

# Temporal and spatio-temporal modelling of infectious diseases

Michael Höhle<sup>1</sup>

<sup>1</sup>Department for Infectious Disease Epidemiology  
Robert Koch Institute, Berlin, Germany

Department of Statistics  
Ludwig-Maximilians-Universität München  
Munich, Germany, 10-13 Oct 2011

ROBERT KOCH INSTITUT





# Outline (1)

- 1 Mathematical models for communicable diseases
- 2 Modelling and monitoring public health surveillance data
- 3 The R package `surveillance`
- 4 Now-casting and back-projection
- 5 Univariate time series detectors



## Outline (2)

- 6 Multivariate surveillance
- 7 Space-time point process modelling
- 8 Discussion and summary



# Outline

- 1 Mathematical models for communicable diseases
  - Discrete time stochastic Reed-Frost model
  - Continuous time deterministic SIR model
  - Stochastic continuous time SIR model
- 2 Modelling and monitoring public health surveillance data
- 3 The R package *surveillance*
- 4 Now-casting and back-projection
- 5 Univariate time series detectors
- 6 Multivariate surveillance



# Definitions and aim of this lecture

## Infectious disease epidemiology

Characterizes the epidemiological analysis of *infectious diseases*. Interest lies in the detection and understanding of epidemics. One possible aim would be the ability to better control outbreaks.

Aims of this lecture:

- Give a taste of how statistical modelling can be of use in infectious disease epidemiology.
- Illustrate this by plenty of examples from theory and practice.



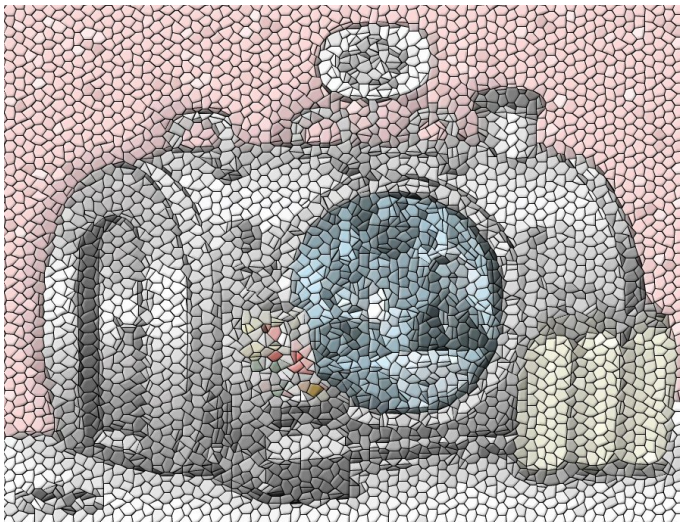
# Statistical modelling of infectious diseases

Three reasons that classical statistical inference is not immediately applicable for infectious disease data:

- ① Data are rarely a result of planned experiments
- ② Individuals are not independent (a case may also be a risk factor)
- ③ The infection process is only partially observable



# A small outbreak experiment...





# Mathematical models for communicable diseases (1)

- Mathematical modeling of infectious diseases has become a key tool in order to understand, predict and control the spread of infections.
- The intention of epidemic modeling is to model the spread of a disease in a population made up of a (possible large) integer number of individuals.
- To simplify the description of the population, it is common to use a compartmental approach to modeling. Here, the population is divided into classes of *susceptible*, *infective* and *recovered* individuals.



# Mathematical models for communicable diseases (2)

- Disease dynamics can then be characterized by a mathematical description of each individual's transitions between classes, subject to the state of the other individuals in the population.
- Mathematical models differ in whether they consider the infection process as *deterministic* or *stochastic*.
- Another distinction between models is whether they operate in continuous-time or discrete-time.



# Outline

- 1 Mathematical models for communicable diseases
  - Discrete time stochastic Reed-Frost model
  - Continuous time deterministic SIR model
  - Stochastic continuous time SIR model



# The Reed-Frost epidemic model

- SIR compartmental model, where individuals are either *Susceptible*, *Infectious* or *Recovered*
- Closed population with initially  $x_0 = n$  susceptible and  $y_0 = m$  infectious individuals
- Dynamics are described in discrete time by evolution of a Markov chain

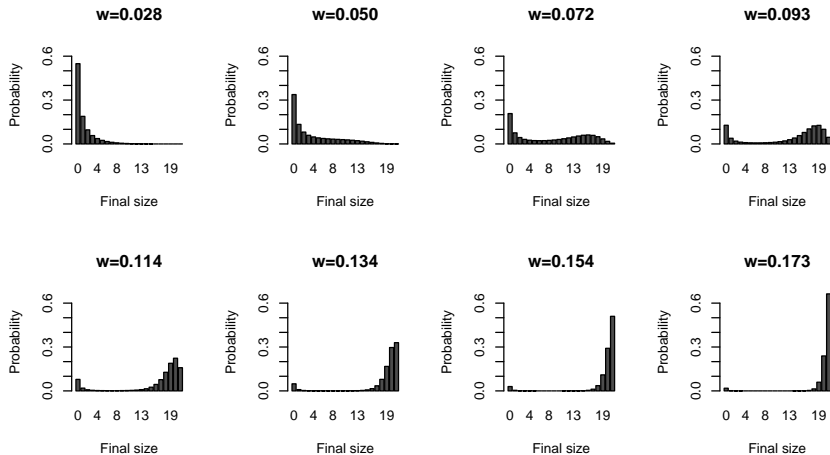
$$Y_{t+1}|x_t, y_t \sim \text{Bin}(x_t, 1 - (1 - w)^{y_t}),$$
$$X_{t+1} = X_t - Y_{t+1},$$

where  $w$  is the probability for an infectious contact of two individuals during one unit of time and  $t = 1, 2, \dots$

- The *final size* of the epidemic is  $Z = Y_1 + Y_2 + Y_3 + \dots$

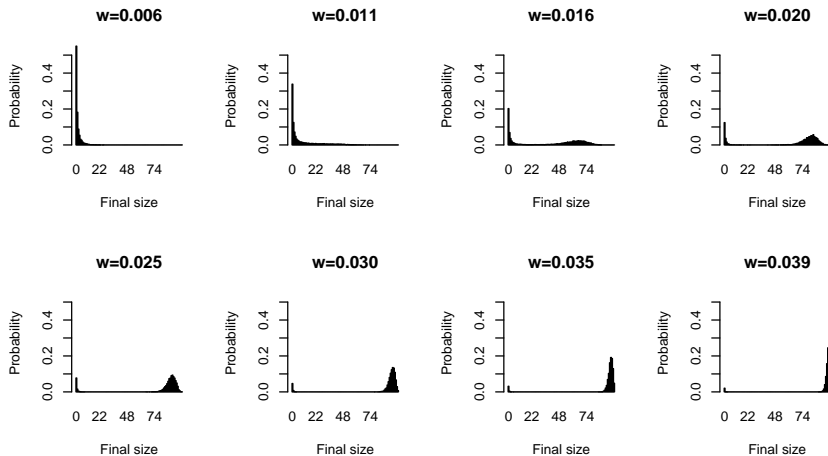


# Final size distribution when $n = 21$





# Final size distribution when $n = 100$





# Estimation of model parameters (1)

- Estimation of  $w$  from data for times  $0, 1, 2, \dots, K$  using maximum likelihood estimation

$$L(w) = \prod_{t=0}^{K-1} \theta_t^{y_{t+1}} (1 - \theta_t)^{x_t - y_{t+1}},$$

where  $\theta_t = 1 - (1 - w)^{y_t}$ .

- Contribution of statistics: Interest is not only in point estimator  $\hat{w}$ , but also in a quantification of the uncertainty of  $\hat{w}$ , e.g. by stating a 95% confidence interval for  $w$ .



## Estimation of model parameters (2)

```

R> #####
R> # Likelihood function for the reed-frost model
R> #
R> # Parameters:
R> # w.logit - logit(w) to have unrestricted parameter space
R> # x       - vector containing the number of susceptibles at each time
R> # y       - vector containing the number of infectious at each time
R> #
R> #####
R>
R> l <- function(w.logit,x,y) {
+   if (length(x) != length(y)) { stop("x and y need to be the same length") }
+
+   K <- length(x)
+   w <- plogis(w.logit)
+   theta <- 1 - (1-w)^y
+   #Compute loglik
+   return(sum(dbinom( y[-1], size=x[-K], prob=theta[-K],log=TRUE)))
+ }
R> #Observed susceptibles and infected
R> y <- c( 1, 4, 5, 6, 4, 0)
R> x <- c(20,16,11, 5, 1, 1)
R> mle <- optim(par=0,fn=l,method="BFGS",x=x,y=y,control=list(fnscale=-1),hessian=TRUE)
R> #Maximum likelihood estimator
R> (w.hat <- plogis(mle$par))

[1] 0.1365412

R> #95% confidence interval
R> (w.95ci <- plogis( mle$par + c(-1,1)*qnorm(0.975)* sqrt(-1/mle$hess)))

[1] 0.08815311 0.20550400

```



# Research questions

- What effect does a vaccination have?
- What effect does an isolation measure have?
- How could the model take different age categories into account?
- Not every infected does actually become infectious.
- The population is not closed, what now?
- It's the rodent, stupid!



# Mathematical challenges

- Mathematical abstractions of real world phenomena → *equations*
- No outbreaks are similar → *stochasticity*
- Different modes of transmission: person-to-person, air-borne, water-borne, food-borne and vector-borne → *direct and indirect transmission*
- Population heterogeneity (e.g. different places of residence, contact behaviour, susceptibility) needs to be taken into account
- Conflict between observation frequency and speed of the epidemic → *time scale of a model*
- Not all relevant events for the course of the epidemic are observable → *partial observability*



# Statistical challenges

## Statistics in a nutshell:

Stochastic model + data  $\rightarrow$

Parameter estimation + quantification of uncertainty

- Only one realization of the epidemic is observed.
- The data used for estimation can contain serious problems, e.g. under-reporting, changes in the test behaviour.
- The analysis is conducted using all available covariables, but the central risk covariates might be missing in the analysis.



# Outline

- 1 Mathematical models for communicable diseases
  - Discrete time stochastic Reed-Frost model
  - Continuous time deterministic SIR model
  - Stochastic continuous time SIR model



# The basic SIR model (1)

- When the considered population is large, it can be sufficient to disregard the stochasticity of the epidemic process and use deterministic models.
- Can formulate a continuous time deterministic *SIR model* by using ordinary differential equations (ODEs).
- The deterministic system intends to model the mean behaviour of the underlying stochastic system.
- We assume a closed population (i.e. no demographics turnover) of size  $N$ .

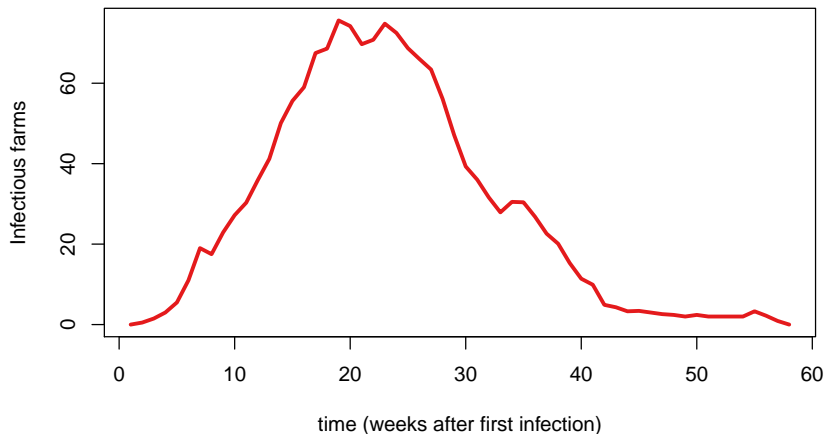


## Example: CSFV in The Netherlands (1)

- Classical swine fever virus (CSFV) is a highly contagious disease of pigs and wild boar.
- Characteristics of the disease are
  - ▶ Symptoms after infection: dullness and anorexia.
  - ▶ Acute form: rapid mortality often without clinical symptoms.
  - ▶ Secondary symptoms: diarrhea or respiratory problems.
- A huge outbreak in the Netherlands lasted from 4 February 1997 to May 1998.
  - ▶ 429 infected farms detected and stamped out ( $\sim 700,000$  pigs)
  - ▶ 1286 herds pre-emptively-slaughtered ( $\sim 1.1$  million pigs)



## Example: CSFV in the Netherlands (2)





## The basic SIR model (2)

- Divide population into three groups (*S*)usceptibles, (*I*)nfectious, and (*R*)ecovered. At all times  $S(t) + I(t) + R(t) = N + a$ .
- Describe dynamics using an ordinary differential equation system

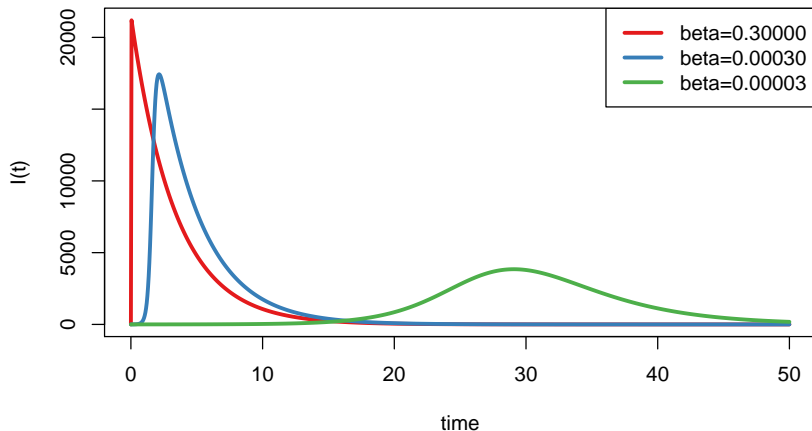
$$\begin{aligned}\frac{dS(t)}{dt} &= -\beta S(t)I(t) \\ \frac{dI(t)}{dt} &= \beta S(t)I(t) - \gamma I(t) \\ \frac{dR(t)}{dt} &= \gamma I(t)\end{aligned}$$

- Solve ODE with initial condition  $(N, a, 0)$  using numerical routines, e.g. an Euler or Runge-Kutta scheme.



## The basic SIR model (3)

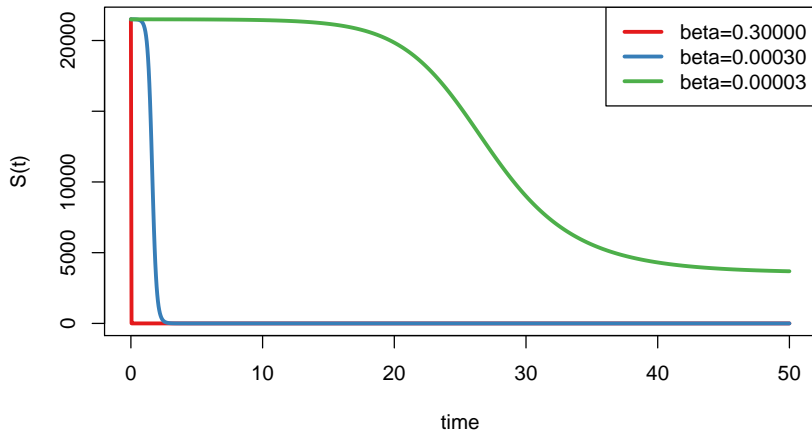
- Number of infected  $I(t)$  as a function of  $\beta$  when  $\gamma = 0.3$  and  $N = 21500$ .





## The basic SIR model (4)

- Number of susceptibles  $S(t)$  as a function of  $\beta$  when  $\gamma = 0.3$  and  $N = 21500$ .





# The basic reproduction rate $R_0$

- Definition:  $R_0$  is the number of new infections produced by one infection in a virgin population, i.e. the initial growth rate.
- If  $R_0 < 1$  the number of infected is expected to fade out right after introduction. If  $R_0 > 1$  an epidemic will result.
- In a simple SIR model

$$R_0 = \frac{\beta N}{\gamma}.$$



# The final size of an epidemic

- In a closed population the number of susceptibles can only decrease, hence it must have a limit for  $t \rightarrow \infty$ .
- Is the limit zero? Or will some fraction of the population escape from ever getting infected?
- Let  $f$  be the fraction of the (initially susceptible) population that got infected. This is also called the *final size* of the epidemic.
- It can be found as the solution to the equation

$$1 - f = \exp(-R_0 f).$$



# How to estimate parameters from data?

Parameter estimation depends on the available data from the epidemic.

- Final size data  $\Rightarrow$  use Equation (9), i.e.

$$R_0 = -\frac{\log(1-f)}{f}$$

- Some function of recovery and infection times is observed at  $k$  discrete time points  $\Rightarrow$  formulate a likelihood of the available observations by characterizing the distributional family of the observations and assume that the ODE system determines the expectation



## Estimating parameters (1) – Gaussian observations

- We have  $k$  observations of type  $\mathbf{y}_i = g(\mathbf{x}(t_i, \boldsymbol{\theta}))$ , where  $\boldsymbol{\theta} = (\beta, \gamma)'$ ,  $\mathbf{x}(t) = (S(t), I(t))'$  and  $g(\cdot)$  is a function indicating that we might only observe part of the state.
- Least squares aims at finding  $\boldsymbol{\theta}$ , which minimizes the function

$$l(\boldsymbol{\theta}) = \sum_{i=1}^k (\mathbf{y}_i - g(\mathbf{x}(t_i, \boldsymbol{\theta})))^2,$$

- Solution  $\hat{\boldsymbol{\theta}}$  is found using numerical optimizing routines.
- If  $g((S(t), I(t))') = I(t)$  least squares corresponds to MLE for Gaussian observations with

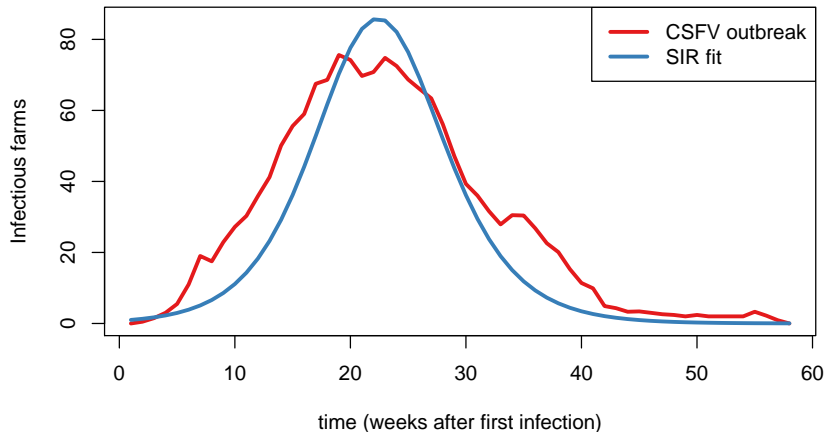
$$y_i \sim N(I(t), \sigma^2).$$

where  $\sigma$  is variance of the observation noise (kept fixed).



# Estimating parameters (2) – MLE for CSFV Data

- Example: SIR model fitted to CSFV curve by Gaussian likelihood





## Estimating parameters (3) – Poisson observations

- Assuming Gaussian observation ignores the fact that we actually observe count data. For small counts this may become problematic.
- An alternative is to use a count data distribution:

$$y_i \sim \text{Po}(I(t_i))$$

- As a consequence the log-likelihood is

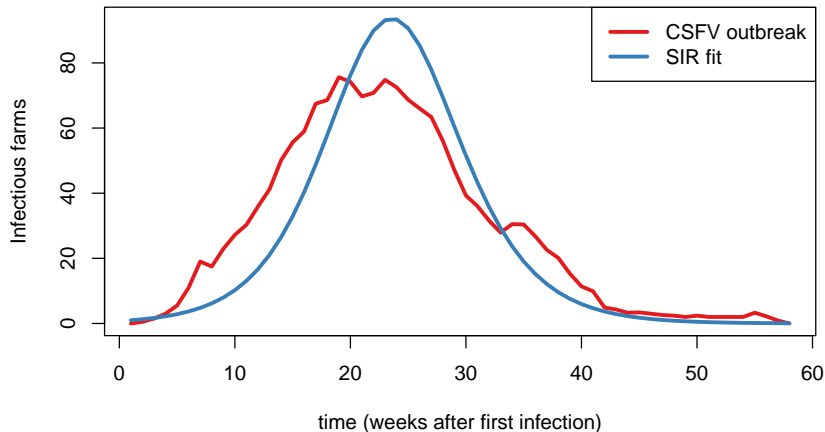
$$\log(L(\theta)) = \sum_{i=1}^k y_i \log(I(t_i)) - I(t_i),$$

- Since for the Poisson distribution  $E(y_i) = \text{Var}(y_i)$ , it might be necessary to address additional over-dispersion in the data using, e.g. a negative binomial distribution.



## Estimating parameters (4) – MLE for CSFV Data

- Example: SIR model fitted to CSFV curve by Poisson likelihood





# Outline

- 1 Mathematical models for communicable diseases
  - Discrete time stochastic Reed-Frost model
  - Continuous time deterministic SIR model
  - Stochastic continuous time SIR model



# Stochastic continuous time SIR model (1)

- If the population under study is large enough, deterministic approximations are reasonably valid to obtain an understanding of the disease.
- In small populations, however, stochasticity plays an important role for extinction, which cannot be ignored.
- Stochastic epidemic modeling is described e.g. in Becker (1989), Daley and Gani (1999) and Andersson and Britton (2000), who all rely heavily on the theory of stochastic processes.



# Stochastic continuous time SIR model (2)

- The *stochastic SIR model* can be described as a birth and death process, where the event rates for infection and removal are:

Event	Rate
$(S(t), I(t)) \rightarrow (S(t) - 1, I(t) + 1)$	$\beta S(t)I(t)$
$(S(t), I(t)) \rightarrow (S(t), I(t) - 1)$	$\gamma I(t)$

- Again,  $R(t)$  is implicitly given, because a fixed population of size  $S(0) + I(0)$  is assumed.
- The integer size of the population is now taken into account: Once  $I(t) = 0$ , the epidemic ceases.
- Point process viewpoint: piecewise constant conditional intensities for the process of infection, while the length of the infective period is given by independent and identically distributed exponential variates.



# The basic reproduction number (1)

- A further important difference between deterministic and stochastic modeling is the interpretation of  $R_0$ :

## The *basic reproductive ratio* in the stochastic setting

Average number of secondary cases directly caused by an infectious case in an entirely susceptible population.

- For the simple stochastic SIR model  $R_0$  can be calculated as

$$R_0 = \frac{\beta}{\gamma} \cdot S(0).$$

- As in the deterministic setting, the epidemic goes extinct if  $R_0 \leq 1$ .



## The basic reproduction number (2)

- In the deterministic setting one can show that an outbreak can only occur if  $R_0 > 1$ .
- In the stochastic SIR model setting the formulation is different: When  $R_0 > 1$ , a major outbreak occurs with probability

$$p = 1 - \left( \frac{R_0}{S(0)} \right)^{I(0)},$$

and with probability  $1 - p$  the epidemic goes extinct (Andersson and Britton, 2000).

- When assessing the risk of an infectious disease the difference between the deterministic and the stochastic interpretation of  $R_0$  can have important consequences.



# Likelihood inference (1)

- Assume that the epidemic process is completely observed over the interval  $(0, \tau]$ , where  $\tau$  is the duration of the epidemic.
- Let the  $K$  exposure times in this interval be  $T_E^1, \dots, T_E^K$  and the joint PDF of incubation and infectious shedding duration  $f_{T_D, T_S}(t_D, t_S)$ .
- Likelihood of the data  $\{(t_E^i, t_D^i, t_S^i), i = 1, \dots, k\}$  is

$$L = \left[ \prod_{i=1}^k f_{T_D, T_S}(t_D^i, t_S^i) \right] \left[ \prod_{i=1}^k \lambda(t_E^i | \mathcal{H}_{t_E^i}) \right] \exp \left( - \int_0^\tau \lambda(u | \mathcal{H}_u) du \right),$$

where  $\lambda(t | \mathcal{H}_t) = \beta I(t^-) S(t^-)$  is the conditional intensity function (CIF) and  $t^-$  denotes the time just prior to  $t_i$  (left-continuous CIF).



## Likelihood inference (2)

- The exposure times  $t_E^i, i = 1, \dots, k$ , are unlikely to be observed, i.e. the previous likelihood can not be constructed since  $S(t)$  is unknown.
- To make inference tractable assume that the incubation period is a constant  $\mu_D$  (known or to be estimated) and denote by  $t_I^i$  the observed start of an individual's infectious period.
- In this case  $t_E^i = t_I^i - \mu_D$  and hence

$$S(t^-) = S(0) - \sum_{i=1}^k \mathbb{1}_{(t_I^i - \mu_D, \infty)}(t)$$

- The likelihood is now

$$L = \left[ \prod_{i=1}^k f_{T_S}(t_S^i) \right] \left[ \prod_{i=1}^k \lambda(t_E^i | \mathcal{H}_{t_E^i}) \right] \exp \left( - \int_0^\tau \lambda(u | \mathcal{H}_u) du \right).$$



## Likelihood inference (3)

- Further assumptions are needed if also the  $t_S^i$  are unobservable, i.e. if only the removal times  $t_R^i = t_E^i + t_D^i + t_S^i$  are observed.
- Let also  $T_S$  be equal to a constant, say  $\mu_S$ . Now,  $t_E^i = t_R^i - \mu_D - \mu_S$  and hence

$$I(t^-) = \mathbb{1}_{(t_R^i - \mu_S, t_R^i)}(t)$$

$$S(t^-) = S(0) - \sum_{i=1}^k \mathbb{1}_{(t_R^i - \mu_D - \mu_S, \infty)}(t)$$

- The likelihood is

$$L = \left[ \prod_{i=1}^k \lambda(t_E^i | \mathcal{H}_{t_E^i}) \right] \exp \left( - \int_0^\tau \lambda(u | \mathcal{H}_u) du \right).$$



## Likelihood inference (4)

- A complication of the presented equations is that the CIF has to be integrated over time. However, for the simple SIR model the CIF is a piecewise constant function and hence integration is tractable.
- A GLM approximation exists to cast inference into the framework of available GLM software → exercise class
- Under certain regularity conditions the classical asymptotic normality of the likelihood also applies in the point-process setup



# Outline

- 1 Mathematical models for communicable diseases
- 2 Modelling and monitoring public health surveillance data**
- 3 The R package `surveillance`
- 4 Now-casting and back-projection
- 5 Univariate time series detectors
- 6 Multivariate surveillance
- 7 Space-time point process modelling



# 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.



# 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)

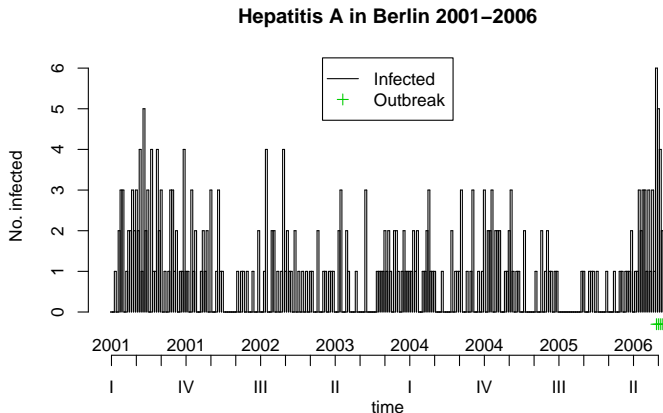
## In veterinary epidemiology

- 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 of surveillance data

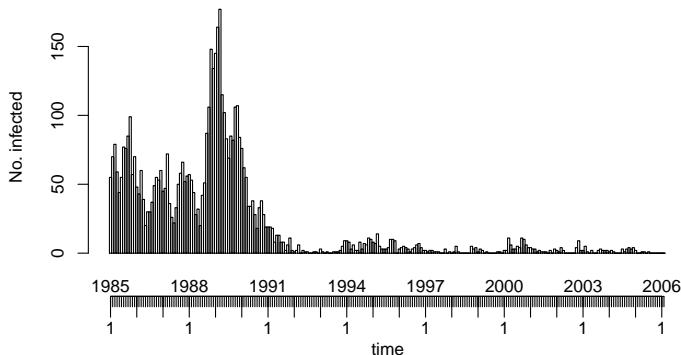
- Weekly number of adult male hepatitis A cases in the federal state of Berlin during 2001-2006
- During summer 2006 health authorities noticed an increased amount of cases (Robert Koch Institute, 2006).





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

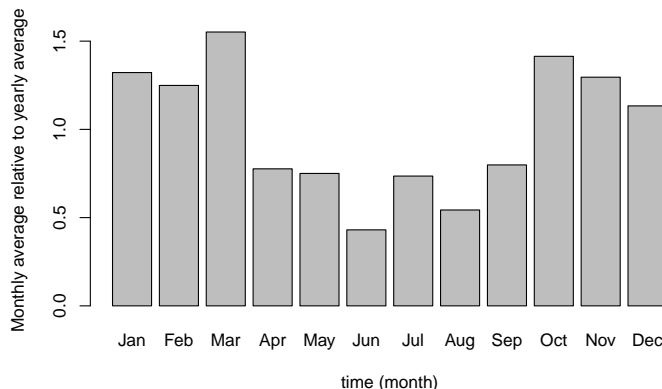
Monthly counts are provided by the WHO Collaboration Centre for Rabies Surveillance and Research. Thanks to Christoph Staubach, Federal Research Institute for Animal Health, Germany.



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

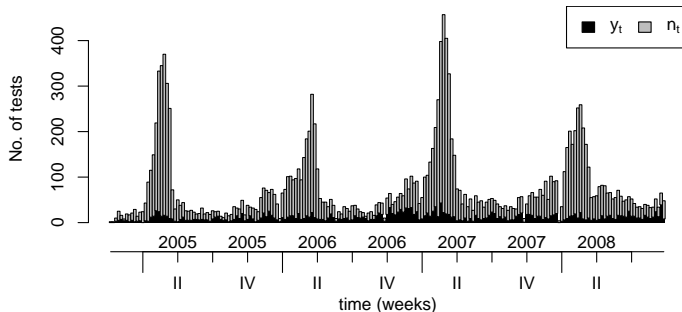


## Surveillance of acute respiratory diseases (2)

- 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.



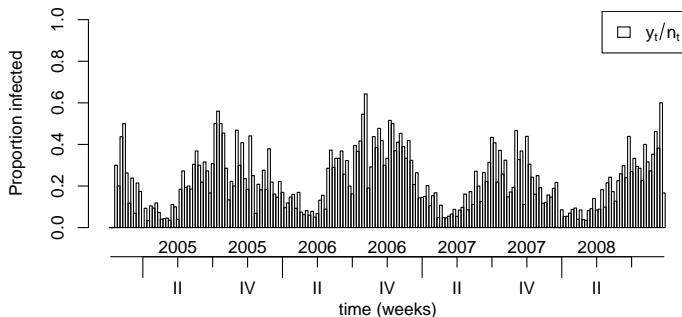


## Surveillance of acute respiratory diseases (2)

- 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.





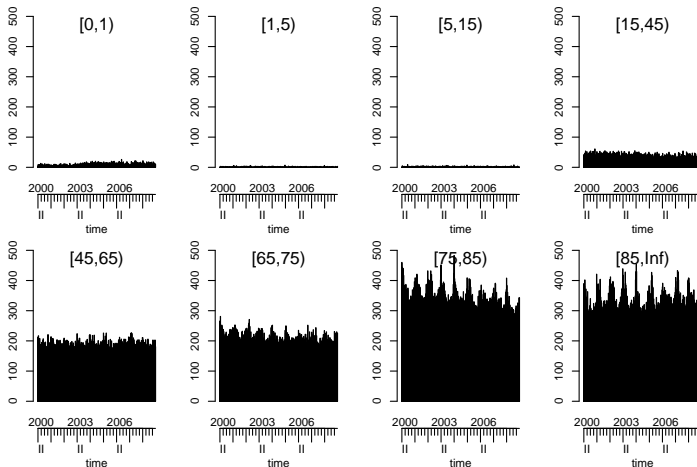
## Example – The EuroMOMO project (1)

- 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)

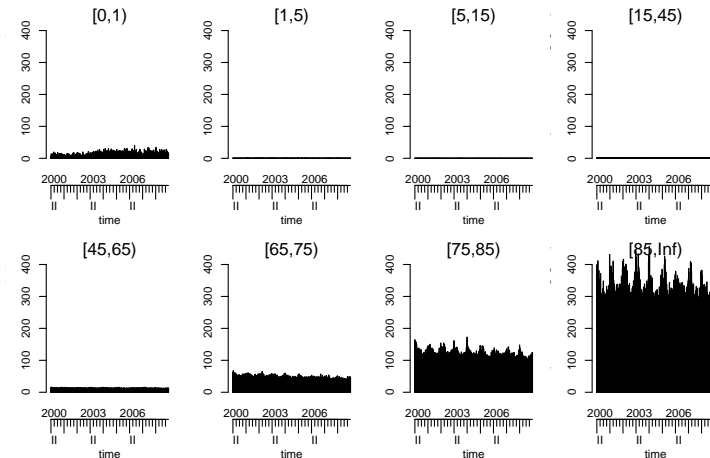
Weekly number of deaths in six age groups (alternatively incidence per 100,000 persons in age group)





## 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 Mathematical models for communicable diseases
- 2 Modelling and monitoring public health surveillance data
- 3 The R package surveillance**
- 4 Now-casting and back-projection
- 5 Univariate time series detectors
- 6 Multivariate surveillance
- 7 Space-time point process modelling



# 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., 2011)<sup>1</sup>
  - ▶ `stcd` – continuous space - continuous time cluster detection (Assunção and Correa, 2009)

---

<sup>1</sup>Work in progress, not fully available in the package yet.



# What is surveillance? (3)

- Interpretating the epidemiological curve of an outbreak<sup>2</sup>:
  - ▶ backprojNP – Non-parametric back-projection (Becker et al., 1991)
  - ▶ nowcast – Now-casting to adjust for reporting delays during an outbreak (an der Heiden et al., 2011)

---

<sup>2</sup>Work in progress, not fully available in the package yet.



## What is surveillance? (4)

- 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 2.13:  
`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)
```

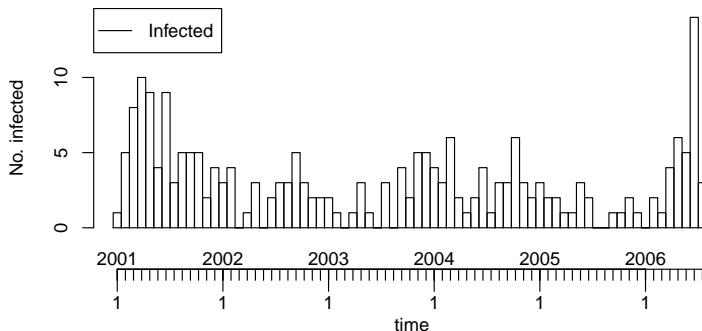
```
[1] 73  6
```



## Visualizing sts objects (1)

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

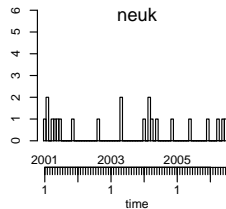
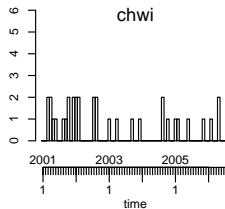
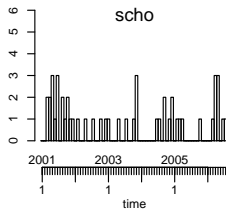
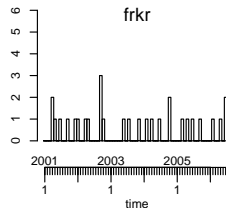
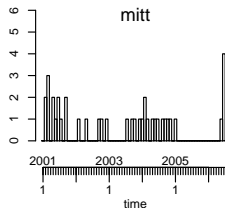
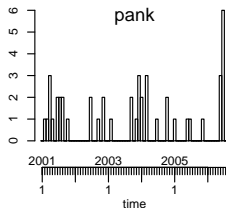
```
R> plot(ha4, type = observed ~ time)
```





# Visualizing sts objects (2)

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

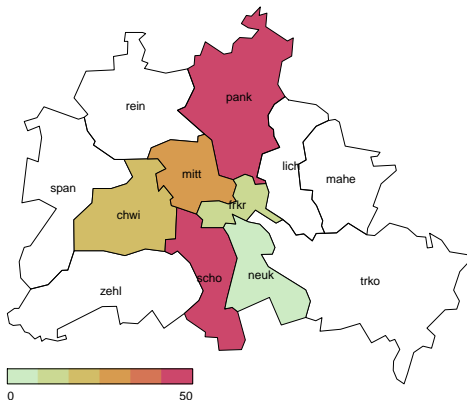




## Visualizing sts objects (3)

Using the `maptools` package `shapefiles` provides map visualizations

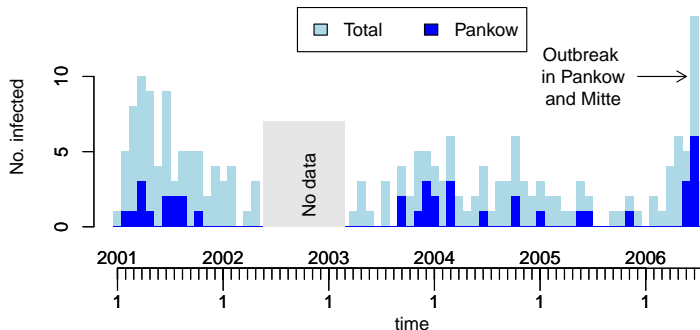
```
R> plot(ha4, type = observed ~ 1 | unit)
```





## 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





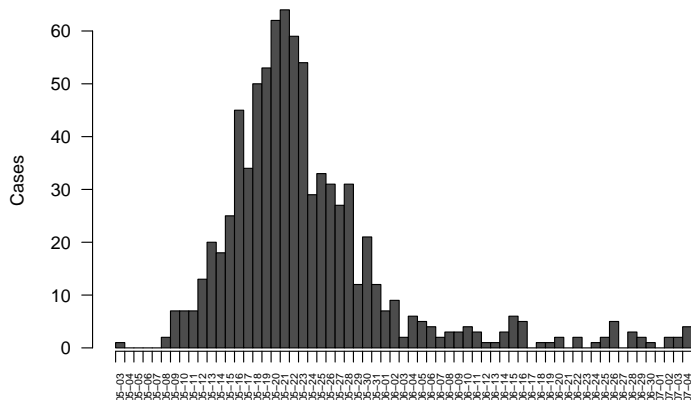
# Outline

- 1 Mathematical models for communicable diseases
- 2 Modelling and monitoring public health surveillance data
- 3 The R package *surveillance*
- 4 Now-casting and back-projection
  - Now-Casting
  - Back-projection
    - Non-parametric back-projection
    - Parametric back-projection
- 5 Univariate time series detectors
- 6 Multivariate surveillance



## Example: EHEC/HUS Outbreak in Germany 2011 (1)

- Large outbreak of haemolytic uraemic syndrome, Germany, May-June 2011 with 854 cases.
- Retrospective epicurve illustrating the onset of diarrhea of the patients (where available, 809 cases)





## Example: EHEC/HUS Outbreak in Germany 2011 (2)

- However, *during* the outbreak the situation is not as clear. Incubation period, time before reporting and other reporting delays complicate decision making.
- Illustration: Day of hospitalization of HUS cases (available for 635 cases) and the day the HUS case arrives at the RKI.

[Animated epidemic curve]

- ▶ Gray boxes – Actual epidemic curve of the hospitalization dates.
  - ▶ Red crosshair – denotes “now”, i.e. the current day.
  - ▶ Black boxes – shows all observations which were available at the RKI at the present day.
- Conclusion: Automatic online outbreak detection is not of use here.



# Outline

## 4 Now-casting and back-projection

- Now-Casting
- Back-projection
  - Non-parametric back-projection
  - Parametric back-projection



# What's the situation now?

- Opposite to the more sophisticated job of forecasting, we would be happy to know at specific point in time what the situation is now, i.e. in an ideal setup of no reporting delay → *now-casting*.

## Now-Casting

Extrapolate currently observed counts by taking the reporting delay from the past into account. Add uncertainty indication to this extrapolation.

- Relies on an assumption that the reporting delay is relatively stable over time.



## Now casting (1)<sup>3</sup>

- Basic idea of now-casting: The  $y_t$  cases on day  $t$  do not all arrive on day  $t$ , they are delayed according to some delay distribution having CDF  $F_D$ .
- The number of reports on time  $t$  available at time  $s \geq t$  is thus only a fraction of the total reports:

$$y_{t,s} = F_D(s - t) \cdot y_t$$

- Hence an estimate for  $y_t = \sum_{s=t}^{\infty} y_{t,s}$  at time  $s$  is

$$\hat{y}_t^s = \frac{y_{t,s}}{F_D(s - t)}.$$

---

<sup>3</sup>Joint work with Matthias an der Heiden, RKI



## Now casting (2)

- Compute confidence interval for  $\hat{y}_t$  by taking uncertainty in the estimation of  $F_D(s - t)$  into account, i.e. by pointwise confidence intervals or a simultaneous confidence band.
- Re-estimate delay distribution for each time point based on all currently available observations.

[Now casting – Animated]

- For time points close to now, i.e  $s - t$  small, the extrapolation is quite unreliable, as a consequence we limit now-casts to  $s - t \geq 3$ .



# Outline

## 4 Now-casting and back-projection

- Now-Casting
- Back-projection
  - Non-parametric back-projection
  - Parametric back-projection



# Motivation for back-projection

- There is a time delay between time of infection and the onset of the disease. This time delay is often denoted *incubation time*.
- Usually, only onset of disease can be observed. Examples:
  - ▶ Time to AIDS onset after HIV infection
  - ▶ Onset of diarrhea after consumption of sprouts (EHEC/HUS)
- Let  $D$  be a discrete random variable describing the delay in number of time units. Assuming this delay is constant over time let  $f(d), d = 0, 1, 2, \dots$ , be the PMF of  $D$ .

## Back-projection

Interest is often in the time of exposure of individuals, but data is only available about their time of disease onset.

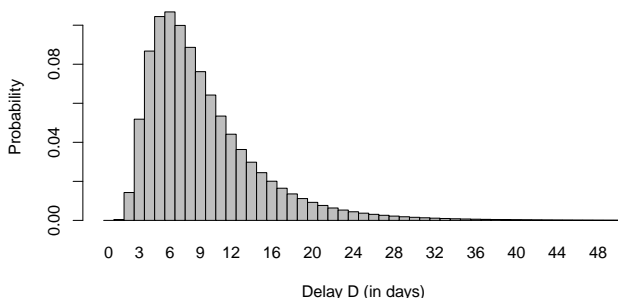


# Incubation time as a random variable

- Assume that the PMF for  $D$  is generated by interval censoring a continuous random variable  $D^*$  with positive support and CDF  $F_{D^*}$ .
- It can be computationally convenient to assume that  $D$  has finite support  $1, \dots, d_{\max}$ . Altogether,

$$f_D(d) = \frac{F_{D^*}(d) - F_{D^*}(d-1)}{F_{D^*}(d_{\max})}, \quad d = 1, 2, \dots, d_{\max}.$$

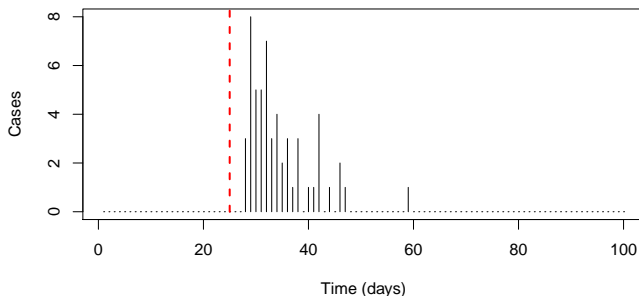
- Example:  $D^*$  log-normal with  $\log \mu = 2$ ,  $\log \sigma = 0.6$  and  $d_{\max} = 50$ .





## Example 1: Point source outbreak at time $t_0$ (2)

- Assume a point source is active on day  $t_0 = 25$  infecting a total of  $n = 55$  individuals.
- The following time series for disease onsets is observed:

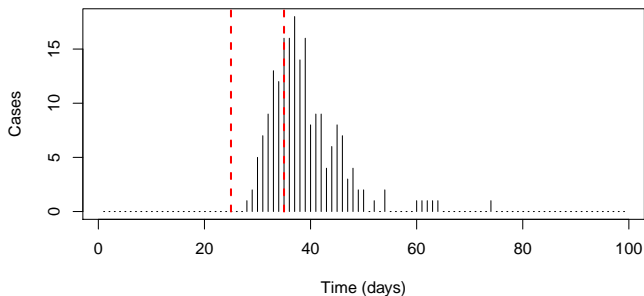


- To identify the possible source, interest is in inferring infection times from the onset times.



## Example 2: Point source during interval $[t_0, t_0 + l - 1]$

- Assume a point source is active from day  $t_0$  until time  $t_0 + l - 1$  infecting a total of  $n$  individuals, where individuals are equally likely to be infected within  $[t_0, t_0 + l - 1]$ .
- Example  $t_0 = 25$ ,  $l = 10$  and  $n = 200$ .





# Simple back-projection methods (1)

- Method 1: Determine the exposure interval by subtracting the shortest incubation time from the first case and the longest incubation from the last case of the epidemic curve
- R-code for outbreak Examples 1 & 2

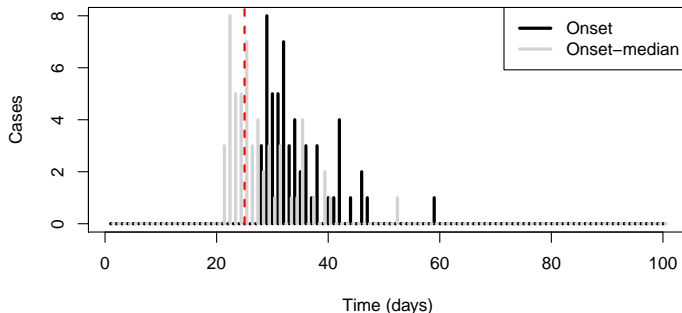
```
R> subtract.minmax <- function(y, d.pmf, eps = 0.001) {  
+   exposure.left <- head(which(y > eps), n = 1) - ((0:d.max)[head(which(d.pmf >  
+     eps), n = 1)])  
+   exposure.right <- tail(which(y > eps), n = 1) - ((0:d.max)[tail(which(d.pmf >  
+     eps), n = 1)])  
+   structure(c(exposure.left, exposure.right - exposure.left), names = c("t0",  
+     "l"))  
+ }  
R> subtract.minmax(y.ts, d.pmf)  
  
t0  1  
26  1  
  
R> subtract.minmax(y.l.ts, d.pmf)  
  
t0  1  
26 16
```



## Simple back-projection methods (2)

- Method 2: Subtract the median incubation time from each onset.

```
R> subtract.median <- function(y, d.pmf) {
+   d.median <- (0:length(d.pmf) - 1)[which(cumsum(d.pmf) > 0.5)][1]
+   structure(c(tail(y, n = -d.median), rep(0, d.median)), names = names(y))
+ }
R> subtract.median(y.ts, d.pmf)
```



- This method is not recommendable since it ignores the order of events in the epidemic curve.



# Non-parametric back-projection by Becker et al. (1991)

- Becker et al. (1991) proposed a non-parametric back-projection method for discrete time interval data.
- Their motivating application was a back-projection of AIDS cases to HIV incidence (before the use of antiretroviral therapy).
- The method differs from the the individual based continuous time parametric back-calculation of Brookmeyer and Gail (1988).
- However, it equally presumes a fixed and known incubation time distribution.



## Model and notation (1)

$N_{t,d}$  – Number of individuals exposed in interval  $t = 1, \dots, T$  having an incubation of time  $d$  (i.e. observed at time  $t + d$ )

$\vdots$				
$N_{1,4}$	$\vdots$			
$N_{1,3}$	$N_{2,3}$	$\vdots$		
$N_{1,2}$	$N_{2,2}$	$N_{3,2}$	$\vdots$	
$N_{1,1}$	$N_{2,1}$	$N_{3,1}$	$N_{4,1}$	$\vdots$
$N_{1,0}$	$N_{2,0}$	$N_{3,0}$	$N_{4,0}$	$N_{5,0}$
$N_1$	$N_2$	$N_3$	$N_4$	$N_5$

$Y_t$  – The observed number of incident cases

$$Y_t = \sum_{i=1}^t N_{i,t-i} = \sum_{i=0}^{t-1} N_{t-i,i}, \quad t = 1, \dots, T.$$



## Model and notation (2)

$N_t$  – Number of individuals infected in interval  $t$ , i.e.

$$N_t = \sum_{d=0}^{\infty} N_{t,d}.$$

- Assume  $N_{t,d} \sim \text{Po}(f(d)\lambda_t)$ , i.e.  $N_t \sim \text{Po}(\lambda_t)$ .
- As a consequence  $Y_t \sim \text{Po}(\mu_t)$ , where

$$\mu_t = \sum_{i=1}^t E(N_{i,t-i}) = \sum_{i=1}^t f(t-i)\lambda_i.$$



# EM Algorithm (1)

- Let  $\mathbf{y} = (y_1, \dots, y_T)'$  be the incomplete data and let  $\mathbf{x} = (\mathbf{N}_1, \dots, \mathbf{N}_T)'$  be the complete data with  $\mathbf{N}_t = (N_{t,0}, \dots, N_{t,\infty})'$
- Interest is in estimating  $\boldsymbol{\theta} = (\lambda_1, \dots, \lambda_T)'$
- Q-function for the EM algorithm:

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E(l(\boldsymbol{\theta}, \mathbf{x})|\mathbf{y}, \boldsymbol{\theta}^{(k)}),$$

where  $l(\boldsymbol{\theta}, \mathbf{x})$  is the loglikelihood of the complete data

$$l(\boldsymbol{\theta}, \mathbf{x}) = \sum_{t=1}^T \sum_{i=0}^{t-1} [N_{t-i,i} \log(\lambda_{t-i} f_i) - \lambda_{t-i} f_i].$$



## EM Algorithm (2)

- Then,

$$E(l(\theta, \mathbf{x}) | \mathbf{y}, \theta^{(k)}) = \sum_{t=1}^T \sum_{i=0}^{t-1} \left[ y_t \frac{\lambda_{t-i}^{(k)} f_i}{\sum_{j=0}^{t-1} \lambda_{t-j}^{(k)} f_j} \log(\lambda_{t-i} f_i) - \lambda_{t-i} f_i \right]$$

- Hence, for  $t \in \{1, \dots, T\}$  the update is

$$\lambda_t^{(k+1)} = \frac{\lambda_t^{(k)}}{F(T-t)} \sum_{d=0}^{T-t} \frac{y_{t+d} f_d}{\sum_{j=1}^{t+d} \lambda_j^{(k)} f_{t+d-j}},$$

where  $F(T-t) = \sum_{d=0}^{T-t} f_d$  is the CDF of the incubation time.



## EM Algorithm (3)

- Iterations proceed until absolute or relative convergence for the parameter values, e.g.

$$\frac{\|\boldsymbol{\lambda}_{1:T}^{(k+1)} - \boldsymbol{\lambda}_{1:T}^{(k)}\|}{\|\boldsymbol{\lambda}_{1:T}^{(k)}\|} < \epsilon \quad (1)$$

- In some cases, where the observations near  $T$  are known to be incomplete, one can not expect the  $\lambda_t$ 's near  $T$  to be estimated with great precision.
- In these cases one may use  $T' \leq T$  in (1) instead of  $T$ .



## EMS Algorithm (1)

- To stabilize the estimation a smoothing step is introduced after each EM step, i.e. let

$$\phi_t^{(k+1)} = \frac{\lambda_t^{(k)}}{F(T-t)} \sum_{d=0}^{T-t} \frac{y_{t+d} f_d}{\sum_{j=1}^{t+d} \lambda_j^{(k)} f_{t+d-j}},$$

and then let

$$\lambda_t^{(k+1)} = \sum_{i=0}^k w_i \cdot \phi_{t+i-k/2}^{(k+1)}.$$

- In other words,  $\lambda_t^{(k+1)}$  is weighted average of the new parameter values the EM step produces.
- Symmetric binomial weights are chosen

$$w_i = \frac{\binom{k}{i}}{2^k}, \quad i = 0, 1, \dots, k.$$



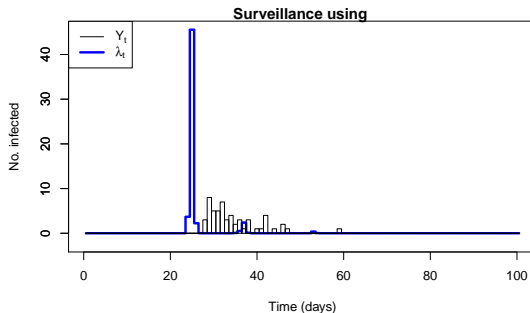
# Implementation in surveillance (1)

- Code:

```
R> #Create vector with incubation time PMF values on (0,...,d_max)
R> incu.pmf <- c(0, (plnorm(1:d.max,logmu,logsd) - plnorm(0:(d.max-1),logmu,logsd))/plnorm(d.max,logmu,logsd))
R> #Create sts object
R> require("surveillance")
R> sts <- new("sts",epoch=1:length(y.ts),observed=matrix(y.ts,ncol=1))
R> #Backproject using the method by Becker et al. (1991)
R> bp.control <- list(k=0,eps=1e-3,iter.max=100,verbose=TRUE)
R> sts.bp.k0 <- backprojNP(sts, incu.pmf.vec=incu.pmf, control=bp.control)
```

- Plotting code (result is saved in upperbound slot):

```
R> plot(sts.bp.k0, xaxis.years = FALSE)
```

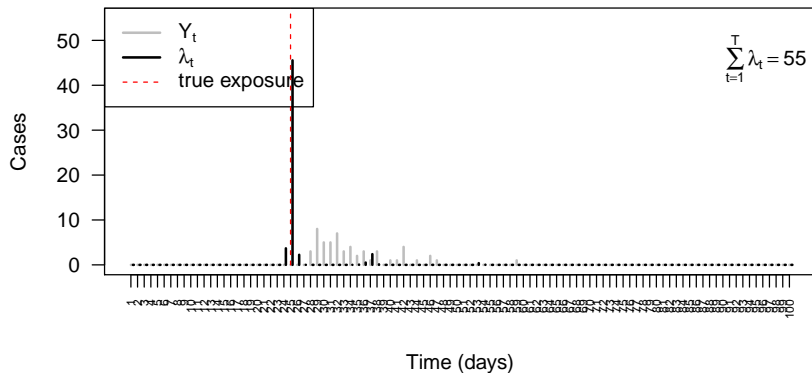




# Back-projection for outbreak Example 1

- Effect of the smoothing parameter  $k$ :

**$k=0$**

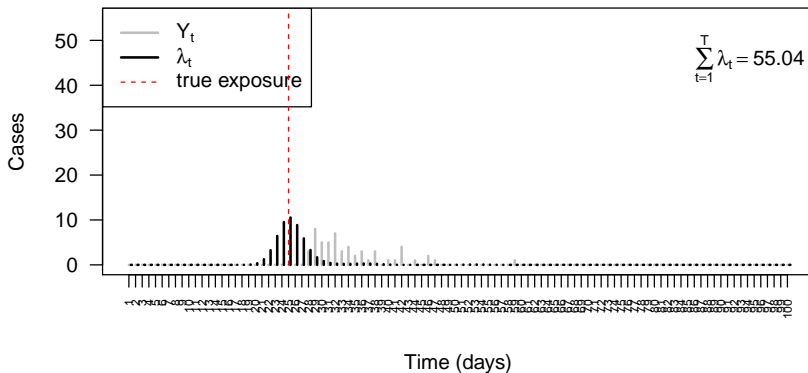




# Back-projection for outbreak Example 1

- Effect of the smoothing parameter  $k$ :

**$k=2$**

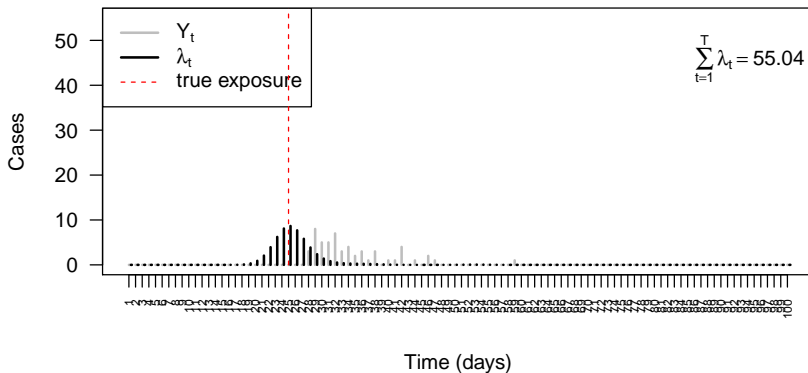




# Back-projection for outbreak Example 1

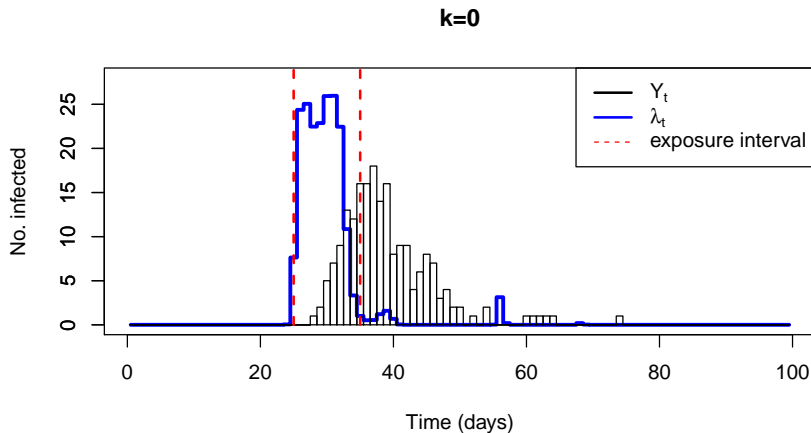
- Effect of the smoothing parameter  $k$ :

**$k=4$**



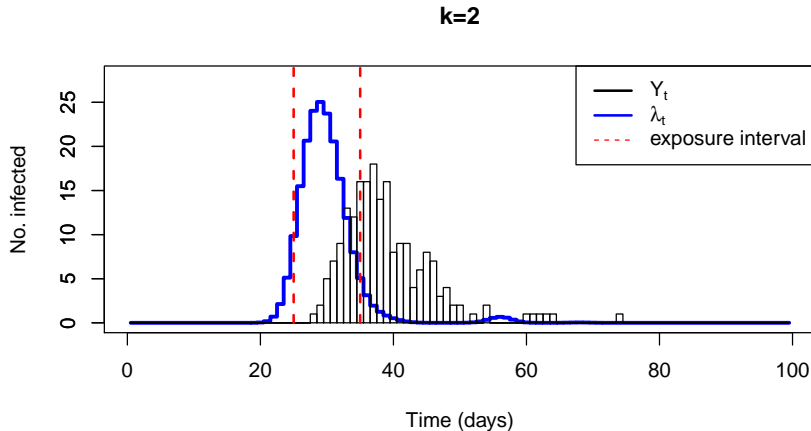


# Back-projection for outbreak Example 2



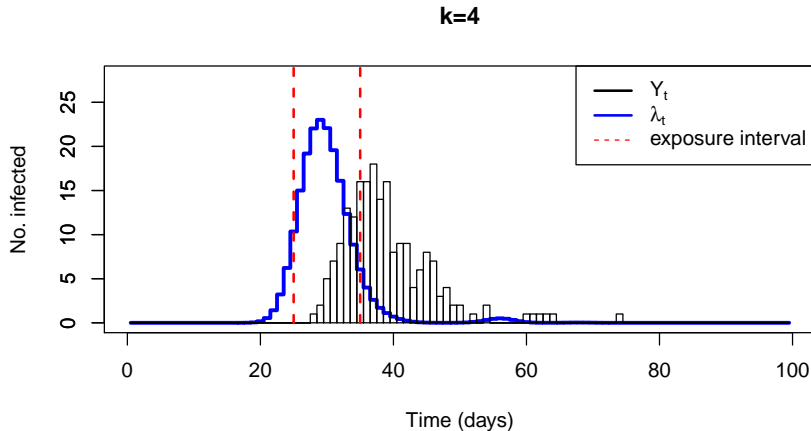


## Back-projection for outbreak Example 2





## Back-projection for outbreak Example 2





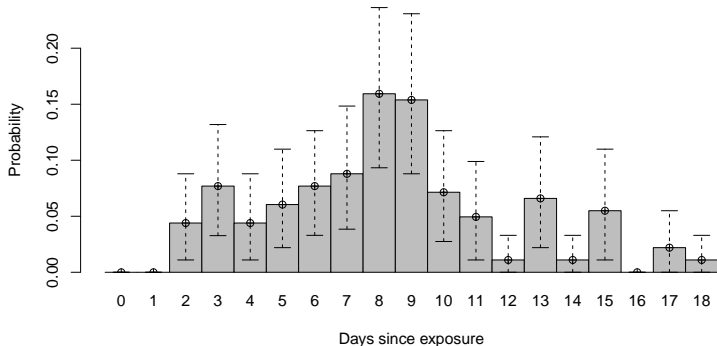
# Uncertainty

- The non-parametric back-projection does not provide any measures of uncertainty for  $\hat{\lambda}$ .
- One possibility to propagate uncertainty from the estimation of the incubation PMF is to use bootstrap:
  - ▶ Generate dataset  $i$  by sampling with replacement the individuals used to construct the PMF. Then estimate  $\hat{f}_D^{(i)}$ .
  - ▶ Apply the EMS algorithm in order to back-project the onset times using  $\hat{f}_D^{(i)}$ . This yields  $\hat{\lambda}^{(i)}$ .
  - ▶ Construct an appropriate measure of uncertainty for  $\hat{\lambda}$  using the  $m$  bootstrap samples  $\hat{\lambda}^{(1)}, \dots, \hat{\lambda}^{(m)}$ , e.g. quantile based point-wise confidence intervals.



## Back-projection for the 2011 EHEC/HUS outbreak (1)

- Determination of the incubation time PMF from 91 cases with a well known exposure time (foreign cases, restaurant cluster, etc.)



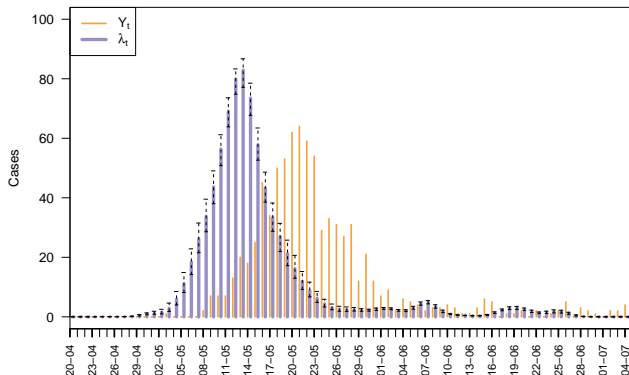
Source: Robert Koch Institute (2011)

- The figure shows point-wise 95% confidence intervals obtained by bootstrapping the PMF estimation.



## Back-projection for the 2011 EHEC/HUS outbreak (2)

- This estimated incubation PMF is then used in the Becker et al. (1991) procedure to back-project the epidemic curve formed by the 809 HUS cases with available disease onset time.



Source: Robert Koch Institute (2011)



# Discussion of the non-parametric back-projection method

- My experiences with the method is that it is sensitive to an appropriate choice of the convergence criterion threshold  $\epsilon$ .
- Especially for  $k = 0$  one should ensure that enough iterations are made.
- The EMS method is subject to instability in the latter stages of the epidemic. Marschner and Watson (1994) suggest a small improvement in the recursion to stabilize the method.
- During an outbreak one should choose  $T$  such that the incidence cases observed at time  $y_T$  are reliable (i.e. sufficiently complete), i.e.  $T$  should not be too close to “now”.



# Parametric back-projection

- Alternatives to the non-parametric back-projection use a parametric or semi-parametric model to model the expected number of incidence cases, i.e.  $\lambda(t)$ .
- Examples from the literature: Brookmeyer and Gail (1988), Bacchetti et al. (1993)
- An alternative to the non-parametric method is to formulate a parametric model for the hazard function in a discrete time survival model.



# Discrete time survival model (1)

- Let  $T_E$  be a discrete random variable describing the duration from origin to exposure by the disease of an individual

## Discrete time hazard function

$$\lambda(t_E|\mathbf{x}) = P(T_E = t_E | T_E \geq t_E, \mathbf{x}), \quad t_E = 1, 2, \dots$$

Here,  $\lambda(\cdot)$  is parametrized by covariates  $\mathbf{x}$  and parameters  $\theta$ .

- The probability of an event at time  $t_E$  is then

$$P(T_E = t_E) = \lambda(t_E) \prod_{s=1}^{t_E-1} \{1 - \lambda(s)\}.$$



## Discrete time survival model (2)

- However, the exposure time  $t_E$  of an individual is not observed, only its onset time  $t_O$ .
- Onset and exposure are related as follows:  $T_O = T_E + D$ , where  $D$  is the incubation time.
- Discrete time convolution of two discrete random variables yields that

$$P(T_O = t_O) = \sum_{d=0}^{d_{\max}} P(T_E = t_O - d)P(D = d).$$

- Hence, the likelihood of an individual with onset at time  $t_O^i$  is

$$L_i = \sum_{d=0}^{d_{\max}} \left\{ \lambda(t_O^i - d) \prod_{s=1}^{t_O^i - d - 1} (1 - \lambda(s)) \right\} f_D(d).$$



## Discrete time survival model (3)

- The log-likelihood of all  $n$  individuals is then

$$\log(L) = \sum_{i=1}^n \log(L_i).$$

- Since we have aggregated data, the likelihood of the  $n = \sum_{t=1}^T y_t$  individuals can be given in grouped form

$$\log(L) = \sum_{t=1}^T y_t \log(L_t),$$

where  $L_t$  is the likelihood of an individual with onset at time  $t$ .



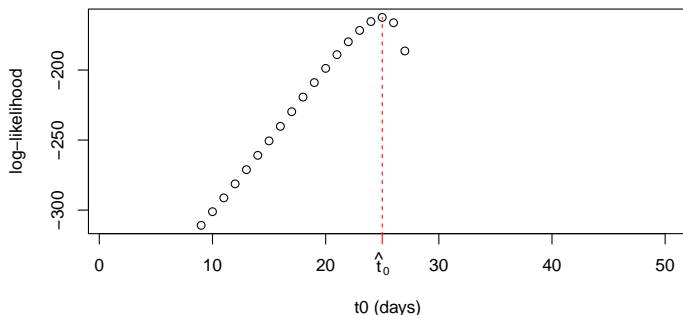
# Parametric model for outbreak Example 1

- Let  $\theta = t_0$  with the hazard model being

$$\lambda(t) = \mathbb{1}_{\{t_0\}}(t), \quad t = 1, 2, \dots$$

where  $\mathbb{1}_A(t)$  is the 0/1 indicator function which is one if  $t \in A$ .

- In this simple case  $P(T_E = t_E) = I(t_E = t_0)$ .
- Computing the log-likelihood on a grid of possible  $t_0$  values one obtains an MLE of  $\hat{t}_0 = 25$ .



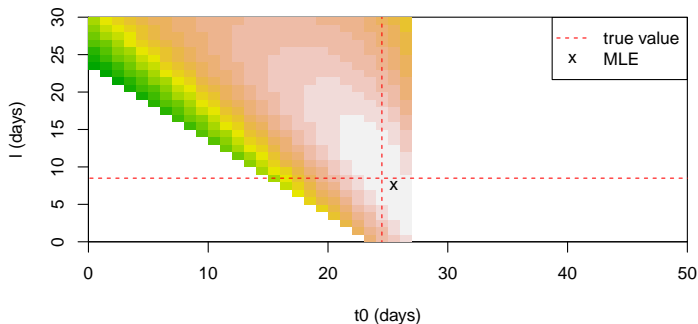


## Parametric model for outbreak Example 2

- Let  $\theta = (t_0, I)'$  and  $\lambda(t) = \frac{1}{I - t + t_0} \mathbb{1}_{[t_0, t_0 + I - 1]}(t)$  and hence

$$P(T_e = t_e) = \frac{1}{I} \cdot \mathbb{1}_{[t_0, t_0 + I - 1]}(t_e)$$

- Computing the log-likelihood on a matrix grid of possible  $(t_0, I)$  values one obtains an MLE of  $\hat{t}_0 = 26$  and  $\hat{I} = 8$ .





# Discussion

- Method does not allow for an immediate quantification of uncertainty, since the likelihood is maximized for a discrete parameter set.
- One way to provide uncertainty is to cast estimation into a Bayesian framework and thus obtain a discrete posterior distribution for  $\theta = (t_0, I)'$ .
- More complicated form of the hazard function include the use of individual (and possible time-varying) covariates, e.g. to denote position, factors influencing the transmission, etc.



# Outline

- 1 Mathematical models for communicable diseases
- 2 Modelling and monitoring public health surveillance data
- 3 The R package *surveillance*
- 4 Now-casting and back-projection
- 5 Univariate time series detectors
  - Farrington algorithm
  - Negative Binomial CUSUM
  - Binomial CUSUM
  - Evaluating performance
  - Likelihood ratio detectors



# Statistical Framework for Aberration Detection

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

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

- 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.



# Outline

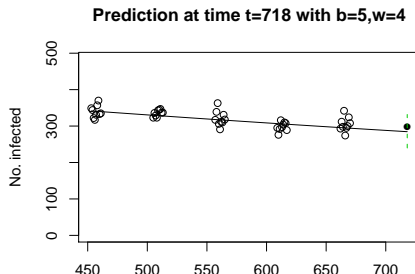
## 5 Univariate time series detectors

- Farrington algorithm
- Negative Binomial CUSUM
- Binomial CUSUM
- Evaluating performance
- Likelihood ratio detectors



# 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  $\phi > 0$ .



## 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.



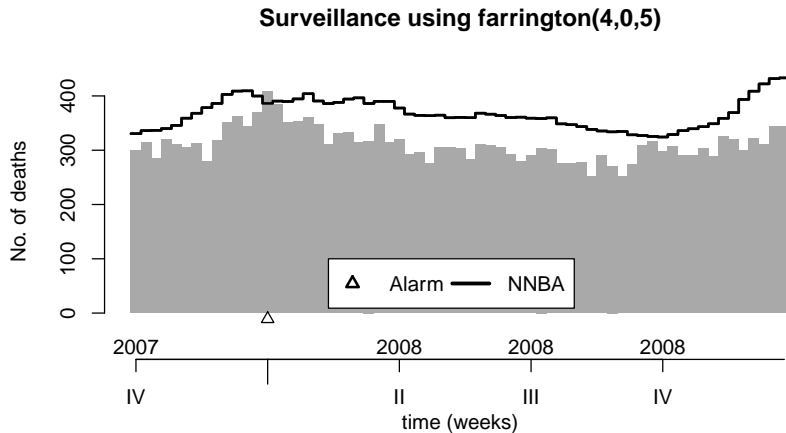
# Farrington algorithm in surveillance (1)

- 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)

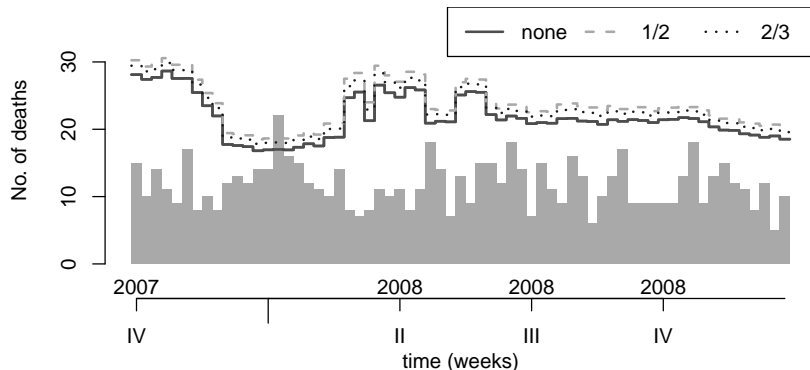
- Results for  $w = 4$ ,  $b = 5$  and  $\alpha = 0.01$  starting at W40-2007:





## 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  $\omega_t$  as

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

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

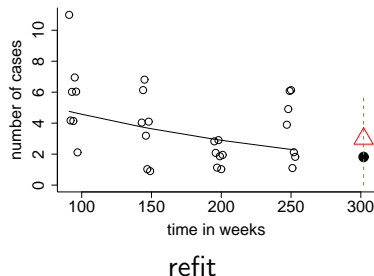
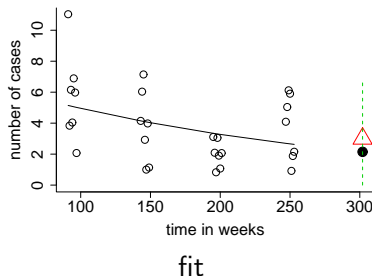


## Correcting for past outbreaks (2)

- Refit the GLM using the  $\omega_t$  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

## 5 Univariate time series detectors

- Farrington algorithm
- Negative Binomial CUSUM
- Binomial CUSUM
- Evaluating performance
- Likelihood ratio detectors



# 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  $\theta$ .

- 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  $\alpha$  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  $\text{pop}_t$  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")
```

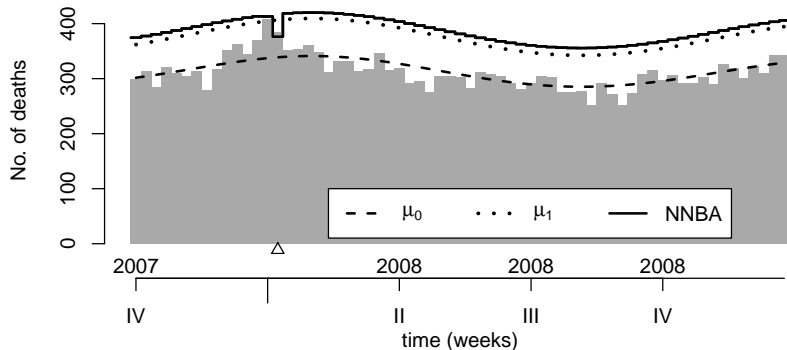
- Aim: to optimally detect a 20% increase in the mean, i.e.  $\kappa = 1.2$ .  
Use  $g = 4.75$  – consequences?

```
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:

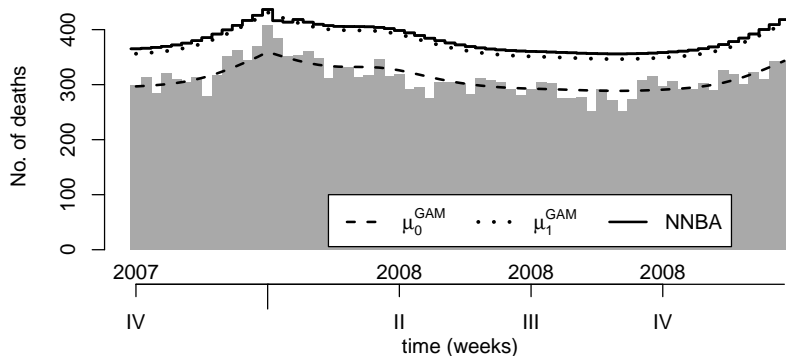


- Also shown is the number needed before alarm (NNBA), i.e. given  $r(\mathbf{y}_{s-1})$  find the minimum  $y_s$  such that  $r(\mathbf{y}_s) > g$ .



## 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

## 5 Univariate time series detectors

- Farrington algorithm
- Negative Binomial CUSUM
- Binomial CUSUM
- Evaluating performance
- Likelihood ratio detectors



## 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  $\pi_0$  to an out-of-control  $\pi_1$  by sequentially computing

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

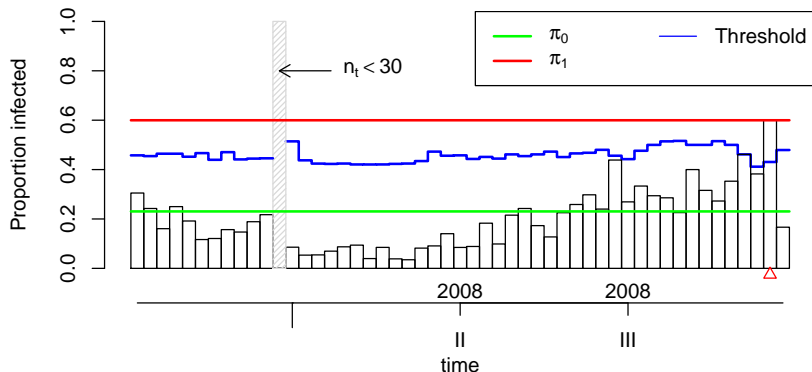
with  $C_0 = 0$  and  $k = \log \left( \frac{\pi_1(1 - \pi_0)}{\pi_0(1 - \pi_1)} \right) - \log \left( \frac{1 - \pi_1}{1 - \pi_0} \right)$ .

- 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.

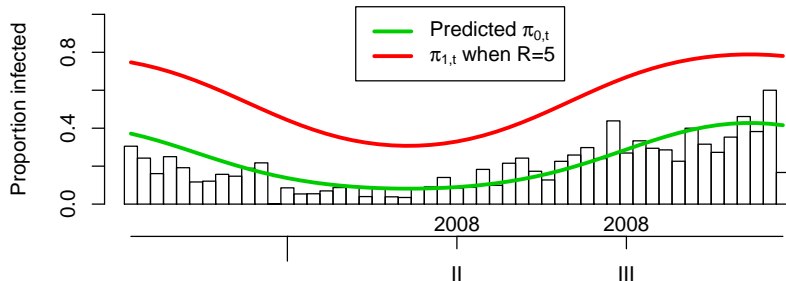


# Time varying proportion Binomial CUSUM (1)

- 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  $\beta$  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).

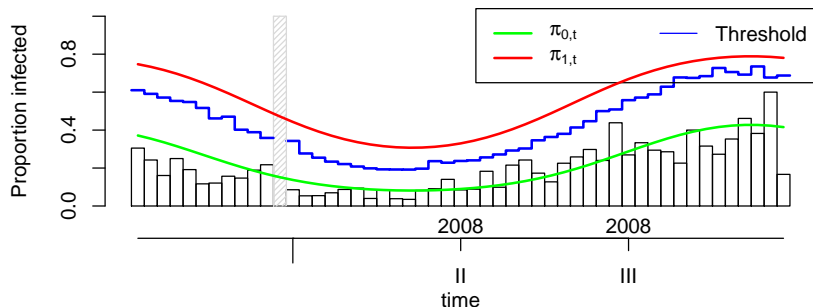


## Time varying proportion Binomial CUSUM (2)

- 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

## 5 Univariate time series detectors

- Farrington algorithm
- Negative Binomial CUSUM
- Binomial CUSUM
- Evaluating performance
- Likelihood ratio detectors



# 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:

State 0  $r_t = 0$

State  $i$   $r_t \in \left((i-1) \cdot \frac{g}{M}, i \cdot \frac{g}{M}\right]$ ,  $i = 1, 2, \dots, M$

State  $M+1$   $r_t > g$

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

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

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:

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

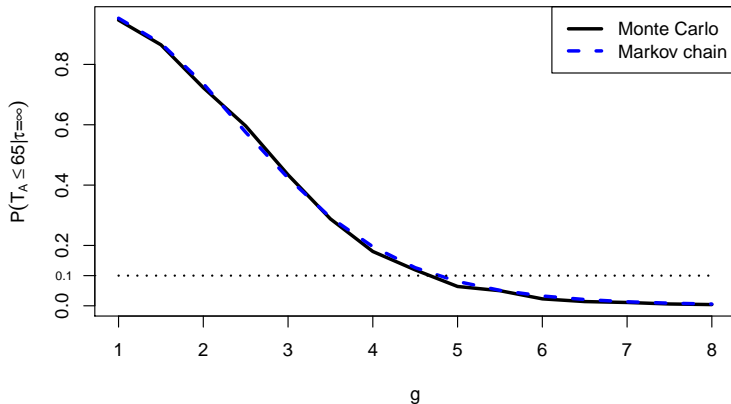
- The PMF of  $T_A$  can thus be determined by subtraction
- Now: Choose  $g$  such that  $P(T_A \leq 65 | \tau = \infty)$  is below some acceptable value, e.g. 10%.

```
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  $\alpha_t$ , 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

## 5 Univariate time series detectors

- Farrington algorithm
- Negative Binomial CUSUM
- Binomial CUSUM
- Evaluating performance
- Likelihood ratio detectors



## Generalized likelihood ratio detector (1)

- A problem of the LR scheme is that detection is only optimal for pre-specified  $\theta_1$ .
- Generalization where  $\theta_1$  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:

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

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  $GLR(s) = \max_{1 \leq k \leq s} l_{s,k}$



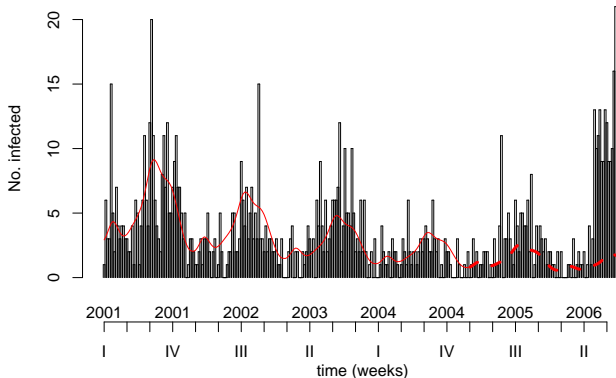
## GLR detector (3) – Poisson and negative Binomial

- 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  $l_{s,k}$  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")
```

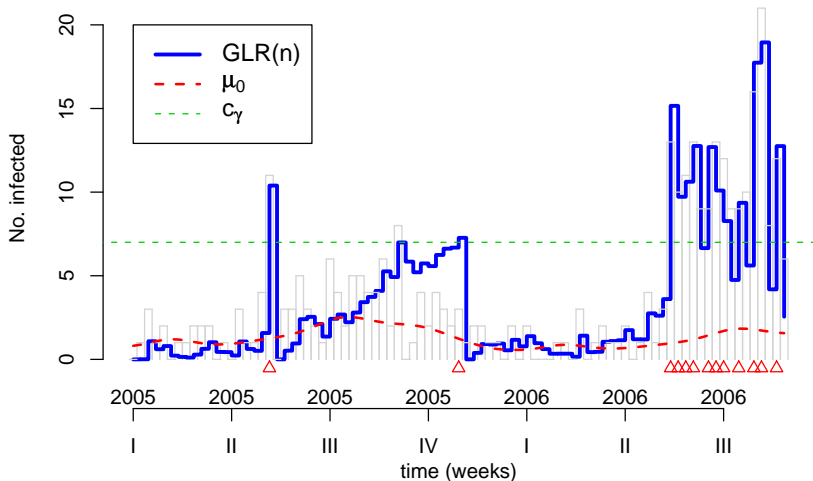
Running the detector:

```
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)
```



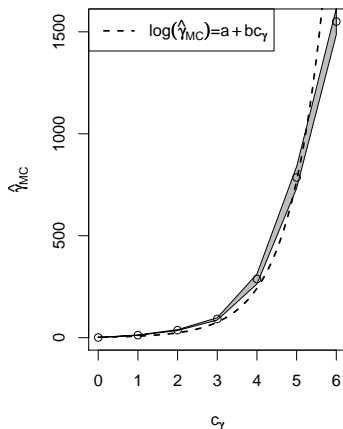
# Applying the GLR detector to salmonella hadar (3)

## Analysis of shadar using glrnb: intercept

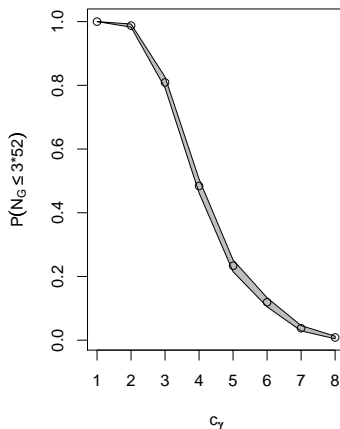




# Average run length and probability of false alarm



(a) Monte-Carlo estimated  $ARL_0(c_\gamma)$



(b)  $P(T_A \leq 3 \cdot 52 | \tau = \infty)$



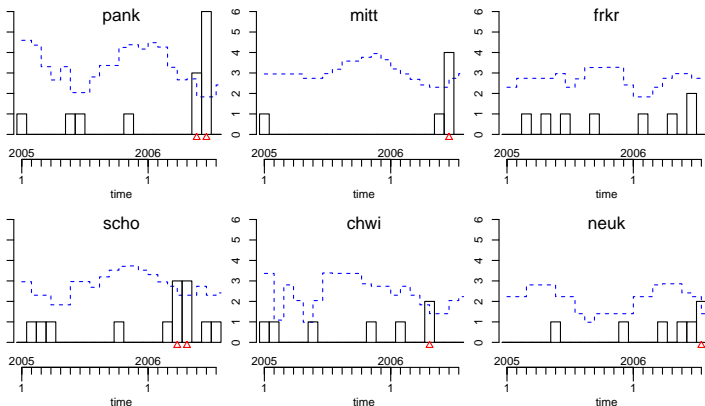
# Outline

- 1 Mathematical models for communicable diseases
- 2 Modelling and monitoring public health surveillance data
- 3 The R package `surveillance`
- 4 Now-casting and back-projection
- 5 Univariate time series detectors
- 6 Multivariate surveillance**
  - Case Study: Rabies in Hesse
  - The HHH model and its spatial extensions
- 7 Space-time point process modelling



# 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

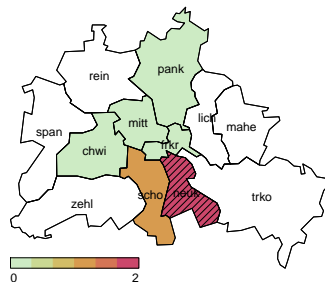
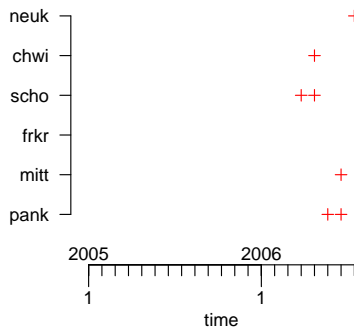


## Towards multivariate surveillance (3)

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

### Surveillance using farrington(2,0,3)

August 2006





# Outline

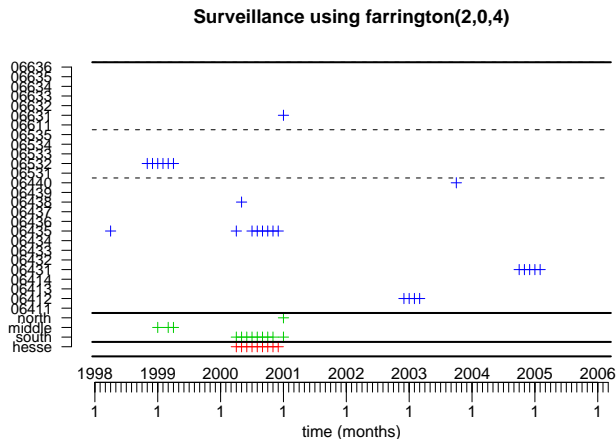
## 6 Multivariate surveillance

- Case Study: Rabies in Hesse
- The HHH model and its spatial extensions



# Rabies surveillance in Hesse

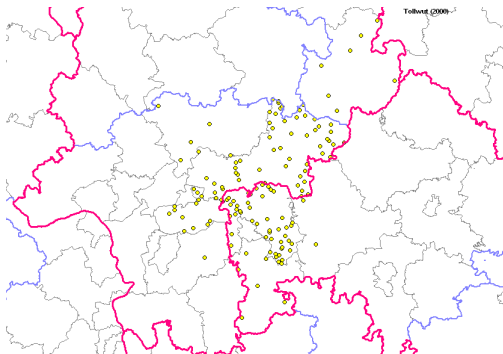
- Alarm plot created by applying the Farrington algorithm to each of 1 federal state, 3 administrative regions and 26 districts time series





## 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.

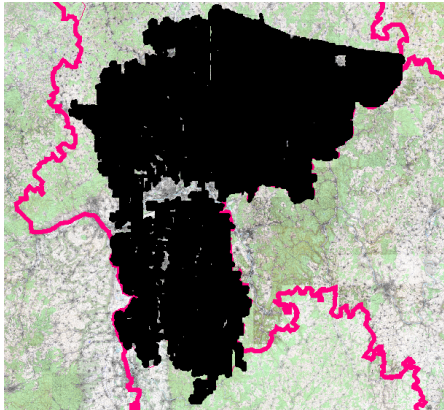


- Source of the figure: C. Staubach, FLI



## Examination of the increased number of cases (2)

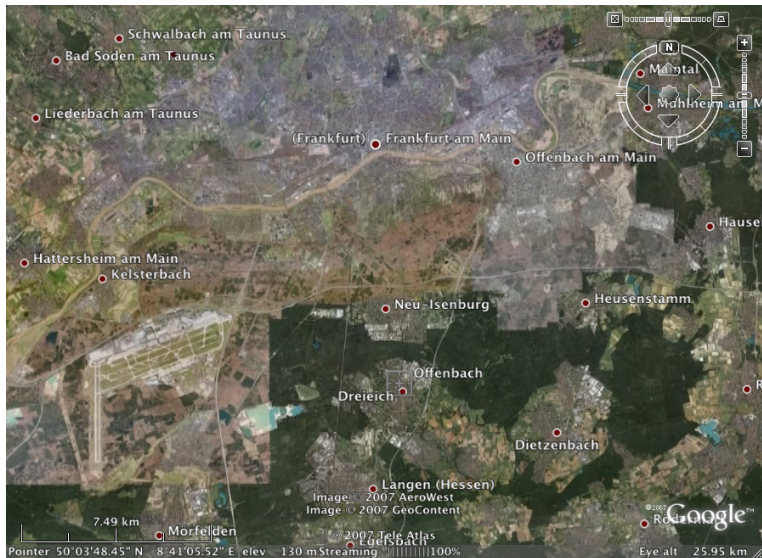
- A map with the coordinates of the baits with vaccine dropped from plane shows the problem:



- Source of the figure: T. Müller, FLI



# Examination of the increased number of cases (3)





# Outline

## 6 Multivariate surveillance

- Case Study: Rabies in Hesse
- The HHH model and its spatial extensions



## Model-based surveillance<sup>4</sup>

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))

---

<sup>4</sup>Slides 159–169 and 171–196 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  $\rightarrow$  *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:  $\psi$ ,  $\nu_t$ ,  $\lambda$  and  $\phi$  may also depend on  $i$

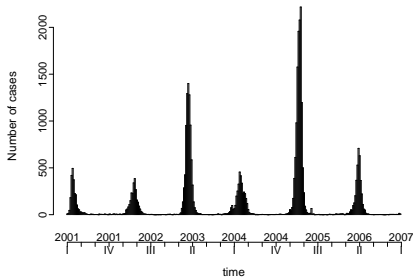


## Example: Influenza and meningococcal disease (1)

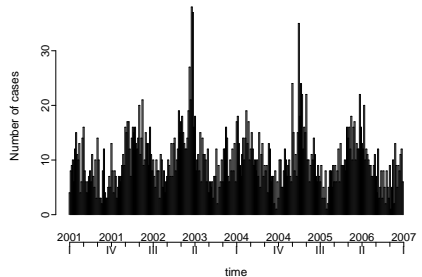
- Interdependencies between disease cases caused by *different pathogens* might be of particular interest to further understand the dynamics of such diseases
- For example, several studies describe an association between *influenza* and *meningococcal disease* (Cartwright et al., 1991; Hubert et al., 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)	$\log L(\mathbf{y}, \boldsymbol{\theta})$	$ \boldsymbol{\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)	$\log L(\mathbf{y}, \boldsymbol{\theta})$	$ \boldsymbol{\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



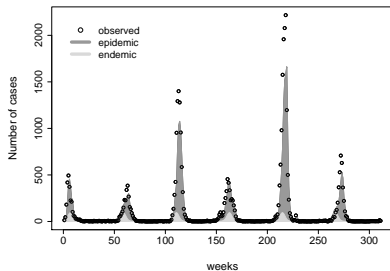
# Multivariate analyses

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)

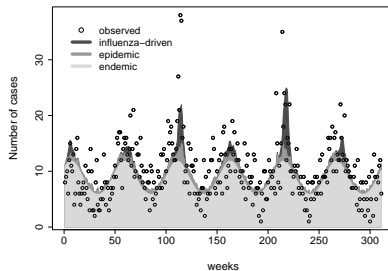
Model	$\hat{\psi}_{ML}$ (se)		$\log L(\mathbf{y}, \boldsymbol{\theta})$	$ \boldsymbol{\theta} $	AIC
	flu	men			
1	3.39 (0.41)	20.34 (4.83)	-1889.7	14	3807.5
2	3.39 (0.41)	20.34 (4.83)	-1889.7	15	3809.5
3	3.39 (0.41)	25.32 (6.98)	-1881.0	15	3791.9
4	3.40 (0.41)	25.32 (6.98)	-1881.0	16	3793.9



# 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  $\nu$ ,  $\lambda$ ,  $\phi$ , may also depend on unit  $r$  by using

- unit-specific fixed effects:

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

↪ 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  $\nu$ ,  $\lambda$ ,  $\phi$ , may also depend on unit  $r$  by using

- unit-specific fixed effects:

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

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

- linking parameters with known explanatory variables:

$$\log(\lambda_{rt}) = \beta_0 + x_{rt}\beta_1$$

↪ for instance  $x_{rt}$  = vaccination coverage in unit  $r$  at time  $t$ .



# Addressing unit-specific heterogeneity

Each of the three unknown quantities  $\nu$ ,  $\lambda$ ,  $\phi$ , may also depend on unit  $r$  by using

- unit-specific fixed effects:

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

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

- linking parameters with known explanatory variables:

$$\log(\lambda_{rt}) = \beta_0 + x_{rt}\beta_1$$

↪ 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

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

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

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

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



# 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

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



# 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  $P$ .

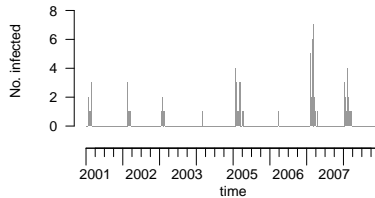
- To compare two models  $A$  and  $B$  compute  $n$  individual scores for both models and use a Monte Carlo test to assess if difference

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

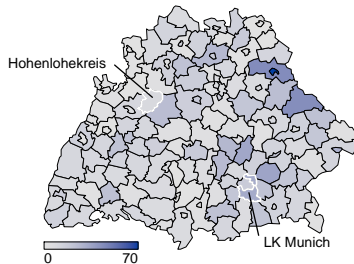
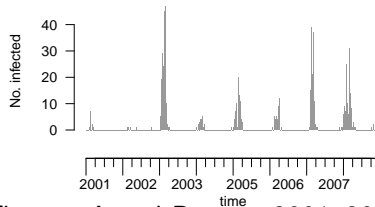
is significant.



## LK Hohenlohekreis



## LK Munich



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 $\lambda$ and random $\phi$

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

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

```
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: $\lambda$	neighbor-driven: $\phi$	$\overline{\log S}$
constant	random	<b>.563</b>
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: $\lambda$	neighbor-driven: $\phi$	$\overline{\log S}$	$p$ -value
constant	random	<b>.563</b>	
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)
```

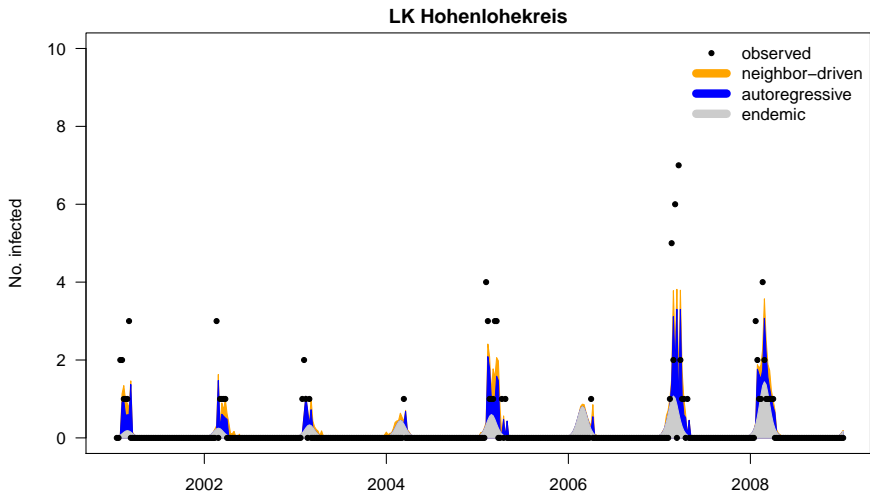
autoregressive: $\lambda$	neighbor-driven: $\phi$	$\overline{\log S}$	$p$ -value
constant	random	<b>.563</b>	
random	random	.564	.5979
random	constant	.565	.0830
constant	constant	.565	.0353
random	—	.569	.0018
constant	—	.569	.0006
—	random	.588	.0001
—	constant	.591	.0001
—	—	.599	.0001

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  $\nu_{rt}$ .

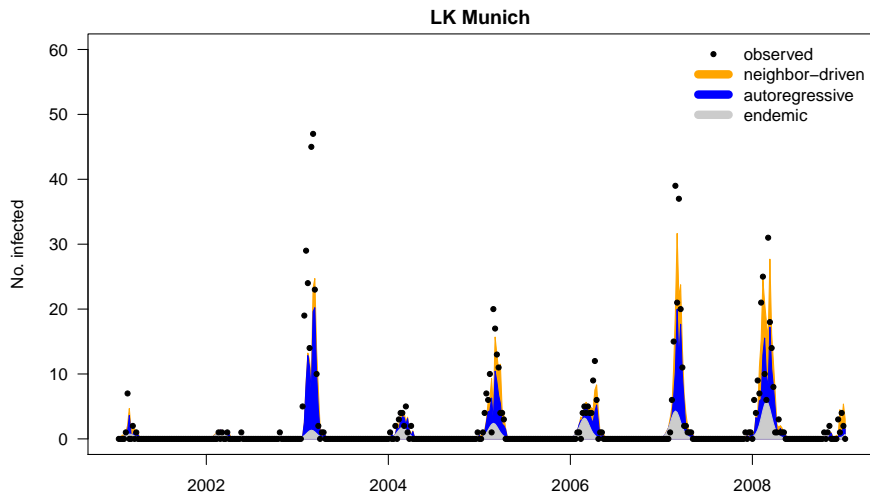


# Fitted incidence





# 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 Mathematical models for communicable diseases
- 2 Modelling and monitoring public health surveillance data
- 3 The R package `surveillance`
- 4 Now-casting and back-projection
- 5 Univariate time series detectors
- 6 Multivariate surveillance
- 7 Space-time point process modelling
  - Maximum Likelihood Inference
  - Data Analysis



# 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

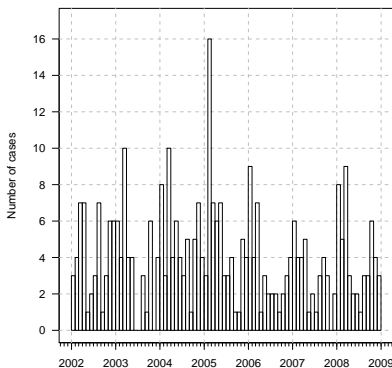
Use this regression framework as building block for model based prospective space-time aberration detection, e.g. to detect disease clusters while adjusting for trend, seasonality and other covariates



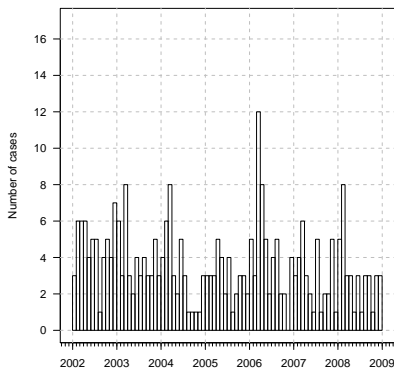
## Example: Invasive meningococcal disease (IMD)

- 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

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



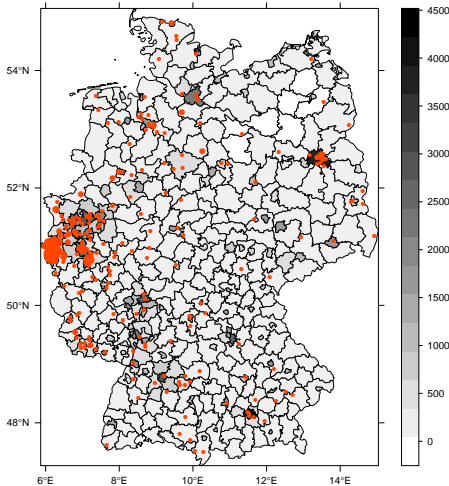
*C:P1.5,2:F3-3*



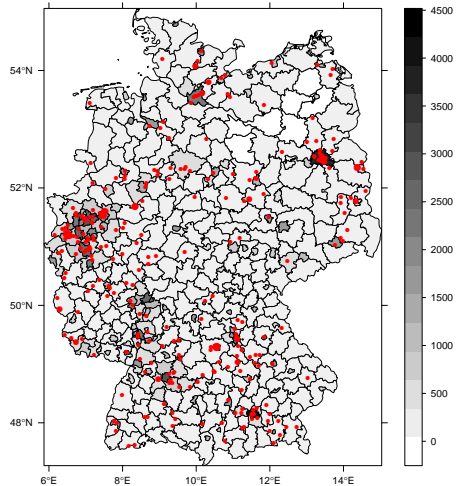


# Spatial distribution

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



C:P1.5,2:F3-3



**Scientific question:** Do the finetypes spread differently?



# Spatio-temporal animation

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

C:P1.5,2:F3-3



## 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)



# 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  $\psi$  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  $\varepsilon$  and  $\delta$



## 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

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

- 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  $\eta_j = \gamma' \mathbf{m}_j$ ,  $\kappa_j$  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  $\eta_j = \gamma' \mathbf{m}_j$ ,  $\kappa_j$  is part of  $\mathbf{m}_j$
- Type-specific interaction functions, e.g. variances  $\sigma_{\kappa}^2$



# 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

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

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



# Outline

- 7 Space-time point process modelling
  - Maximum Likelihood Inference
  - Data Analysis
  - Prospective space-time monitoring



# 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)

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

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

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

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



## Further details

- The score function is determined analytically but requires numerical integration for  $\int \frac{\partial}{\partial \sigma_l} f_{\sigma}(\mathbf{s}|\kappa) d\mathbf{s}$
- Wald confidence intervals can be computed using the asymptotic variance matrix  $\hat{\mathcal{I}}^{-1}(\hat{\theta}_{ML})$  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., 2011)



# Outline

- 7 Space-time point process modelling
  - Maximum Likelihood Inference
  - Data Analysis
  - Prospective space-time monitoring



# 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  $\geq 19$  years), gender, finetype and age-finetype interaction
- $\varepsilon = 30$  days,  $\delta = 200$  km
- Spatial interaction function  $f$ : Gaussian or constant

Resulting best AIC model:

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



# Selected joint model (1)

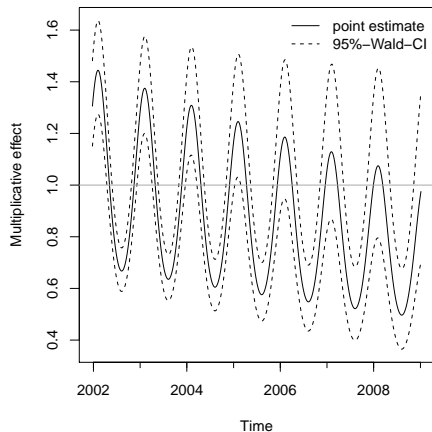
```
R> fit <- twinstim(endemic = ~1 + offset(log(popdensity)) + I(start/365) +
  sin(start * 2 * pi/365) + cos(start * 2 * pi/365),
  epidemic = ~1 + agegrp + type
  s1af = s1af_1, data = imdepi, subset = allEpiCovNonNA,
  optim.args = optim.args, method = "nlminb",
  control = list(fnscale = -10000)), nCub = 36,
  typeSpecificEndemicIntercept = 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*1*pi/365)	0.26184	0.06493	4.032	$5.52 \cdot 10^{-05}$
h. cos(start*2*1*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. s1af	2.82866	0.08191		
AIC:	18968			
Log-likelihood:	-9475			

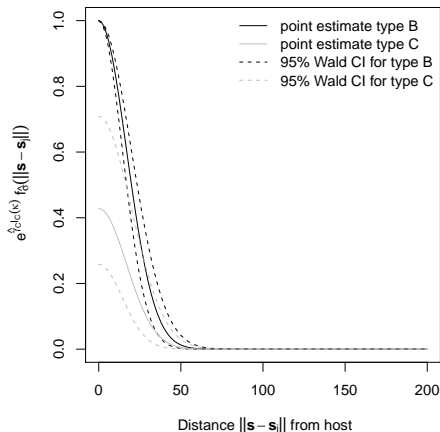


## Selected joint model

- 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  $\approx 50$  km



## Selected joint model (3) – residual analysis

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

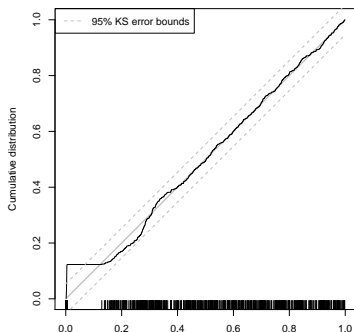
$$Y_i = \hat{\Lambda}^*(t_i) - \hat{\Lambda}^*(t_{i-1}), \quad i = 2, \dots, N,$$

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

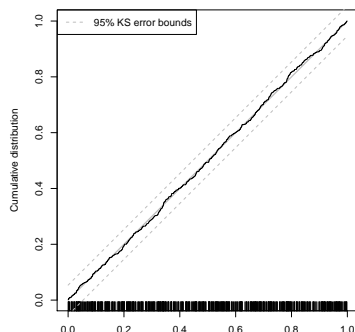
- If the estimated CIF describes the true CIF well, then

$$U_i = 1 - \exp(-Y_i) \stackrel{\text{iid}}{\sim} U(0, 1)$$

$\epsilon = 0.01$  tie breaking



$U(0, 1)$  tie breaking





# Outline

- 7 Space-time point process modelling
  - Maximum Likelihood Inference
  - Data Analysis
  - Prospective space-time monitoring



# 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 (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

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

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

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



## 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  $[t_j, t_k]$

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

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

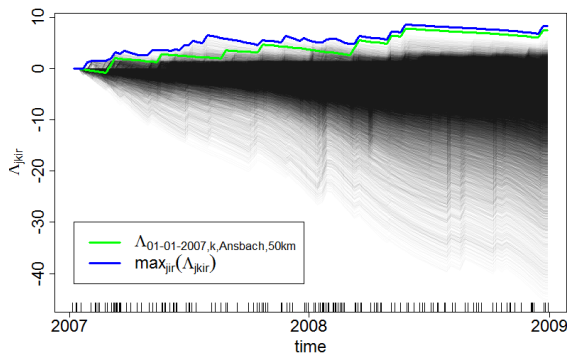
$$T_\gamma = \min_k \{SR_k > \gamma\}, \quad SR_k = \sum_{j=1}^k \exp(\Lambda_{j,k})$$

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



# Simulation example (1)

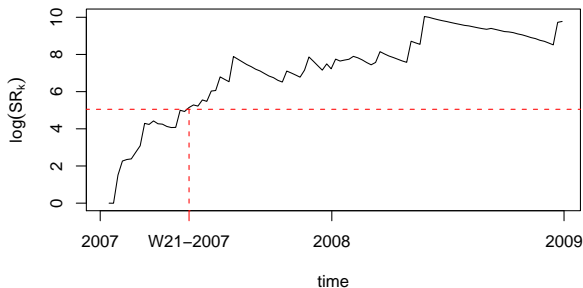
- Simulated epidemic from best AIC model with  $\delta_C = 50$  km cluster around Ansbach region starting on 01 Jan 2007 having  $\phi = \log(5)$
- Cluster detection using  $\delta_C \in \{25\text{km}, 50\text{km}, 75\text{km}\}$  and  $t_j$  in two-week intervals after 01 Jan 2007





## Simulation example (2)

- Resulting Shiryaev-Roberts statistic

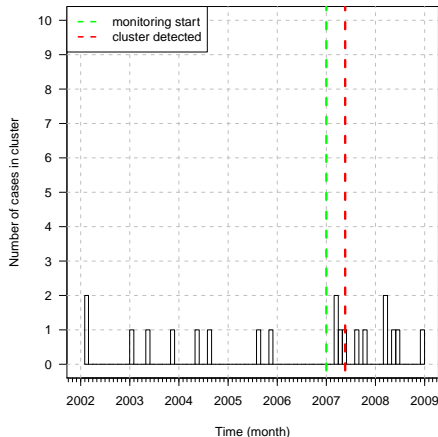
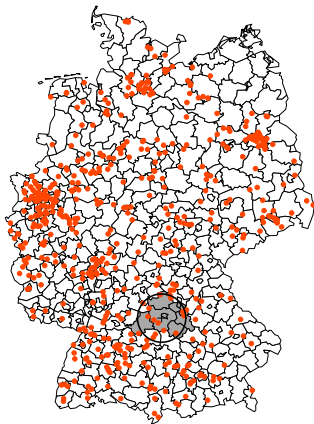


- Using  $\gamma = 52 \cdot 3$  results in an alarm at  $t_{20}$  (W21-2007) with cluster location defined as the cluster producing  $\max_{j=1}^{20} \exp(\Lambda_{j,20})$ , i.e. here  $C=(\text{Ansbach}, 50\text{km}, \text{W09-2007})$



## Simulation example (3)

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



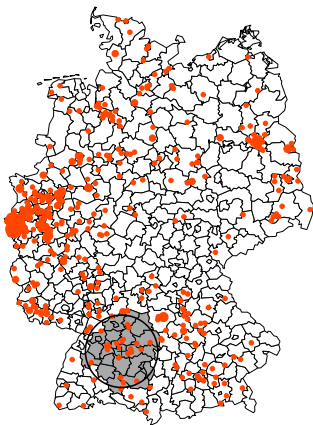
Location of cluster (grey) at W21-2007

Time series with cases in cluster

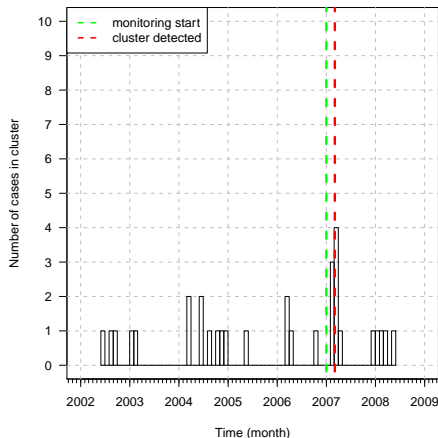


## Cluster detection for IMD data

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. (2011) 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 Mathematical models for communicable diseases
- 2 Modelling and monitoring public health surveillance data
- 3 The R package `surveillance`
- 4 Now-casting and back-projection
- 5 Univariate time series detectors
- 6 Multivariate surveillance
- 7 Space-time point process modelling



# 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

- an der Heiden, M., Wadl, M., and Höhle, M. (2011). Now-casting during a huge outbreak of haemolytic-uremic syndrome in Germany, 2011. In *Proceedings of the European Scientific Conference on Applied Infectious Disease Epidemiology (ESCAIDE)*. To appear.
- Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and their Statistical Analysis*, volume 151 of *Springer Lectures Notes in Statistics*. Springer-Verlag.
- 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.
- Bacchetti, P., Segal, M. R., and Jewell, N. P. (1993). Backcalculation of HIV infection rates. *Statistical Science*, 2:82–119.
- Becker, N. G. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall/CRC.
- Becker, N. G., Watson, L. F., and Carlin, J. B. (1991). A method of non-parametric back-projection and its application to AIDS data. *Statistics in Medicine*, 10:1527–1542.
- 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.



## Literature II

- 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.
- Brookmeyer, R. and Gail, M. (1988). A method for obtaining short-term projections and lower bounds on the size of the aids epidemic. *Journal of the American Statistical Association*, 83:301–308.
- 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.
- 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.
- Daley, D. J. and Gani, J. (1999). *Epidemic Modelling: An introduction*. Cambridge University Press.
- Diggle, P. J., Rowlingson, B., and Su, T.-L. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, 16:423–34.



## Literature III

- 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.
- 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.



# Literature IV

- Höhle, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics. *Biometrical Journal*, 51(6):961–978.
- Höhle, M. (2010). Change-point 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.
- 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.



# Literature V

- 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.
- Marschner, I. and Watson, L. (1994). An improved ems algorithm for back-projection of aids incidence data. *Journal of Statistical Computation and Simulation*, 50:1–20.
- Meyer, S., Elias, J., and Höhle, M. (2011). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*. To appear.
- 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.
- 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.



## Literature VI

- 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.
- Robert Koch Institute (2011). Abschließende Darstellung und Bewertung der epidemiologischen Erkenntnisse im EHEC O104:H4 Ausbruch. Technical report, Robert Koch Institute. Available from <http://www.rki.de>.
- 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/>.
- 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.