L06 Outbreak Detection[1]

Michael Höhle[1]

[1]Department of Mathematics, Stockholm University, Sweden
m_hoehle

STA427 FS2021
Statistical Methods in Infectious Disease Epidemiology
Epidemiology, Biostatistics and Prevention Institute
University of Zurich, Switzerland

University of Zurich^UZH

[1] LaMo: 2021-03-31 @ 12:49:15

Outline

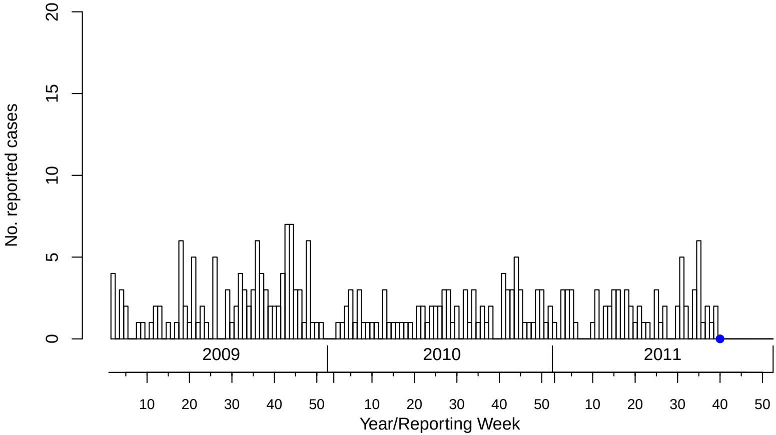
1. Monitoring of univariate count data time series
    - Statistical Framework for Aberration Detection
    - Simple Algorithm for Ad-Hoc Detection
    - Farrington algorithm and beyond
2. Multivariate Methods
    - Univariate Methods in Parallel
    - Kulldorff's scan statistic
    - Case Study: Meningococcal disease in Germany
3. A System for Automated Outbreak Detection
4. Discussion

Outline

1. Monitoring of univariate count data time series
   - Statistical Framework for Aberration Detection
   - Simple Algorithm for Ad-Hoc Detection
   - Farrington algorithm and beyond
2. Multivariate Methods
3. A System for Automated Outbreak Detection
4. Discussion

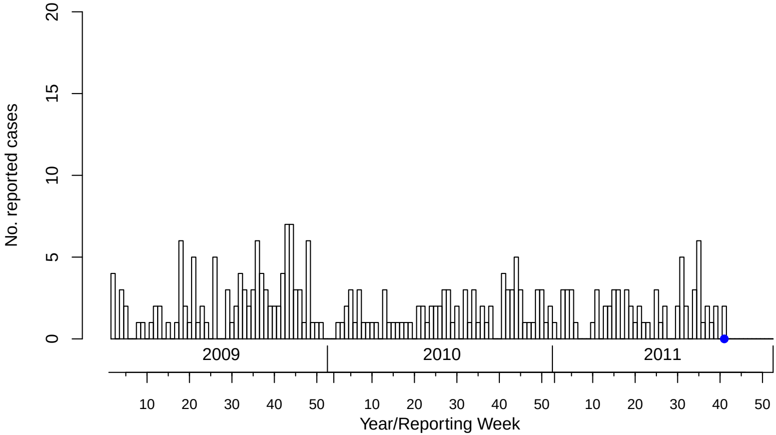
Example: Monitoring German Salmonella Newport Cases

German Infection Protection Act (IfSG) data from the Robert Koch Institute (up to W40-2011):



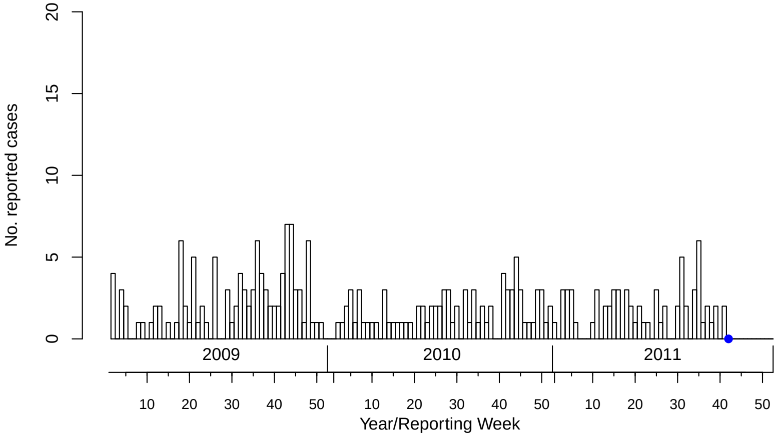
Example: Monitoring German Salmonella Newport Cases

German Infection Protection Act (IfSG) data from the Robert Koch Institute (up to W40-2011):



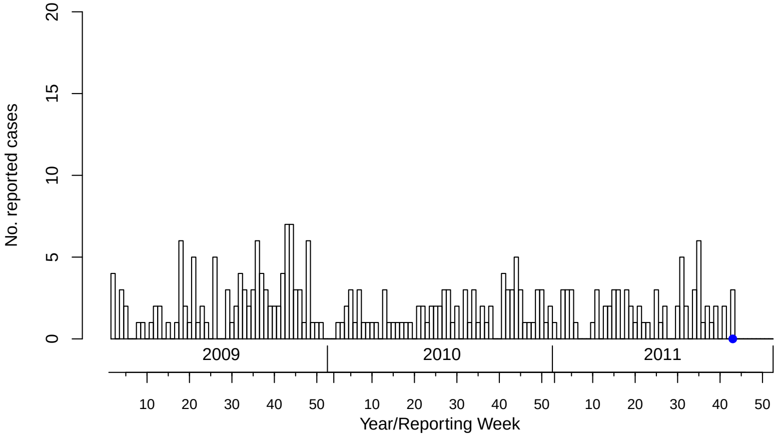
Example: Monitoring German Salmonella Newport Cases

German Infection Protection Act (IfSG) data from the Robert Koch Institute (up to W40-2011):



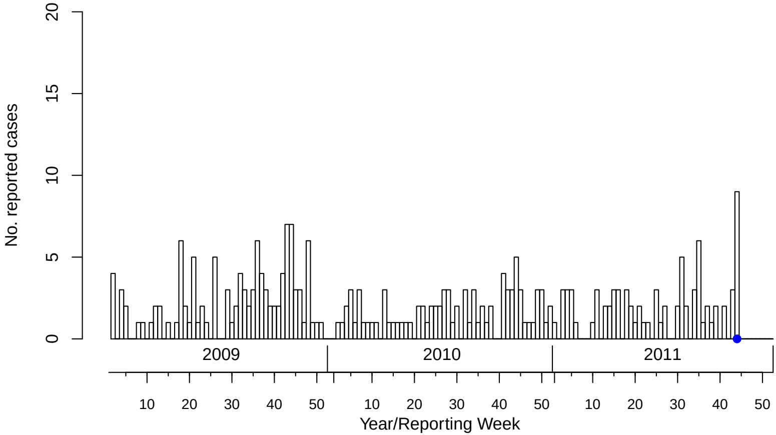
Example: Monitoring German Salmonella Newport Cases

German Infection Protection Act (IfSG) data from the Robert Koch Institute (up to W40-2011):



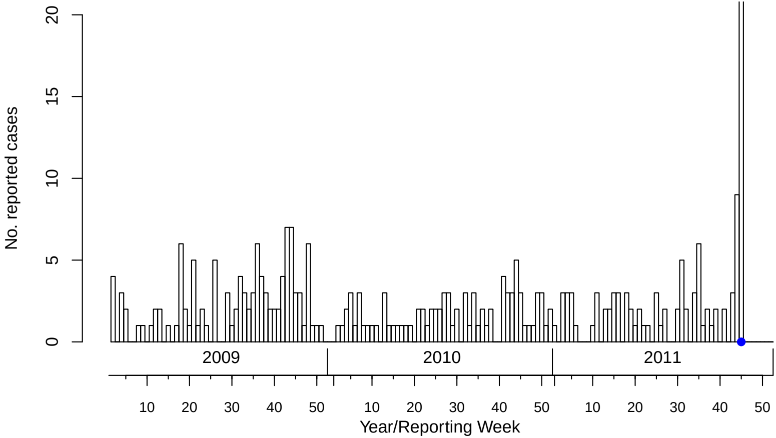
Example: Monitoring German Salmonella Newport Cases

German Infection Protection Act (IfSG) data from the Robert Koch Institute (up to W40-2011):



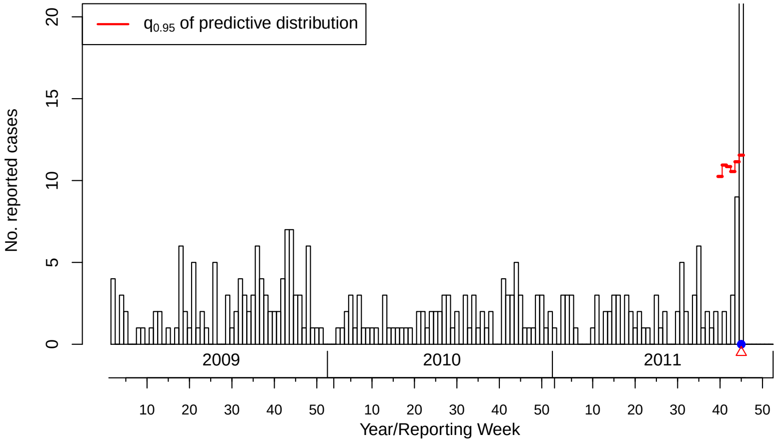
Example: Monitoring German Salmonella Newport Cases

German Infection Protection Act (IfSG) data from the Robert Koch Institute (up to W40-2011):



Example: Monitoring German Salmonella Newport Cases

German Infection Protection Act (IfSG) data from the Robert Koch Institute (up to W40-2011):



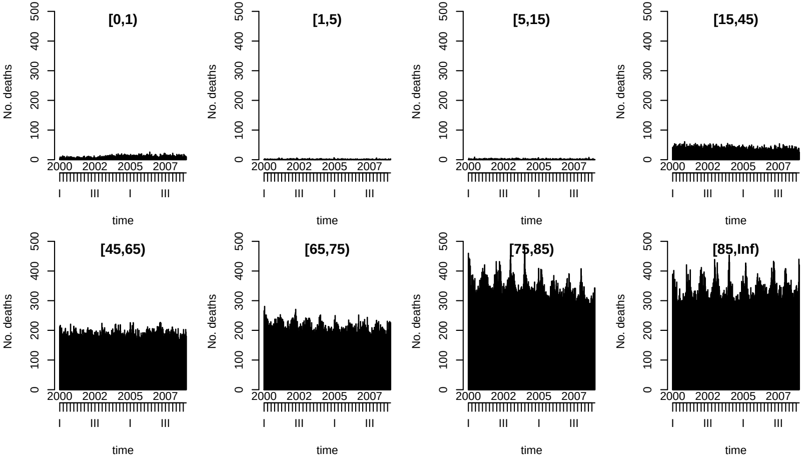
During Oct-Nov 2011 there was an outbreak associated with mung bean sprouts (RKI 2012)

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: Focus on monitoring aspect

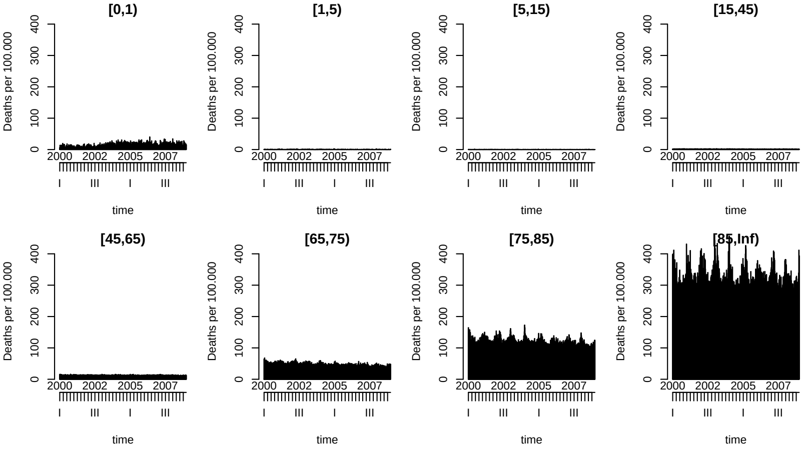
Example – The EuroMOMO project (2)

Weekly danish mortalities 2000-2008 in 8 age-groups as provided by Statens Serum Institute (Höhle et al. 2010).



Example – The EuroMOMO project (2)

Weekly danish mortalities 2000-2008 in 8 age-groups as provided by Statens Serum Institute (Höhle et al. 2010).



Statistical Framework for Aberration Detection (1)

- Univariate time series $\{y_t,\ t = 1, 2, \ldots\}$ to monitor
- For each time t we differentiate between two underlying states: in-control (everything is fine) or out-of-control (something is wrong).
- At time $s \geq 1$, the available information is $\boldsymbol{y}_s = \{y_t\,;\ t \leq s\}$.
- Based on $\boldsymbol{y}_s$ an automatic detection procedure has to decide if there is unusual activity at time s (or not).

Statistical Framework for Aberration Detection (2)

- The detectors are initially only based on the one-step-ahead predictive distribution at each time point (Shewhart-like control chart):
  - Let $G(y_s|y_1, \ldots, y_{s-1}; \boldsymbol{\theta})$ be the distribution of Y_s in case everything is in-control.
  - If the actual observed value $Y_s = y_s$ is extreme in G, this is evidence against things being in-control.
  - The alarm threshold $a_{1-\alpha,s}$ at each time point is calculated as the $(1-\alpha)$'th quantile of the predictive distribution. If $y_s > a_{1-\alpha,s}$ then we have an alarm.
- This can be generalized to more sequential control charts accumulating information, e.g. cumulative sum (CUSUM) methods.

Intermezzo: Estimation, prediction and uncertainty

- Data $\boldsymbol{y}$ are the observed value of a random variable $\boldsymbol{Y}$ characterized by a parametric model with density $f(\boldsymbol{y};\boldsymbol{\theta})$.
- Aim: predict the value of a random variable $\boldsymbol{Z}$, which, conditionally on $\boldsymbol{Y} = \boldsymbol{y}$ has distribution function $G(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})$, *depending on* $\boldsymbol{\theta}$.
- Simplest form of the prediction problem:

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\boldsymbol{\theta}),$$

  and the task is to predict $Z = Y_{n+1}$.
- In *time series 1-step-ahead prediction* the observations are correlated and the aim is to predict $\boldsymbol{Z} = Y_{n+1}$.

Example: Predicting a new $N(\mu, \sigma^2)$ observation (1)

- Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with unknown μ and σ^2. Then

$$\frac{Y_{n+1} - \overline{Y}}{s\sqrt{1 + \frac{1}{n}}} \sim t(n-1),$$

  where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ are the sample mean and sample variance of $\boldsymbol{Y}$, respectively.

- A $(1-2\alpha)\cdot 100\%$ two-sided **prediction interval** (PI) is thus given by

$$\overline{Y} \pm t_{1-\alpha}(n-1) \cdot s \cdot \sqrt{1 + \frac{1}{n}}.$$

Example: Predicting a new $N(\mu, \sigma^2)$ observation (2)

- A *plug-in* $(1-2\alpha)\cdot 100\%$ two-sided **prediction interval** for Y_{n+1} is:

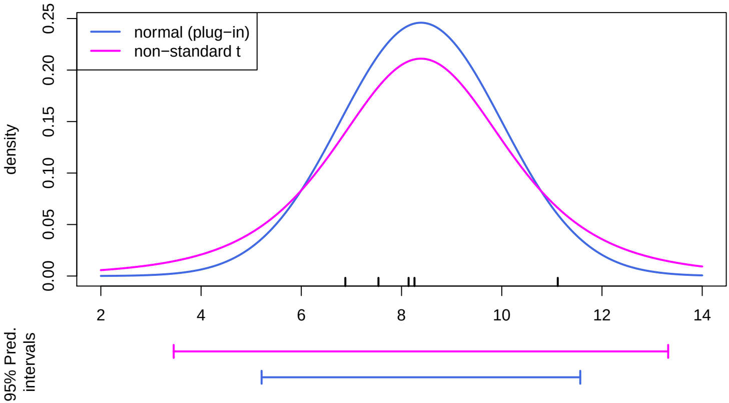
$$\overline{Y} \pm z_{1-\alpha} \cdot s.$$

- Both of these are not to be confused with a $(1-2\alpha)\cdot 100\%$ two-sided **confidence interval** for μ:

$$\overline{Y} \pm z_{1-\alpha} \cdot \frac{s}{\sqrt{n}}.$$

Example: Predicting a new $N(\mu, \sigma^2)$ observation (3)

- Illustration: PIs based on $n = 5$ observations from $N(\mu, \sigma^2)$.



- For $n = 5$ the 95% plug-in PI corresponds to a 85% PI. The 95% CI for μ is 7.2–9.6, which only corresponds to a 46% PI.

Summary: Ad-Hoc Outbreak Detection Algorithm

- Predict value y_s at time $s = (s^w, s^y)$ using a set of reference values from window of size $2w+1$ up to b years back.
- Let $n = b(2w+1)$ and compute threshold as the upper 97.5% quantile of the predictive distribution for y_s, i.e.

$$a_{0.975,s} = \overline{y} + t_{0.975}(n-1) \cdot s \cdot \sqrt{1 + \frac{1}{n}}.$$

- Sound alarm, if $y_s > a_{0.975,s}$.

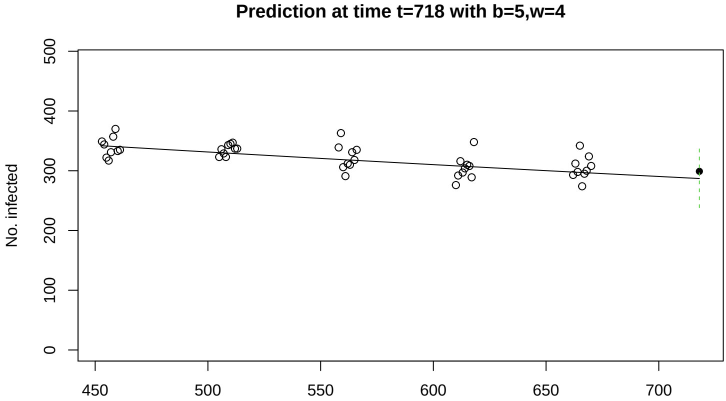
Challenges of surveillance data

Issues making the statistical modelling and monitoring of surveillance time series a challenge:

- Lack of clear case definitions
- Under-reporting and reporting delays
- Often no denominator data
- Seasonality
- Low number of reported cases
- Presence of past outbreaks
- Existence of concurrent “explanatory” processes

Farrington algorithm (1) – basic model

- Predict value y_s at time $s = (s^w, s^y)$ 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 $\mathrm{E}(y_t) = \mu_t$, $\mathrm{Var}(y_t) = \phi \cdot \mu_t$ with $\log \mu_t = \alpha + \beta t$ and $\phi > 0$.

Farrington algorithm (2) – outbreak detection

Predict and compare:

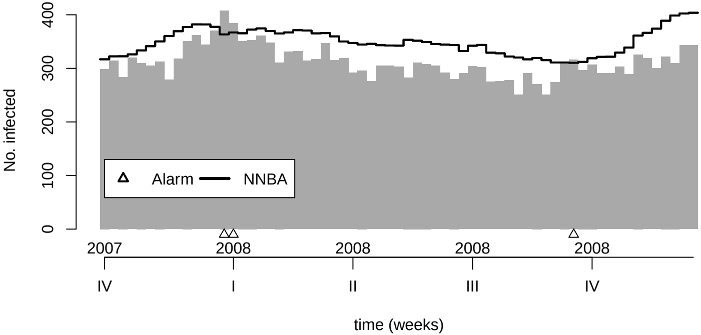
- An approximate $(1-\alpha)$ one-sided prediction interval for y_s based on the GLM has upper limit $a_{1-\alpha,s} = \hat{\mu}_s + z_{1-\alpha} \cdot \sqrt{\mathrm{Var}(y_s - \hat{\mu}_s)}$
- If the oserved y_s is greater than $a_{1-\alpha,s}$, then flag s as outbreak

Refinements of the algorithm include:

- Computation of the prediction interval on a transformed scale
- Use a re-weighted fit with weights based on Anscombe residuals in order to correct for outliers
- Low count protection

Application: Danish mortality data (age group 75-84 years)

- Results of the Farrington algorithm, respectively, with $w = 4, b = 5$ and $\alpha = 0.005$ starting at W40-2007:



Outline

1. Monitoring of univariate count data time series
2. Multivariate Methods
   - Univariate Methods in Parallel
   - Kulldorff's scan statistic
   - Case Study: Meningococcal disease in Germany
3. A System for Automated Outbreak Detection
4. Discussion

Setup

- Instead of a univariate time series $\{Y_t; t = 1, 2, \}$ as in the previous section the observation at each time point consists of a p-variate vector $\boldsymbol{Y}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,p})'$
- Each component $Y_{t,i}$ could represent the disease incidence (as a count) of a given region/age-group/gender/serotype/pathogen combination at time t
- Aim is to monitor the p time series simultaneously. The hope is that this gains strengths to detect vague signals

Univariate Methods in Parallel

- Simple approach for multiple data streams is to use one of the univariate methods from the previous section to each time series
- Pros:
  - Easy to use, scales linearly
  - Can aggregate results in suitable fashion
- Cons:
  - False positive probability per time point is α per series so probability of raising at least one false alarm will be much greater than α (multiple testing).
  - If one uses a small α this might make outbreaks harder to detect.

Kulldorff's prospective scan statistic (1)

- Kulldorff (2001) proposed a method for prospective spatio-temporal detection in spatial time series data
- The method assumes that

  $$Y_{it} \sim \mathrm{Po}(q_{it} \cdot b_{it}),$$

  where b_{it} is an 'expected count' proportional to the population at risk in region i at time t.
- Note: $q_{it} > 0$ is assumed to be the same $q_{it} = q$ for all i and t provided there is no outbreak (null hypothesis)
- However, for areas with outbreaks the relative risk is higher inside a space-time window $W = Z \times \{T - D + 1, \ldots, T\}$, consisting of a subset of regions $Z \subset \{1, \ldots, N\}$ and stretching over the D most recent time periods.

Kulldorff's prospective scan statistic (2)

- Focus of the method: what W and D combination gives the greatest discrepancy from null-hypothesis?
- Contrast this with the the distribution of such a maximum under the null-hypothesis in order to get P-values,
  1. calculate the MLE of q_W and $q_{\overline{W}}$.
  2. calculate the likelihood ratio of W between H_0 and H_1
  3. calculate likelihood ratio λ_W for all W of interest
  4. the scan statistic is defined $\lambda^* = \max_W \lambda_W$. The corresponding window W^*, often called the most likely cluster
  5. calculate the p-value for W^* and flag alarm if below threshold

Step 1

- Estimation of q_W and $\hat{q}_{\overline{W}}$

$$\hat{q}_W = \frac{Y_W}{B_W},$$
$$\hat{q}_{\overline{W}} = \frac{Y - Y_W}{B - B_W} = \frac{Y_{\overline{W}}}{B_{\overline{W}}},$$

where

$$Y_W = \sum_{(i,t)\notin W} y_{it}, B_W = \sum_{(i,t)\in W} b_{it}, \text{ and}$$
$$Y = \sum_{i=1}^{N}\sum_{t=1}^{T} y_{it} = \sum_{i=1}^{N}\sum_{t=1}^{T} b_{it}.$$

Step 2

- Thus, the likelihood ratio statistic conditional on the window W is then given by

$$\lambda_W = \begin{cases} \left(\frac{Y_W}{B_W}\right)^{Y_W} \left(\frac{Y-Y_W}{Y-B_W}\right)^{Y-Y_W} & \text{if} Y_W > B_W, \\ 1 & \text{otherwise} \end{cases}$$

up to a multiplicative constant not dependent on q_W or $q_{\overline{W}}$.

Hypothesis testing (1)

- No closed formula available for the distribution of λ^*
- Instead: Monte Carlo where new data for each region i and time t are simulated under the null hypothesis using the expected counts b_{it}.
- For Kulldorff's scan statistic, the sampling is made conditional on the total observed count $Y = C$, leading to a multinomial distribution
- Sampling is repeated R times. A Monte Carlo P-value for the observed scan statistic is given by its rank among the simulated values:

$$P = \frac{1 + \sum_{r=1}^{R} \mathbf{1}\{\lambda_r^* > \lambda_{\text{obs}}^*\}}{1 + R}.$$

Hypothesis testing (2)

- Typically, a number such as $R = 999$ or $R = 9999$ is used in order to get a fixed number of digits for the P-value.
- Note: As for univariate investigations one has a multiple testing problem, because one repeats the analyses for every time point

Implementation

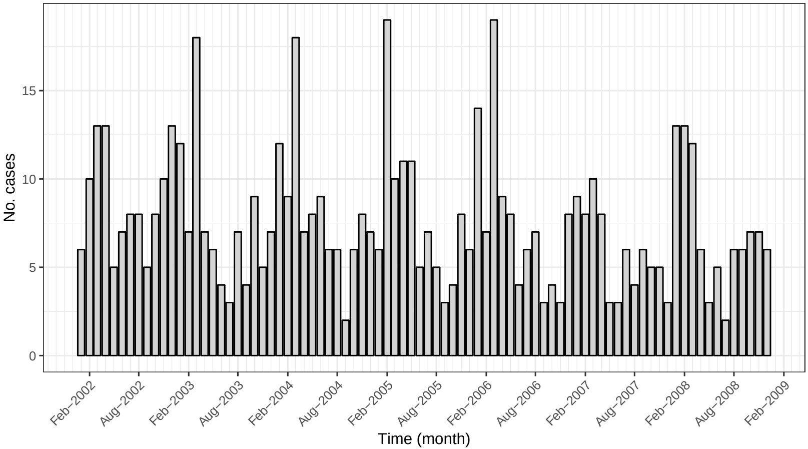
- Kulldorff's scan statistic is implemented in the R package `rsatscan`, which is just a call-through to the SaTScan™ program
- A true open-source alternative is the function `scan_pb_poisson` in the package `scanstatistics`

Case Study: Meningococcal disease in Germany (1)

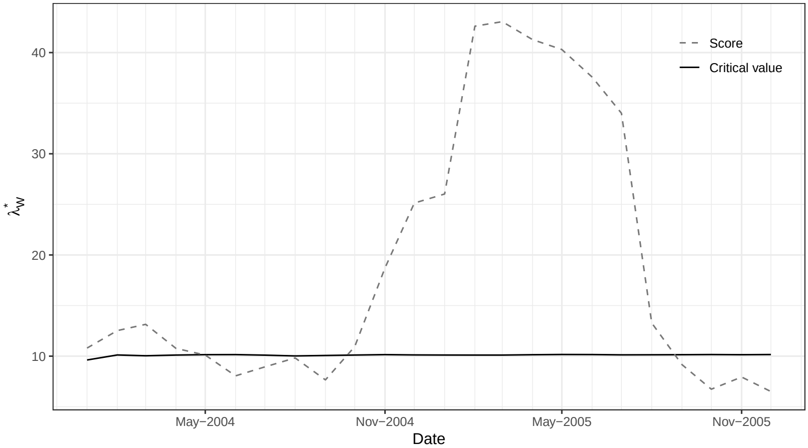
- Application of Kulldorff's prospective scan statistic to German Meningococcal data aggregated to monthly counts for each of Germany's 413 districts
- We show the resulting scan statistics for each month of the study period (2004–2005). At each time step, the statistic was calculated using at most the latest 6 months of data
- The b_{it} for each district and time point was estimated as

$$\hat{b}_{it} = \frac{Y}{T} \cdot \frac{\text{Pop}_i}{\text{Pop}_{\text{total}}}.$$

Case Study: Meningococcal disease in Germany (2)

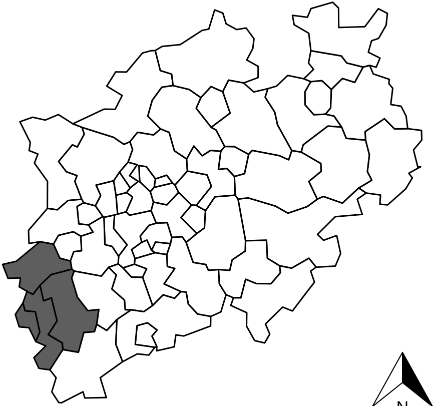


Case Study: Meningococcal disease in Germany (3)



Case Study: Meningococcal disease in Germany (4)

- The core cluster consists of four districts in North Rhine-Westphalia, one of them the city Aachen



Case Study: Meningococcal disease in Germany (5)

- An issue with the scan statistic might be that it is ill-suited for data with an abundance of zeros as the Meningococcal data
- For this type of data, a scan statistic based on e.g. the zero-inflated Poisson distribution (see Allévius et al. 2019) may perform better

Outline

1. Monitoring of univariate count data time series
2. Multivariate Methods
3. A System for Automated Outbreak Detection
4. Discussion

System Design

- Salmon et al. (2016) describes a system integrating outbreak detection algorithms into the epidemiological workflow

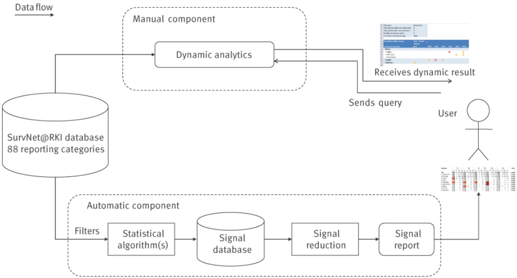


Figure source: Salmon et al. (2016)

- Example of using machine learning methods for the more than 30,000 time series

Application on Salmonella Montevideo 2009-2010

Results from the extended Farrington procedure using last five years as reference values:

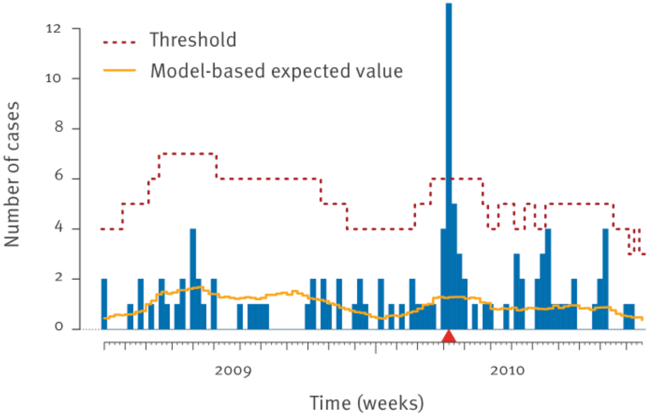
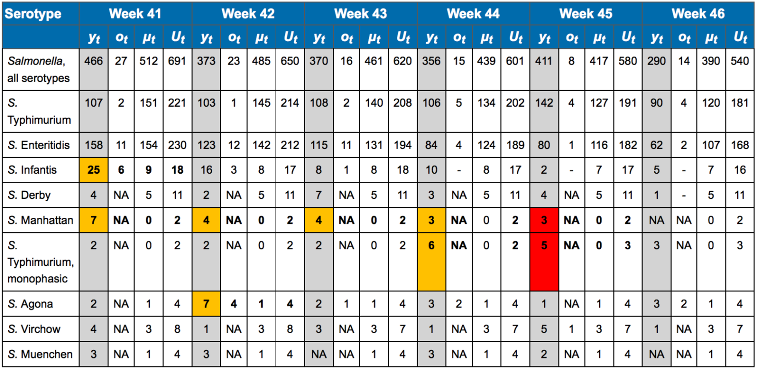


Figure source: Salmon et al. (2016)

Salmonella Report for W41–46 of 2013

Weekly Report at National Level:



Serotype	Week 41				Week 42				Week 43				Week 44				Week 45				Week 46			
	y_t	o_t	μ_t	U_t	y_t	o_t	μ_t	U_t	y_t	o_t	μ_t	U_t	y_t	o_t	μ_t	U_t	y_t	o_t	μ_t	U_t	y_t	o_t	μ_t	U_t
Salmonella, all serotypes	466	27	512	691	373	23	485	650	370	16	461	620	356	15	439	601	411	8	417	580	290	14	390	540
S. Typhimurium	107	2	151	221	103	1	145	214	108	2	140	208	106	5	134	202	142	4	127	191	90	4	120	181
S. Enteritidis	158	11	154	230	123	12	142	212	115	11	131	194	84	4	124	189	80	1	116	182	62	2	107	168
S. Infantis	**25**	**6**	**9**	**18**	16	3	8	17	8	1	8	18	10	-	8	17	2	-	7	17	5	-	7	16
S. Derby	4	NA	5	11	2	NA	5	11	7	NA	5	11	3	NA	5	11	4	NA	5	11	1	-	5	11
S. Manhattan	**7**	**NA**	**0**	**2**	**4**	**NA**	**0**	**2**	**4**	**NA**	**0**	**2**	**3**	**NA**	0	**2**	**3**	**NA**	**0**	**2**	NA	NA	0	2
S. Typhimurium, monophasic	2	NA	0	2	2	NA	0	2	2	NA	0	2	**6**	**NA**	0	**2**	**5**	**NA**	**0**	**3**	3	NA	0	3
S. Agona	2	NA	1	4	**7**	**4**	**1**	**4**	2	1	1	4	3	2	1	4	1	NA	1	4	3	2	1	4
S. Virchow	4	NA	3	8	1	NA	3	8	3	NA	3	7	1	NA	3	7	5	1	3	7	1	NA	3	7
S. Muenchen	3	NA	1	4	3	NA	1	4	NA	NA	1	4	3	NA	1	4	2	NA	1	4	NA	NA	1	4

Table source: Salmon et al. (2016)

Outline

1. Monitoring of univariate count data time series
2. Multivariate Methods
3. A System for Automated Outbreak Detection
4. Discussion

Discussion

- The presented methods are implemented in the R package `surveillance` (Salmon et al. 2016)
- Developing, maintaining and improving automatic outbreak detection systems is an interdisciplinary activity!
  - Even more work could be put into user adaptation.
  - Delay adjusted monitoring (Salmon et al. 2015)
- The system proved to be a good insurance against missing anything important – see e.g. Gertler et al. (2015)

Literature I

Allévius, B., and M. Höhle. 2019. "An unconditional space–time scan statistic for ZIP-distributed data". Preprint available as `http://bit.ly/2rFUdpR`, Scandinavian Journal of Statistics 46 (1): 142–159. doi:`10.1111/sjos.12341`.

Gertler, Maximilian, et al. 2015. "Outbreak of cryptosporidium hominis following river flooding in the city of Halle (Saale), Germany, August 2013". BMC Infectious Diseases 15 (1): 88. ISSN: 1471-2334. doi:`10.1186/s12879-015-0807-1`. `http://www.biomedcentral.com/1471-2334/15/88`.

Höhle, M., and A. Mazick. 2010. "Aberration detection in R illustrated by Danish mortality monitoring". In Biosurveillance: A Health Protection Priority, ed. by T. Kass-Hout and X. Zhang, 215–238. CRC Press. `https://staff.math.su.se/hoehle/pubs/hoehle_mazick2009-preprint.pdf`.

Literature II

Kulldorff, Martin. 2001. "Prospective time periodic geographical disease surveillance using a scan statistic". Journal of the Royal Statistical Society Series a-Statistics in Society 164:61–72.

RKI. 2012. "Salmonella Newport-Ausbruch in Deutschland und den Niederlanden, 2011". Available as `http://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2012/Ausgaben/20_12.pdf`, Epidemiologisches Bulletin, no. 20: 177–184.

Salmon, M., D. Schumacher, and M. Höhle. 2016. "Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance". Also available as vignette of the R package surveillance. Journal of Statistical Software 70 (10). doi:10.18637/jss.v070.i10.

Salmon, M., D. Schumacher, K. Stark, and M. Höhle. 2015. "Bayesian outbreak detection in the presence of reporting delays". `http://dx.doi.org/10.1002/bimj.201400159`, Biometrical Journal 57 (6): 1051–1067.