A new method for real-time outbreak detection

Michaela Paul

Biostatistics Unit
Institute for Social and Preventive Medicine
University of Zurich

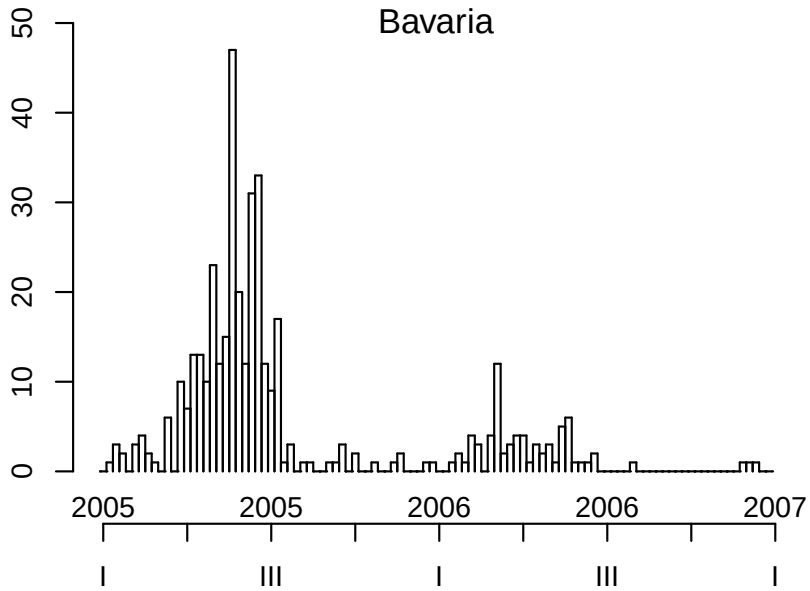
July 2 2009

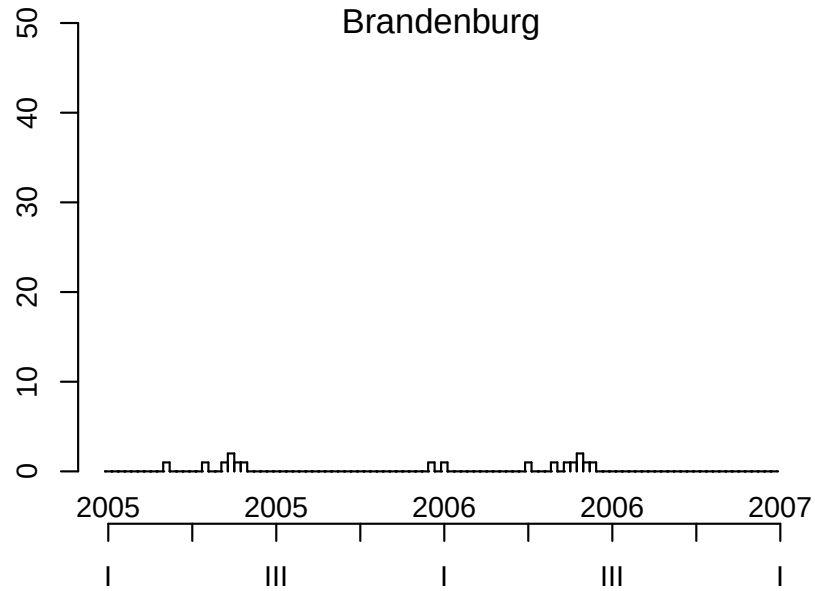
This talk summarises work done during a research visit with Paddy Farrington at the Open University in Milton Keynes, UK.

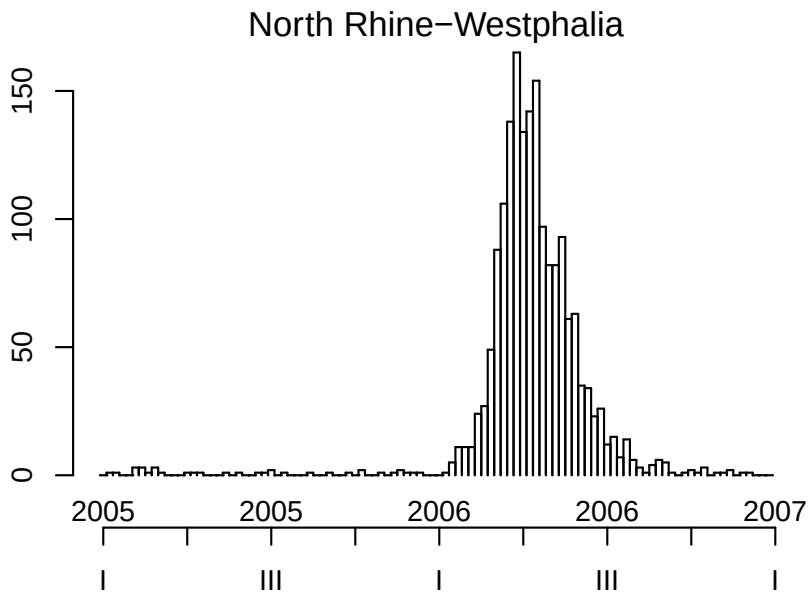
Outline

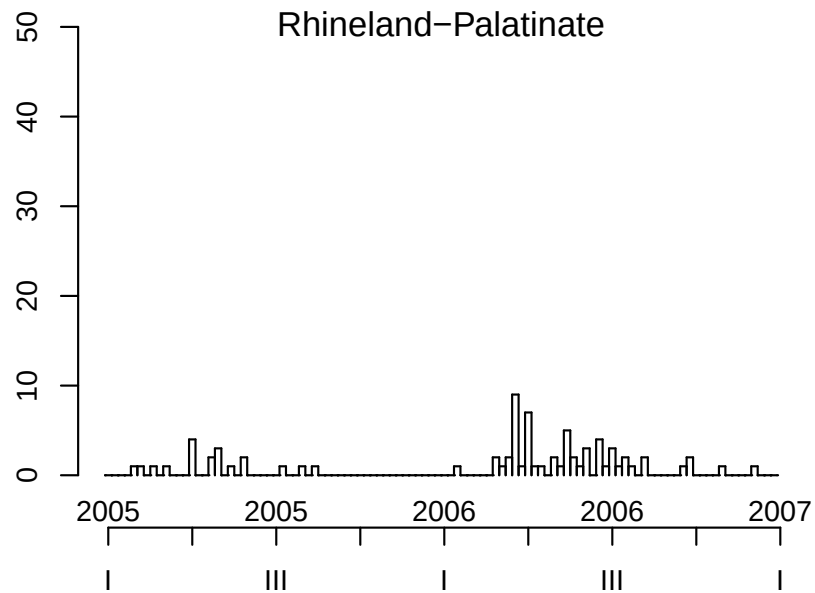
Measles in Germany

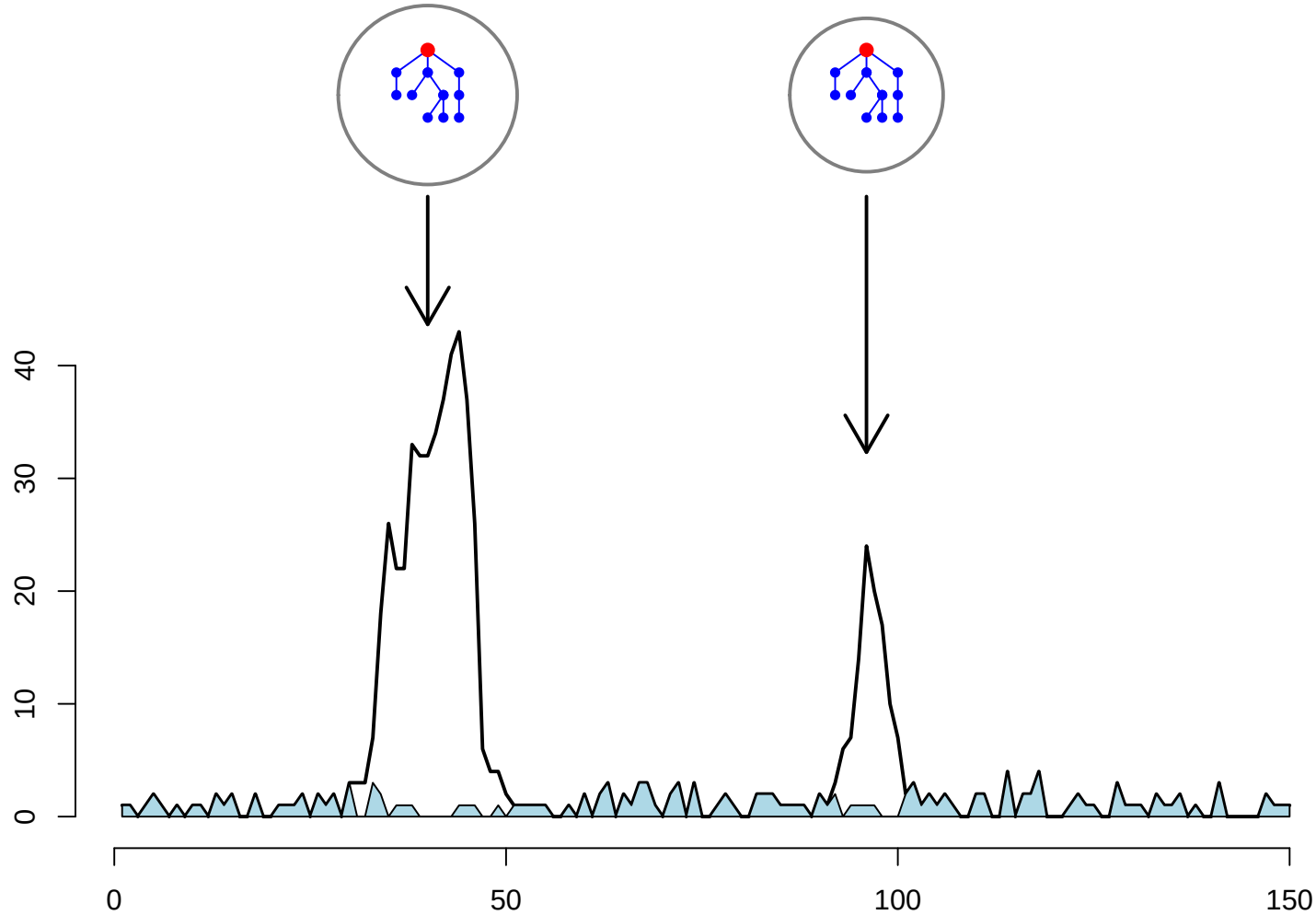
- Measles is a highly contagious disease.
- The introduction of the measles vaccines has considerably lowered the incidence level in Germany to a historical low of 0.2 cases per 100 000 inhabitants in 2004.
- However, large local outbreaks occurred in some of the federal states in recent years.
- This is most likely due to very heterogeneous vaccination coverage rates: e.g. rates range from 58% to 96% in the 77 districts of Bavaria.


Bavaria
0
10
20
30
40
50
2005
2005
2006
2006
2007
I
III
I
III
I


Brandenburg
0
10
20
30
40
50
2005
2005
2006
2006
2007
I
III
I
III
I


North Rhine−Westphalia
0
50
100
150
2005
2005
2006
2006
2007
I
III
I
III
I


Rhineland−Palatinate
0
10
20
30
40
50
2005
2005
2006
2006
2007
I
III
I
III
I



The importation of a case may lead to local outbreaks in which the transmission rate is (much) higher than usual.

Model

$$y_t | y_{t-1} \sim \text{Po}(\lambda y_{t-1} + \nu)$$

where

y_t number of cases at time t

λy_{t-1} "outbreak" cases

ν "endemic" cases

Idea

- Let the autoregressive parameter λ "jump" between two values λ_0 and λ_1 depending on the absence or presence of an outbreak at time t.
- Construct a method for prospective outbreak detection using this model.

Algorithm

Fix a threshold value $h > 0$ and proceed iteratively through "training data" $y_1, \ldots, y_n$:

At time t we have

- estimate $\hat{\theta}_t(h) = (\hat{\lambda}_{0t}, \hat{\lambda}_{1t}, \hat{\nu}_t)$ of $\theta = (\lambda_0, \lambda_1, \nu)$
- status $S_t = \begin{cases} 0 & \text{if no outbreak at time } t \\ 1 & \text{if outbreak at time } t \end{cases}$

Algorithm

Do until $t = n$, the end of the training set:

1. Let $\hat{\mu}^0_{t+1} = \hat{\lambda}_{0t} y_t + \hat{\nu}_t, \quad \hat{\mu}^1_{t+1} = \hat{\lambda}_{1t} y_t + \hat{\nu}_t$
2. Observe y_{t+1} and set
   - $S_{t+1}(h) = \begin{cases} 0 & \text{if } y_{t+1} - \hat{\mu}^0_{t+1} \leq h \\ 1 & \text{if } y_{t+1} - \hat{\mu}^0_{t+1} > h \end{cases}$
   - $\hat{\mu}_{t+1} = \hat{\mu}^{S_{t+1}(h)}_{t+1}$
3. Maximise the conditional likelihood of $y_1, \ldots y_{t+1} | S_1(h), \ldots, S_{t+1}(h)$ to obtain

$$\hat{\theta}_t(h) = \arg\max_{\theta} \sum_{i=1}^{t+1} \log \mathrm{L}(\theta; y_i | s_i(h))$$

Choice of threshold h

Repeat the algorithm for a grid of threshold values and "optimise" the fit with respect to h.

The threshold should be chosen such that (ideally)

- all outbreaks are detected
- alarms are given at the beginning of an outbreak
- no false alarms are given

Performance measures

sensitivity $\frac{\text{number of outbreaks that are flagged}}{\text{total number of outbreaks}}$

specificity $\frac{\text{number of non-outbreaks that are not flagged}}{\text{total number of non-outbreaks}}$

PPV $\frac{\text{number of correctly flagged outbreaks}}{\text{total number of alarms}}$

delay time to detection, i.e. the period of time between the start of an outbreak and the first alarm

Note: Sensitivity depends on the definition of an outbreak

1. outbreak := week with aberrant cases
2. outbreak := period of consecutive weeks with aberrant cases

Characteristics of the method

Outbreak detection methods have to deal with the presence of outbreaks in the 'training data' when determining a 'null model' for the non-outbreak situation.

Here we attempt to model outbreaks instead of treating them just as outliers in a null model.

- Past outbreaks should have less impact on the null model.
- Does this lead to a better detection of outbreaks?

- Criteria of fit are available.
- Are those criteria helpful for choosing threshold parameters?

Evaluation

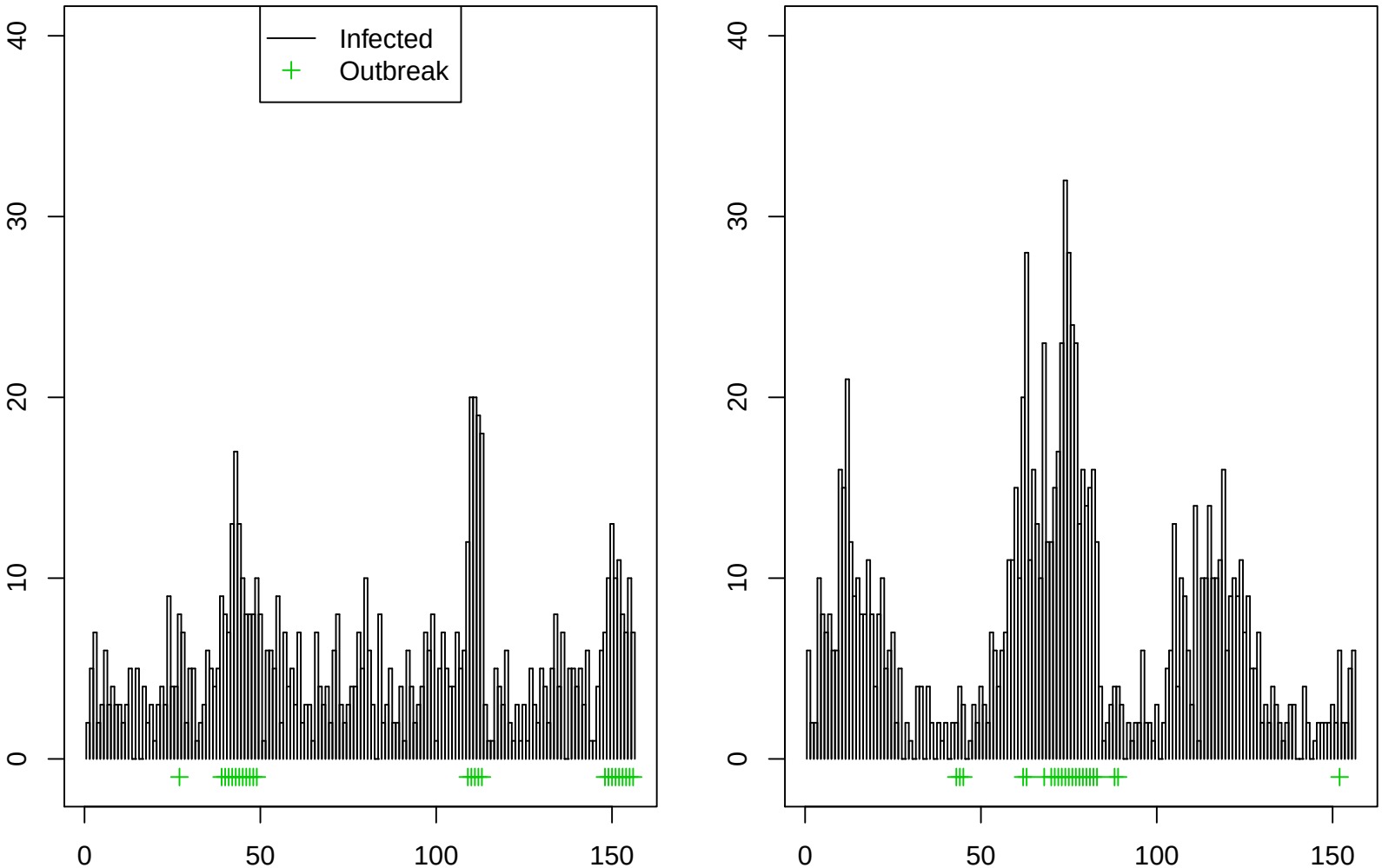
To choose a threshold and evaluate the performance, data $y_t = x_t + z_t$, $t = 1, \ldots, T$ are simulated:

- background counts: $x_t \sim \text{Po}\left(\alpha + A\sin(\frac{2\pi(t+\varphi)}{52})\right)$
- outbreak counts: $z_t \sim \text{Po}(\lambda_t y_{t-1})$,

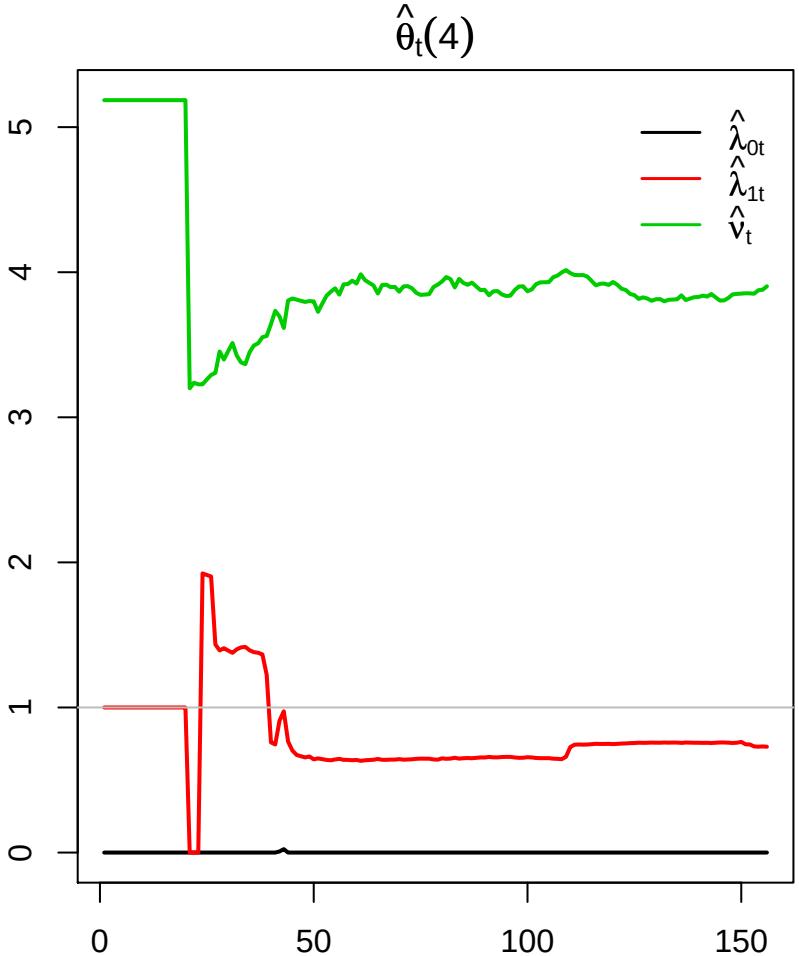
  where the value of λ_t is randomly assigned to be either $\lambda_0 = 0$ or $\lambda_1 = 0.8$.

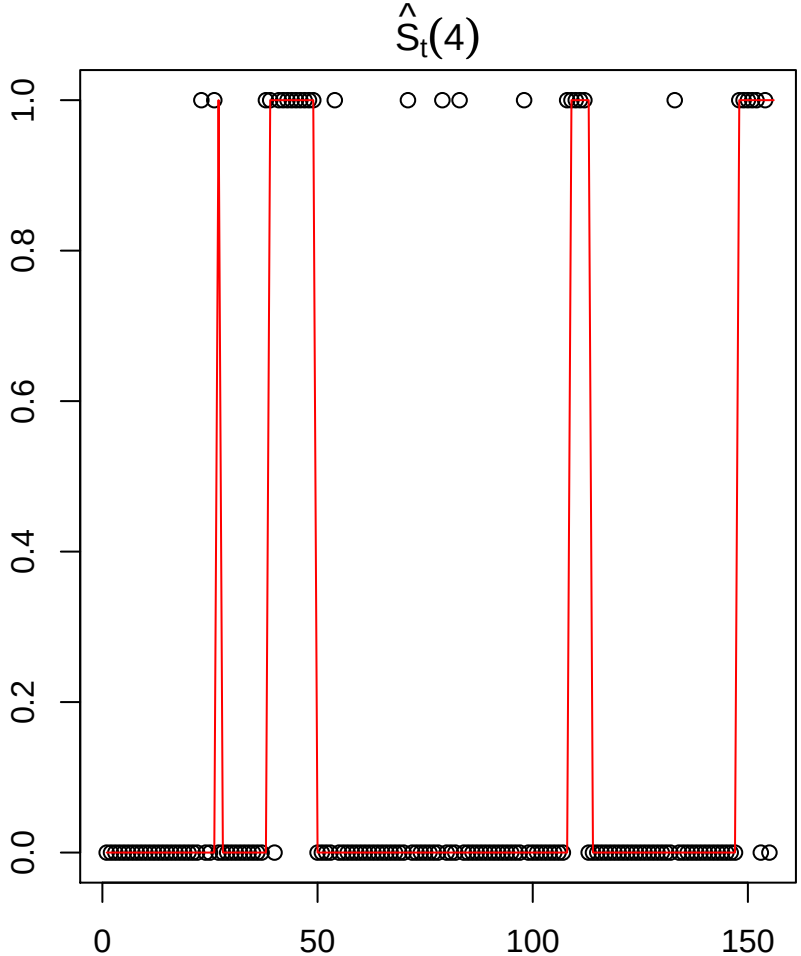
If $\lambda_t = \lambda_1$ and $z_t > 0$, week t is defined to be an outbreak week, otherwise week t is a non-outbreak week.

Examples of simulated data

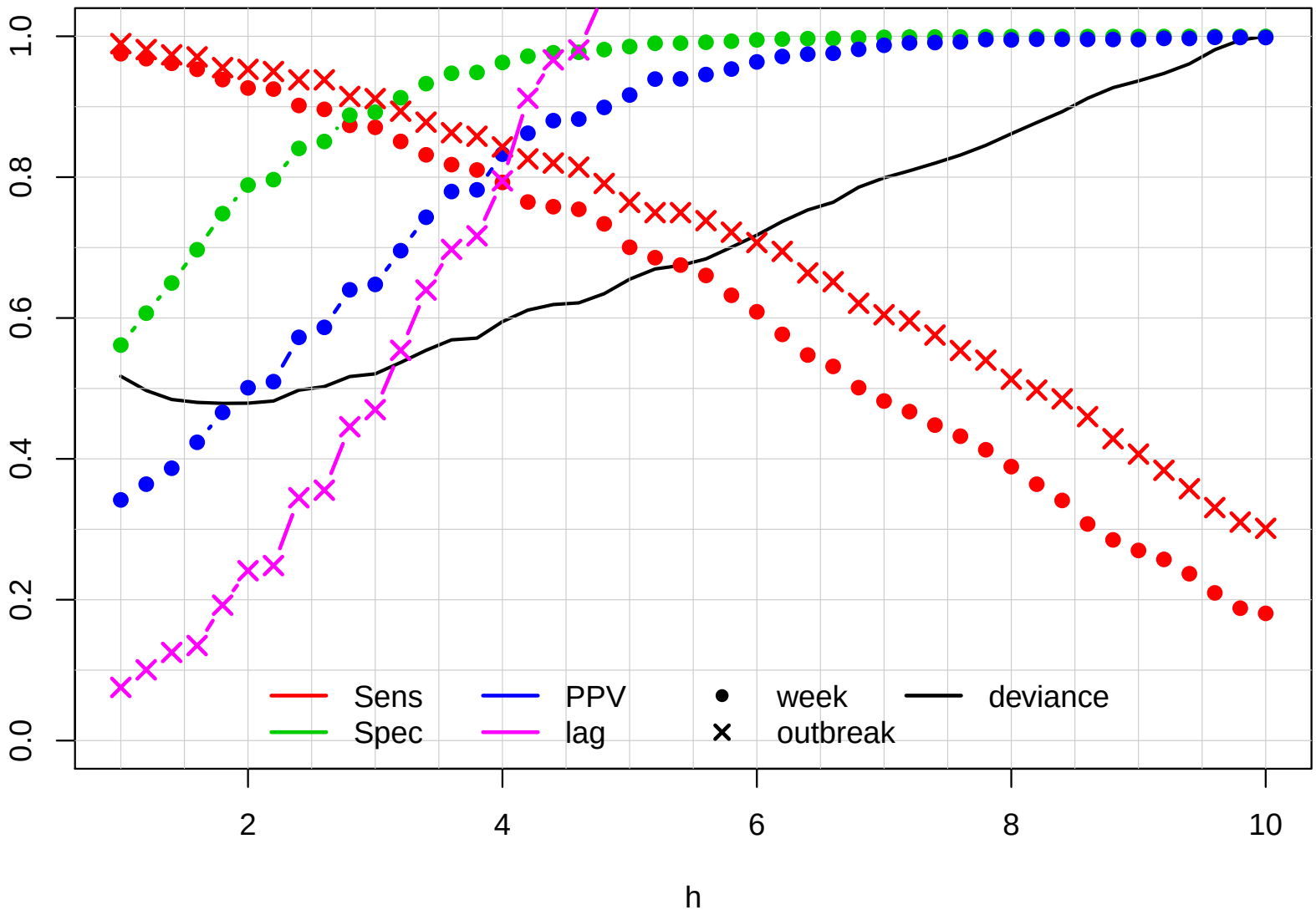


Performance for $y_t \sim \text{Po}(\lambda_t y_{t-1} + 4)$, $h = 4$

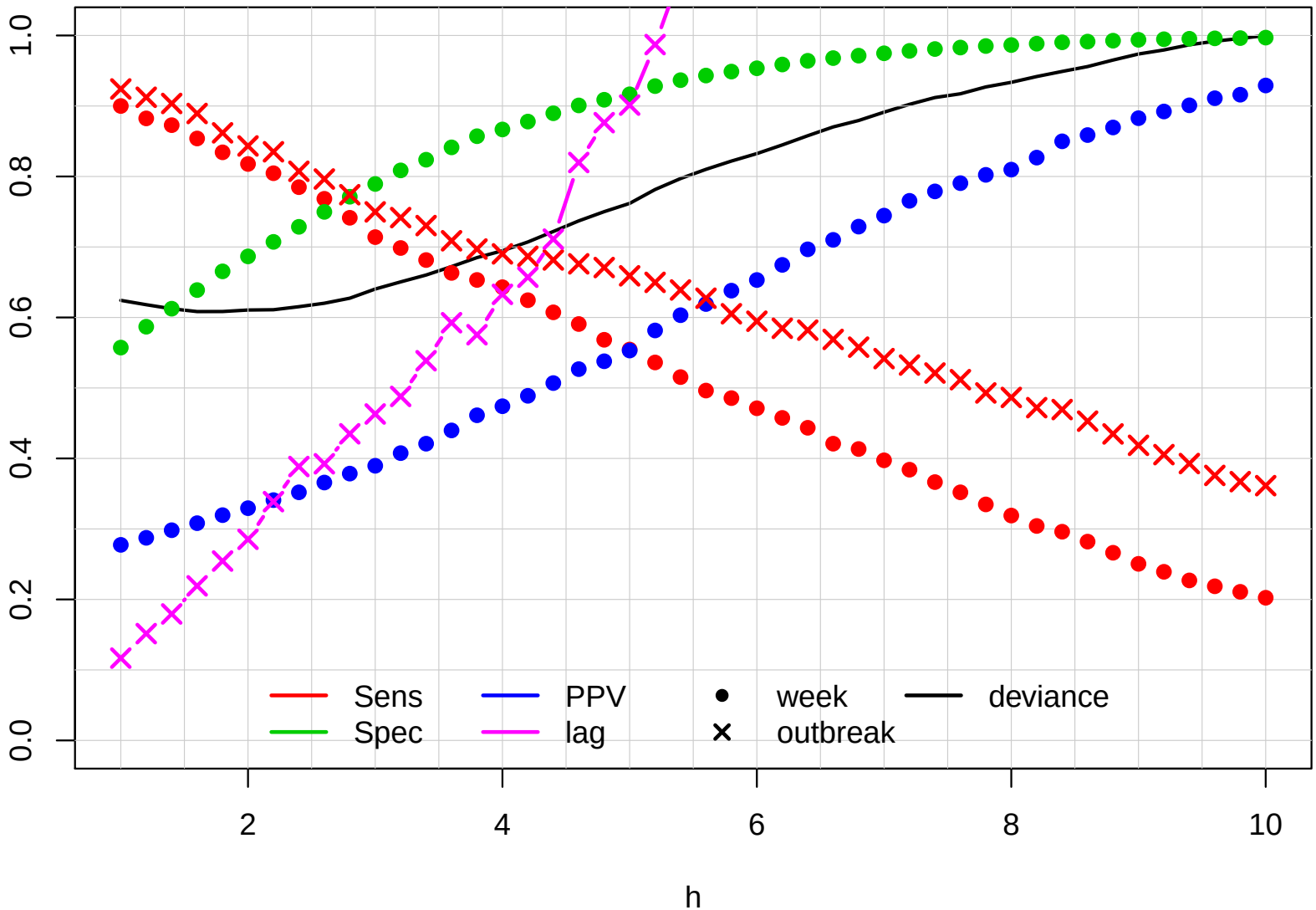




Choice of threshold: $y_t \sim \text{Po}(\lambda_t y_{t-1} + 4)$



$$y_t \sim \text{Po}(\lambda_t y_{t-1} + 4 + \sin(2\pi t/52))$$



Exemplary comparison to CDC method

The CDC method compares the number of cases in the 4 most recent periods with historical incidence data based on reference values from the preceding 5 years

- $y_t \sim \text{Po}(\lambda_t y_{t-1} + 4)$
- simulation of 10 years, last 5 years are used for evaluation
- threshold $h = 4$

Results (using outbreak week definition):

	Sens	Spec	PPV	lag
h=4	55.5	96.9	77.3	5.7
CDC(5y,4w)	37.4	97.0	73.0	12.8

To-do's

Issues that should be further investigated comprise

- comparing the performance to other methods
- assessing the performance with respect to outbreak type
- checking the robustness with respect to the data generating process
- dealing with trends, seasonality, overdispersion
- assessing the degree of "automation"