Automated and Early Detection of Disease Outbreaks

Kasper Schou Telkamp

Supervisors: Jan Kloppenborg Møller, Lasse Engbo Christiansen

Master Thesis Defence

14th of August 2023

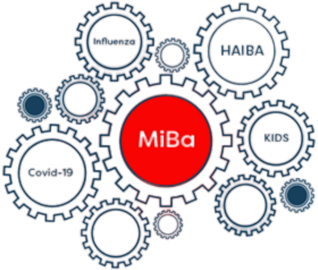
Technical University of Denmark

DTU Compute

Department of Applied Mathematics and Computer Science

Motivation

- Establishment of the Danish Microbiology Database (MiBa) by Statens Serum Institut (SSI) in 2010
- Great opportunity for data analysis
- No fully automated procedures in place at SSI



Research goals

- Review of existing literature on statistical methods for detecting disease outbreaks
- Identification and implementation of state-of-the-art methods for detection of disease outbreaks
- Formulation of hierarchical models for the individually notifiable diseases
- Development of an automated method, based on the hierarchical models, for automated and early detection of disease outbreaks
- Comparison of the developed method and state-of-the-art methods in one or more case study
- Comparison of the developed method and state-of-the-art methods in a simulation study

Research goals

- Review of existing literature on statistical methods for detecting disease outbreaks
- Identification and implementation of state-of-the-art methods for detection of disease outbreaks
- Formulation of **hierarchical** models for the individually notifiable diseases
- Development of an **automated method**, based on the hierarchical models, for automated and early detection of disease outbreaks
- Comparison of the developed method and state-of-the-art methods in **one** or more case study
- Comparison of the developed method and state-of-the-art methods in a **simulation study**

State-of-the-art algorithms

State-of-the-art algorithms for aberration detection is presented in Salmon, Schumacher, and Höhle 2016 and implemented in the R package **surveillance**. The R package includes the method introduced by Farrington et al. 1996 together with the subsequently improved method proposed by Noufaily et al. 2013.

Novel algorithm

The novel algorithm utilizes a generalized mixed effects model or a hierarchical mixed effects model as a modeling framework to model the count case observations $\boldsymbol{y}$ and assess the unobserved random effects $\boldsymbol{u}$. These random effects are used directly to characterize an outbreak.

Step 1: Modeling framework

- Assume a hierarchical Poisson Normal or Poisson Gamma model to reference data using a log link
- Incorporate covariates by supplying a model formula on the form

$$\log(\lambda_{it}) = \boldsymbol{x}_{it}\boldsymbol{\beta} + \log(n_{it}), \quad i = 1, \dots, m, \quad t = 1, \dots, T \tag{1}$$

- Account for structural changes in the time series using a rolling window of width k

Step 2: Inference of random effects

- Infer one-step ahead random effects u_{it_1} for each group using the fitted model
- Define outbreak detection threshold U_{t_0} as a quantile of the second stage model's random effects distribution
- Use either a Gaussian or Gamma distribution with respective plug-in estimates

Step 3: Parameter estimations and outbreak detection

- Compare inferred random effects u_{it_1} to an threshold U_{t_0}
- Raise and alarm if the inferred random effect exceeds the threshold, i.e. $u_{it_1} > U_{t_0}$
- Omit outbreak related observations from future parameter estimation

Formulation of hierarchical models

Poisson Normal

$$\boldsymbol{Y}|\boldsymbol{u} \sim \mathrm{Pois}\left(\boldsymbol{\lambda}\exp(\boldsymbol{u})\right)$$
$$\boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, I\sigma^2)$$

Poisson Gamma

$$\boldsymbol{Y}|\boldsymbol{u} \sim \mathrm{Pois}(\boldsymbol{\lambda}\boldsymbol{u})$$
$$\boldsymbol{u} \sim \mathrm{G}(\boldsymbol{1}/\phi, \phi)$$

Case study

Shiga toxin (verotoxin)-producing *Escherichia coli* (STEC)

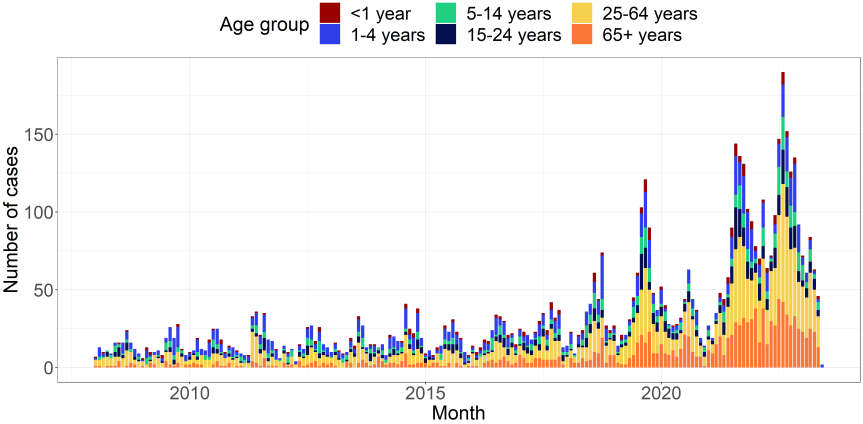


Figure: A stacked bar graph illustrating the number of monthly STEC cases observed in the period from 2008 to 2022 for the six age groups.

Constant model

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \log(n_{it}) \tag{2}$$

- λ_{it} is the outbreak intensity at time t for age group i
- $\beta(ageGroup_i)$ is the fixed effect specific to age group i
- $\log(n_{it})$ acts as an offset, accounting for the population size at time t for age group i

Trend model

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \beta_{trend}t + \log(n_{it}) \tag{3}$$

- In addition to constant model, includes a trend component
- β_{trend} quantifies the rate of change in the outbreak intensity over time

Case study

Seasonality model

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \sin\left(\frac{2\pi \cdot \tau_t}{12}\right)\beta_{\sin} + \cos\left(2\frac{\pi \cdot \tau_t}{12}\right)\beta_{\cos} + \log(n_{it}) \tag{4}$$

- In addition to constant model, incorporates an annual seasonality pattern
- τ_t represents the time period t within a year (1-12)
- $\beta_{\sin}$ and $\beta_{\cos}$ capture the effect of the seasonal pattern

Combined trend and seasonality model

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \beta_{trend}t + \sin\left(\frac{2\pi \cdot \tau_t}{12}\right)\beta_{\sin} + \cos\left(\frac{2\pi \cdot \tau_t}{12}\right)\beta_{\cos} + \log(n_{it}) \quad (5)$$

- Builds upon previous models, combining trend and seasonality components
- Includes both β_{trend}, $\beta_{\sin}$, and $\beta_{\cos}$ parameters

Estimated one-step ahead random effects

- Upper bound U_{t_0} is based on the 90% quantile of the random effects distribution
- If the one-step ahead random effects u_{it_1} exceeds U_{t_0} an alarm is raised
- In the Poisson Normal model (left), the random effects are exponentiated to transform them into the same domain as the Poisson Gamma model (right)

- 30 alarms are generated using the hierarchical Poisson Normal framework, while 31 alarms are generated using the hierarchical Poisson Gamma framework.
- A great number of alarms are generated in the period from March 2021 to March 2022

Case study

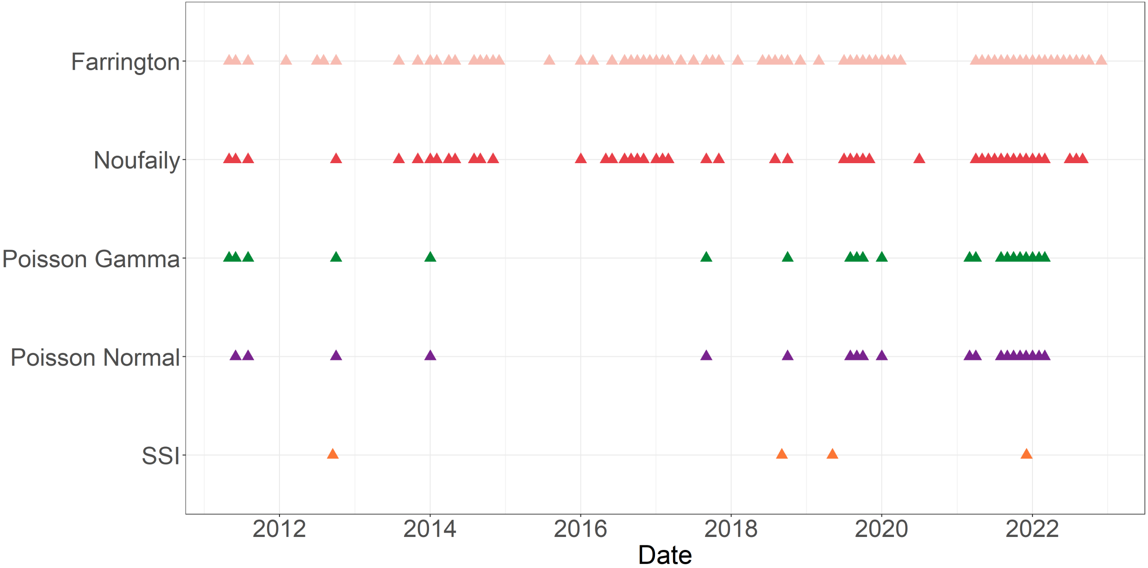
Estimated one-step ahead random effects

- Upper bound U_{t_0} is based on the 90% quantile of the random effects distribution
- If the one-step ahead random effects u_{it_1} exceeds U_{t_0} an alarm is raised
- In the Poisson Normal model (left), the random effects are exponentiated to transform them into the same domain as the Poisson Gamma model (right)

- 30 alarms are generated using the hierarchical Poisson Normal framework, while 31 alarms are generated using the hierarchical Poisson Gamma framework.
- A great number of alarms are generated in the period from March 2021 to March 2022

Case study

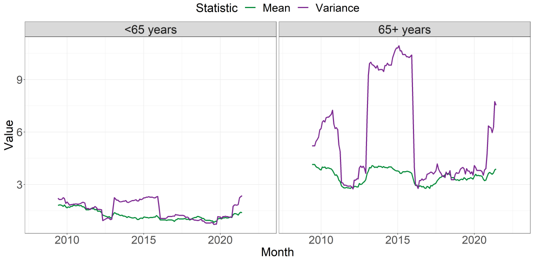
Performance of statistical outbreak detection algorithms



Case study

Challenges in statistical outbreak detection

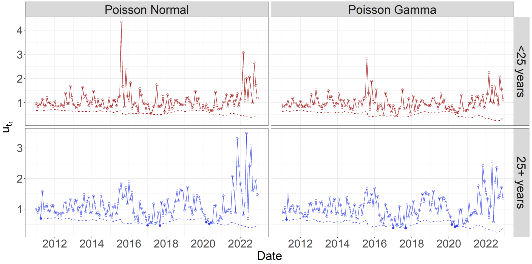
- Role of overdispersion in statistical outbreak detection
- The impact of context and observational bias
- Handling diseases with frequent outbreaks



Case study

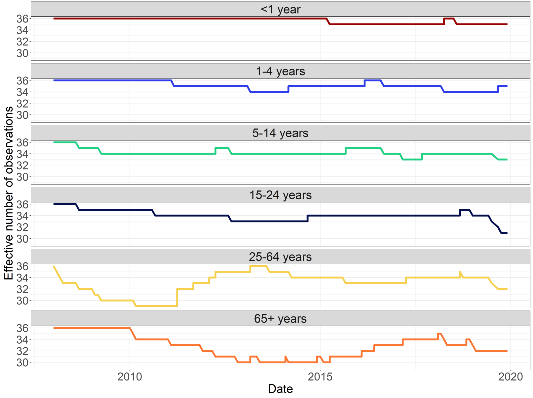
Challenges in statistical outbreak detection

- Role of overdispersion in statistical outbreak detection
- The impact of context and observational bias
- Handling diseases with frequent outbreaks



Challenges in statistical outbreak detection

- Role of overdispersion in statistical outbreak detection
- The impact of context and observational bias
- Handling diseases with frequent outbreaks



Baseline data

Simulated baseline data is generated according to a Negative Binomial distribution with mean μ and a variance parameter $\phi\mu$. The equation for the mean $\mu(t)$ is given as:

$$\mu(t) = \exp\left(\theta + \beta_t + \sum_{j=1}^{m}\left(\gamma_1 \cos\left(\frac{2\pi jt}{52}\right) + \gamma_2 \sin\left(\frac{2\pi jt}{52}\right)\right)\right) \tag{6}$$

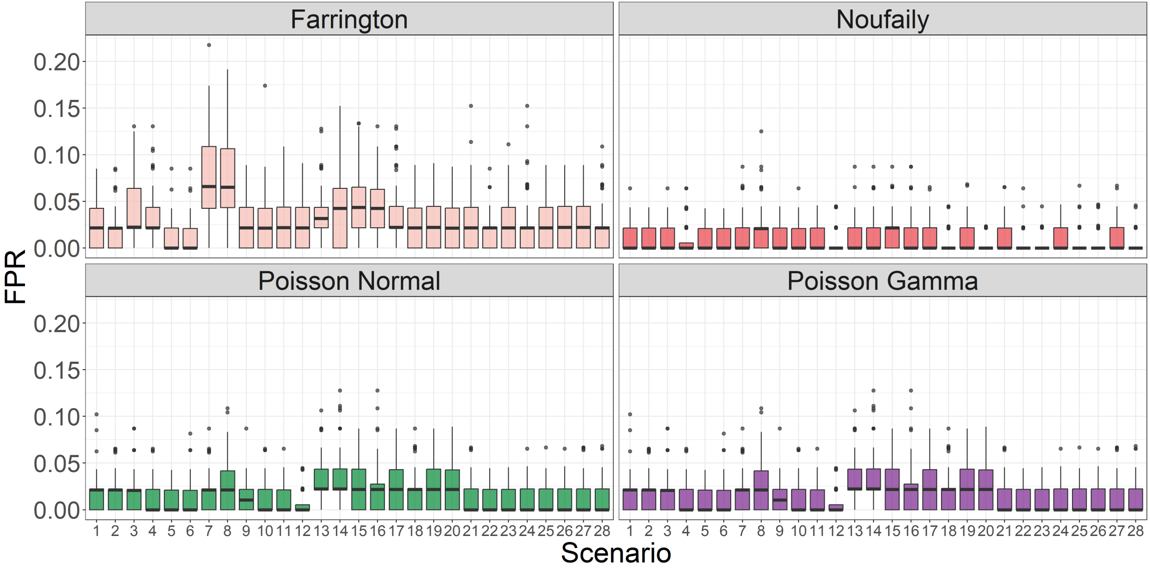
Refer to Table 6.1 in the thesis to see the 28 different scenarios

Outbreaks

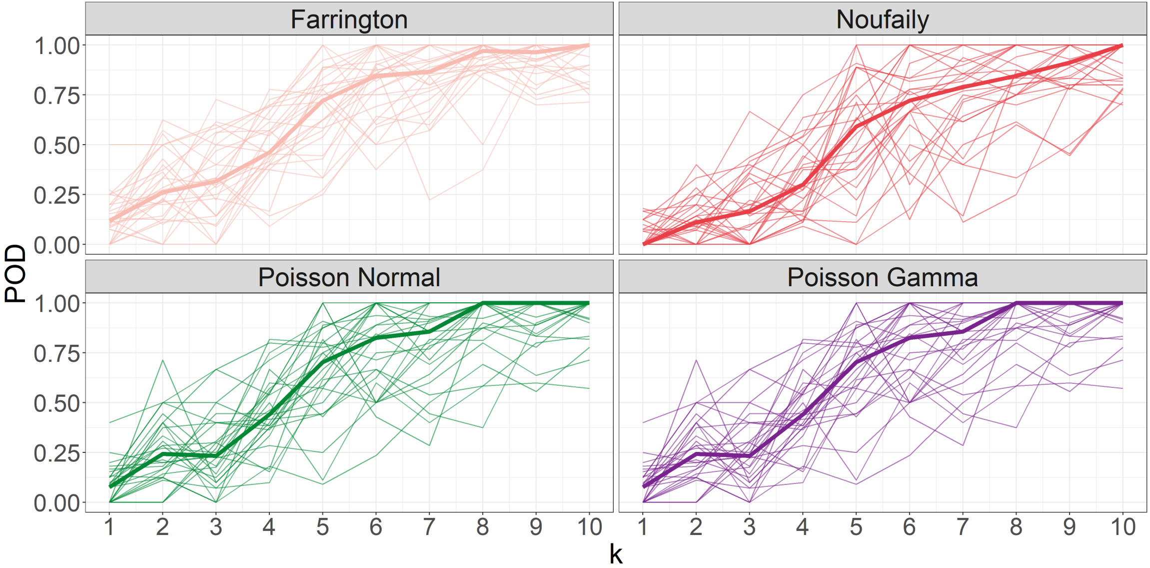
- Random constant value k is chosen
- Outbreak size v is generated from a Poisson distribution with mean equal to k times the standard deviation from the baseline data
- The v outbreak cases are distributed randomly in time according to a discretized log-normal distribution represented as $Z \sim \lfloor \mathrm{LN}(0, 0.5^2) \rfloor$

Simulation study

False Positive Rates



Probability an outbreak is detected



Campylobacter

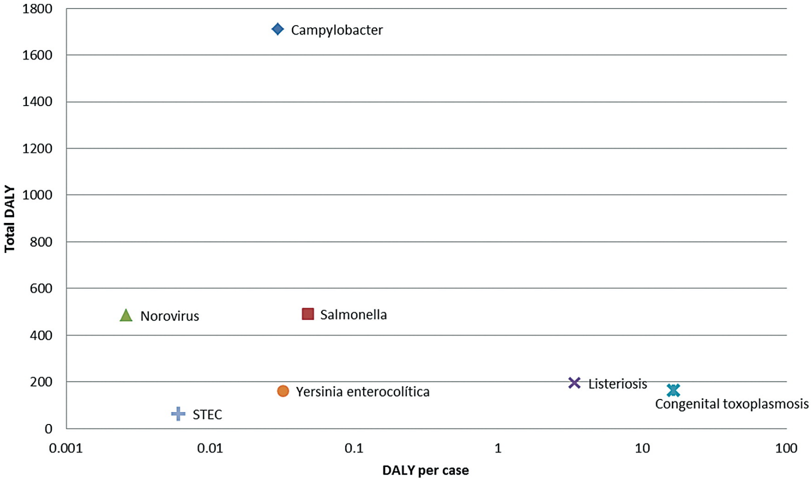


Figure: Disability adjusted life years (DALY) at the population level and at individual level. Reprinted from Pires et al. 2020.

Campylobacter

LIVE DEMONSTRATION

Summary

- Easy incorporation of **covariates**
- Estimates are **consistent** across the two modeling frameworks
- Effectively **control the number of "false alarms"**
- Great potential in utilizing **MiBa-based surveillance**

References

Farrington, C. P. et al. (1996). "A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159.3, pp. 547–563. ISSN: 09641998, 1467985X. URL: http://www.jstor.org/stable/2983331 (visited on 01/27/2023).

Noufaily, Angela et al. (2013). "An Improved Algorithm for Outbreak Detection in Multiple Surveillance Systems". en. In: *Online Journal of Public Health Informatics* 32.7, pp. 1206–1222.

Pires, Sara Monteiro et al. (2020). "Burden of Disease Estimates of Seven Pathogens Commonly Transmitted Through Foods in Denmark, 2017". English. In: *Foodborne Pathogens and Disease* 17.5. ISSN: 1535-3141. DOI: 10.1089/fpd.2019.2705.

Salmon, Maëlle, Dirk Schumacher, and Michael Höhle (2016). "Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance". In: *Journal of Statistical Software* 70.10, pp. 1–35. DOI: 10.18637/jss.v070.i10. URL: https://www.jstatsoft.org/index.php/jss/article/view/v070i10.

Probability function for Y

$$
\begin{aligned}
P[Y = y] &= g_Y(y; \boldsymbol{\beta}, \phi) \\
&= \frac{\lambda^y}{y!\Gamma(1/\phi)\phi^{1/\phi}} \frac{\phi^{y+1/\phi}\Gamma(y+1/\phi)}{(\lambda\phi+1)^{y+1/\phi}} \\
&= \frac{\Gamma(y+1/\phi)}{\Gamma(1/\phi)y!} \frac{1}{(\lambda\phi+1)^{1/\phi}} \left(\frac{\lambda\phi}{\lambda\phi+1}\right)^y \\
&= \binom{y+1/\phi-1}{y} \frac{1}{(\lambda\phi+1)^{1/\phi}} \left(\frac{\lambda\phi}{\lambda\phi+1}\right)^y, \quad \text{for } y = 0, 1, 2, \dots
\end{aligned}
\tag{7}
$$

where the following convention is used

$$
\binom{z}{y} = \frac{\Gamma(z+1)}{\Gamma(z+1-y)y!}
\tag{8}
$$

The marginal distribution of Y is a negative binomial distribution, $Y \sim \mathrm{NB}\left(1/\phi, 1/(\lambda\phi+1)\right)$

Proof

The probability function for the conditional distribution of Y for given u

$$f_{Y|u}(y; u, \boldsymbol{\beta}) = \frac{(\lambda u)^y}{y!} \exp(-\lambda u) \tag{9}$$

and the probability density function for the distribution of u is

$$f_u(u; \phi) = \frac{1}{\phi \Gamma(1/\phi)} \left(\frac{u}{\phi}\right)^{1/\phi - 1} \exp(-u/\phi) \tag{10}$$

Proof

Given (9) and (10), the probability function for the marginal distribution of Y is determined from

$$
\begin{aligned}
g_Y(y;\beta,\phi) &= \int_{u=0}^{\infty} f_{Y|u}(y;u,\beta) f_u(u;\phi)\, du \\
&= \int_{u=0}^{\infty} \frac{(\lambda u)^y}{y!} \exp(-\lambda u) \frac{1}{\phi\Gamma(1/\phi)} \left(\frac{u}{\phi}\right)^{1/\phi-1} \exp(-u/\phi)\, du \\
&= \frac{\lambda^y}{y!\Gamma(1/\phi)\phi^{1/\phi}} \int_{u=0}^{\infty} u^{y+1/\phi-1} \exp\big(-u(\lambda\phi+1)/\phi\big)\, du
\end{aligned}
\tag{11}
$$

Proof

In (11) it is noted that the integrand is the *kernel* in the probability density function for a Gamma distribution, $\mathrm{G}\left(y+1/\phi, \phi/(\lambda\phi+1)\right)$. As the integral of the density shall equal one, we find by adjusting the norming constant that

$$\int_{u=0}^{\infty} u^{y+1/\phi-1} \exp\Big(-u/(\phi/(\lambda\phi+1))\Big)\, du = \frac{\phi^{y+1/\phi}\Gamma(y+\mathbf{1}/\phi)}{(\lambda\phi+1)^{y+1/\phi}} \tag{12}$$

and then (7) follows