# Automatic Signal Detection

Lectures in Infectious-Disease Epidemiology
Robert Koch Institute
21 January 2019

Stéphane Ghozzi

RKI – Signale/Unit 31 Infectious-Disease Data Science
ghozzis@rki.de
https://www.rki.de/signale-project

**1. Motivation**

- 1.1. Signal detection for infectious epidemiology
- 1.2. Filter, quantify, disentangle
- 1.3. Use cases, setting

**2. Applications: Some Examples**

- 2.1. Retrospective
- 2.2. Prospective: Seasons
- 2.3. Prospective: Clusters
- 2.4. Technological Implementations

**3. Statistical Approaches**

- 3.1. Regression on univariate time-series
- 3.2. Scan statistics
- 3.3. Other approaches

**4. Evaluation**

- 4.1. Goodness of fit
- 4.2. Evaluation of binary classification

**5. Conclusion and Outlook**

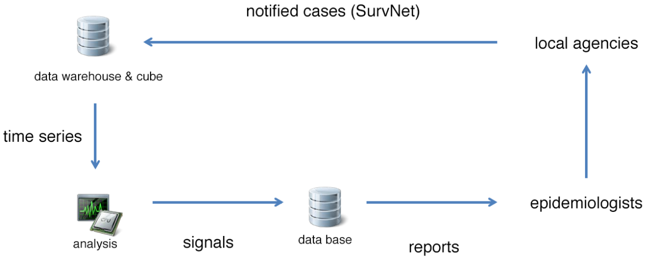
- 5.1. Routine but not standard
- 5.2. Methodological priorities
- 5.3. Usability

# 1. Motivation

# 1.1. Signal detection for infectious epidemiology

find anomalies in surveillance data that may suggest an outbreak

(mostly syndromic data outside Germany)



# 1.2. Filter, quantify, disentangle

Why automatic/algorithmic detection?

**Filter** the many combinations of “what, who, where”... Which ones are interesting?
$\sim$ food-borne diseases

**Quantify** the anomaly: Remove human bias; communicate homogeneously and reliably
$\sim$ seasonal diseases

**Disentangle** the contributing factors: Remove artefacts, find determinants
$\sim$ vector-borne diseases

Always "just" an indication for further action/investigation!

# 1.3. Use cases, setting

Think: salmonella, influenza, dengue, borreliosis, MRSA... not so much HIV or TB

Either *retrospective* or *prospective*

What is an outbreak? “Noticeably many infection cases”

Data: weekly aggregated cases

Prospective: one week ahead

Definitions: “signal”/“alarm” = indication, “alert” = official notice
Public Health England: “signal” = variable being observed

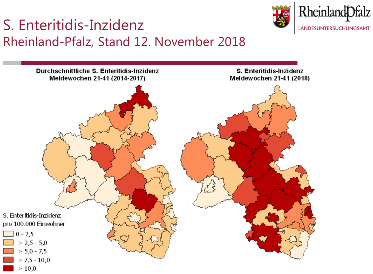
**Not treated here:**

- outbreak/infection-chain reconstructions
- clustering of genetic sequences

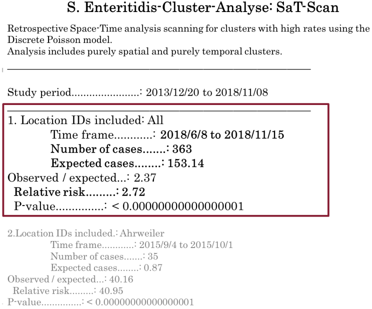
# 2. Applications: Some Examples

# 2.1. Retrospective

## Rhineland-Palatinate Investigation Office



S. Enteritidis-Inzidenz
Rheinland-Pfalz, Stand 12. November 2018

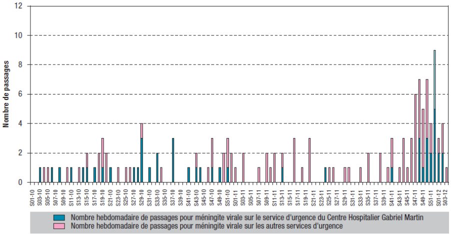


### S. Enteritidis-Cluster-Analyse: SaT-Scan

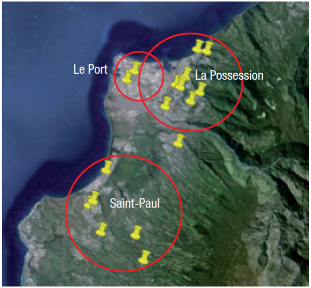
```
Retrospective Space-Time analysis scanning for clusters with high rates using the
Discrete Poisson model.
Analysis includes purely spatial and purely temporal clusters.
____________________________________________________________

Study period......................: 2013/12/20 to 2018/11/08
____________________________________________________________
1. Location IDs included: All
          Time frame............: 2018/6/8 to 2018/11/15
          Number of cases.......: 363
          Expected cases........: 153.14
Observed / expected...: 2.37
 Relative risk.........: 2.72
 P-value..............: < 0.00000000000000001

2.Location IDs included.: Ahrweiler
          Time frame............: 2015/9/4 to 2015/10/1
          Number of cases.......: 35
          Expected cases........: 0.87
Observed / expected...: 40.16
 Relative risk.........: 40.95
P-value..............: < 0.00000000000000001
```

# Santé publique France

## Viral Meningitis in the Réunion



Time series:
CW46 2011 – CW02 2012 significantly more cases than expected



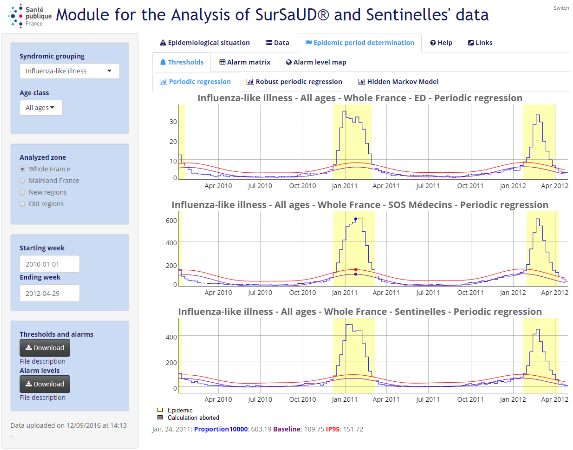
SaTScan:
Three clusters for the weeks CW48 2011 – CW02 2012

Vilain et al (2014) Bull Epidémiol Hebd. (3-4):53-7 http://invs.santepubliquefrance.fr//beh/2014/3-4/2014_3-4_3.html

# 2.2. Prospective: Seasons

Santé publique France

MASS: among other things detection of influenza epidemic season



Module for the Analysis of SurSaUD® and Sentinelles' data

Syndromic grouping: Influenza-like illness

Age class: All ages

Analyzed zone
- ● Whole France
- ○ Mainland France
- ○ New regions
- ○ Old regions

Starting week: 2010-01-01

Ending week: 2012-04-29

Data uploaded on 12/09/2016 at 14:13

Influenza-like illness - All ages - Whole France - ED - Periodic regression

Influenza-like illness - All ages - Whole France - SOS Médecins - Periodic regression

Influenza-like illness - All ages - Whole France - Sentinelles - Periodic regression

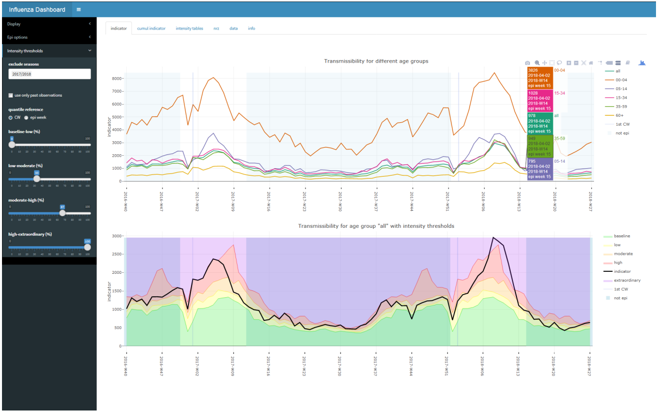
Jan, 24, 2011: **Proportion10000**: 603.19 **Baseline**: 109.75 **IP95**: 151.72

https://cpelat.shinyapps.io/mass/

Pelat et al (2017) Euro Surveillance 22(32) 30593 https://doi.org/10.2807/1560-7917.ES.2017.22.32.30593

# Robert Koch Institute

## Influenza Dashboard: detection and severity of influenza epidemic season



Influenza Dashboard

Display

Epi options

Intensity thresholds

**exclude seasons**

2017/2018

□ use only past observations

**quantile reference**

● CW ○ epi week

**baseline-low (%)** 0

**low-moderate (%)** 33

**moderate-high (%)** 67

**high-extraordinary (%)** 100

# 2.3. Prospective: Clusters

Bureau of Communicable Disease

New York City Department of Health and Mental Hygiene

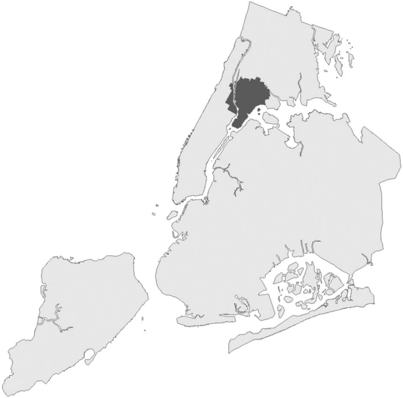
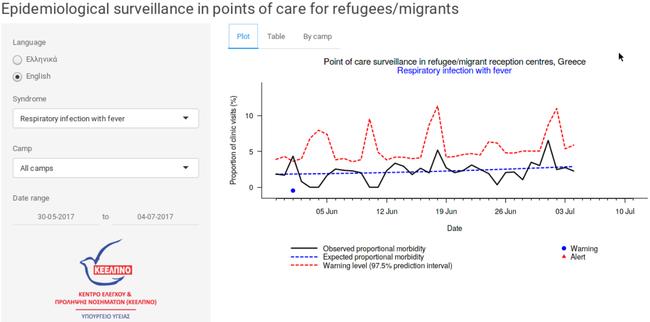


**Figure.** Automated output from spatiotemporal analysis on July 17, 2015, indicating a cluster (dark gray) of 8 legionellosis cases over 8 days centered in the South Bronx, New York City, New York, USA. In subsequent days, this cluster expanded in space and time into the second largest US outbreak of community-acquired legionellosis.

Greene et al (2016) Emerging Infectious Diseases 22(10) 1808 https://doi.org/10.3201/eid2210.160097

# Hellenic Centre for Disease Control and Prevention



Epidemiological surveillance in points of care for refugees/migrants

Language

○ Ελληνικά

● English

Syndrome

Respiratory infection with fever

Camp

All camps

Date range

30-05-2017 to 04-07-2017

Plot
Table
By camp
Point of care surveillance in refugee/migrant reception centres, Greece
Respiratory infection with fever
Proportion of clinic visits (%)
10
5
0
05 Jun
12 Jun
19 Jun
26 Jun
03 Jul
10 Jul
Date
Observed proportional morbidity
Expected proportional morbidity
Warning level (97.5% prediction interval)
● Warning
▲ Alert
ΚΕΕΛΠΝΟ
ΚΕΝΤΡΟ ΕΛΕΓΧΟΥ &
ΠΡΟΛΗΨΗΣ ΝΟΣΗΜΑΤΩΝ (ΚΕΕΛΠΝΟ)
ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ

https://github.com/thlytras/syndroCampsGR

# Robert Koch Institute

## Signals Reports

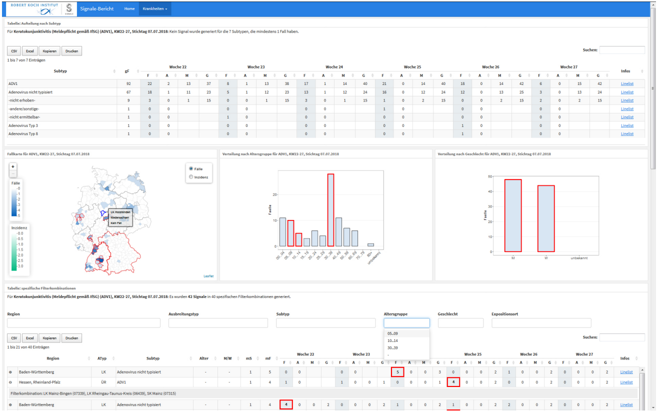


Tabelle: Aufteilung nach Subtyp

Für **Keratokunjunktivitis (Meldepflicht gemäß IfSG) (ADV1), KW22-27, Stichtag 07.07.2018:** Kein Signal wurde generiert für die 7 Subtypen, die mindestens 1 Fall haben.

1 bis 7 von 7 Einträgen

| Subtyp | gF | Woche 22 | | | | Woche 23 | | | | Woche 24 | | | | Woche 25 | | | | Woche 26 | | | | Woche 27 | | | | Infos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F | A | M | G | F | A | M | G | F | A | M | G | F | A | M | G | F | A | M | G | F | A | M | G | |
| ADV1 | 92 | 22 | 2 | 13 | 37 | 8 | 1 | 13 | 38 | 17 | 1 | 14 | 40 | 21 | 0 | 14 | 40 | 18 | 0 | 14 | 42 | 6 | 0 | 15 | 42 | Linelist |
| Adenovirus nicht typisiert | 67 | 18 | 1 | 11 | 23 | 5 | 1 | 12 | 23 | 13 | 1 | 12 | 24 | 16 | 0 | 12 | 24 | 12 | 0 | 13 | 25 | 3 | 0 | 13 | 24 | Linelist |
| -nicht erhoben- | 9 | 3 | 0 | 1 | 15 | 0 | 0 | 1 | 15 | 3 | 0 | 1 | 15 | 1 | 0 | 2 | 15 | 0 | 0 | 2 | 15 | 2 | 0 | 2 | 15 | Linelist |
| -andere/sonstige- | 1 | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | | 1 | 0 | | | 0 | 0 | | | 0 | 0 | | | Linelist |
| -nicht ermittelbar- | 1 | 0 | 0 | | | 1 | 0 | | | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | | Linelist |
| Adenovirus Typ 3 | 1 | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | | 1 | 0 | | | 0 | 0 | | | Linelist |
| Adenovirus Typ 8 | 1 | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | | 1 | 0 | | | 0 | 0 | | | Linelist |

Fallkarte für ADV1, KW22-27, Stichtag 07.07.2018

Verteilung nach Altersgruppe für ADV1, KW22-27, Stichtag 07.07.2018

Verteilung nach Geschlecht für ADV1, KW22-27, Stichtag 07.07.2018

Tabelle: spezifische Filterkombinationen

Für **Keratokunjunktivitis (Meldepflicht gemäß IfSG) (ADV1), KW22-27, Stichtag 07.07.2018:** Es wurden **42 Signale** in 40 spezifischen Filterkombinationen generiert.

1 bis 21 von 40 Einträgen

| | Region | ATyp | Subtyp | Alter | M/W | mS | mF | Woche 22 | | | | Woche 23 | | | | | | | | Woche 25 | | | | Woche 26 | | | | Woche 27 | | | | Infos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | F | A | M | G | F | A | M | G | F | A | M | G | F | A | M | G | F | A | M | G | F | A | M | G | |
| ⊕ | Baden-Württemberg | LK | Adenovirus nicht typisiert | - | - | 1 | 5 | 0 | 0 | | | 0 | 0 | | | 5 | 0 | 0 | 3 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | Linelist |
| ⊖ | Hessen, Rheinland-Pfalz | ÜR | ADV1 | - | - | 1 | 4 | 1 | 0 | | | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 4 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | Linelist |
| Filterkombination: LK Mainz-Bingen (07339), LK Rheingau-Taunus-Kreis (06439), SK Mainz (07315) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| ⊕ | Baden-Württemberg | LK | Adenovirus nicht typisiert | - | - | 1 | 4 | 4 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 2 | 1 | 0 | 0 | 2 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | Linelist |

Salmon et al (2016) Euro Surveillance 21(13) 30180 https://doi.org/10.2807/1560-7917.ES.2016.21.13.30180

# 2.4. Technological Implementations

. . . very diverse, but R is emerging as a standard:

- Analysis: R, in particular *surveillance* package; free software SaTScan
- Reports: R-Markdown
- Interactive web sites: R-Shiny; commercial solutions

# 3. Statistical Approaches

# 3.1. Regression on univariate time-series

**Idea**

- filter cases (age, place, sex, ...) and aggregate weekly = 1 time series

- compare the **observed case count** this week with what is **expected**

- define a **threshold** above which a count is so unexpected that it warrants an alarm

**Strategy**

- **threshold** = upper bound of **confidence interval**

e.g. "If less than 1% chance of seeing such a high case count, that's suspect! Let's generate a signal."

N.B. another strategy, threshold = mean + *n* standard deviations, is intuitive but problematic

http://staff.math.su.se/hoehle/blog/2018/10/29/gauss.html

**Non-parametric approach:**

Upper bound $U_t$ = maximum over the last *n* observations (assuming no ties), with confidence (n - 1)/(n + 1)
e.g. n = 199 for a one-sided 99% confidence interval

$$U_t = \max(y_{t-n}, \ \ldots \ , y_{t-1})$$

Problems:
- needs many observations, especially at low counts
- no structural changes considered (trend, seasonality)

http://staff.math.su.se/hoehle/blog/2018/10/29/gauss.html

https://en.wikipedia.org/wiki/Prediction_interval#Non-parametric_methods

**Parametric approach:**

- fit a given distribution

- compute p-value of observing a given count under that distribution

e.g. signal if p-value $< 1\%$

Choices of distribution:

- **Poisson:** natural for count data, but only one parameter: rigid / too narrow (standard deviation = mean)

- **Quasi-Poisson:** Poisson with supplementary parameter: over dispersion $\phi =$ variance/mean

- **Negative Binomial:** natural for picking samples of one in two categories (Bernoulli trials); also two parameters

Problems:

- assumption on distribution
- doesn't account for structural changes

**Sliding window:**
Fit your distribution on the last data points

Problem:
- discards most of the available information

**Generalised linear models (GLM):**

Model the dependency of distribution on given factors, here on **time**

$$y_t \sim P(\mathrm{mean} = \mu_t, \mathrm{variance} = \phi \times \mu_t)$$

$$\log \mu_t = \beta_0 + \beta_1 \times t + \beta_2 \cos(2\pi\ t/52) + \beta_3 \sin(2\pi\ t/52)$$

Problem:

- past outbreaks skew the expectation

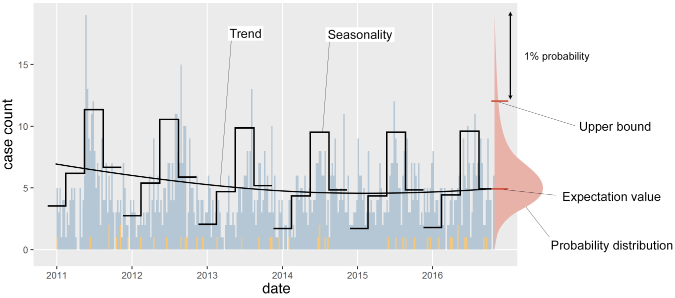
**"Farrington modified"** $\approx$ GLM with:

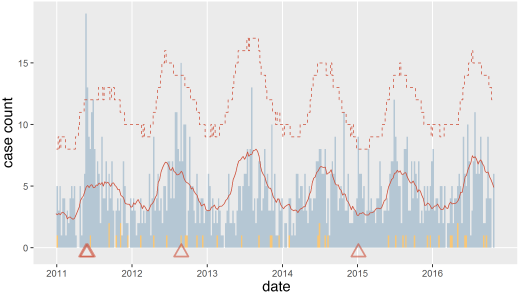
- past aberrations removed (reweighting)
- ignore last weeks
- ignore low counts

. . . used a lot, especially in Europe

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


Trend
Seasonality
1% probability
Upper bound
Expectation value
Probability distribution
case count
0
5
10
15
2011
2012
2013
2014
2015
2016
date


case count
0
5
10
15
2011
2012
2013
2014
2015
2016
date

# 3.2. Scan statistics

**Idea**

- observe *regions* over *periods* of time: Does one stand out?

**Strategy** (flavour: “Space–Time Permutation Scan Statistic”)

- define space-time observation windows
- compute a likelihood for current observation
- identify the most unlikely cluster
- how unlikely is it?
- threshold on the p-value

Space-time **observation windows** $\{A\}$: "Cylinder" $=$ set $\{z\}$ of administrative units (zip code) with centroid in a base circles $\times$ last $\{d\}$ time points (days)

(Stratify for day of week)

Expected count in $A$: $\mu_A = \sum_{z',d' \in A} \sum_z c_{zd'} \sum_d c_{z'd}/C$, with $C$ the total count

Case count $c_A$ in $A \sim \mathrm{Poisson}(\mu_A)$ if $C \gg c_A$

Poisson generalised **likelihood ratio** $\mathrm{GLR} = (c_A/\mu_A)^{c_A} \times ((C - c_A)/(C - \mu_A))^{C - c_A}$

Compute GLR for many different base circles and durations, keep the one with maximum GLR $A^*$

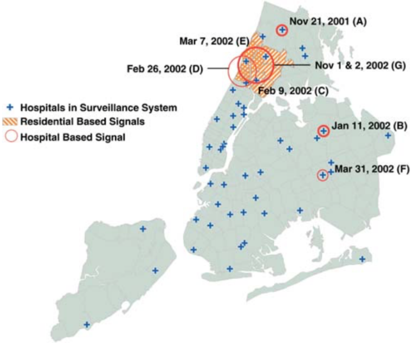
Correct for **multiple testing**:

- random permutations of $z$ and $d$ for each case

- for each permutation $p$, cylinder with largest GLR is $A_p^*$

- Monte Carlo hypothesis testing: p-value $= R/(S+1)$ with $R$ the rank of $A^*$ among $A_p^*$ and $S$ the number of permutations

**Threshold** on p-value to generate a signal



No modelling, but testing of many combinations: accounts for spatial and temporal structural differences

Implementations:
- SaTScan (free software made specifically for these analyses)
- R package `scanstatistics`

Kulldorff et al (2005) PLoS Medicine 2(3) e59 http://doi.org/10.1371/journal.pmed.0020059

Greene et al (2016) Emerging Infectious Diseases 22(10) 1808 http://doi.org/10.3201/eid2210.160097

Allévius, Höhle (2017) arXiv 1711.08960 http://arxiv.org/abs/1711.08960

`scanstatistics` vignette: https://cran.r-project.org/web/packages/scanstatistics/vignettes/introduction.html

# 3.3. Other approaches

Autoregressive models: ARIMA, INAR (generalise random walks)

Bayesian inference, e.g. Hidden Markov Models:
GLM + binary hidden state = “outbreak: yes/no”

Control charts, e.g. cumulative sums (CUSUM)

GLMs with delay

Spatial GLMs, spatial CUSUMS

. . . and many more

+ in principle all modelling approaches could be used for signal detection (there’s a lot of them)

Unkel et al (2012) J Royal Statistical Society A 175(1) 49 http://doi.org/10.1111/j.1467-985X.2011.00714.x

Allévius, Höhle (2017) arXiv 1711.08960 http://arxiv.org/abs/1711.08960

# 4. Evaluation

# 4.1. Goodness of fit

If an explicit model is used: How good does it reproduce the data?

Standard scores for goodness of fit, 2 examples:

- Normalised Squared Error Score $= ((y_t - \mu_t)/\sigma_t)^2$

$y_t$ = observed count, $\mu_t$ = expectation value, $\sigma_t$ = estimated standard deviation

- Bayesian Information Criterion (BIC) $= -2 \sum_t \log(p_t(y_t)) + \log(n_{\mathrm{eff}})\, df$

$p_t(y_t)$ = probability of observing $y_t$ at time $t$ under the model, $n_{\mathrm{eff}}$ = number of data points, $df$ = number of parameters

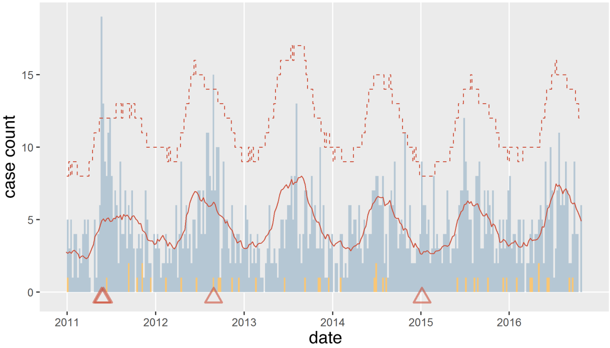
Liboschik (2016) PhD Thesis, TU Dortmund University, page 19

Salmon (2016) PhD Thesis, Ludwig–Maximilians–Universität, pages 89-90

# 4.2. Evaluation of classification

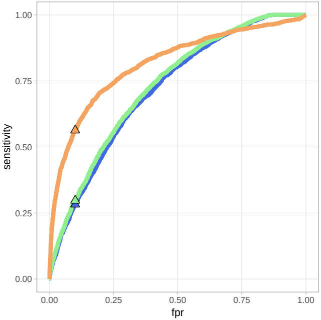
Signals vs. week/place with outbreaks

**Confusion matrix** of true positives, true negatives, false positives and false negatives



⇒ scores, e.g. **sensitivity = TP/P** and **specificity = 1 - false positive rate = TN/N**

ROC curve: sensitivity vs. false positive rate with varying threshold
for campylobacter and 3 detection algorithms



But also: probability of detection, timeliness, size before detection, etc.

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

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

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

Real data: Ghozzi, Ullrich, in preparation

# 5. Conclusion and Outlook

## 5.1. Routine but not standard

Many **statistical approaches** exist, with two types the most common:

- model + regression on univariate time series ~ **Farrington**

- spatio-temporal clusters ~ **SaTScan**

Many different ways of **evaluating**:

- the modelling

- the detection itself

. . . but no clear picture yet

**Communication:**

- Complexity of results: Too much vs. too little. . . Is interaction/exploration (dashboards) a solution?

- Use many different algorithms?

- Signals crossing administrative boundaries?

# 5.2. Methodological priorities

Use real outbreak data to reach **conclusions** and make **recommendations**
- gold-standard real data set
- hyperparameter optimisation
- model selection/combination (stacking)

Busche, Ullrich, Ghozzi, in preparation

Use labelled data to improve detection (**supervised learning**)

Ghozzi, Ullrich, in preparation

Zacher, Czogiel, in preparation

Adapt **epidemiological models** for signal detection:
- space-time dynamics, including delays (nowcasting)
- propagation models (SIR), including networks

Höhle, an der Heiden (2014) Biometrics 70(4) 993 https://doi.org/10.1111/biom.12194

Salmon et al (2015) Biometrical Journal 57(6) 1051 https://doi.org/10.1002/bimj.201400159

Manitz et al (2014) PLoS Currents Outbreaks 1–31 http://currents.plos.org/outbreaks/index.html%3Fp=36515.html

Integrate **secondary data sources**, e.g.
- medical (vaccination)
- online activity (social networks, internet searches)
- socio-environmental (holidays, economics, weather, geography)
- mass gatherings

Ma et al (2015) Epidemiology and Infection 143(11) 2390 https://doi.org/10.1017/S0950268814003240

Routine **integration of molecular** and epidemiological information

Ashton et al (2015) bioRxiv https://www.biorxiv.org/content/early/2015/11/29/033225

**Case-based** detection (clustering of individual cases)

**Epidemiologically relevant score:** space-time extension, measure of severity, case based

overall, in references cited, 31 scores... let's add a 32nd!

Continuous user **feedback**: Evaluate signals and tweak models (**reinforcement learning**)

## 5.3. Usability

Signals other than outbreaks? **“anomaly detection”**

Publish code and data. . . but also consult epidemiologists and evaluate tools: **include community**

User needs as starting point: **user-oriented development** rather than method driven

Organisation: **Data-science projects** with epidemiology + statistics + software dev

**Inspirations:**

- data journalism
- self-tracking apps & virtual assistants