

Short course on
Temporal and spatio-temporal monitoring of infectious diseases

Michael Höhle¹

¹Department for Infectious Disease Epidemiology
Robert Koch Institute, Berlin, Germany

Department of Statistics
“G. Parenti” University of Florence
Florence, Italy, 09-10 Feb 2011



Outline

- 1 Introduction
- 2 The R package *surveillance*
- 3 Univariate time series detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling
- 6 Discussion and Summary

Outline

- 1 Introduction
- 2 The R package *surveillance*
- 3 Univariate time series detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling
- 6 Discussion and Summary

Introduction

- This short course is about the statistical analysis of routinely collected surveillance data seen as
 - ▶ multivariate time series of counts
 - ▶ realizations of spatio-temporal point processes
- Course aim is to explain concepts behind retrospective modelling and prospective monitoring in infectious disease epidemiology.
- The statistical methods of this talk are implemented in the R-package *surveillance* available from the Comprehensive R Archive Network (CRAN) (Höhle, 2007).

Aims of statistical surveillance

Public health surveillance

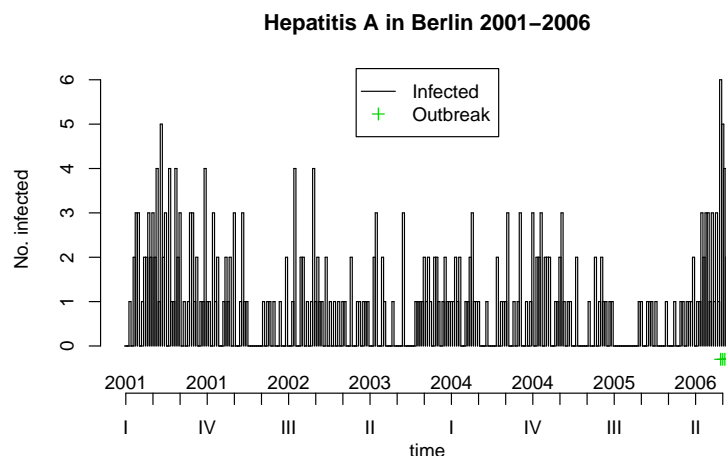
Ongoing systematic collection, analysis, interpretation and dissemination of health data for the purpose of preventing and controlling disease, injury, and other health problems (Thacker, 2000).

Course view:

- Real-time online monitoring within a setting of statistical process control.
- Detect aberrations for public health events in a statistical setting with a little less heuristics involved than sometimes applied at the moment.
- Provide formal tool as a supplement to gut instinct.

Example of surveillance data

- Weekly number of adult male hepatitis A cases in the federal state of Berlin during 2001-2006
- During summer 2006 health authorities noticed an increased amount of cases (Robert Koch Institute, 2006).



Examples of disease surveillance applications

In human epidemiology

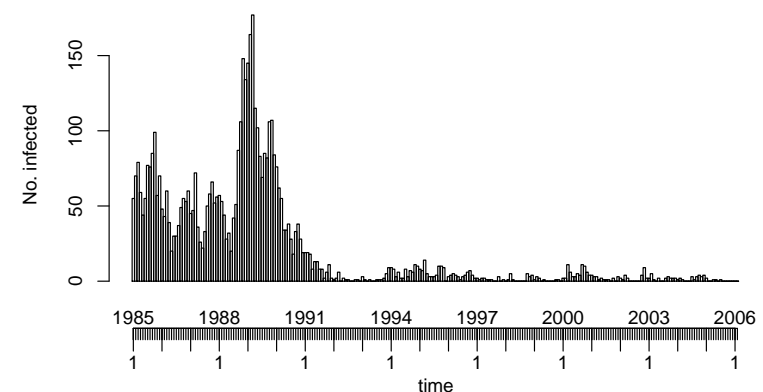
- Monitoring of congenital malformations (Chen, 1978)
- Surveillance of notifiable diseases (Robert Koch Institute, 2009; Widdowson et al., 2003)
- Monitoring surgical outcomes (Steiner et al., 2000)

In veterinary epidemiology

- Salmonella in livestock reports, Veterinary Laboratories Agency, UK (Kosmider et al., 2006)
- Rabies Surveillance (WHO Collaboration Centre for Rabies Surveillance and Research, 2007)
- Monitoring of abortions in dairy cattle (Carpenter et al., 2007)

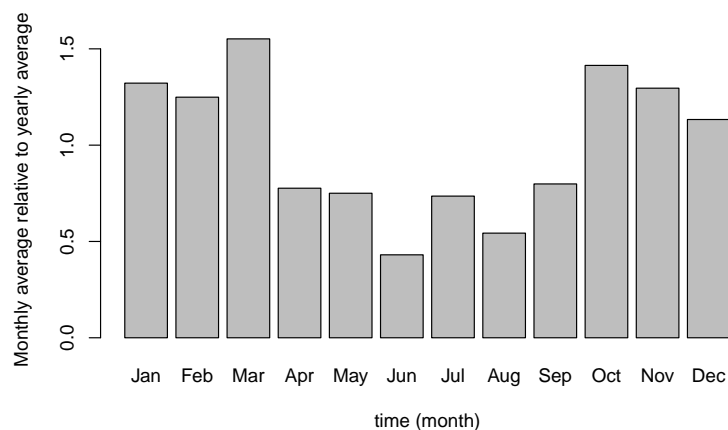
Example – Rabies among foxes in Hesse 1985-2006 (1)

Monthly counts are provided by the WHO Collaboration Centre for Rabies Surveillance and Research. Thanks to Christoph Staubach, Federal Research Institute for Animal Health, Germany.



The observed count time series is $\{y_t\}_{t=1}^{254} = \{y_{1:1985}, \dots, y_{2:2006}\}$.

Example – Rabies among foxes in Hesse 1985-2006 (2)



To illustrate seasonality:

- 1 divide monthly cases by the respective yearly average
- 2 compute monthly mean of this detrended time series

Surveillance of acute respiratory diseases (1)

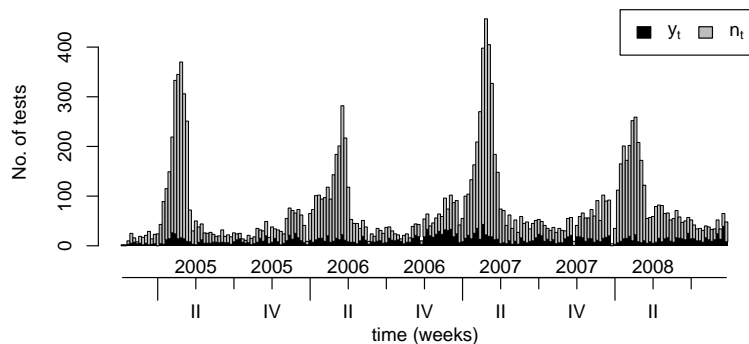
- Since autumn 2004 the Governmental Institute of Public Health of Lower Saxony carries out a surveillance of acute respiratory diseases (Beyrer et al., 2006)
- The surveillance consists of two modules
 - 1 Voluntary reporting module for daycare facilities
 - 2 Module containing the investigation of throat swabs from selected medical practices (pediatrists and general practitioners)
- Focus on module 2, where each throat swab is tested for five viral agents: influenza virus, respiratory syncytial virus (RSV), adeno virus, picorna virus and metapneumo virus

Surveillance of acute respiratory diseases (2)

- For each agent one has a binomial time series

$$y_t \sim \text{Bin}(n_t, \pi_t).$$

- Example: Positive picorna virus tests during surveillance.

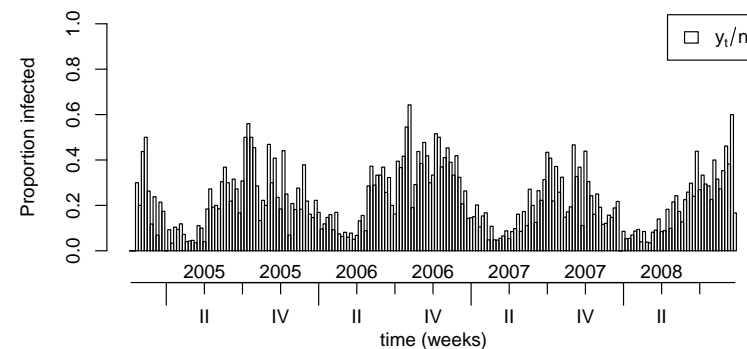


Surveillance of acute respiratory diseases (2)

- For each agent one has a binomial time series

$$y_t \sim \text{Bin}(n_t, \pi_t).$$

- Example: Positive picorna virus tests during surveillance.

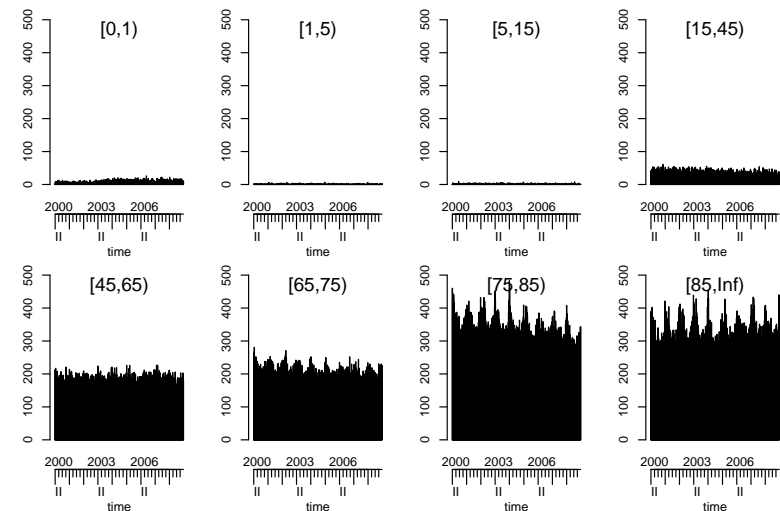


Example – The EuroMOMO project (1)

- 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: Surveillance aspect illustrated by Danish mortality data provided by Statens Serum Institut, Denmark

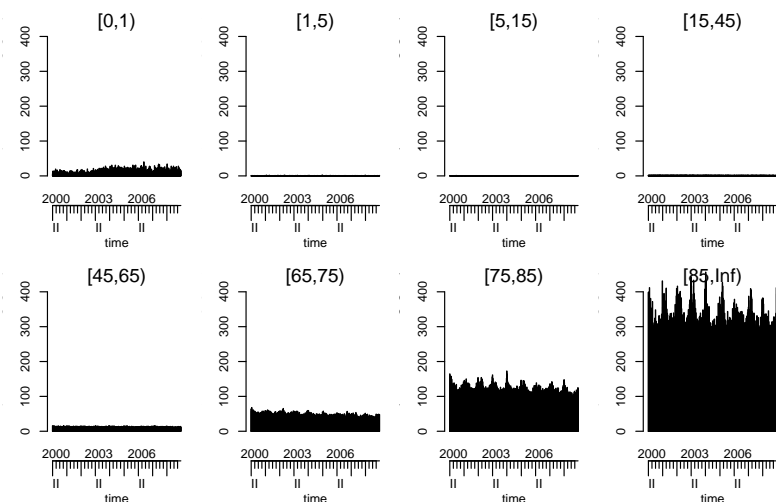
Example – The EuroMOMO project (2)

Weekly number of deaths in six age groups (alternatively incidence per 100,000 persons in age group)



Example – The EuroMOMO project (2)

Weekly number of deaths in six age groups (alternatively incidence per 100,000 persons in age group)



The quality of surveillance data


Issues complicating statistical analysis of the time series

- Lack of clear case definition
- Under-reporting and reporting delays
- Lack of denominator data
- Seasonality
- Low number of disease cases
- Presence of past outbreaks
- Heterogeneity caused by factors such as age, sex, vaccination status, environmental factors

Outline

- 1 Introduction
- 2 The R package surveillance
- 3 Univariate time series detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling
- 6 Discussion and Summary

What is surveillance? (1)

An open source  package for the visualization, modeling and monitoring of routinely collected public health surveillance data

- Prospective monitoring for univariate count data time series:
 - ▶ `farrington` – Farrington et al. (1996)
 - ▶ `cusum` – Rossi et al. (1999) and extensions
 - ▶ `rogerson` – Rogerson and Yamada (2004)
 - ▶ `bayes` – Höhle (2007)
 - ▶ `glrnb` – Höhle and Paul (2008)
- Prospective changepoint detection for categorical time series:
 - ▶ `pairedbinCUSUM` – surgical performance (Steiner et al., 2000)
 - ▶ `categoricalCUSUM` – binomial-, beta-binomial-, multinomial logit- and Bradley-Terry modelling (Höhle, 2010)

What is surveillance? (2)

- Retrospective count data time series models:
 - ▶ `hhh` – Held et al. (2005); Paul et al. (2008)
 - ▶ `hhh4` – Paul and Held (2011)
 - ▶ `twins` – Held et al. (2006)
- Spatio-Temporal point process modelling and monitoring:
 - ▶ `twinSIR` – discrete space - continuous time modelling (Höhle, 2010)
 - ▶ `twins` – continuous space - continuous time modelling (Meyer et al., 2010)
 - ▶ `stcd` – continuous space - continuous time cluster detection (Assunção and Correa, 2009)

What is surveillance? (3)

- Motivation: Provide data structure and implementational framework for methodological developments
- Spin-off: Tool for epidemiologists and others working in applied disease monitoring
- Availability: CRAN, current development version from


```
http://surveillance.r-forge.r-project.org/
```
- To install the development version under R version 2.12:


```
install.packages("surveillance",repos="http://r-forge.r-project.org")
```
- Package is available under the GNU General Public License (GPL) v. 2.0.

Data structure: The sts class (1)

- A surveillance time series $\{y_{it}; t = 1, \dots, n, i = 1, \dots, m\}$ is represented using objects of class `sts` (surveillance time series)
- The `sts` S4 class has the following form


```
setClass("sts", representation(epoch = "numeric",
                                freq = "numeric",
                                start = "numeric",
                                observed = "matrix",
                                state = "matrix",
                                alarm = "matrix",
                                upperbound = "matrix",
                                neighbourhood = "matrix",
                                populationFrac = "matrix",
                                map = "SpatialPolygonsDataFrame",
                                control = "list",
                                epochAsDate = "logical",
                                multinomialTS = "logical"))
```
- Old S3 class `disProg` objects can be converted to `sts` objects using the function `disProg2sts`.

Data structure: The sts class (2)

- observed** A $n \times m$ matrix of counts representing y_{it}
- start** A vector of length two containing the origin of the time series as `c(year, week)`.
- freq** A numeric specifying the period of the time series, i.e. 52 for weekly data, 12 for monthly data, etc.
- alarm** A $n \times m$ matrix of Booleans containing the result of applying a surveillance algorithm to the time series
- upperbound** A $n \times m$ matrix containing the number of cases which would result in an alarm (specific interpretation is algorithm dependent)
- control** List with control arguments used for the surveillance algorithm

Data structure: The sts class (3)

- populationFrac** Population data, either population data or denominator data
- map** `SpatialPolygonsDataFrame` from package `sp` containing geographical locations
- neighbourhood** A $m \times m$ matrix of Booleans indicating neighbourhood relationships between regions
- epochAsDate** Boolean, if TRUE then the epoch vector is interpreted as a vector of class `Date`, i.e. dates in ISO 8601 date standard
- multinomialTS** If TRUE the `populationFrac` slot is interpreted as denominator data (binomial, multinomial)

Data I/O

- To import data into R one can use `read.table/read.csv`, package `foreign` (SAS, SPSS, Stata, Systat, dBase) or the RODBC database interface (Access, Excel, SQL databases).
- An `sts` object is then created from the resulting matrix of counts.

```
R> ha.counts <- as.matrix(read.csv("../data/ha.csv"))
R> ha <- new("sts", epoch = 1:nrow(ha.counts), start = c(2001,
+ 1), freq = 52, observed = ha.counts, state = matrix(0,
+ nrow(ha.counts), ncol(ha.counts)))
```

- All plotting, accessing, aggregating and application of surveillance algorithms works on `sts` objects.

Accessing sts objects (1)

- Printing provides basic information about the time series:

```
R> print(ha)
```

```
-- An object of class sts --
freq:          52
start:         2001 1
dim(observed): 290 12

Head of observed:
  chwi frkr lich mahe mitt neuk pank rein span zehl scho trko
[1,]   0   0   0   0   0   0   0   0   0   0   0   0

map:
[1] chwi frkr lich mahe mitt neuk pank rein scho span trko zehl
12 Levels: chwi frkr lich mahe mitt neuk pank rein scho span ... zehl

head of neighbourhood:
  chwi frkr lich mahe mitt neuk pank rein span zehl scho trko
chwi  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

Accessing sts objects (2)

- Matrix like accessing such as `ha[1:52,]` or `ha[, "mitt"]` results in sts objects containing the respective sub time series.

- Functions such as `dim`, `nrow` and `ncol` are also defined:

```
R> dim(ha)
[1] 290 12
```

- The time series can be aggregated temporally and spatially:

```
R> dim(aggregate(ha, by = "unit"))
```

```
[1] 290 1
```

```
R> dim(aggregate(ha, by = "time"))
```

```
[1] 1 12
```

- Currently, the slots of sts objects are accessed directly:

```
R> head(ha@observed, n = 1)
```

```
      chwi frkr lich mahe mitt neuk pank rein span zehl scho trko
[1,]   0   0   0   0   0   0   0   0   0   0   0   0
```

Accessing sts objects (3)

- Aggregation can also be of subsets.
- Example: Aggregate weekly data into 4 week blocks (corresponding to 13 observations per year)

```
R> ha4 <- aggregate(ha[, c("pank", "mitt", "frkr", "scho",
+ "chwi", "neuk")], nfreq = 13)
```

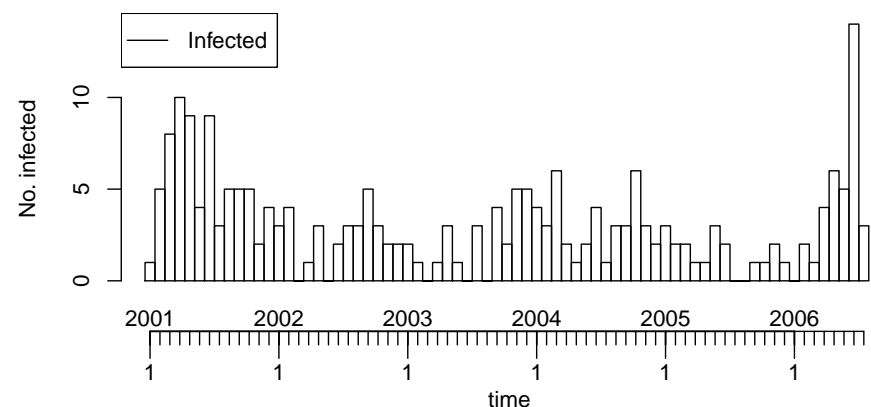
```
R> dim(ha4)
```

```
[1] 73 6
```

Visualizing sts objects (1)

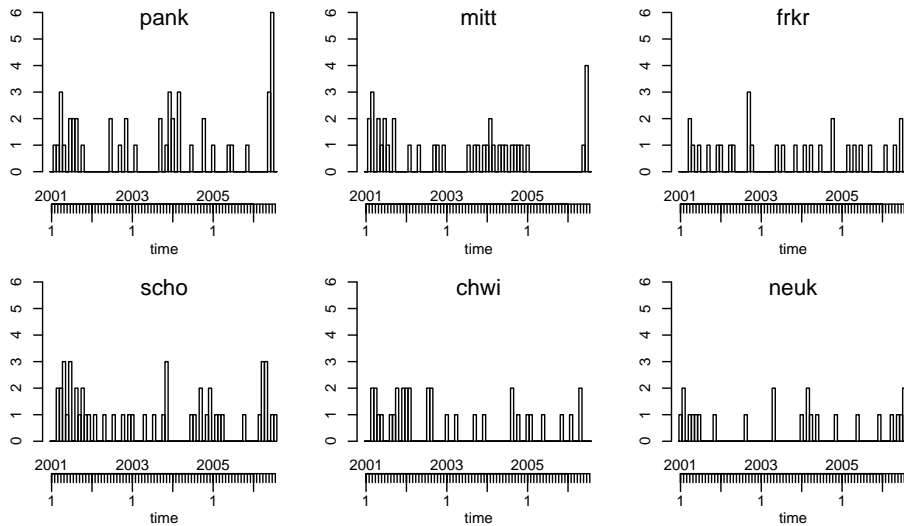
- The `plot` function provides an interface to several visual representations controlled by the `type` argument.

```
R> plot(ha4, type = observed ~ time)
```



Visualizing sts objects (2)

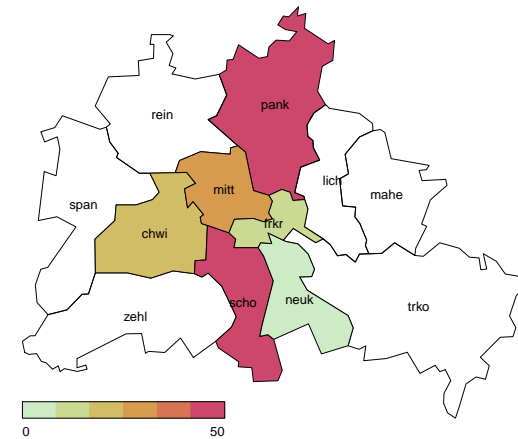
```
R> plot(ha4, type = observed ~ time | unit)
```



Visualizing sts objects (3)

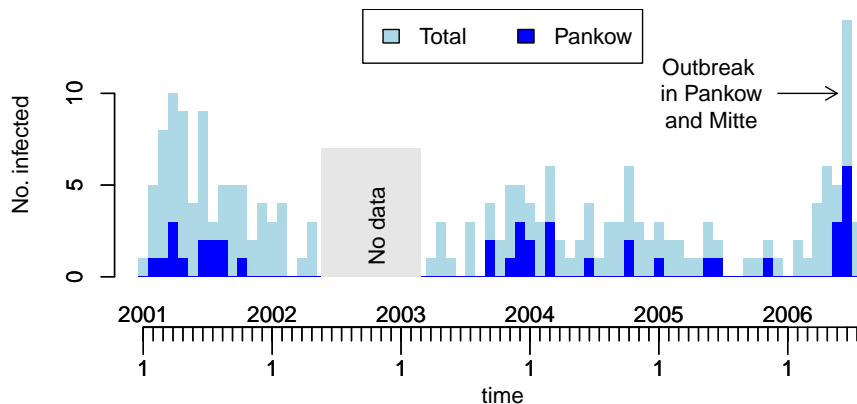
Using the `maptools` package shapefiles provides map visualizations

```
R> plot(ha4, type = observed ~ 1 | unit)
```



Visualizing sts objects (4)

- Using `type = observed~1|time*unit` one would have created an animation of pictures for each time index
- Plotting functionality is customizable as in R-graphics



Outline

- 1 Introduction
- 2 The R package `surveillance`
- 3 Univariate time series detectors
 - Farrington algorithm
 - Negative Binomial CUSUM
 - Binomial CUSUM
 - Evaluating performance
 - Likelihood ratio detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling

Statistical Framework for Aberration Detection

- Univariate time series $\{y_t, t = 1, 2, \dots\}$ to monitor
- At the unknown time τ , an important change in the process occurs. For each time t we differentiate between two-states:

$$x_t = \begin{cases} 0 & \text{if } t < \tau \quad (\text{in-control}), \\ 1 & \text{otherwise} \quad (\text{out-of-control}). \end{cases}$$

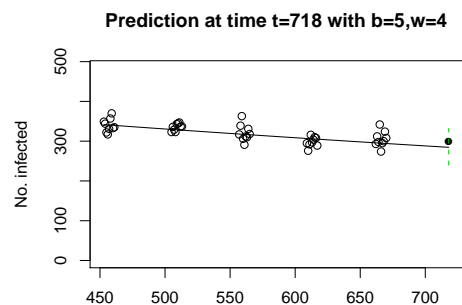
- At time $s \geq 1$, the available information is $\mathbf{y}_s = \{y_t; t \leq s\}$.
- Detection is based on a statistic $r(\cdot)$ with resulting alarm time

$$T_A = \min\{s \geq 1 : r(\mathbf{y}_s) > g\},$$

where g is a known threshold.

Farrington algorithm (1) – basic model

- Predict value y_{t_0} at time $t_0 = (t_0^m, t_0^y)$ using a set of reference values from window of size $2w + 1$ up to b years back.



- 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 $\phi > 0$.

Outline

- 1 Introduction
- 2 The R package *surveillance*
- 3 Univariate time series detectors
 - Farrington algorithm
 - Negative Binomial CUSUM
 - Binomial CUSUM
 - Evaluating performance
 - Likelihood ratio detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling

Farrington algorithm (2) – outbreak detection

Predict and compare:

- An approximate $(1 - \alpha)\%$ prediction interval for y_{t_0} based on the GLM has upper limit $U = \hat{\mu}_{t_0} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\text{Var}(y_{t_0} - \hat{\mu}_{t_0})}$
- If observed y_{t_0} is greater than U , then flag t_0 as outbreak

Remarks:

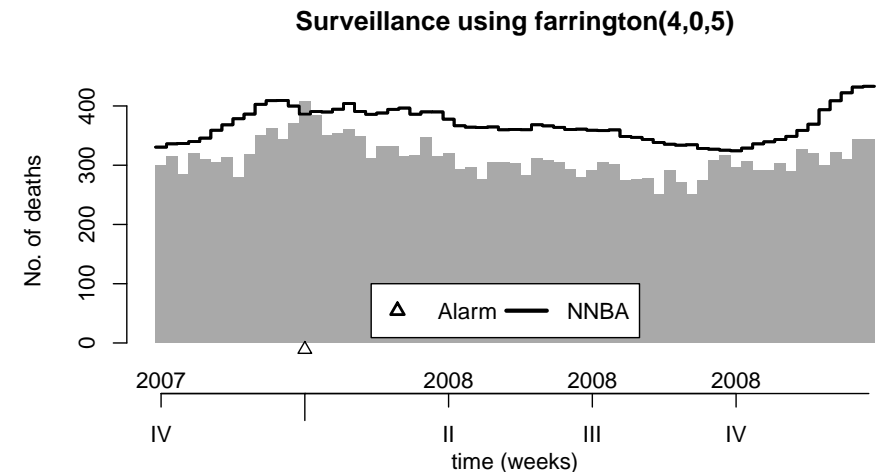
- Linear trend is only included if significant at 5% level, $b \geq 3$ and no over-extrapolation occurs.
- Automatic correction for past outbreaks by computing Anscombe residuals for reference values and re-fit GLM assigning lower weights to values with large residuals.
- Low count protection – the algorithm raises an alarm only if more than 5 cases in past 4 weeks.

Farrington algorithm in surveillance (1)

- Function `farrington` takes an `sts` and a `control` object as arguments
- `control` is a list with the following components:
 - `range` Specifies the index of all timepoints in `sts` to monitor.
 - `b` Number of years to go back in time
 - `w` Window size
 - `reweight` Boolean stating whether to perform reweight step using Anscombe residuals
 - `trend` If TRUE a trend is included in first fit and kept in case the conditions are met. Otherwise no trend.
 - `alpha` An approximate two-sided $(1 - \alpha)\%$ prediction interval is calculated

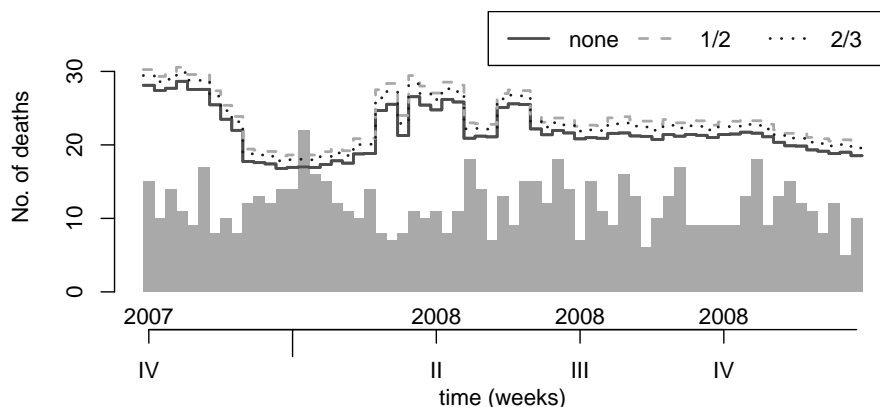
Farrington algorithm in surveillance (2)

- Results for $w = 4, b = 5$ and $\alpha = 0.01$ starting at W40-2007:



Farrington algorithm in surveillance (4)

- Argument `powertrans` in `control` indicates which power transformation to use:
 - "2/3" skewness correction in low count scenario
 - "1/2" variance stabilizing square-root transformation
 - "none" no transformation



Correcting for past outbreaks (1)

- Problems arise when base-line counts contain outbreaks. A reweighting procedure is used to downweight such observation.
- Compute standardized Anscombe residuals for Poisson distribution:

$$s_t = \frac{r_t}{\hat{\phi}\sqrt{1-h_{tt}}}, \quad \text{where } r_t = \frac{3(y_t^{\frac{2}{3}} - \hat{\mu}_t^{\frac{2}{3}})}{2\hat{\mu}_t^{\frac{1}{6}}}$$

- Define weights ω_t as

$$\omega_t = \begin{cases} \gamma \frac{1}{s_t^2} & \text{if } s_t > 1 \\ \gamma & \text{otherwise} \end{cases},$$

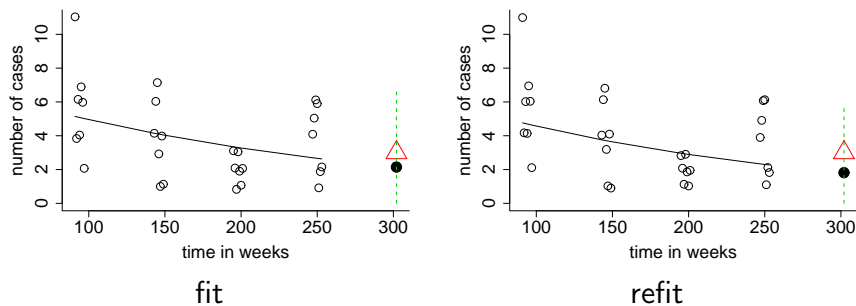
where γ ensures $\sum_{i=1}^k \omega_t = n$.

Correcting for past outbreaks (2)

- Refit the GLM using the ω_t weights, i.e.

$$\text{Var}(y_t) = \frac{\phi \mu_t}{\omega_t}$$

- Effect of weights is to downweight large positive outliers in the data:



Outline

- 1 Introduction
- 2 The R package `surveillance`
- 3 Univariate time series detectors
 - Farrington algorithm
 - **Negative Binomial CUSUM**
 - Binomial CUSUM
 - Evaluating performance
 - Likelihood ratio detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling

Theory: Negative Binomial CUSUM (1)

- Likelihood ratio between the out-of-control and in-control models at time s given that $\tau = t$:

$$L(s, t) = \frac{f(\mathbf{y}_s | \tau = t)}{f(\mathbf{y}_s | \tau > s)} = \prod_{i=t}^s \frac{f(y_i; \theta_1)}{f(y_i; \theta_0)},$$

where $f(\cdot; \theta)$ is the negative binomial PMF with parameter vector θ .

- Cumulative Sum (CUSUM) procedure advantageous for detecting sustained shifts:

$$r(\mathbf{y}_s) = \max\{1 \leq t \leq s : \log L(s, t)\}.$$

Theory: Negative Binomial CUSUM (2)

- The computation of $r(\mathbf{y}_s)$ in recursive form:

$$r_0 = 0, \\ r_s = \max\left(0, r_{s-1} + \log \left\{ \frac{f(y_s; \theta_1)}{f(y_s; \theta_0)} \right\}\right), \quad s \geq 1.$$

- When there is evidence against in-control, the LLR contributions are added up.
- No credit in the direction of the in-control is given because r_s cannot get below zero.

Theory: Negative Binomial CUSUM (3)

- Negative-binomial response with fixed dispersion parameter α and in-control mean modeled using a GLM with log-link

$$y_t \sim \text{NegBin}(\mu_{0,t}, \alpha),$$

$$\log(\mu_{0,t}) = \log(\text{pop}_t) + \beta_0 + \beta_1 \cdot t + c_t,$$

where c_t is a cyclic function with period 52 or 53 depending on the number of ISO weeks in the year of t and pop_t denotes the population size in the respective age group at time t .

- As a consequence, $E(y_t) = \mu_{0,t}$ and $\text{Var}(y_t) = \mu_{0,t} + \alpha \cdot \mu_{0,t}^2$
- Out-of-control model for given $\kappa > 1$:

$$\mu_{1,t} = \kappa \cdot \mu_{0,t}.$$

Application: Negative Binomial CUSUM (1)

- Monitoring example: Age group 75-84 starting from week 40 in 2007 (i.e. 1st October 2007) using past 5 years as reference:

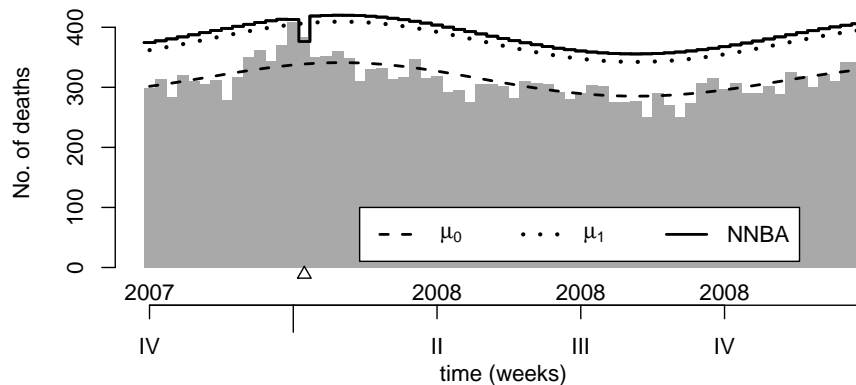
```
R> m <- glm.nb(`observed.[75,85]` ~ 1 + epoch + sin(2*pi*epochInPeriod) +
+ cos(2*pi*epochInPeriod) + offset(log(`population.[75,85]`)),
+ data=momo.df[phase1,])
R> mu0 <- predict(m, newdata=momo.df[phase2,], type="response")
```

- Aim: to optimally detect a 20% increase in the mean, i.e. $\kappa = 1.2$. Use $g = 4.75$ – consequences?

```
R> kappa <- 1.2
R> s.nb <- glrnb(momo[, "[75,85]"], control = list(range = phase2,
+ alpha = 1/m$theta, mu0 = mu0, c.ARL = 4.75, theta = log(kappa),
+ ret = "cases"))
```

Application: Negative Binomial CUSUM (2)

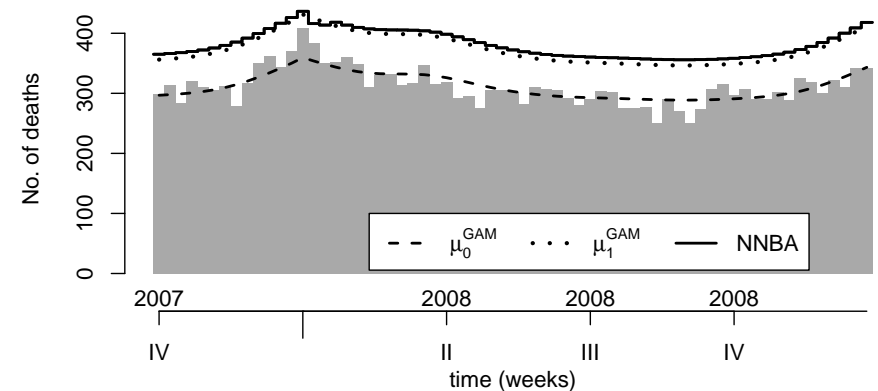
- For week 2 in 2008 an alarm is generated:



- Also shown is the number needed before alarm (NNBA), i.e. given $r(\mathbf{y}_{s-1})$ find the minimum y_s such that $r(\mathbf{y}_s) > g$.

Application: Negative Binomial CUSUM (2)

- For week 2 in 2008 an alarm is generated:



- Also shown is the number needed before alarm (NNBA), i.e. given $r(\mathbf{y}_{s-1})$ find the minimum y_s such that $r(\mathbf{y}_s) > g$.

Outline

- 1 Introduction
- 2 The R package *surveillance*
- 3 Univariate time series detectors
 - Farrington algorithm
 - Negative Binomial CUSUM
 - **Binomial CUSUM**
 - Evaluating performance
 - Likelihood ratio detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling

Binomial CUSUM (1)

- Reweighted CUSUM originally developed by Rogerson and Yamada (2004) for Poisson data.
- Adopted to the binomial situation where $y_t \sim \text{Bin}(n_t, \pi_0)$, $t = 1, 2, \dots$ denote the observations
- Optimal detection from an in-control proportion π_0 to an out-of-control π_1 by sequentially computing

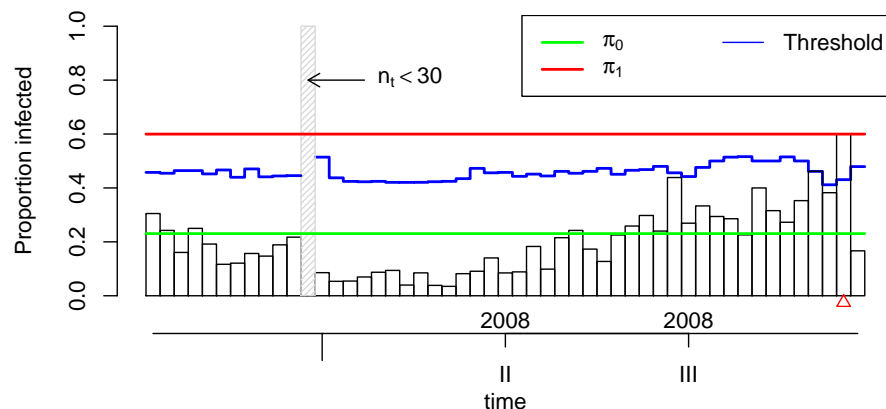
$$C_t = \max(0, C_{t-1} + y_t - n_t k), \quad t = 1, 2, \dots,$$

$$\text{with } C_0 = 0 \text{ and } k = \log \left(\frac{\pi_1(1 - \pi_0)}{\pi_0(1 - \pi_1)} \right) - \log \left(\frac{1 - \pi_1}{1 - \pi_0} \right).$$

- An alarm is sounded the first time where $C_t > h$, and h is a known threshold determining the properties of the detector.
- Given h , one can compute the average time until the first false alarm (ARL_0) using e.g. the algorithm of Hawkins (1992).

Binomial CUSUM (2)

- Detection in the picorna time series for a change from $\pi_0 = 0.23$ to $\pi_1 = 0.60$ corresponding to $OR(\pi_1, \pi_0) = 5$.



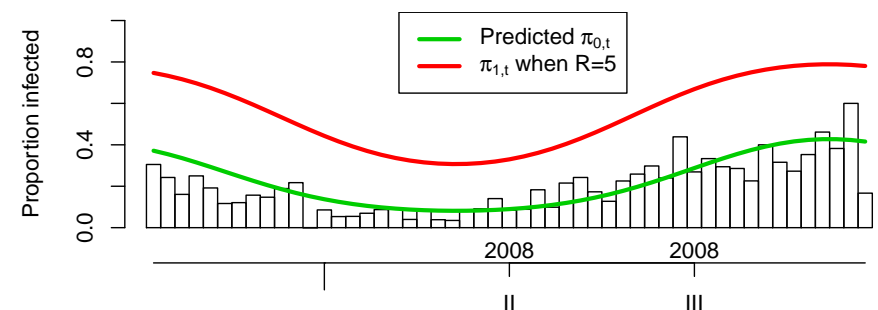
- CUSUM begins monitoring in week 41/2007 and is prospective, i.e. only information up to the time point is used.

Time varying proportion Binomial CUSUM (1)

- Time varying proportion in a logistic regression context

$$\text{logit}(\pi_{0,t}) = \beta_0 + \beta_1 \cdot t + \beta_2 \cos \left(\frac{2\pi}{52} \cdot t \right) + \beta_3 \sin \left(\frac{2\pi}{52} \cdot t \right)$$

- Estimate β from past and predict $\pi_{0,t}$ for future time points.



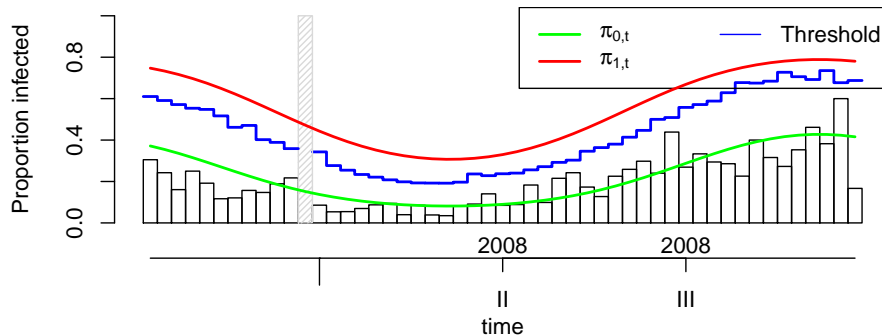
- Develop optimal detector for a change from odds $\frac{\pi_{0,t}}{1 - \pi_{0,t}}$ to odds $R \cdot \frac{\pi_{0,t}}{1 - \pi_{0,t}}$ similar to Steiner et al. (2000).

Time varying proportion Binomial CUSUM (2)

- New: Reweight CUSUM contributions in order to maintain a fixed average time until first false alarm ARL_0 :

$$C_t = \max \left\{ 0, C_{t-1} + \frac{h}{h_t} (y_t - n_t k_t) \right\},$$

where h_t is computed as the threshold giving the desired ARL_0 in a setup with $\pi_{0,t}$ and $\pi_{1,t}$.



Time varying proportion Binomial CUSUM (3)

```
R> phase1 <- 1:(52 * 3)
R> phase2 <- 1:nrow(oPic) %without% phase1
R> m.logit <- glm(cbind(observed, population - observed) ~
+ 1 + I(epoch - mean(epoch)) + I(sin(epoch * 2 * pi/freq)) +
+ I(cos(epoch * 2 * pi/freq)), family = binomial,
+ data = as.data.frame(oPic)[phase1, ])
R> theta0 <- matrix(predict(m.logit, newdata = data.frame(epoch = phase2,
+ freq = 52), type = "response"), ncol = 1)
R> R <- 5
R> theta1 <- R * theta0 / (1 - theta0 + R * theta0)
R> control <- list(range = phase2, distribution = "binomial",
+ ARLO = 10 * 52, digits = 1, s = R, theta0t = theta0,
+ limit = 0)
R> s.binomCUSUM <- rogerson(oPic, control = control)
```

Outline

- 1 Introduction
- 2 The R package `surveillance`
- 3 Univariate time series detectors
 - Farrington algorithm
 - Negative Binomial CUSUM
 - Binomial CUSUM
 - Evaluating performance
 - Likelihood ratio detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling

Evaluating the performance of a surveillance algorithm

Choice of threshold in surveillance algorithms should be based on performance measure:

- Location parameters of the run length distribution, e.g. the ARLs $E(T_A | \tau = 0)$ or $E(T_A | \tau = \infty)$
- Conditional expected delay $E(T_A - \tau | \tau, T_A \geq \tau)$
- Probability of false alarm within first m time points, i.e. $P(T_A \leq m | \tau = \infty)$
- Sensitivity, Specificity, ROC-Curves

Computation of measures rarely available as closed formulas. Instead Monte-Carlo sampling is used.

Run-length of CUSUM detectors

- Among all procedures with the same in-control ARL, the CUSUM has the smallest expected time until it signals a change in the case, where the process shifts to the out-of-control state (Moustakides, 1986).
- In practice no single out-of-control state exists. Thus we select a state where we want detection to be optimal and count on a robust performance in case of another shift.
- For further details see e.g. Hawkins and Olwell (1998) or Frisén (2003)

Run-length of NegBin CUSUM (1)

- Interest is in the PMF of T_A . Compute this either by Monte Carlo simulation or by using a Markov chain approximation.
- Generalization of Bissell (1984) to time varying count data CUSUMs: dynamics of r_t described by a Markov chain:

$$\begin{aligned} \text{State } 0 & r_t = 0 \\ \text{State } i & r_t \in \left((i-1) \cdot \frac{g}{M}, i \cdot \frac{g}{M} \right], i = 1, 2, \dots, M \\ \text{State } M+1 & r_t > g \end{aligned}$$

- Calculation of the $(M+2) \times (M+2)$ transition matrix \mathbf{P}_t with elements

$$p_{t,i,j} = P(r_t \in \text{State } j | r_{t-1} \in \text{State } i), i, j = 0, 1, \dots, M+1$$

by approximations suggested in Hawkins and Olwell (1998)

Run-length of NegBin CUSUM (2)

- State $M+1$ is absorbing.
- The cumulative probability of an alarm at any step up to time n , $n \geq 1$, is:

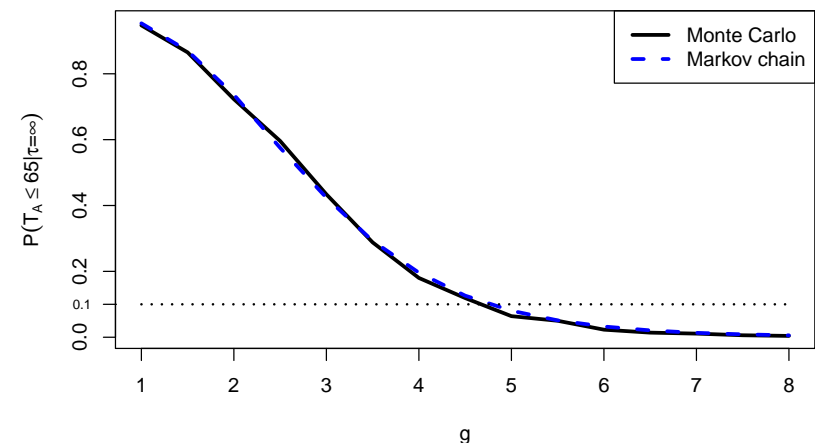
$$P(T_A \leq n) = \left[\prod_{t=1}^n \mathbf{P}_t \right]_{0, M+1}$$

- The PMF of T_A can thus be determined by subtraction
- Now: Choose g such that $P(T_A \leq 65 | \tau = \infty)$ is below some acceptable value, e.g. 10%.

```
R> pMarkovChain <- sapply(g.grid, function(g) {
+   TA <- LRCUSUM.runlength(mu = t(mu0), mu0 = t(mu0),
+     mu1 = kappa * t(mu0), h = g, dfun = dY, n = rep(600,
+     length(mu0)), alpha = 1/m$theta)
+   return(tail(TA$cdf, n = 1))
+ })
```

Run-length of NegBin CUSUM (3)

- $P(T_A \leq 65 | \tau = \infty)$ as a function of g – computed by both Monte Carlo simulation and the Markov chain approximation ($M = 5$).



- The Markov chain approximation is 6.8 times faster than Monte Carlo based on 1000 samples.

Comparison with the Farrington algorithm

- Fitted negative binomial model with mean $\mu_{0,t}$ and dispersion α_t , matching the quasi-Poisson model, as true model.
- Based on 1000 realizations of $I(T_A \leq 65 | \tau = \infty)$ for the Farrington et al. (1996) algorithm with $\frac{2}{3}$ -power transform, $b = 5$, $w = 4$ and $\alpha = 0.001$, we obtain

$$P(T_A \leq 65 | \tau = \infty) \approx 0.19.$$

- A rough estimate of this number would have been

$$1 - \left(1 - \frac{\alpha}{2}\right)^{65} = 0.03.$$

- Note: Using `farrington` without reweighting and always including a trend, we obtain the Monte Carlo estimate 0.04.

Outline

- 1 Introduction
- 2 The R package `surveillance`
- 3 Univariate time series detectors
 - Farrington algorithm
 - Negative Binomial CUSUM
 - Binomial CUSUM
 - Evaluating performance
 - Likelihood ratio detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling

Generalized likelihood ratio detector (1)

- A problem of the LR scheme is that detection is only optimal for pre-specified θ_1 .
- Generalization where θ_1 is estimated for each instance:

Generalized likelihood ratio (GLR) based stopping rule

$$T_A = \inf \left\{ s \geq 1 : \max_{1 \leq k \leq s} \sup_{\theta_1 \in \Theta_1} \left[\sum_{t=k}^s \log \left\{ \frac{f_{\theta_1}(y_t | z_t)}{f_{\theta_0}(y_t | z_t)} \right\} \right] \geq c_\gamma \right\}$$

- No recursive updating as in LR-CUSUM possible: worst case number of operations to determine if $T_A \leq m$ is $O(m^3)$
- Lai and Shan (1999) show for the Gaussian case how it is possible to reduce this complexity by recursive least squares and clever treatment of the sums and sups

Generalized likelihood ratio detector (2)

The GLR detector rephrased:

$$\begin{aligned} I_{s,k} &= \sup_{\theta_1 \in \Theta_1} \left[\sum_{t=k}^s \log \left\{ \frac{f_{\theta_1}(y_t | z_t)}{f_{\theta_0}(y_t | z_t)} \right\} \right] \\ &= \left[\sup_{\theta_1 \in \Theta_1} \sum_{t=k}^s \log f_{\theta_1}(y_t | z_t) \right] - \left[\sum_{t=k}^s \log f_{\theta_0}(y_t | z_t) \right] \\ &= \sum_{t=k}^s \log \left\{ \frac{f_{\hat{\theta}_{s,k}}(y_t | z_t)}{f_{\theta_0}(y_t | z_t)} \right\}, \end{aligned}$$

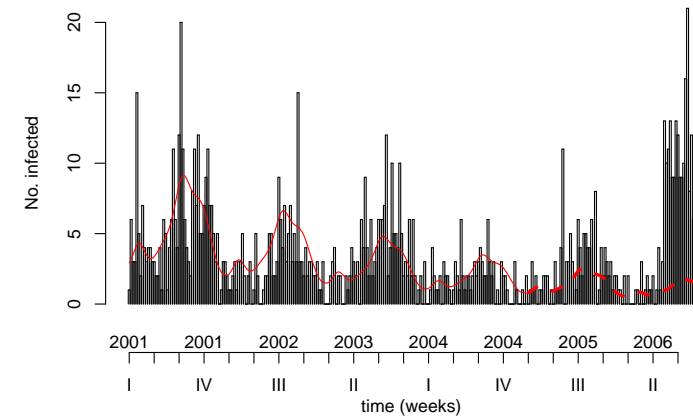
where $\hat{\theta}_{s,k} = \arg \sup_{\theta_1 \in \Theta_1} \sum_{t=k}^s \log f_{\theta_1}(y_t | z_t)$. Now $GLR(s) = \max_{1 \leq k \leq s} I_{s,k}$

GLR detector (3) – Poisson and negative Binomial

- For the Poisson case with $\log \mu_{1,t} = \log \mu_{0,t} + \kappa$, efficient computations are possible since an efficient computation of $\hat{\kappa}_{s,k}$ and $I_{s,k}$ is available.
- For the NegBin case with $\log \mu_{1,t} = \log \mu_{0,t} + \kappa$ the MLE $\hat{\kappa}_{s,k}$ has to be found by iterative methods
- Speedup the GLR detector by using a *window-limited* approach as proposed by Willsky and Jones (1976). Maximization only for a moving window of $k \in \{s - M, \dots, s\}$, where $M \geq 1$
- For details about the GLR detector see Höhle and Paul (2008)

Applying the GLR detector to salmonella hadar (1)

- A seasonal negative binomial GLM is fitted to the training period.



- The fitted model is used to predict $\mu_{0,t}$ of the test period.

Applying the GLR detector to salmonella hadar (2)

Predicting $\mu_{0,t}$ using mgcv:

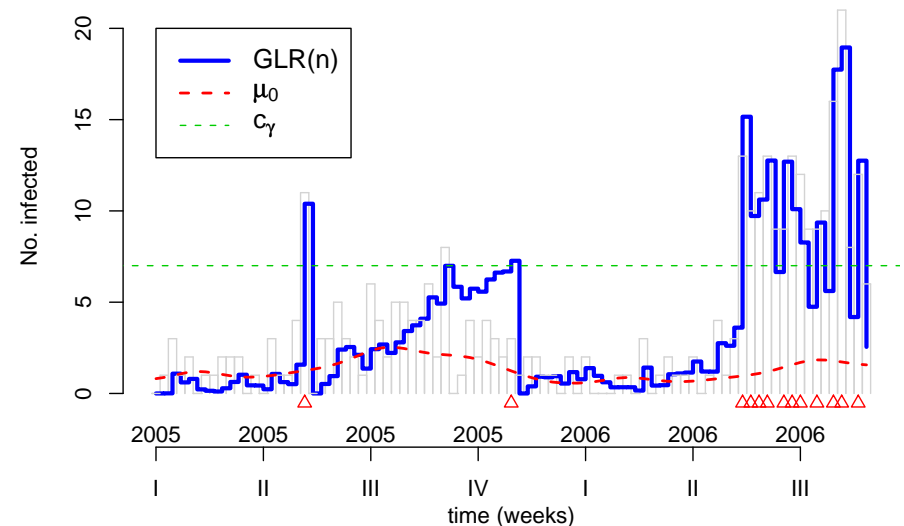
```
R> train <- 1:(4 * 52)
R> test <- (max(train) + 1):nrow(shadar)
R> m.hadar <- gam(observed ~ 1 + epoch + s(epoch%%52, bs = "cc",
+   fx = FALSE), family = negbin(theta = c(0.1, 1/0.2 *
+   2)), data = as.data.frame(shadar[train, ]))
R> alpha.hat <- 1/m.hadar$family$getTheta()
R> mu0.hat <- predict(m.hadar, newdata = data.frame(epoch = test),
+   type = "response")
```

Running the detector:

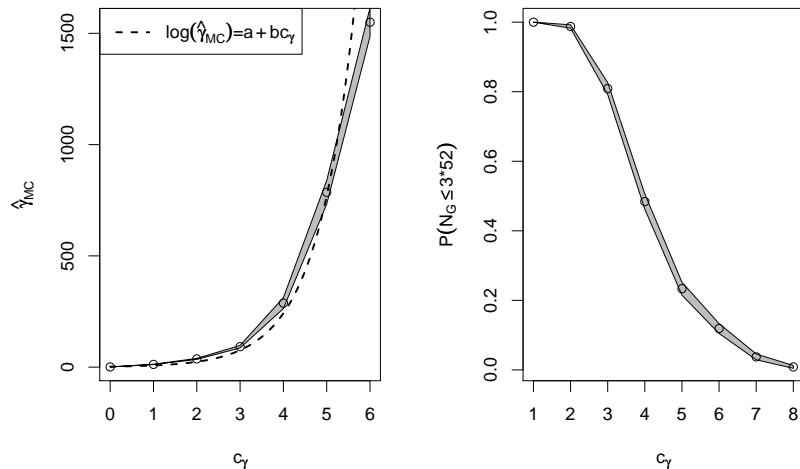
```
R> cntrl = list(range = test, mu0 = mu0.hat, alpha = alpha.hat,
+   c.ARL = 7, Mtilde = 1, change = "intercept")
R> shadar.surv <- glrnb(shadar, control = cntrl)
```

Applying the GLR detector to salmonella hadar (3)

Analysis of shadar using glrnb: intercept



Average run length and probability of false alarm

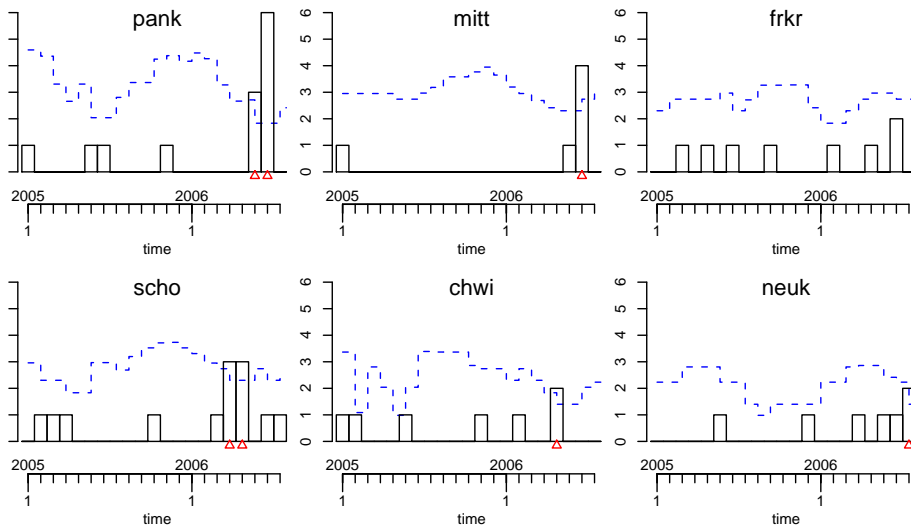
(a) Monte-Carlo estimated $ARL_0(c_\gamma)$ (b) $P(T_A \leq 3 \cdot 52 | \tau = \infty)$

Outline

- 1 Introduction
- 2 The R package surveillance
- 3 Univariate time series detectors
- 4 **Multivariate surveillance**
 - Case Study: Rabies in Hesse
 - The HHH model and its spatial extensions
- 5 Space-Time Point Process Modelling
- 6 Discussion and Summary

Towards multivariate surveillance (1)

- A simple way to perform surveillance for a number of time series is to monitor each independently



Towards multivariate surveillance (2)

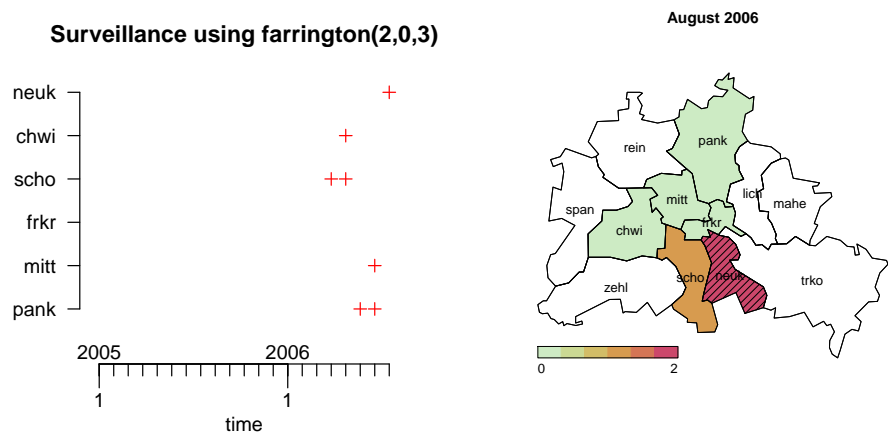
- Results for current month (say August 2006) are easily accessed for further report generation

```
R> control <- list(b = 3, w = 2, range = 53:73, alpha = 0.01,
+   limit54 = c(0, 1))
R> ha4.surv <- farrington(ha4, control = control)
R> sapply(c("observed", "upperbound", "alarm"), function(str) {
+   slot(ha4.surv, str)[nrow(ha4.surv), ]
+ })
```

	observed	upperbound	alarm
pank	0	2.42	0
mitt	0	2.97	0
frkr	0	2.74	0
scho	1	2.42	0
chwi	0	2.23	0
neuk	2	1.40	1

Towards multivariate surveillance (3)

- An alarm plot gives an overview of alarms for the different time series
- Shaded regions indicate alarms for the current month

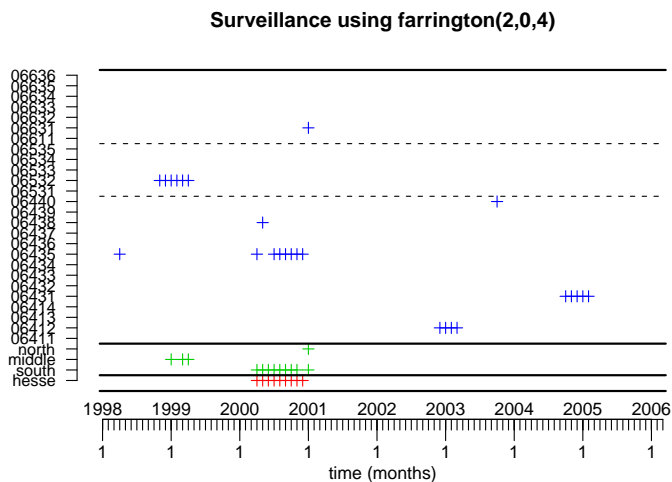


Outline

- 1 Introduction
- 2 The R package surveillance
- 3 Univariate time series detectors
- 4 **Multivariate surveillance**
 - Case Study: Rabies in Hesse
 - The HHH model and its spatial extensions
- 5 Space-Time Point Process Modelling
- 6 Discussion and Summary

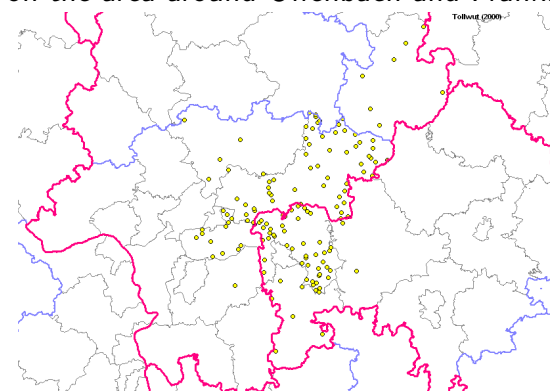
Rabies surveillance in Hesse

- Alarm plot created by applying the Farrington algorithm to each of 1 federal state, 3 administrative regions and 26 districts time series



Examination of the increased number of cases (1)

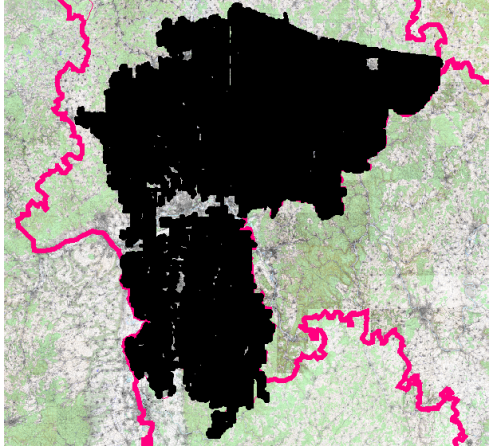
- An inspection of the cases in year 2000 showed that problems centered on the area around Offenbach and Frankfurt.



- Source of the figure: C. Staubach, FLI

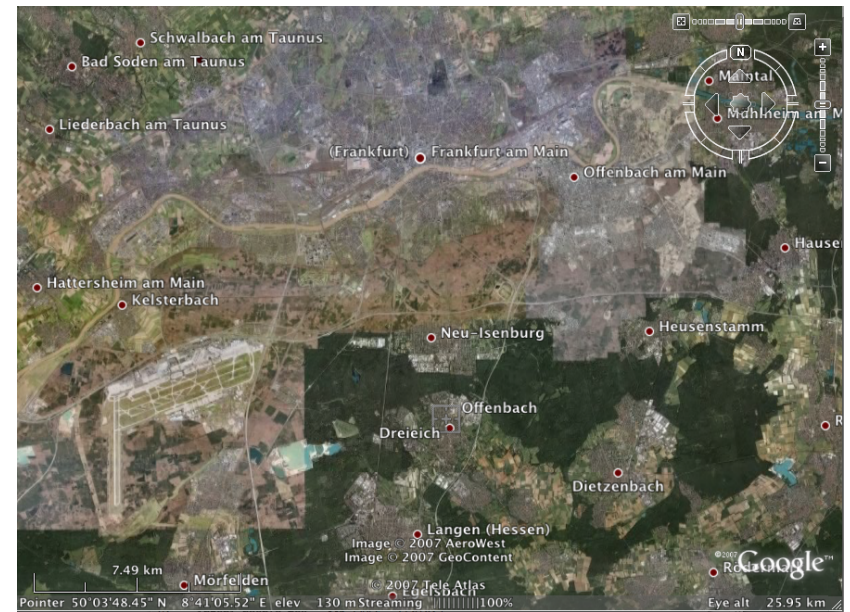
Examination of the increased number of cases (2)

- A map with the coordinates of the baits with vaccine dropped from plane shows the problem:



- Source of the figure: T. Müller, FLI

Examination of the increased number of cases (3)



Outline

- 1 Introduction
- 2 The R package *surveillance*
- 3 Univariate time series detectors
- 4 Multivariate surveillance
 - Case Study: Rabies in Hesse
 - The HHH model and its spatial extensions
- 5 Space-Time Point Process Modelling
- 6 Discussion and Summary

Model-based surveillance¹

So far the philosophy has been

- Use of a simple statistical model to describe the incidence, e.g. using a Poisson GLM
- No modelling of epidemic behaviour
- Comparison of observed cases with expected cases for the current time point
- Attempt to *detect* outbreaks instead of *predicting* them
- Implicit assumption that *no outbreak* has happened in the past (except the ad-hoc adjustment in Farrington et al. (1996))

¹Slides 80–90 and 92–117 are slightly revised versions of work kindly provided by L. Held and M. Paul, respectively

The HHH model (1)

- Approach in Held et al. (2005) and (Paul et al., 2008): Development of a *realistic* stochastic model for the statistical analysis of surveillance data of infectious disease counts
- Compromise between *mechanistic* and *empirical* modelling
- Model is based on a generalized *branching process* with immigration
- Note: Branching process is a useful approximation of SIR-models in the absence of information on susceptibles
- Explicit decomposition of the incidence in *endemic* and *epidemic* component
- Past counts act *additively* on disease incidence → model is not a GLM

The HHH model (2)

- For $t = 1, 2, \dots$ we have $y_t \sim \text{Po}(\mu_t)$, where

$$\begin{aligned}\mu_t &= \nu_t + \lambda y_{t-1} \\ \log(\nu_t) &= \alpha + \sum_{s=1}^S (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t))\end{aligned}$$

- Autoregressive coefficient $0 < \lambda < 1$ determines stationarity of y_t , can be interpreted as *epidemic proportion*
- $\log \nu_t$ is modelled parametrically as in log-linear Poisson regression; includes terms for *seasonality*
- Adjustments for *overdispersion* straightforward: Replace $\text{Po}(\mu_t)$ by $\text{NegBin}(\mu_t, \psi)$ -Likelihood
- Model can be fitted by Maximum-Likelihood using function `algo.hhh` in *surveillance*

Multivariate HHH modelling

- Suppose now *multiple* time series $i = 1, \dots, n$ are available over the same time horizon $t = 1, \dots, T$
- Notation: $y_{i,t}$ is the number of disease cases observed in the i -th time series at time t
- Examples:
 - ▶ Incidence in *different age groups*
 - ▶ Incidence of *related diseases*
 - ▶ Incidence in *different geographical regions*
- Idea: Include now also the number of counts from other time series as autoregressive covariates → *multi-type branching process*

Bivariate modelling

Joint analysis of two time series $i = 1, 2$

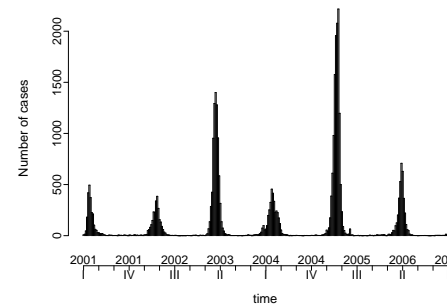
$$\begin{aligned}y_{i,t} &\sim \text{NegBin}(\mu_{i,t}, \psi) \\ \mu_{i,t} &= \nu_t + \lambda y_{i,t-1} + \phi y_{j,t-1} \quad \text{where } j \neq i\end{aligned}$$

Note: ψ , ν_t , λ and ϕ may also depend on i

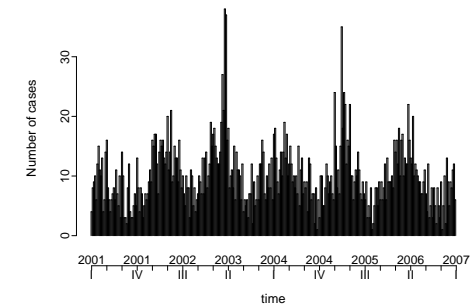
Example: Influenza and meningococcal disease (1)

- Interdependencies between disease cases caused by *different pathogens* might be of particular interest to further understand the dynamics of such diseases
- For example, several studies describe an association between *influenza* and *meningococcal disease* (Cartwright et al., 1991; Hubert et al., 1992; Makras et al., 2001; Jensen et al., 2004)
- Analysis of routinely collected surveillance data from Germany, 2001-2006, from SurvStat@RKI (Robert Koch Institute, 2009)

Example: Influenza and meningococcal disease (2) – Data



(a) influenza



(b) meningococci

Univariate analysis of influenza infections

- Results from analysing the influenza time series with HHH models using the Poisson, Negative Binomial and an increasing number of seasonal components

S	$\hat{\lambda}_{ML}$ (se)	$\hat{\psi}_{ML}$ (se)	$\log L(\mathbf{y}, \theta)$	$ \theta $	AIC
0	0.99 (0.01)	-	-4050.9	2	8105.9
0	0.98 (0.05)	2.41 (0.27)	-1080.2	3	2166.5
1	0.86 (0.05)	2.74 (0.31)	-1064.1	5	2138.2
2	0.76 (0.05)	3.12 (0.37)	-1053.3	7	2120.6
3	0.74 (0.05)	3.39 (0.41)	-1044.1	9	2106.3
4	0.74 (0.05)	3.44 (0.42)	-1042.2	11	2106.3

Univariate analysis of meningococcal infections

- Results from analysing the meningococcal time series with HHH models using the Poisson, Negative Binomial and a increasing number of seasonal components

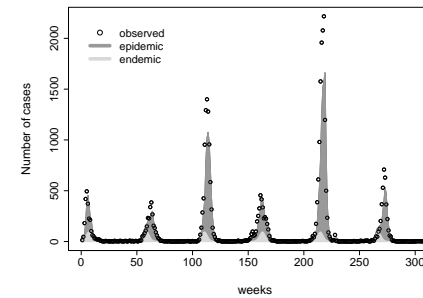
S	$\hat{\lambda}_{ML}$ (se)	$\hat{\psi}_{ML}$ (se)	$\log L(\mathbf{y}, \theta)$	$ \theta $	AIC
0	0.50 (0.04)	-	-919.2	2	1842.4
0	0.48 (0.05)	11.80 (2.09)	-880.5	3	1767.0
1	0.16 (0.06)	20.34 (4.83)	-845.6	5	1701.2
2	0.16 (0.06)	20.41 (4.86)	-845.5	7	1705.0

Multivariate analyses

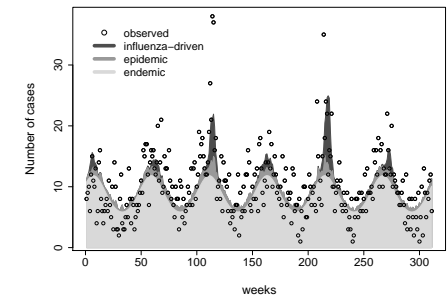
Model	S		$\hat{\lambda}_{ML}$ (se)		$\hat{\phi}_{ML}$ (se)	
	flu	men	flu	men	flu	men
1	3	1	0.74 (0.05)	0.16 (0.06)	-	-
2	3	1	0.74 (0.05)	0.16 (0.06)	0.000 (0.000)	-
3	3	1	0.74 (0.05)	0.10 (0.06)	-	0.005 (0.001)
4	3	1	0.74 (0.05)	0.10 (0.06)	0.000 (0.000)	0.005 (0.001)

Model	$\hat{\psi}_{ML}$ (se)		$\log L(\mathbf{y}, \theta)$	$ \theta $	AIC
	flu	men			
1	3.39 (0.41)	20.34 (4.83)	-1889.7	14	3807.5
2	3.39 (0.41)	20.34 (4.83)	-1889.7	15	3809.5
3	3.39 (0.41)	25.32 (6.98)	-1881.0	15	3791.9
4	3.40 (0.41)	25.32 (6.98)	-1881.0	16	3793.9

Fitted time series



(a) influenza



(b) meningococci

Figure: Results from a multivariate analysis influenza and meningococcal infections in Germany, 01/2001 – 52/2006 using HHH

HHH in surveillance

```
R> # weekly counts of influenza and meningococcal infections
R> # in Germany, 2001-2006
R> data("influMen")
R> # specify model with two autoregressive parameters lambda_i, overdispersion
R> # parameters psi_i, an autoregressive parameter phi for meningococcal infections
R> # (i.e. nu_flu,t = lambda_flu * y_flu,t-1
R> # and nu_men,t = lambda_men * y_men,t-1 + phi_men*y_flu,t-1 )
R> # and S=(3,1) Fourier frequencies
R> model <- list(lambda=c(TRUE,TRUE), neighbours=c(FALSE,TRUE),
+               linear=FALSE,nseason=c(3,1),negbin="multiple")
R> #Fit the model
R> res.hhh <- algo.hhh(influMen, control=model)
```

Algorithm claims to have converged

```
R> AIC(res.hhh)
```

```
[1] 3791.938
```

Model formulation

Suppose now multiple time series are available:

y_{rt} number of cases in unit $r = 1, \dots, R$ at time $t = 1, \dots, T$

$$y_{rt} | \mathbf{y}_{t-1} \sim \text{NegBin}(\mu_{rt}, \psi) \quad (\psi > 0)$$

$$\mu_{rt} = \nu_{rt} + \lambda y_{r,t-1} \quad (\nu_{rt}, \lambda > 0)$$

Model formulation

Suppose now multiple time series are available:

y_{rt} number of cases in unit $r = 1, \dots, R$ at time $t = 1, \dots, T$

$$y_{rt} | \mathbf{y}_{t-1} \sim \text{NegBin}(\mu_{rt}, \psi) \quad (\psi > 0)$$

$$\mu_{rt} = \nu_{rt} + \lambda y_{r,t-1} \quad (\nu_{rt}, \lambda > 0)$$

The unknown quantities are given e.g. by

- $\log(\nu_{rt}) = \log(e_{rt}) + \alpha_0 + \alpha_1 \sin\left(\frac{2\pi}{52}t\right) + \alpha_2 \cos\left(\frac{2\pi}{52}t\right)$

e_{rt} : offset, e.g. population numbers

Model formulation

Suppose now multiple time series are available:

y_{rt} number of cases in unit $r = 1, \dots, R$ at time $t = 1, \dots, T$

$$y_{rt} | \mathbf{y}_{t-1} \sim \text{NegBin}(\mu_{rt}, \psi) \quad (\psi > 0)$$

$$\mu_{rt} = \nu_{rt} + \lambda y_{r,t-1} \quad (\nu_{rt}, \lambda > 0)$$

The unknown quantities are given e.g. by

- $\log(\nu_{rt}) = \log(e_{rt}) + \alpha_0 + \alpha_1 \sin\left(\frac{2\pi}{52}t\right) + \alpha_2 \cos\left(\frac{2\pi}{52}t\right)$

e_{rt} : offset, e.g. population numbers

- $\log(\lambda) = \beta_0$

Model formulation

Suppose now multiple time series are available:

y_{rt} number of cases in unit $r = 1, \dots, R$ at time $t = 1, \dots, T$

$$y_{rt} | \mathbf{y}_{t-1} \sim \text{NegBin}(\mu_{rt}, \psi) \quad (\psi > 0)$$

$$\mu_{rt} = \nu_{rt} + \lambda y_{r,t-1} + \phi \sum_{q \neq r} w_{qr} y_{q,t-1} \quad (\nu_{rt}, \lambda, \phi > 0)$$

The unknown quantities are given e.g. by

- $\log(\nu_{rt}) = \log(e_{rt}) + \alpha_0 + \alpha_1 \sin\left(\frac{2\pi}{52}t\right) + \alpha_2 \cos\left(\frac{2\pi}{52}t\right)$

e_{rt} : offset, e.g. population numbers

- $\log(\lambda) = \beta_0$

- neighbor-driven component:** $\log(\phi) = \gamma_0$

w_{qr} : known weights, e.g. $\mathbb{1}(q \sim r)$, travel intensities

Addressing unit-specific heterogeneity

Each of the three unknown quantities ν, λ, ϕ , may also depend on unit r by using

- unit-specific fixed effects:

$$\log(\phi_r) = \gamma_r$$

\rightsquigarrow this allows us to explore interdependencies between different pathogens (e.g. influenza and meningococcal disease)

Addressing unit-specific heterogeneity

Each of the three unknown quantities ν , λ , ϕ , may also depend on unit r by using

- unit-specific fixed effects:
 $\log(\phi_r) = \gamma_r$
 \rightsquigarrow this allows us to explore interdependencies between different pathogens (e.g. influenza and meningococcal disease)
- linking parameters with known explanatory variables:
 $\log(\lambda_{rt}) = \beta_0 + x_{rt}\beta_1$
 \rightsquigarrow for instance x_{rt} = vaccination coverage in unit r at time t .

Addressing unit-specific heterogeneity

Each of the three unknown quantities ν , λ , ϕ , may also depend on unit r by using

- unit-specific fixed effects:
 $\log(\phi_r) = \gamma_r$
 \rightsquigarrow this allows us to explore interdependencies between different pathogens (e.g. influenza and meningococcal disease)
- linking parameters with known explanatory variables:
 $\log(\lambda_{rt}) = \beta_0 + x_{rt}\beta_1$
 \rightsquigarrow for instance x_{rt} = vaccination coverage in unit r at time t .
- unit-specific random effects:
 $\log(\nu_r) = \alpha_0 + a_r$, $a_r \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma_\nu^2)$, $r = 1, \dots, R$

Random effects specification

Consider the model

$$\mu_{rt} = \nu_{rt} + \phi_r \sum_{q \neq r} w_{qr} y_{q,t-1}$$

- $\log(\nu_{rt}) = \alpha_0 + a_r + (\text{season}) + \dots$
- $\log(\phi_r) = \gamma_0 + c_r$

where the random effects $\mathbf{a} = (a_1, \dots, a_R)^\top$ and $\mathbf{c} = (c_1, \dots, c_R)^\top$ are assumed to be

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{c} \end{pmatrix} \sim \text{N} \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_\nu^2 & \\ & \sigma_\phi^2 \end{pmatrix} \otimes \mathbf{I}_R \right)$$

Random effects specification

Consider the model

$$\mu_{rt} = \nu_{rt} + \phi_r \sum_{q \neq r} w_{qr} y_{q,t-1}$$

- $\log(\nu_{rt}) = \alpha_0 + a_r + (\text{season}) + \dots$
- $\log(\phi_r) = \gamma_0 + c_r$

where the random effects $\mathbf{a} = (a_1, \dots, a_R)^\top$ and $\mathbf{c} = (c_1, \dots, c_R)^\top$ are assumed to be

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{c} \end{pmatrix} \sim \text{N} \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_\nu^2 & \rho\sigma_\nu\sigma_\phi \\ \rho\sigma_\nu\sigma_\phi & \sigma_\phi^2 \end{pmatrix} \otimes \mathbf{I}_R \right)$$

Random effects specification

Consider the model

$$\mu_{rt} = \nu_{rt} + \phi_r \sum_{q \neq r} w_{qr} y_{q,t-1}$$

- $\log(\nu_{rt}) = \alpha_0 + \mathbf{a}_r + (\text{season}) + \dots$
- $\log(\phi_r) = \gamma_0 + \mathbf{c}_r$

where the random effects $\mathbf{a} = (a_1, \dots, a_R)^\top$ and $\mathbf{c} = (c_1, \dots, c_R)^\top$ are assumed to be

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{c} \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_\nu^2 & \rho\sigma_\nu\sigma_\phi \\ \rho\sigma_\nu\sigma_\phi & \sigma_\phi^2 \end{pmatrix} \otimes \mathbf{I}_R \right)$$

Alternatively, a conditional autoregressive (CAR) model (Besag et al., 1991) may be adopted for \mathbf{a} , say.

Estimation

- Model does not belong to the class of GL(M)Ms
- Fixed effects model: maximum likelihood estimates are obtained via a (globally convergent) Newton Raphson type algorithm.
- Random effects model: estimation involves a multidimensional integral without closed form solution.

Estimation – random effects model

We adopt a penalized likelihood approach (Breslow and Clayton (1993); Kneib and Fahrmeir (2007)) with alternating steps:

- 1 Estimate regression parameters for given variance components.
- 2 Estimate variance components for given regression parameters based on an approximate marginal likelihood (using a first order Laplace approximation).

Note: CAR effects require reparameterization

Estimation – random effects model

We adopt a penalized likelihood approach (Breslow and Clayton (1993); Kneib and Fahrmeir (2007)) with alternating steps:

- 1 Estimate regression parameters for given variance components.
- 2 Estimate variance components for given regression parameters based on an approximate marginal likelihood (using a first order Laplace approximation).

Note: CAR effects require reparameterization

All methods are incorporated in `surveillance` as function `hhh4`.

Model choice

- Classical model choice criteria such as AIC can be problematic in the presence of random effects.
- Models are validated based on probabilistic **one-step-ahead predictions**.
- The often used mean squared prediction error does not incorporate prediction uncertainty.
- We use **strictly proper scoring rules** (Gneiting and Raftery (2007); Czado et al. (2009))
 - ▶ evaluate a model based on the predictive distribution and the later observed true value
 - ▶ simultaneously address sharpness and calibration
 - ▶ are negatively oriented (the smaller the better)

Intermezzo: Scoring rules (1)

- A scoring rule $S(P, y)$ measures the predictive quality of a stated predictive distribution P by comparing it with the actual observed value y
- Denote the expectation of $S(P, \cdot)$ under distribution Q by $S(P, Q)$. A scoring rule is called *proper* if $S(P, Q)$ is minimal if y is indeed a realization from P . If the minimum is unique the scoring rule is called *strictly proper*.
- In practice scores are reported as averages over suitable sets of forecasts

$$\bar{S} = \frac{1}{n} \sum_{i=1}^n S(P^{(i)}, y^{(i)}),$$

where $P^{(i)}$ and $y^{(i)}$ refer to the i 'th predictive distribution and i 'th observation, respectively

Intermezzo: Scoring rules (2)

- The most popular strictly proper scoring rule for count data is the *logarithmic score*

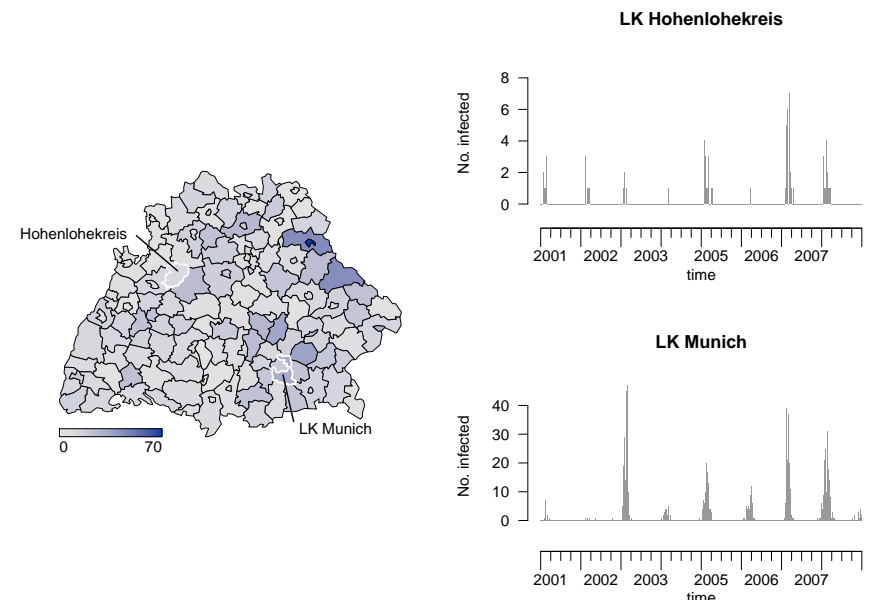
$$\log S(P, y) = -\log(f_P(Y = y)),$$

where $f_P(Y = y)$ is the PMF of the predictive distribution P .

- To compare two models A and B compute n individual scores for both models and use a Monte Carlo test to assess if difference

$$\Delta_{A,B} = \bar{S}_A - \bar{S}_B$$

is significant.



Number of laboratory confirmed influenza A and B cases 2001–2008 in 140 administrative districts in Southern Germany (RKI, 2009)

Case study: Influenza in Southern Germany

- We considered several negative binomial models, which differ depending on whether and how the autoregression is specified.
- The endemic components always includes
 - ▶ population fractions as offset
 - ▶ linear trend and seasonal terms
 - ▶ iid random intercepts
- Model choice using the logarithmic score
 - ▶ one-step-ahead predictions for the last two years
 - ▶ average scores are based on these predictions
 - ▶ differences in mean scores may be tested e.g. via a Monte Carlo permutation test

Results for model with constant λ and random ϕ

$$\mu_{rt} = \nu_{rt} + \lambda y_{r,t-1} + \phi_r \sum_{q \neq r} w_{qr} y_{q,t-1} \quad \text{with}$$

$$\log(\phi_r) = \gamma_0 + c_r \text{ and } \log(\nu_{rt}) = \alpha_0 + a_r + \dots$$

```
R> #Load influenza data in Baden-Wuerttemberg and Bavaria
R> data("flu-BYBW")
R> # specify components of the model and fit it using hhh4
R> phi <- ~ -1 + ri(type = "iid", corr = "all")
R> nu <- addSeason2formula(~ -1+ri(type = "iid",corr = "all")+I((t-208)/100),S=3)
R> model <- list(end = list(f = nu, offset = population(sts.flu)),
+               ar = list(f = ~ 1),
+               ne = list(f = phi, weights = wji),
+               family = "NegBin1")
R> result <- hhh4(sts.flu, model)
```

Parameter estimates:

$\hat{\alpha}_0$ (se)	$\hat{\lambda}$ (se)	$\hat{\phi}$ (se)	$\hat{\sigma}_\nu^2$	$\hat{\sigma}_\phi^2$	$\hat{\rho}_{\nu\phi}$
0.22 (0.10)	0.41 (0.02)	0.22 (0.02)	0.51	0.96	0.56

One-step-ahead predictive validation for 2007–2008

```
> pred <- oneStepAhead(result, nrow(sts.flu) - 2*52)
> scores(pred)
```

One-step-ahead predictive validation for 2007–2008

```
> pred <- oneStepAhead(result, nrow(sts.flu) - 2*52)
> scores(pred)
```

autoregressive: λ	neighbor-driven: ϕ	$\overline{\log S}$
constant	random	.563
random	random	.564
random	constant	.565
constant	constant	.565
random	—	.569
constant	—	.569
—	random	.588
—	constant	.591
—	—	.599

One-step-ahead predictive validation for 2007–2008

```
> pred <- oneStepAhead(result, nrow(sts.flu) - 2*52)
> scores(pred)
```

autoregressive: λ	neighbor-driven: ϕ	$\overline{\log S}$	p -value
constant	random	.563	
random	random	.564	.5979
random	constant	.565	.0830
constant	constant	.565	.0353
random	—	.569	.0018
constant	—	.569	.0006
—	random	.588	.0001
—	constant	.591	.0001
—	—	.599	.0001

Monte Carlo p -values based on 9999 permutations

One-step-ahead predictive validation for 2007–2008

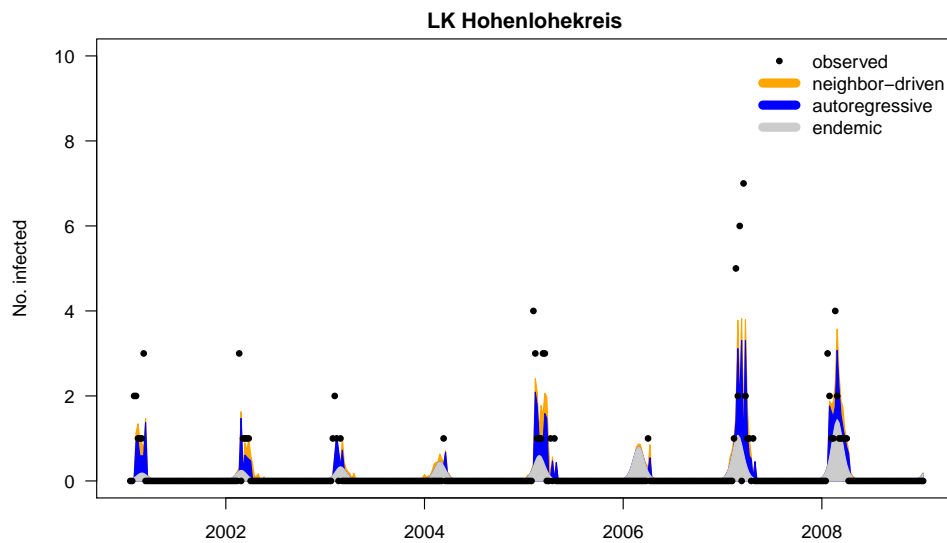
```
> pred <- oneStepAhead(result, nrow(sts.flu) - 2*52)
> scores(pred)
```

autoregressive: λ	neighbor-driven: ϕ	$\overline{\log S}$	p -value
constant	random	.563	
random	random	.564	.5979
random	constant	.565	.0830
constant	constant	.565	.0353
random	—	.569	.0018
constant	—	.569	.0006
—	random	.588	.0001
—	constant	.591	.0001
—	—	.599	.0001

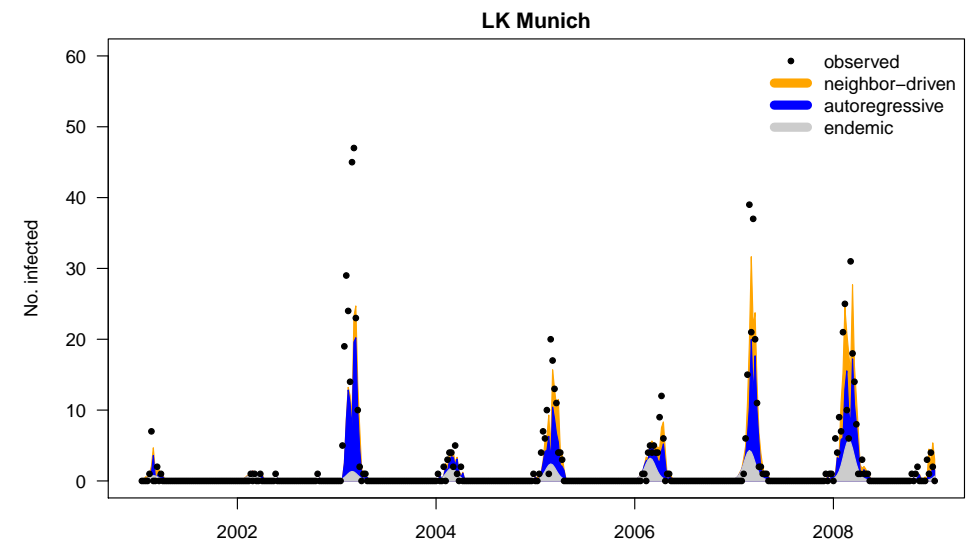
Monte Carlo p -values based on 9999 permutations

For comparison: $\overline{\log S} = 0.564$ for the best model with CAR instead of iid random effects in the endemic component ν_{rt} .

Fitted incidence



Fitted incidence



Summary

- A flexible modelling framework was developed to identify outbreaks and spatio-temporal patterns in infectious disease surveillance data.
- Different types of variation and correlation can be incorporated within a single model.
- Random effects formulation enables a realistic analysis of a large number of parallel time series.
- Methods are particularly well suited for model validation based on one-step-ahead predictions and strictly proper scoring rules.
- For further details see Paul and Held (2011).

Outline

- 1 Introduction
- 2 The R package *surveillance*
- 3 Univariate time series detectors
- 4 Multivariate surveillance
- 5 **Space-Time Point Process Modelling**
 - Maximum Likelihood Inference
 - Data Analysis
 - Prospective space-time monitoring
- 6 Discussion and Summary

Motivation and Aims (1)

- Public health *surveillance* of infectious diseases is an essential instrument in the attempt to control and prevent their spread
- Vast amounts of data resulting from routine surveillance demands the development of *automated algorithms* for the detection of *abnormalities*
- The spatial and temporal resolution of routine collected infectious disease data is becoming better and better
- Interest in developing models and aberration detection methods taking this spatio-temporal aspect better into account

Motivation and Aims (2)

Aim 1

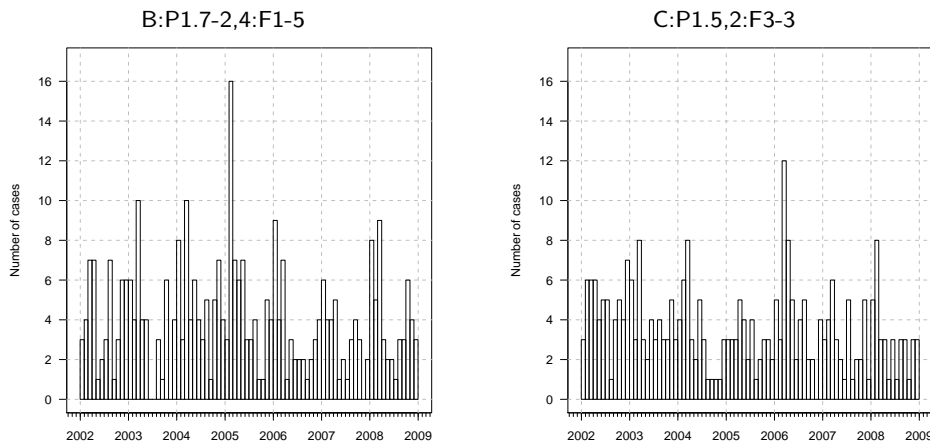
Establish a *regression framework* for point referenced infectious disease surveillance data, where the transmission dynamics and its dependency on covariates can be quantified within a *spatio-temporal stochastic process* context

Aim 2

Use this regression framework as building block for model based prospective space-time aberration detection, e.g. to detect disease clusters while adjusting for trend, seasonality and other covariates

Example: Invasive meningococcal disease (IMD)

- IMD is a life-threatening infectious disease triggered by the bacterium *Neisseria meningitidis* (aka *meningococcus*)
- Two most common finetypes in Germany in 2002–2008: 336 cases of *B:P1.7-2,4:F1-5*, 300 cases of *C:P1.5,2:F3-3*
- Case variables: date, residence postcode, age, gender



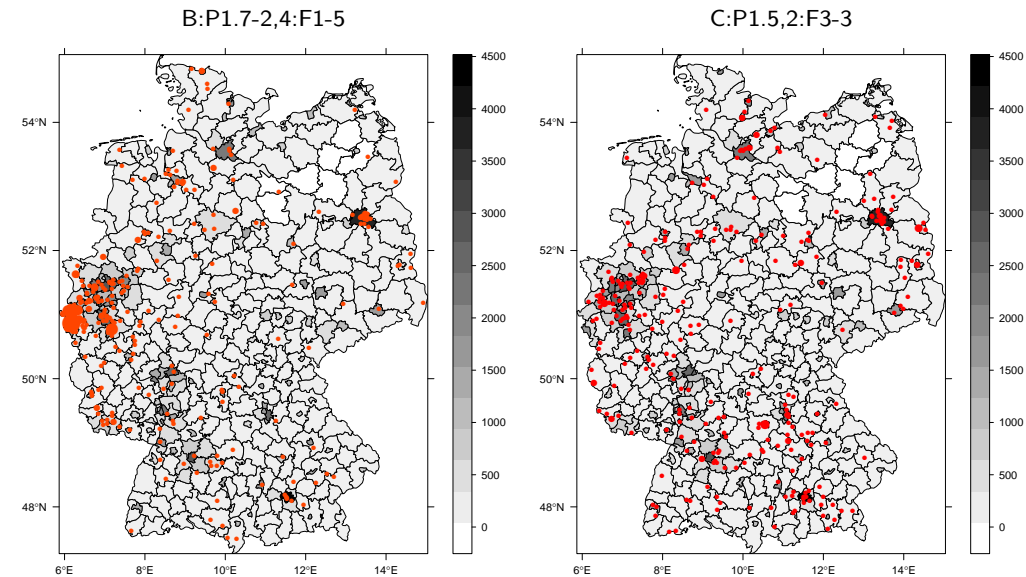
M. Höhle

Monitoring of infectious diseases

106 / 146

Point Process Modelling

Spatial distribution



Scientific question: Do the finetypes spread differently?

M. Höhle

Monitoring of infectious diseases

107 / 146

Point Process Modelling

Spatio-temporal animation

Conditional intensity function (CIF)

A regular spatio-temporal point process N on $\mathbb{R}_+ \times \mathbb{R}^2$ can be uniquely characterised by its left-continuous CIF $\lambda^*(t, \mathbf{s})$.

Definition

$$\lambda^*(t, \mathbf{s}) = \lim_{\Delta t \rightarrow 0, |\mathbf{ds}| \rightarrow 0} \frac{\mathbb{P}\left(N([t, t + \Delta t] \times \mathbf{ds}) = 1 \mid \mathcal{H}_{t-}\right)}{\Delta t |\mathbf{ds}|}$$

- Instantaneous event rate at (t, \mathbf{s}) given all past events
- Key to modelling, likelihood analysis and simulation of evolutionary point processes
- In application, N is only defined on a subset $(0, T] \times W \subset \mathbb{R}_+ \times \mathbb{R}^2$ (observation period and region)

B:P1.7-2,4:F1-5

C:P1.5,2:F3-3

M. Höhle

Monitoring of infectious diseases

108 / 146

M. Höhle

Monitoring of infectious diseases

109 / 146

Sources of inspiration (1)

Temporal self-exciting process (Hawkes, 1971)

$$\begin{aligned}\lambda^*(t) &= \psi + \int_{(-\infty, t)} g(t-u) dN(u) \\ &= \psi + \sum_{j:t_j < t} g(t-t_j)\end{aligned}$$

- Constant rate ψ of immigration independent of \mathcal{H}_{t-}
- Birth rate $g(t)$ for offspring events, e.g. exponential decay $g(t) = \alpha_0 e^{-\alpha_1 t}$
- Interpretation: branching process with immigration, cluster process (immigrants & offspring)

Sources of inspiration (2)

Spatio-temporal ETAS model (Ogata, 1998)

$$\lambda^*(t, \mathbf{s}) = \psi(\mathbf{s}) + \sum_{j:t_j < t} \underbrace{\kappa(m_j) g(t-t_j) f(\mathbf{s} - \mathbf{s}_j | m_j)}_{\text{"triggering function"}}$$

- $\psi(\mathbf{s})$ Inhomogeneous background seismicity rate
- $\kappa(m_j)$ Magnitude-dependent impact factor, e.g. $\kappa(m_j) = e^{\gamma m_j}$
- $g(t)$ Aftershock rate, e.g. hyperbolic decay $g(t) = K(t+c)^{-p}$
- $f(\mathbf{s}|m)$ Spatial kernel, e.g. elliptic bivariate normal density

Sources of inspiration (3)

Additive-multiplicative SIR compartmental model (Höhle, 2009)

$$\lambda_i^*(t) = Y_i(t) \cdot \left\{ h_i(t) + e_i^*(t) \right\} \quad (i = 1, \dots, n)$$

- Fixed, finite population with locations $\mathbf{s}_1, \dots, \mathbf{s}_n$
- At-risk indicator $Y_i(t)$
- Superposition of endemic (h) and epidemic (e) rates:
 - ▶ Multiple outbreaks initiated by "imported" cases

$$h_i(t) = \exp\left(h_0(t) + \mathbf{z}_i(t)' \boldsymbol{\beta}\right)$$

- ▶ Infectious ("self-exciting") character of the process based on the set $I^*(t)$ of current infectives, e.g.

$$e_i^*(t) = \sum_{j \in I^*(t)} f(\|\mathbf{s}_i - \mathbf{s}_j\|)$$

Additive-multiplicative continuous space-time intensity model proposed

$$\lambda^*(t, \mathbf{s}) = h(t, \mathbf{s}) + e^*(t, \mathbf{s})$$

Additive-multiplicative continuous space-time intensity model proposed

$$\lambda^*(t, \mathbf{s}) = h(t, \mathbf{s}) + e^*(t, \mathbf{s})$$

Multiplicative endemic component

$$h(t, \mathbf{s}) = \exp\left(o_{\xi(\mathbf{s})} + \beta' \mathbf{z}_{\tau(t), \xi(\mathbf{s})}\right)$$

- Piecewise constant function on a spatio-temporal grid $\{B_1, \dots, B_D\} \times \{A_1, \dots, A_M\}$ with time interval index $\tau(t)$, region index $\xi(\mathbf{s})$
- Region-specific offset $o_{\xi(\mathbf{s})}$, e.g. the log-population density
- Endemic linear predictor $\beta' \mathbf{z}_{\tau(t), \xi(\mathbf{s})}$ includes discretised time trend and exogenous effects, e.g. the influenza cases

Additive-multiplicative continuous space-time intensity model proposed

$$\lambda^*(t, \mathbf{s}) = h(t, \mathbf{s}) + e^*(t, \mathbf{s})$$

Additive epidemic (self-exciting) component

$$e^*(t, \mathbf{s}) = \sum_{j \in I^*(t, \mathbf{s}; \varepsilon, \delta)} e^{\eta_j} g_{\alpha}(t - t_j) f_{\sigma}(\mathbf{s} - \mathbf{s}_j)$$

- Individual infectivity weighting through linear predictor $\eta_j = \gamma' \mathbf{m}_j$ based on the vector of unpredictable marks
- Positive parametric interaction functions, e.g. $f_{\sigma}(\mathbf{s}) = \exp\left(-\frac{\|\mathbf{s}\|^2}{2\sigma^2}\right)$ and $g_{\alpha}(t) = e^{-\alpha t}$
- Set of active infectives depends on fixed maximum temporal and spatial interaction ranges ε and δ

Marked extension with event type

- Motivation: joint modelling of both finetypes of IMD
- Additional dimension $\mathcal{K} = \{1, \dots, K\}$ for event type $\kappa \in \mathcal{K}$

Marked CIF

$$\lambda^*(t, \mathbf{s}, \kappa) = \exp\left(\beta_{0, \kappa} + o_{\xi(\mathbf{s})} + \beta' \mathbf{z}_{\tau(t), \xi(\mathbf{s})}\right) + \sum_{j \in I^*(t, \mathbf{s}; \varepsilon, \delta)} q_{\kappa_j, \kappa} e^{\eta_j} g_{\alpha}(t - t_j | \kappa_j) f_{\sigma}(\mathbf{s} - \mathbf{s}_j | \kappa_j)$$

Marked extension with event type

- Motivation: joint modelling of both finetypes of IMD
- Additional dimension $\mathcal{K} = \{1, \dots, K\}$ for event type $\kappa \in \mathcal{K}$

Marked CIF

$$\lambda^*(t, \mathbf{s}, \kappa) = \exp\left(\beta_{0, \kappa} + o_{\xi(\mathbf{s})} + \beta' \mathbf{z}_{\tau(t), \xi(\mathbf{s})}\right) + \sum_{j \in I^*(t, \mathbf{s}; \varepsilon, \delta)} q_{\kappa_j, \kappa} e^{\eta_j} g_{\alpha}(t - t_j | \kappa_j) f_{\sigma}(\mathbf{s} - \mathbf{s}_j | \kappa_j)$$

- Type-specific endemic intercept

Marked extension with event type

- Motivation: joint modelling of both finetypes of IMD
- Additional dimension $\mathcal{K} = \{1, \dots, K\}$ for event type $\kappa \in \mathcal{K}$

Marked CIF

$$\lambda^*(t, \mathbf{s}, \kappa) = \exp\left(\beta_{0,\kappa} + o_{\xi(\mathbf{s})} + \beta' \mathbf{z}_{\tau(t), \xi(\mathbf{s})}\right) + \sum_{j \in I^*(t, \mathbf{s}; \varepsilon, \delta)} q_{\kappa_j, \kappa} e^{\eta_j} g_{\alpha}(t - t_j | \kappa_j) f_{\sigma}(\mathbf{s} - \mathbf{s}_j | \kappa_j)$$

- Type-specific endemic intercept
- **Type-specific transmission**, $q_{k,l} \in \{0, 1\}$, $k, l \in \mathcal{K}$

Marked extension with event type

- Motivation: joint modelling of both finetypes of IMD
- Additional dimension $\mathcal{K} = \{1, \dots, K\}$ for event type $\kappa \in \mathcal{K}$

Marked CIF

$$\lambda^*(t, \mathbf{s}, \kappa) = \exp\left(\beta_{0,\kappa} + o_{\xi(\mathbf{s})} + \beta' \mathbf{z}_{\tau(t), \xi(\mathbf{s})}\right) + \sum_{j \in I^*(t, \mathbf{s}; \varepsilon, \delta)} q_{\kappa_j, \kappa} e^{\eta_j} g_{\alpha}(t - t_j | \kappa_j) f_{\sigma}(\mathbf{s} - \mathbf{s}_j | \kappa_j)$$

- Type-specific endemic intercept
- Type-specific transmission, $q_{k,l} \in \{0, 1\}$, $k, l \in \mathcal{K}$
- **Type-specific effect modification** $\eta_j = \gamma' \mathbf{m}_j$, κ_j is part of \mathbf{m}_j

Marked extension with event type

- Motivation: joint modelling of both finetypes of IMD
- Additional dimension $\mathcal{K} = \{1, \dots, K\}$ for event type $\kappa \in \mathcal{K}$

Marked CIF

$$\lambda^*(t, \mathbf{s}, \kappa) = \exp\left(\beta_{0,\kappa} + o_{\xi(\mathbf{s})} + \beta' \mathbf{z}_{\tau(t), \xi(\mathbf{s})}\right) + \sum_{j \in I^*(t, \mathbf{s}; \varepsilon, \delta)} q_{\kappa_j, \kappa} e^{\eta_j} g_{\alpha}(t - t_j | \kappa_j) f_{\sigma}(\mathbf{s} - \mathbf{s}_j | \kappa_j)$$

- Type-specific endemic intercept
- Type-specific transmission, $q_{k,l} \in \{0, 1\}$, $k, l \in \mathcal{K}$
- Type-specific effect modification $\eta_j = \gamma' \mathbf{m}_j$, κ_j is part of \mathbf{m}_j
- **Type-specific interaction functions**, e.g. variances σ_{κ}^2

Basic reproduction number

- An important quantity in epidemic modelling is the mean number of offspring each case generates
- Since offspring are generated in time according to an inhomogeneous Poisson process we define

Basic reproduction number

$$\mu_i = e^{\eta_i} \cdot \left[\int_0^{\varepsilon} g_{\alpha}(t) dt \right] \cdot \left[\int_{b(0, \delta)} f_{\sigma}(\mathbf{s}) d\mathbf{s} \right], \quad i = 1, \dots, N.$$

- Type specific reproduction numbers are obtained by averaging the μ_i 's for each type.

Outline

- 1 Introduction
- 2 The R package `surveillance`
- 3 Univariate time series detectors
- 4 Multivariate surveillance
- 5 **Space-Time Point Process Modelling**
 - Maximum Likelihood Inference
 - Data Analysis
 - Prospective space-time monitoring
- 6 Discussion and Summary

Log-likelihood of proposed model (1)

- Observed spatio-temporal marked point pattern:

$$\mathbf{x} = \left\{ (t_i, \mathbf{s}_i, \mathbf{m}_i) : i = 1, \dots, N \right\}$$

- No modelling of the unpredictable marks being part of \mathbf{m}_i , e.g. age and gender
- Endemic covariate information on a spatio-temporal grid

$$G = \left\{ \mathbf{z}_{\tau, \xi} : \tau \in \{1, \dots, D\}, \xi \in \{1, \dots, M\} \right\}$$

- Unknown parameters:

$$\boldsymbol{\theta} = \left(\beta'_0, \beta', \gamma', \sigma', \alpha' \right)'$$

Log-likelihood of proposed model (2)

$$l(\boldsymbol{\theta}) = \left[\sum_{i=1}^N \log \lambda_{\boldsymbol{\theta}}^*(t_i, \mathbf{s}_i, \kappa_i) \right] - \int_0^T \int_W \sum_{\kappa \in \mathcal{K}} \lambda_{\boldsymbol{\theta}}^*(t, \mathbf{s}, \kappa) dt ds$$

- Easy integration of piecewise constant endemic rate $h_{\boldsymbol{\theta}}(t, \mathbf{s}, \kappa)$
- Integration of epidemic component $e_{\boldsymbol{\theta}}^*(t, \mathbf{s}, \kappa)$ involves

$$\int_0^{\min\{T-t; \varepsilon\}} g_{\alpha}(t|\kappa_j) dt \quad \text{and} \quad \int_{[W \cap b(\mathbf{s}_j; \delta)]_{-\mathbf{s}_j}} f_{\sigma}(\mathbf{s}|\kappa_j) ds$$

- For the spatial integration we use the *two-dimensional midpoint rule* with adaptive bandwidth choice depending on the value of σ as best trade off between accuracy and speed

Further details

- The score function is determined analytically but requires numerical integration for $\int \frac{\partial}{\partial \sigma_i} f_{\sigma}(\mathbf{s}|\kappa) ds$
- Wald confidence intervals can be computed using the asymptotic variance matrix $\hat{I}^{-1}(\hat{\boldsymbol{\theta}}_{ML})$ where we use an expected Fisher information matrix estimate (Rathbun, 1996)
- To inspect goodness-of-fit residuals based on the cumulative CIF suggested by Rathbun (1996) can be used
- Simulation from the model is possible using an adaption of Ogata's modified thinning algorithm (Meyer et al., 2010)

Outline

- 1 Introduction
- 2 The R package `surveillance`
- 3 Univariate time series detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling
 - Maximum Likelihood Inference
 - Data Analysis
 - Prospective space-time monitoring
- 6 Discussion and Summary

Data representation: `epidataCS` class

IMD data representation in surveillance:

```
R> imdepi <- as.epidataCS(events, stgrid, W = germany, qmatrix = diag(2))
R> print(imdepi, n=5)
```

```
History of an epidemic
Observation period: 0 -- 2562
Observation window (bounding box): [4034.126, 4670.351] x [2686.701, 3543.229]
Spatio-temporal grid (not shown): 366 time blocks, 413 tiles
Types of events: 'B' 'C'
Overall number of events: 636
```

```

              coordinates ID  time  tile type eps.t eps.s age  sex  BLOCK
103 (4112.19, 3202.79)  1  0.99 05554  B   30  200  17  male  1
402 (4122.51, 3076.97)  2  1.00 05382  C   30  200   3  male  1
312 (4412.47, 2915.94)  3  6.00 09574  B   30  200  34  female 1
314 (4202.64, 2879.7)  4  8.00 08212  B   30  200  15  female 2
629 (4128.33, 3223.31)  5 23.00 05554  C   30  200  15  male  4
      start popdensity influenza0 influenza1 influenza2 influenza3
103      0  260.8612          0          0          0          0
402      0  519.3570          0          0          0          0
312      0  209.4464          0          0          0          0
314      7 1665.6117          0          0          0          0
629     21  260.8612          0          0          0          0
[...]
```

IMD model selection

Joint analysis of the two finetypes with model selection by AIC

- Linear effect of weekly number of influenza cases registered in the district of a point (lag 0 – lag 3)
- Linear time trend and 0–2 harmonics for time-of-year effects
- Epidemic predictor with Age (categorized as 0-2, 3-18 and ≥ 19 years), gender, finetype and age-finetype interaction
- $\varepsilon = 30$ days, $\delta = 200$ km
- Spatial interaction function f : Gaussian or constant

Resulting best AIC model:

$$\lambda_{\theta}^*(t, \mathbf{s}, \kappa) = \rho_{\xi(\mathbf{s})} \cdot \exp\left(\beta_0 + \beta_{\text{trend}} \frac{|t|}{365} + \beta_{\text{sin}} \sin\left(|t| \frac{2\pi}{365}\right) + \beta_{\text{cos}} \cos\left(|t| \frac{2\pi}{365}\right)\right) + \sum_{j \in I^*(t, \mathbf{s}, \kappa; \varepsilon, \delta)} q_{\kappa_j, \kappa} e^{\gamma_0 + \gamma_{3-18} \mathbb{1}_{[3,18]}(\text{age}_j) + \gamma_{\geq 19} \mathbb{1}_{[19, \infty)}(\text{age}_j) + \gamma_C \mathbb{1}_{\{C\}}(\kappa_j)} f_{\sigma}(\mathbf{s} - \mathbf{s}_j).$$

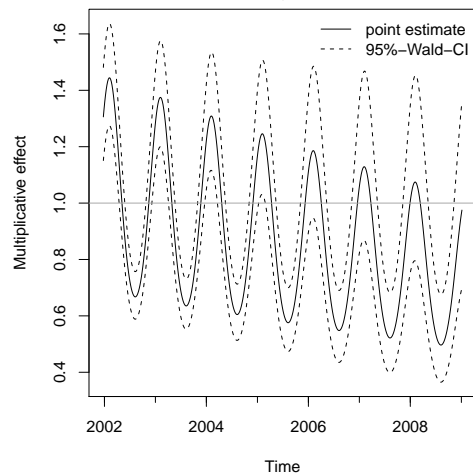
Selected joint model (1)

```
R> fit <- twinstim(endemic = ~1 + offset(log(popdensity)) + I(start/365) +
                 sin(start * 2 * pi/365) + cos(start * 2 * pi/365),
                 epidemic = ~1 + agegrp + type
                 siaf = siaf_1, data = imdepi, subset = allEpiCovNonNA,
                 optim_args = optim.args, method = "nlnmb",
                 control = list(fnscale = -10000), nCub = 36,
                 typeSpecificEpidemicIntercept = FALSE, partial=FALSE)
R> toLatex(summary(fit))
```

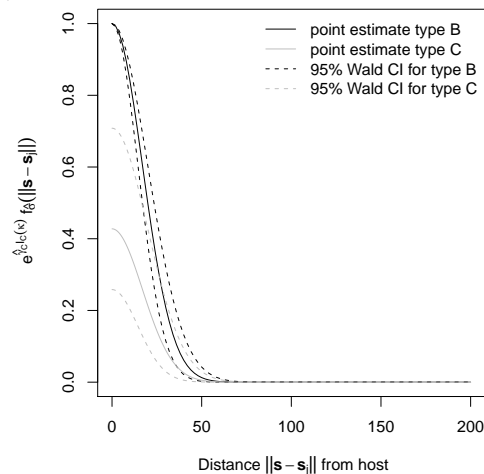
	Estimate	Std. Error	z value	$\mathbb{P}(Z > z)$
h. (Intercept)	-20.36516	0.08721	-233.527	$< 2 \cdot 10^{-16}$
h.I(start/365)	-0.04927	0.02229	-2.210	0.0271
h.sin(start*2*pi/365)	0.26184	0.06493	4.032	$5.52 \cdot 10^{-05}$
h.cos(start*2*pi/365)	0.26682	0.06437	4.145	$3.40 \cdot 10^{-05}$
e. (Intercept)	-12.57459	0.31275	-40.206	$< 2 \cdot 10^{-16}$
e.agegrp[3,19)	0.64632	0.31953	2.023	0.043102
e.agegrp[19,Inf)	-0.18676	0.43210	-0.432	0.665584
e.typeC	-0.84956	0.25742	-3.300	0.000966
e.siaf	2.82866	0.08191		
AIC:	18968			
Log-likelihood:	-9475			

Selected joint model

- Basic reproduction numbers: $\hat{\mu}_B \approx 0.25$ (95%-CI: 0.19-0.33) vs. $\hat{\mu}_C \approx 0.11$ (95%-CI: 0.07-0.18)
- LQ-test for $H_0 : \gamma_C = 0$ vs. $H_1 : \gamma_C \neq 0$ has p -value 0.013



IMD peak in late February



Effective interaction range ≈ 50 km

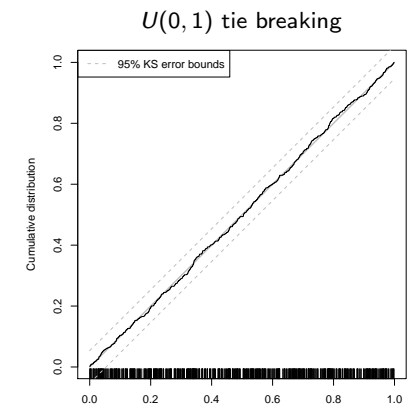
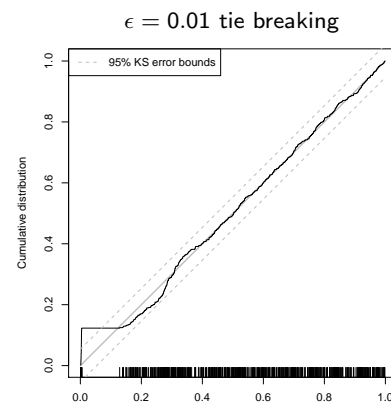
Selected joint model (3) – residual analysis

- To inspect goodness-of-fit Rathbun (1996) uses

$$Y_i = \hat{\Lambda}^*(t_i) - \hat{\Lambda}^*(t_{i-1}), \quad i = 2, \dots, N,$$

where $\hat{\Lambda}^*(t)$ is the cumulative intensity function

- If the estimated CIF describes the true CIF well, then $U_i = 1 - \exp(-Y_i) \stackrel{iid}{\sim} U(0, 1)$



Outline

- 1 Introduction
- 2 The R package `surveillance`
- 3 Univariate time series detectors
- 4 Multivariate surveillance
- 5 **Space-Time Point Process Modelling**
 - Maximum Likelihood Inference
 - Data Analysis
 - Prospective space-time monitoring

- 6 Discussion and Summary

Prospective space-time monitoring (1)

- Idea: Use `twinstim` as model framework for aberration detection within a statistical process control context
- Let $\hat{\theta}_0$ be the MLE for the `twinstim` model m_0 based on all events in a *pre-monitoring period* $[0, T_0]$
- Given the endemic–epidemic nature of the model previous outbreaks are thus taken into account
- After time T_0 new events are actively monitored as they arrive

Prospective space-time monitoring (2)

- Denote the knots in the time grid of G following T_0 by t_1, t_2, \dots and for each $k \geq 1$ compute

$$\Lambda_k^C = l_{m_0}(\hat{\theta}_1^C) - l_{m_0}(\hat{\theta}_0),$$

where the loglikelihoods are computed over all events in $[0, t_k]$

- In the above, $\hat{\theta}_1^C$ denotes $\hat{\theta}_0$, but with endemic intercept

$$\hat{\beta}_{0,\kappa} + \phi \cdot \mathbb{1}_C(t, \mathbf{s})$$

where $\phi > 0$ is a predefined constant and C the cluster

$$C = \{g \in G : \text{centroid}(g) \in [t_c, t_k] \times \text{circle}(\mathbf{s}_c, \delta_c)\}$$

Prospective space-time monitoring (4)

- Typically, one would look through a set of clusters \mathcal{C} with different centroids and radii all having time-length $[t_j, t_k]$

$$\Lambda_{j,k} = \max_{C \in \mathcal{C}} \left\{ l_{m_0}(\hat{\theta}_1^C) - l_{m_0}(\hat{\theta}_0) \right\}$$

- Aberration detection can now be based on, e.g. the Shiryaev-Roberts (SR) method used in Assunção and Correa (2009)

$$T_\gamma = \min_k \{SR_k > \gamma\}, \quad SR_k = \sum_{j=1}^k \exp(\Lambda_{j,k})$$

- An important result is that the SR method has in-control run-length greater or equal to γ

Prospective space-time monitoring (3)

- Other models for the change of the CIF within the cluster are possible, but the suggested intercept change is computationally advantageous
- Log likelihood ratio of endemic intercept change

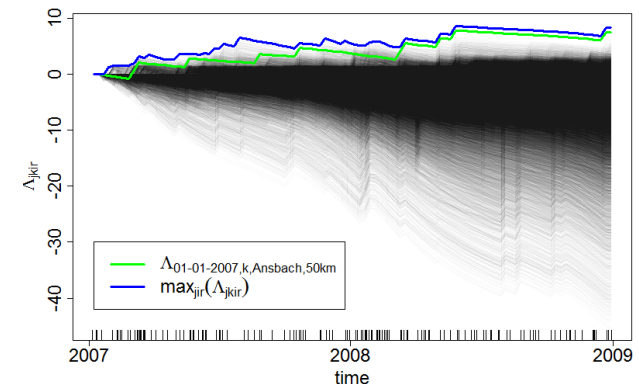
$$\Lambda_k^C = \sum_{i=1}^N \mathbb{1}_{[0,t_k]}(t_i) \left\{ \log(\lambda_{\theta_1}^*(t_i, \mathbf{s}_i, \kappa_i)) - \log(\lambda_{\theta_0}^*(t_i, \mathbf{s}_i, \kappa_i)) \right\} - \sum_{\tau=1}^D \sum_{\xi=1}^M \sum_{\kappa \in \mathcal{K}} \mathbb{1}_{[0,t_k]}(\tau) |B_\tau| |A_\xi| h(\tau, \xi, \kappa) \left[\exp(\phi \mathbb{1}_C(\tau, \xi)) - 1 \right],$$

where $|\cdot|$ denotes area and length, respectively, and

$$h(\tau, \xi, \kappa) = \exp(\alpha_\xi + \beta_{0,\kappa} + \beta' \mathbf{z}_{\tau,\xi})$$

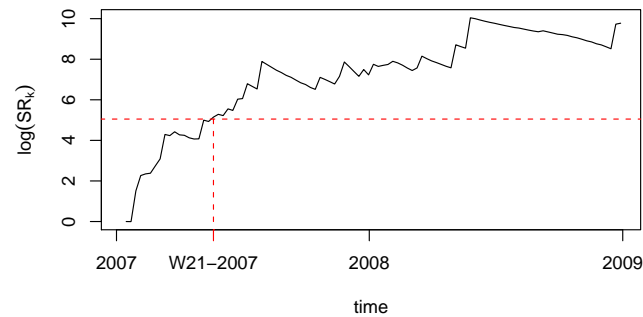
Simulation example (1)

- Simulated epidemic from best AIC model with $\delta_C = 50$ km cluster around Ansbach region starting on 01 Jan 2007 having $\phi = \log(5)$
- Cluster detection using $\delta_c \in \{25\text{km}, 50\text{km}, 75\text{km}\}$ and t_j in two-week intervals after 01 Jan 2007



Simulation example (2)

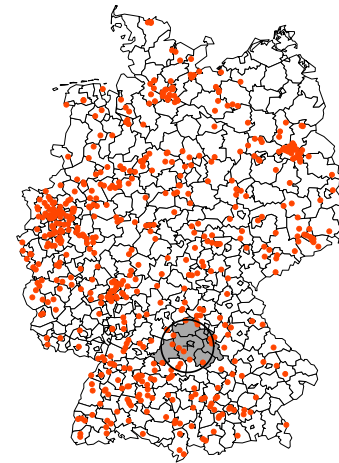
- Resulting Shiryaev-Roberts statistic



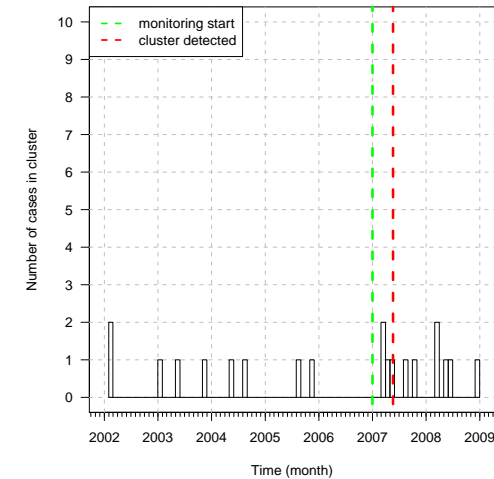
- Using $\gamma = 52 \cdot 3$ results in an alarm at t_{20} (W21-2007) with cluster location defined as the cluster producing $\max_{j=1}^{20} \exp(\Lambda_{j,20})$, i.e. here $C=(\text{Ansbach}, 50\text{km}, \text{W09-2007})$

Simulation example (3)

Illustration of the cluster location and available cases at alarm time (W21-2007) together with the corresponding univariate time series



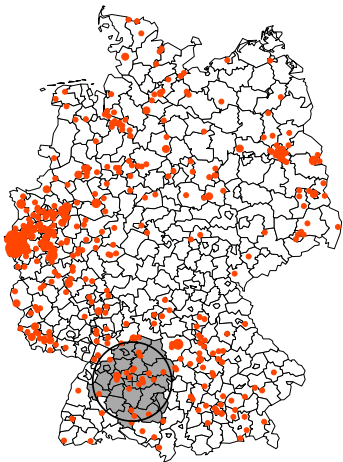
Location of cluster (grey) at W21-2007



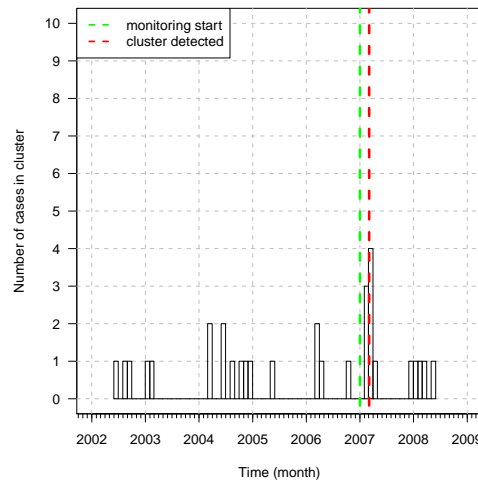
Time series with cases in cluster

Cluster detection for IMD data

Using same parametrization for original IMD data sounds alarm at W10-2007 with cluster $C=(\text{Esslingen}, 75\text{km}, \text{W05-2007})$



Location of cluster (grey) at W10-2007



Time series with cases in cluster

Discussion and Outlook (1)

- twinstim is a comprehensive framework for the modelling, inference and simulation of general self-exciting spatio-temporal point patterns
- An implementation is to be made available in the R package *surveillance* on CRAN
- Edge effects probably result in underestimated epidemic weight
- Full observability of the relevant epidemic events was assumed
- Meyer et al. (2010) contains further details on the twinstim modelling

Discussion and Outlook (2)

- This talk showed preliminary results on how to use `twinstim` for prospective space-time cluster-detection while adjusting for covariates
- Clustering as change in endemic intercept ensures speedy computations, but clusters are limited to a union of cells from the space-time grid G
- Actual run-length behaviour of method needs to be investigated by a simulation study
- Comparison with existing methods, e.g. Kulldorff (2001) or Diggle et al. (2005), of interest

Outline

- 1 Introduction
- 2 The R package `surveillance`
- 3 Univariate time series detectors
- 4 Multivariate surveillance
- 5 Space-Time Point Process Modelling
- 6 Discussion and Summary

Discussion and Summary (1)

- The focus of prospective surveillance is on *outbreak detection*
- Choice of the detection algorithm depends heavily on the epidemiological aims
- Combination of SPC and classical GLMs yielded nice changepoint detector for count time series
- Retrospective surveillance tries to *explain* temporal and spatio-temporal pattern in the data through *statistical modelling*
- Emphasis was on the *time series aspect* of surveillance as an alternative to spatial and spatio-temporal cluster detection methods, e.g. scan statistics

Discussion and Summary (2)

- The `surveillance` package offers a free and open-source implementation of the described algorithms
- Application of methods not restricted to infectious diseases
- Current work:
 - ▶ Robustify code, improve documentation and prepare for R CMD check running without warnings → get new version 1.3 on CRAN
 - ▶ Provide more methods for spatio-temporal cluster detection (also discrete time – discrete space)
 - ▶ Increase knowledge about package and integrate relevant existing code into the `surveillance` framework

Acknowledgements

Persons:

- Sebastian Meyer and Valentin Wimmer, Ludwig-Maximilians-Universität München, Germany, and Mathias Hofmann, Technische Universität München, Germany
- Michaela Paul, Andrea Riebler and Leonhard Held, Institute of Social and Preventive Medicine, University of Zurich, Switzerland
- Doris Altmann, Johannes Dreesman, Johannes Elias, Christopher W. Ryan, Klaus Stark, Christoph Staubach, Stefan Steiner, Yann Le Strat, André Michael Toschke, Mikko Virtanen and Achim Zeileis

Financial Support:

- German Science Foundation (DFG, 2003-2006)
- Swiss National Science Foundation (SNF, 2007–2010)
- Munich Center of Health Sciences (2007–2010)

Literature I

- Assunção, R. and Correa, T. (2009). Surveillance to detect emerging space-time clusters. *Computational Statistics & Data Analysis*, 53(8):2817–2830.
- Assunção, R. and Correa, T. (2009). Surveillance to detect emerging space-time clusters. *Computational Statistics & Data Analysis*, 53(8):2817–2830.
- Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.
- Beyrer, K., Dreesman, J., Heckler, R., Bradt, K., Scharlach, H., Baillot, A., Monazahian, M., Tabeling, D., Holle, I., Pulz, M., and Windorfer, A. (2006). Surveillance akuter respiratorischer Erkrankungen (ARE) in Niedersachsen: Erste Erfahrungen aus den Jahren 2005 - 2006. *Gesundheitswesen* 2006; 68: 679-685, 68:679–685.
- Bissell, A. F. (1984). The Performance of Control Charts and Cusums Under Linear Trend. *Applied Statistics*, 33(2):145–151.
- Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.
- Carpenter, T. E., Chriel, M., and Greiner, M. (2007). An analysis of an early-warning system to reduce abortions in dairy cattle in Denmark incorporating both financial and epidemiologic aspects. *Preventive Veterinary Medicine*, 78:1–11.
- Cartwright, K., Jones, D., Smith, A., Stuart, J., Kaczmarek, E., and Palmer, S. (1991). Influenza A and meningococcal disease. *Lancet*, 338(8766):554–557.

Literature II

- Chen, R. (1978). A surveillance system for congenital malformations. *Journal of the American Statistical Association*, 73:323–327.
- Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.
- Diggle, P. J., Rowlingson, B., and Su, T.-L. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, 16:423–34.
- 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.
- Frisén, M. (2003). Statistical surveillance: Optimality and methods. *International Statistical Review*, 71(2):403–434.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.
- Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.
- Hawkins, D. M. (1992). Evaluation of average run lengths of cumulative sum charts for an arbitrary data distribution. *Communications in Statistics. Simulation and Computation*, 21(4):1001–1020.
- Hawkins, D. M. and Olwell, D. H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement*. Statistics for Engineering and Physical Science. Springer.

Literature III

- 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. (2009). Additive-multiplicative regression models for spatio-temporal epidemics. *Biometrical Journal*, 51(6):961–978.
- Höhle, M. (2010). Changepoint detection in categorical time series. In Kneib, T. and Tutz, G., editors, *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pages 377–397. Springer.
- Höhle, M. and Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Computational Statistics & Data Analysis*, 52(9):4357–4368.
- Hubert, B., Watier, L., Garnerin, P., and Richardson, S. (1992). Meningococcal disease and influenza-like syndrome: a new approach to an old question. *Journal of Infectious Diseases*, 166:542–545.
- Jensen, E., Lundbye-Christensen, S., Samuelson, S., Sorensen, H., and Schonheyder, H. (2004). A 20-year ecological study of the temporal association between influenza and meningococcal disease. *European Journal of Epidemiology*, 19:181–187.

Literature IV

- Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, 34(1):207–228.
- Kosmider, R., Kelly, L., Evans, S., and Gettinby, G. (2006). A statistical system for detecting salmonella outbreaks in British livestock. *Epidemiology and Infection*, 134:952–960.
- Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A*, 164:61–72.
- Lai, T. and Shan, J. (1999). Efficient recursive algorithms for detection of abrupt changes in signals and control systems. *IEEE Transactions on Automatic Control*, 44:952–966.
- Makras, P., Alexiou-Daniel, S., Antoniadis, A., and Hatzigeorgiou, D. (2001). Outbreak of meningococcal disease after an influenza b epidemic at a hellenic air force recruit training center. *Clinical Infectious Diseases*, 33:e48–50.
- Meyer, S., Elias, J., and Höhle, M. (2010). A space-time conditional intensity model for infectious disease occurrence. Technical Report 95, Department of Statistics, Ludwig-Maximilians-Universität München. Available as <http://epub.ub.uni-muenchen.de/11898/>.
- Moustakides, G. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387.
- Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402.

Literature V

- Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*. Epub ahead of print DOI: 10.1002/sim.4177.
- Paul, M., Held, L., and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27:6250–6267.
- Rathbun, S. L. (1996). Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and Inference*, 51(1):55–74.
- Robert Koch Institute (2006). Epidemiologisches Bulletin 33. Available from <http://www.rki.de>.
- Robert Koch Institute (2009). SurvStat@RKI. <http://www3.rki.de/SurvStat>. Last access: 9 June 2009.
- 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.
- Steiner, S. H., Cook, R. J., Farewell, V. T., and Treasure, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*, 1(4):441–452.
- Thacker, S. B. (2000). Principles and practice of public health surveillance. chapter Historical Development, pages 1–16. Oxford University Press, 2nd edition.
- WHO Collaboration Centre for Rabies Surveillance and Research (2007). WHO Rabies - Bulletin - Europe. <http://www.rbe.fli.bund.de/>.

Literature VI

- Widdowson, M.-A., Bosman, A., van Straten, E., Tinga, M., Chaves, S., van Eerden, L., and van Pelt, W. (2003). Automated, laboratory-based system using the internet for disease outbreak detection, the Netherlands. *Emerging Infectious Diseases*, 9(9):1046–1052.
- Willsky, A. and Jones, H. (1976). Generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic control*, 21:108–112.