
```

Message of the day

Packages are on CRAN. Starting points are H. (2007); H. and Feldmann (2007). Maybe they are of help. If adaptation is needed for your problem let me know.

Acknowledgements

Persons:

- Ulrike Feldmann, Sebastian Meyer and Valentin Wimmer, Ludwig-Maximilians-Universität München, Germany
- Michaela Paul and Andrea Riebler, Institute of Social and Preventive Medicine, University of Zurich, Switzerland
- Christoph Staubach, Federal Research Institute for Animal Health (FLI), Germany

Financial Support:

- Munich Center of Health Sciences

Literature I

Anonymous (2003). Sars bulletin. Technical report, Health, Welfare and Food Bureau, Government of the Hong Kong Special Administrative Region.

Aragon, T. (2007). *epitools: Epidemiology Tools*. R package version 0.4-9.

Donelly, C. A., Ghani, A. C., Leung, G. M., and 16 more authors (2003). Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *The Lancet*, 361:1761.

Farrington, C., Andrews, N., Beale, A., and Catchpole, M. (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., Hofmann, M., Höhle, M., and Schmid, V. (2006). A two component model for counts of infectious diseases. *Biostatistics*, 7:422–437.

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

Höhle, M. (2007). surveillance: An R package for the monitoring of infectious diseases. *Computational Statistics*, 22(4):571–582.

Höhle, M. and Feldmann, U. (2007). RLadyBug – an R package for stochastic epidemic models. *Computational Statistics and Data Analysis, Special Issue on Statistical Software*, 52(2).

Literature II

Höhle, M. and Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Computational Statistics and Data Analysis*, 52(9):4357–4368.

Laevens, H., Koenen, F., Deluyker, H., and de Kruif, A. (1999). Experimental infection of slaughter pigs with classical swine fever virus: transmission of the virus, course of the disease and antibody response. *Vet. Rec.*, 145:243–248.

Paul, M., Held, L., and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*. In review.

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

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

Scheike, T. (2006). *timereg: Additive Semiparametric Timevarying Regression*. R package version 1.0-2.

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