Supervised Learning for Automated Infectious-Disease-Outbreak Detection

ISDS Conference 2019
San Diego, 29 January 2019

Benedikt Zacher, Alexander Ullrich, **Stéphane Ghozzi**

Robert Koch Institute, Germany
ghozzis@rki.de

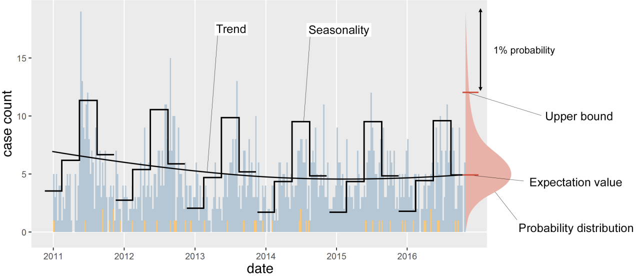
Outline

1. Outbreak Detection as Binary Classification
2. Outbreak Labels: Statistical Description
3. Supervised Learning: Two Simple Approaches
4. Evaluating and Comparing Algorithms
5. Conclusion and Outlook

1. Automated Outbreak Detection as Binary Classification

"Are there too many cases, here and now, compared with expectations?"

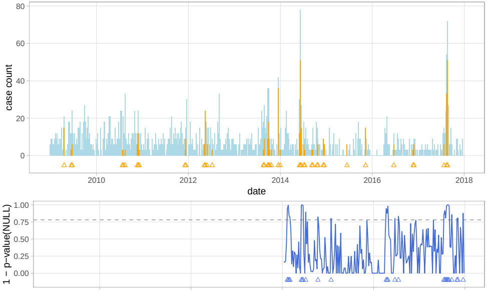
One standard approach: Univariate time series + Regression + Confidence Interval



For example:
farringtonFlexible (from R-package *surveillance*), used here for benchmarking

Noufaily et al (2013) Statistics in Medicine 32(7) 1206 http://doi.org/10.1002/sim.5595

Salmon et al (2016) Journal of Statistical Software 70(10) http://doi.org/10.18637/jss.v070.i10



label △ = week with outbreak

signal △ =
1 - P-value("no outbreak") >
cut-off

Idea 1: learn what's an outbreak from the labels

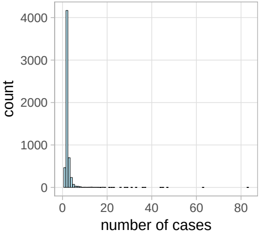
Idea 2: evaluate how good the signals are:
- signal & week with outbreak = true positive **TP**
- signal & week without outbreak = false positive **FP**
- no signal & week without outbreak = true negative **TN**
- no signal & week with outbreak = false negative **FN**

2. Outbreak Labels: Statistical Description

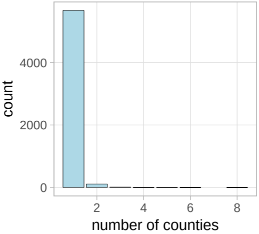
In Germany:
Outbreaks are reported, individual infection **cases are labelled with an outbreak ID**

Reported outbreaks for food-borne diseases are particularly reliable:
campylobacteriosis and salmonellosis

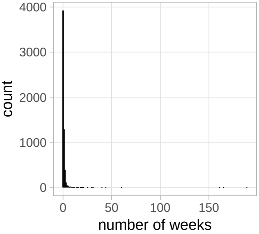
Size of outbreaks:



Extent of outbreaks:



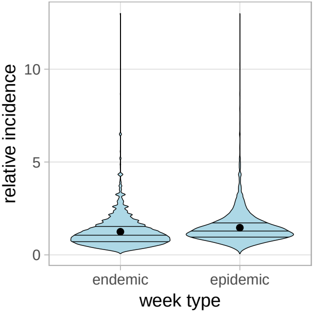
Duration of outbreaks:



Outbreaks are typically **small**, **local**, **short lived** $\Longrightarrow$ point detection might be OK

https://www.rki.de/DE/Content/Infekt/Jahrbuch/Jahrbuch_2017.pdf?__blob=publicationFile

Weekly incidences relative to 13-weeks window (only weeks with cases)



on average: outbreaks are additional cases. . . but *many* outbreaks are subcritical

simple univariate methods might not work well. . . let's use the outbreak information!

3. Supervised Learning: Two Simple Approaches

1. farringtonOutbreak

   farringtonFlexible but outbreak cases removed from training

   cut-off on 1 - P-value("no outbreak")

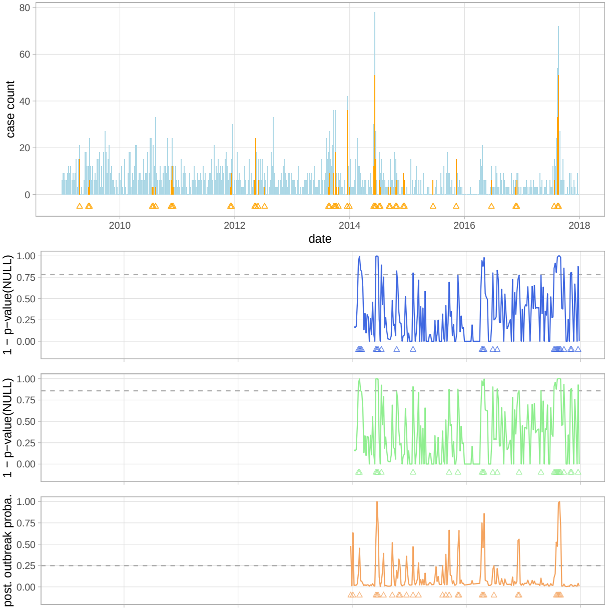
2. hmmOutbreak

   - hidden state $s_t \in \{0,1\}$ ($= 1$ if outbreak in week t, else $= 0$)
   - transition probabilities $a_{ij} = \sum_t \delta_{i\, s_{t-1}} \delta_{j\, s_t} / \sum_t \delta_{i\, s_{t-1}}$
   - emission function $c_t \sim \psi$ NegBin with

     $$\log \mu_t = \beta_0 + \sum_{i=1}^{3} \beta_i\, t^i + \beta_4 \cos\left(\tfrac{2\pi}{52}\, t\right) + \beta_5 \sin\left(\tfrac{2\pi}{52}\, t\right) + \beta_6\, s_t,$$

     and constant over-dispersion
   - posterior outbreak probability (one-week ahead: one-step forward algorithm)

     $$p_t = a_{s_{t-1}1} \cdot \psi(c_t; s_t = 1, t) / \sum_{i=0,1} a_{s_{t-1}i} \cdot \psi(c_t; s_t = i, t)$$
   - **cut-off** on p_t



farringtonFlexible, farringtonOutbreak, hmmOutbreak

4. Evaluating and Comparing Algorithms

- Data:

  weekly reported infection cases and outbreaks for notifiable diseases in Germany

  1 time series for each county

  with frequency of weeks with outbreaks between 2% and 98%

  time range 2009-2017 = 8 years

- Training and test sets = 5 years + 1 week

  training = 5 years

  test on next week (prospective 1 week ahead: data available until last week)

- Scores = functions of *TP*, *FP*, *TN*, *FN*

  sensitivity, specificity, precision, F1. . .

Enki et al (2016) PLOS ONE 11(8) e0160759 http://doi.org/10.1371/journal.pone.0160759

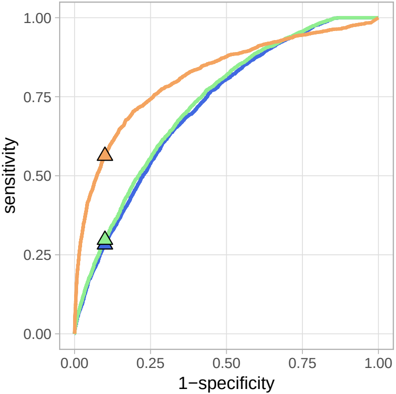
Bédubourg, Le Strat (2017) PLOS ONE 12(7) e0181227 http://doi.org/10.1371/journal.pone.0181227

Hoffmann, Dreesman (2010) PAE-project report, Niedersächsische Landesgesundheitsamt (NLGA) / ESCAIDE poster

Ghozzi, Ullrich, in preparation

Evaluation 1: with varying cut-off

ROC curve (sensitivity vs. 1-specificity): sensitivity $= TP/(TP + FN)$, specificity $= TN/(TN + FP)$



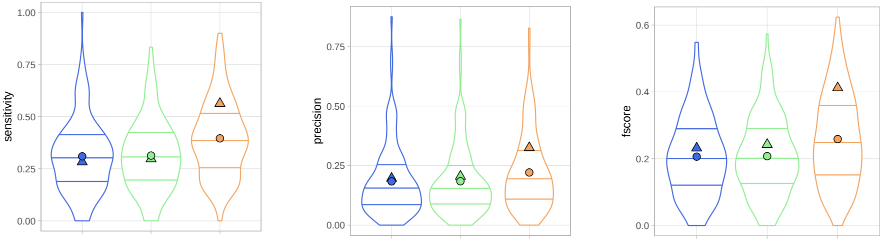
farringtonFlexible, farringtonOutbreak, hmmOutbreak

Evaluation 2:
cut-offs set so that specificity $= 0.9$ on each time series (and overall as well)

sensitivity

precision
$= TP/(TP + TF)$

F1 score
$= 2TP/(2TP + FP + FN)$



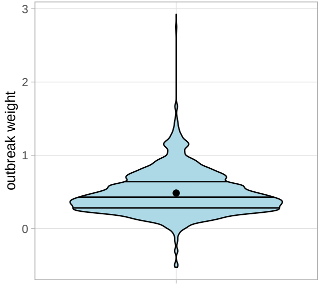
farringtonFlexible, farringtonOutbreak, hmmOutbreak

distributions with 25th, 50th and 75th percentiles; ● = mean, ▲ = overall

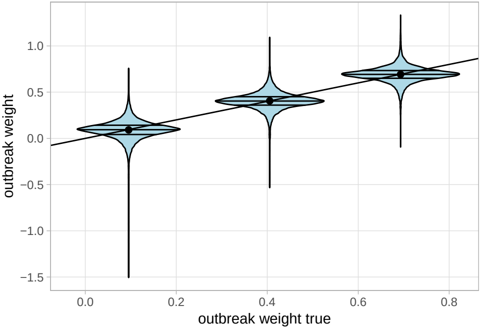
Dynamical properties can be inferred from hmmOutbreak, for example:

Outbreak weight β_6 (weeks with outbreaks have e^{β_6} more cases):

Campylobacteriosis



Simulations



For campylobacteriosis:
- weeks with outbreaks indeed have significantly more cases
- on average $e^{0.5} \approx 1.6$ more cases in outbreak weeks, all other things equal

5. Conclusion and Outlook

- supervised learning is a **promising** venue for outbreak detection!
  - labelled data are available
  - simple HMM more transparent (explicit proba) and performs better
- account for **delays** in reporting and labelling
- hyper-parameter **optimisation** + stacking (combine algorithms)

$\Longrightarrow$ Framework for machine learning:

- devise, optimise, combine algorithms **based on expert knowledge**
- integrate **continuous user feedback:** signal evaluation, reinforcement learning
- towards a **standard data set** (with labels) for outbreak detection

Ghozzi, Ullrich, in preparation

Zacher, Czogiel, in preparation

Busche, Ullrich, Ghozzi, in preparation

Thank you!

see also talk

“Dashboards as strategy to integrate multiple data streams for real time surveillance”

by Alexander Ullrich

Friday, Feb. 1, 2019 / 10:00 am / Rio Vista F room

ROBERT KOCH INSTITUT

SIGNALE

signale@rki.de

rki.de/signale-project