The Mathematics and Statistics of Infectious Disease Outbreaks

Michael Höhle[1]

[1]Department of Mathematics
Stockholm University, Sweden

L10: Multivariate outbreak detection[1]

[1]LaMo: 2020-08-11 @ 22:10:56

Overview

Outline

Setup

- Instead of a univariate time series $\{Y_t; t = 1, 2, \}$ as in L9 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

Outline

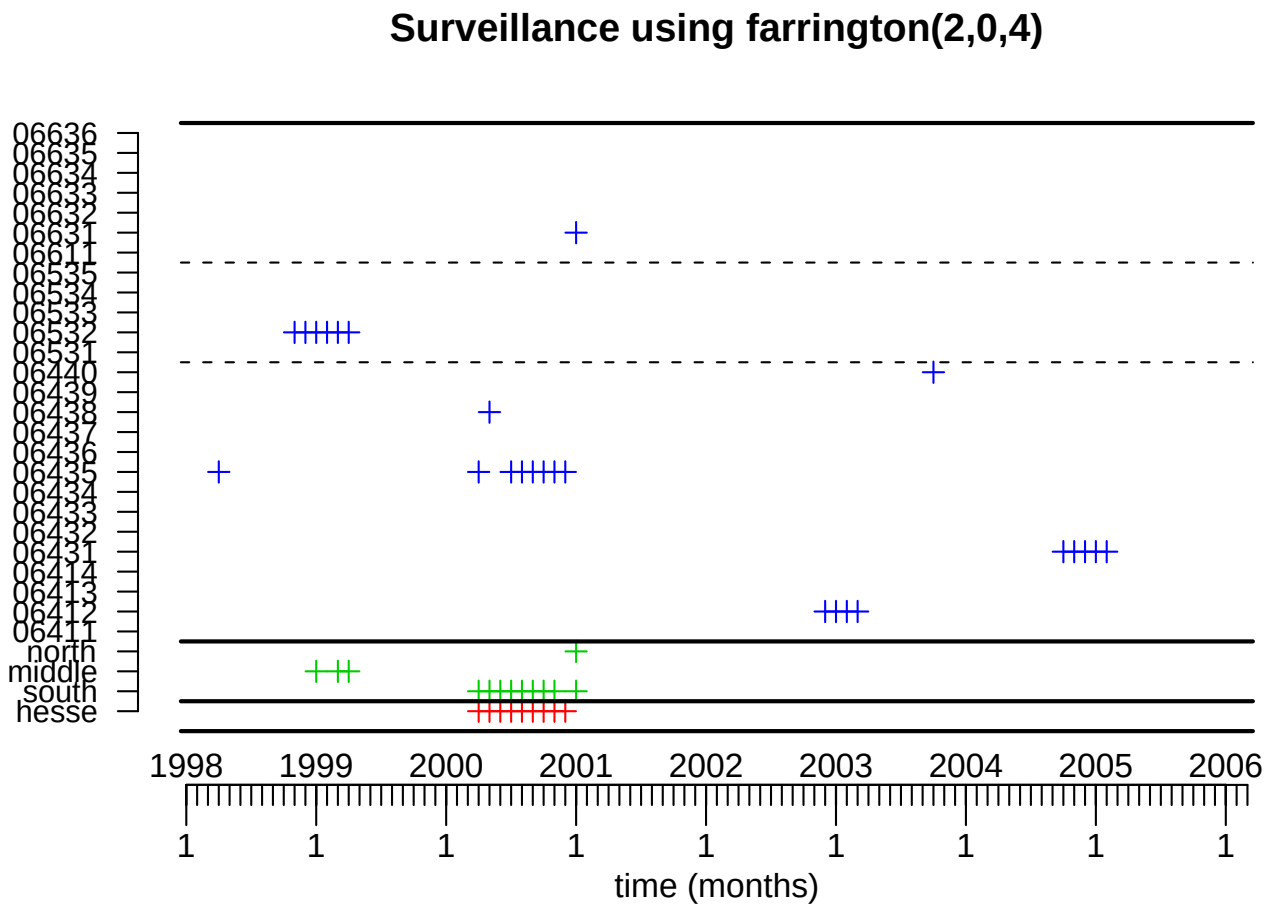
Univariate Methods in Parallel

- Simple approach for multiple data streams is to use one of the univariate methods from L9 to each time series
- Pros:
  - Easy to use, scales linearly
  - Can aggregate results in suitable fashion
- Cons:
  - False positive probability 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.

Outline

Case Study: Rabies surveillance in Hesse (1)

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



Case Study: Rabies surveillance in Hesse (2)

- An inspection of the cases in year 2000 showed that problems centered on the area around Offenbach and Frankfurt.
- A map with the coordinates of the baits with vaccine dropped from plane shows the problem.

Outline

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

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

  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

$$\lambda_W = \left(\frac{Y_W}{B_W}\right)^{Y_W}\left(\frac{Y - Y_W}{Y - B_W}\right)^{Y - Y_W} \mathbf{1}_{\{Y_W > B_W\}},$$

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

Exercises

Exercise 1

Let $B = Y$ and thus the estimate of q used under the null hypothesis will be $q = 1$. Use this to show the given likelihood ratio expression. As part of your solution explain why the equation for λ_W does not contain the term $\left(\frac{Y}{B}\right)^Y$, i.e. why can this term be neglected?

Exercise 2

Explain in your own words why the term $\mathbf{1}_{\{Y_W > B_W\}}$ is needed in the above equation.

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

Exercises

Exercise 3

State the multinomial probability for each space-time region (i, t) in the resampling. Hint: Use intuition to generalize the answer of Stackoverflow Q151272 to the sum of more than two Poisson random variables (no need for a formal proof).

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`

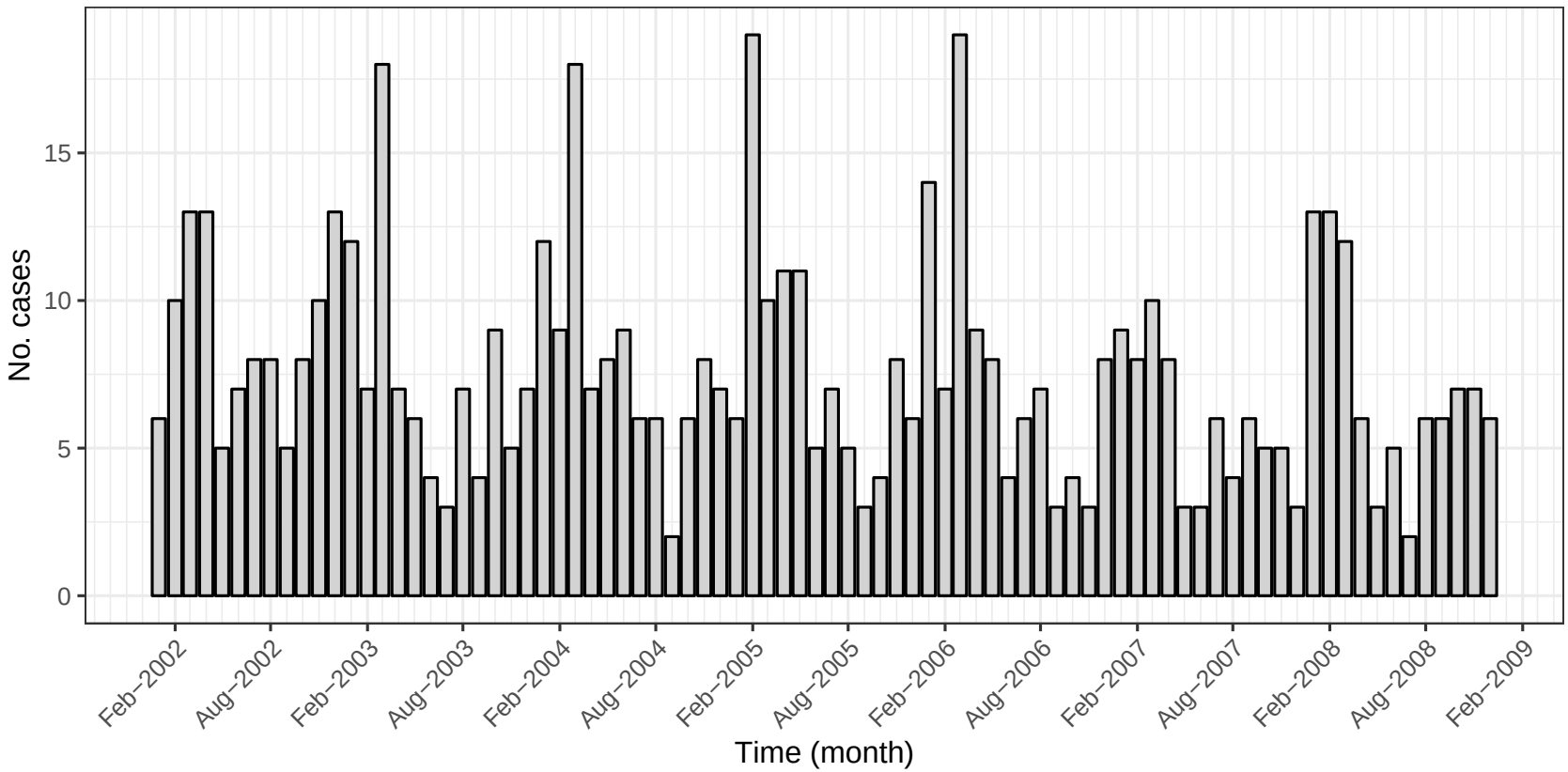
Outline

Case Study: Meningococcal disease in Germany (1)

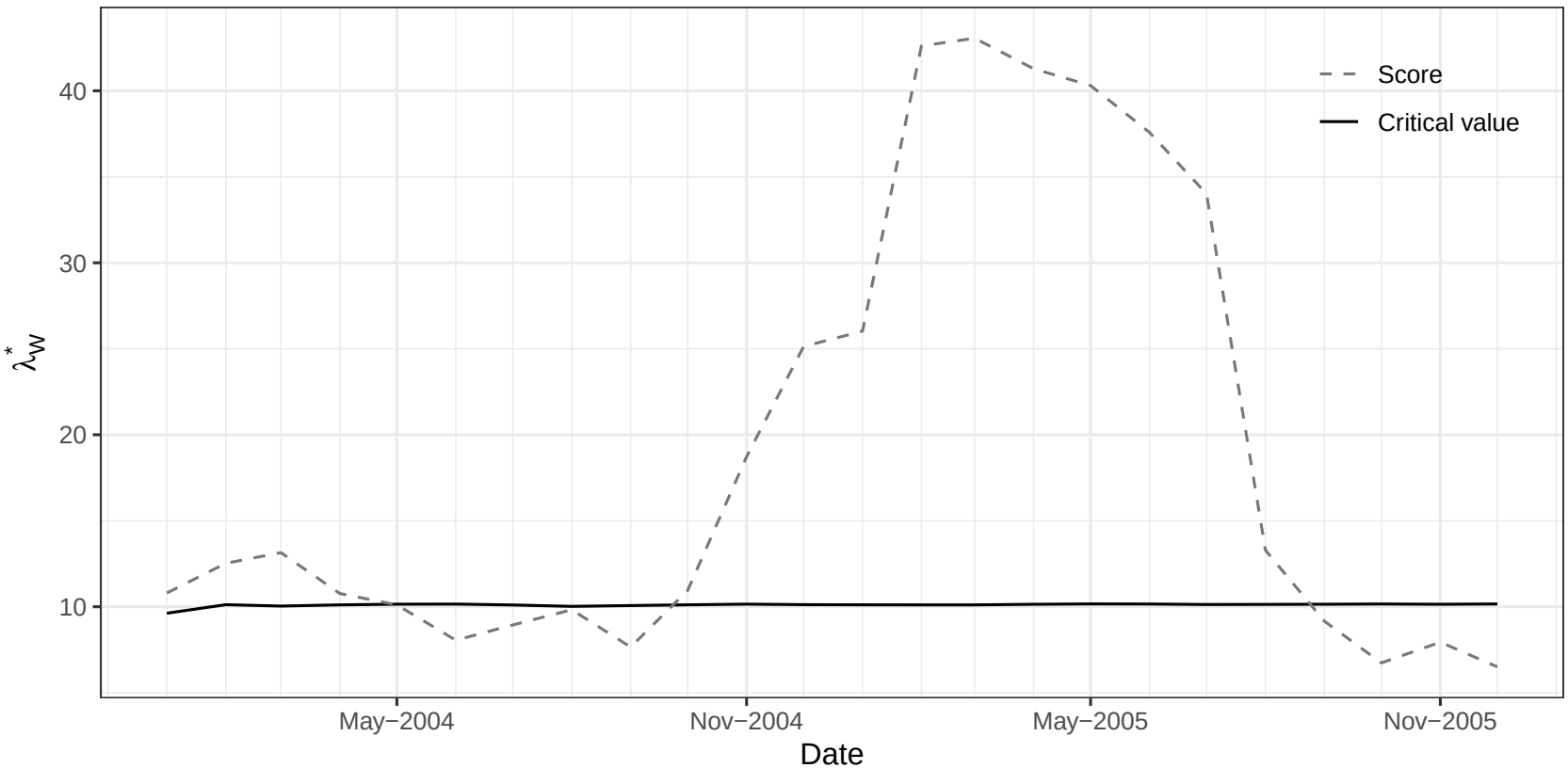
- Application of Kulldorff's prospective scan statistic to German Meningococcal data aggregated to monthly counts for each of Germany's 413 districts
- We show the resulting scan statistics for each month of the study period (2004–2005). At each time step, the statistic was calculated using at most the 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)

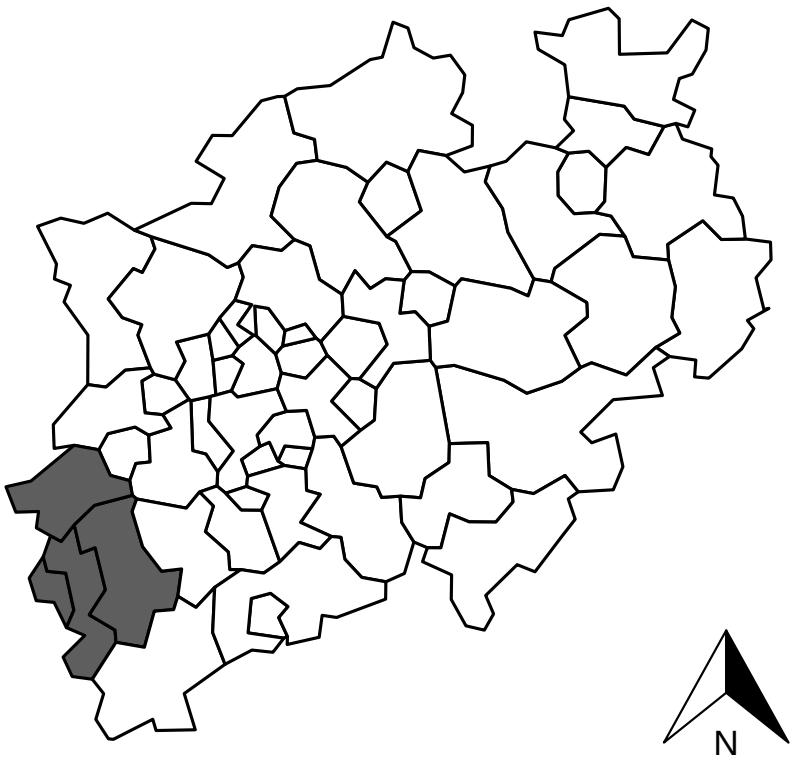


Case Study: Meningococcal disease in Germany (3)



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 and Höhle, 2019) may perform better

Literature I


Allévius, B. and M. Höhle (2019). "An unconditional space–time scan statistic for ZIP-distributed data". In: *Scandinavian Journal of Statistics* 46.1. Preprint available as http://bit.ly/2rFUdpR, pp. 142–159. DOI: 10.1111/sjos.12341.


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