DTU Compute
Department of Applied Mathematics and Computer Science

Automated and Early Detection of Disease Outbreaks
AEDDO

Master Thesis

Automated and Early Detection of Disease Outbreaks
AEDDO

Master Thesis
August, 2023

By
Kasper Schou Telkamp

Cover photo: Vibeke Hempler, 2012
Published by: DTU, Department of Applied Mathematics and Computer Science, Richard Petersens Plads, Building 324, 2800 Kgs. Lyngby Denmark
www.compute.dtu.dk
ISSN: [0000-0000] (electronic version)
ISBN: [000-00-0000-000-0] (electronic version)

ISSN: [0000-0000] (printed version)
ISBN: [000-00-0000-000-0] (printed version)

Approval

This thesis has been prepared over six months at the Section for Dynamical Systems, Department of Applied Mathematics and Computer Science, at the Technical University of Denmark, DTU, in collaboration with Epidemiologisk Forskning / Modelgruppen at Statens Serum Institut, SSI, in partial fulfilment for the degree Master of Science in Engineering, MSc Eng., Quantitative Biology and Disease Modelling.

It is assumed that the reader has a basic knowledge in the areas of statistics.

Kasper Schou Telkamp - s170397

.....

Signature

.....

Date

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Acknowledgements

Lasse Engbo Christiansen, Senior Researcher, Statens Serum Institut
Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Jan Kloppenborg Møller, Associate Professor, Technical University of Denmark
Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Notation

The mathematical notation in this master's thesis is adapted from Madsen and Thyregod 2011. All vectors are column vectors. Vectors and matrices are emphasized using a bold font. Lowercase letters are used for vectors and uppercase letters letters are used for matrices. Transposing is denoted with the upper index ^{T}. Random variables are always written using uppercase letters. Thus, it is not possible to distinguish between a multivariate random variable and a matrix. However, variables and random variables are assigned to letters from the last part of the alphabet $(X, Y, Z, U, V, \dots)$, while constants are assigned to letters from the first part of the alphabet $(A, B, C, D, \dots)$. From the context it should be possible to distinguish between a matrix and a random vector.

Contents

1 Introduction

This master's thesis will primarily focus on diseases where timely detection of outbreaks results in interventions that are feasible to impose. However, the detection algorithms discussed in this master's thesis are not restricted to these diseases and can be used to detect aberrant counts in any type of disease.

Today, outbreaks of foodborne illnesses can be detected in several ways:

1. Reports from doctors: Physicians may report cases of foodborne illnesses they encounter in their practice to the relevant authorities.
2. Citizen reports: Individuals may directly contact food or health authorities to report suspected cases of foodborne illnesses.
3. Cluster identification through laboratory surveillance: Clusters of cases can be identified through routine laboratory testing and surveillance of samples from patients with suspected foodborne illnesses.
4. Identification of identical "fingerprints": When bacteria or viruses are type-tested, the presence of identical fingerprints among multi cases can strongly indicate a common source of infection.

These various methods help in the early detection and investigation of foodborne disease outbreaks, enabling timely intervention and prevention measures. However, …

While generalized linear mixed effects models and hierarchical models have earned a reputation within …, they are fairly unproven in the field of automatic detection of disease outbreaks.

The novel outbreak detection algorithms discussed in this master's thesis are open-source and can be found at https://github.com/telkamp7/AEDDO

2 Literature

In recent years, there has been a notable increase in the interest surrounding statistical methods for automated and early detection of infectious disease outbreaks. These methodologies encompass a wide range of statistical techniques, including regression analysis, time series methodology, methods inspired by statistical process control, approaches incorporating spatial information, and multivariate outbreak detection. A comprehensive review of these methods can be found in Buckeridge (2007) and Unkel et al. (2012).

In addition to the aforementioned studies, state-of-the-art methods for aberration detection are presented in Salmon, Schumacher, and Höhle (2016). These methods have been implemented in the R package called **surveillance**, which can be accessed from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/web/packages/surveillance/index.html. The **surveillance** package provides various techniques for detecting aberrations in disease surveillance data, including the Farrington method initially introduced by Farrington et al. (1996) and subsequent improvements proposed by Noufaily et al. (2013). These methods offer advanced statistical tools for detecting and monitoring disease outbreaks and are currently *the* method choice at European public health institutes (Hulth et al. 2010).

Therefore, in this master's thesis, these established methods will be compared to a novel outbreak detection algorithm based on generalized mixed effects models and hierarchical generalized linear models respectively. The thesis introduces this new algorithm as an innovative approach to outbreak detection and aims to assess its performance in comparison to existing methods.

3 Surveillance data in Denmark

This chapter delves into the data collection methods and quality assurance procedures within the Danish surveillance system. Moreover, it introduces the case studies selected for this master's thesis, which include *Listeriosis*, *Shigellosis*, Shiga toxin (verotoxin)-producing *Escherichia coli* (STEC), and *Salmonellosis*. These diseases will be referred to using either their full name or their related case definition:

- LIST: *Listeriosis*
- SHIG: *Shigellosis*
- STEC: Shiga toxin (verotoxin)-producing *Escherichia coli*
- SALM: *Salmonellosis*

3.1 Data collection and data quality

In Denmark, the surveillance of infectious diseases is conducted by Statens Serum Institut (SSI). This surveillance system plays a pivotal role in national and international disease preparedness. It encompasses more than just the collection and registration of disease data; it also involves the prompt and ongoing dissemination of knowledge to the relevant authorities responsible for treatment, prevention, and control. This comprehensive approach ensures efficient communication and facilitates appropriate measures to address infectious diseases.

The quality of the Danish surveillance registers is maintained at a high standard, thanks to The National Board of Health Statutory Order on Physicians' Notification of Infectious Diseases (https://www.retsinformation.dk/eli/lta/2000/277). This order specifies that several diseases[1] are individually notifiable by physicians and general practitioners. Notifications consist of essential patient information and are submitted in paper form to both the Ministry of Health and to SSI. This rigorous notification process ensures accurate and comprehensive data collection for disease surveillance purposes in Denmark.

In addition to the individually notifiable diseases, SSI has implemented a laboratory notification system for numerous microorganisms. Clinical-microbiological laboratories are obligated to report the identification of specific microorganisms, along with relevant patient information. These data are then stored in the Danish Microbiology Database (MiBa), which was established by SSI in 2010. MiBa is a nationwide and automatically updated database specifically designed to collect and store microbiological test results. In order to utilize the data from MiBa, the information in the test results needs to have a standardized structure with common codes and terminology. MiBa employs national standards to harmonize the data, which initially may be structured in diverse formats. The standards currently used are XRPT05, which is widely employed for the exchange of microbiological test results in the healthcare system, and a specific standard called XRPT06. These standards are regularly revised and exist in various versions[2]. The national surveillance system focuses on diseases of a severe nature, those that are highly contagious, and the majority of vaccine-preventable diseases.

[1] For a full list of diseases see https://www.ssi.dk/sygdomme-beredskab-og-forskning/anmeldelse-af-sygdomme/lovpligtige-meldesystemer/individ_anmeldelses_sygdomme

[2] Information on national standards and codes within the healthcare domain can be found on MedCom's website (https://medcom.dk/).

3.2 Introducing the case studies

For the scope of this master's thesis, only a specific subset of diseases from the mandatory notification system will be considered. This subset consists of *Listeriosis*, *Shigellosis*, Shiga toxin (verotoxin)-producing *Escherichia coli* (STEC), and *Salmonellosis*. These diseases have been chosen for analysis and investigation based on various factors, such as seasonality, incidence, and severity. Additionally, these diseases have been associated with documented outbreaks observed by SSI (https://www.ssi.dk/sygdomme-beredskab-og-forskning/sygdomsudbrud/arkiv) in the past decade, which adds to their relevance for the study. An epidemic curve graph for each of the diseases considered in this master's thesis is shown in Figure 3.1.

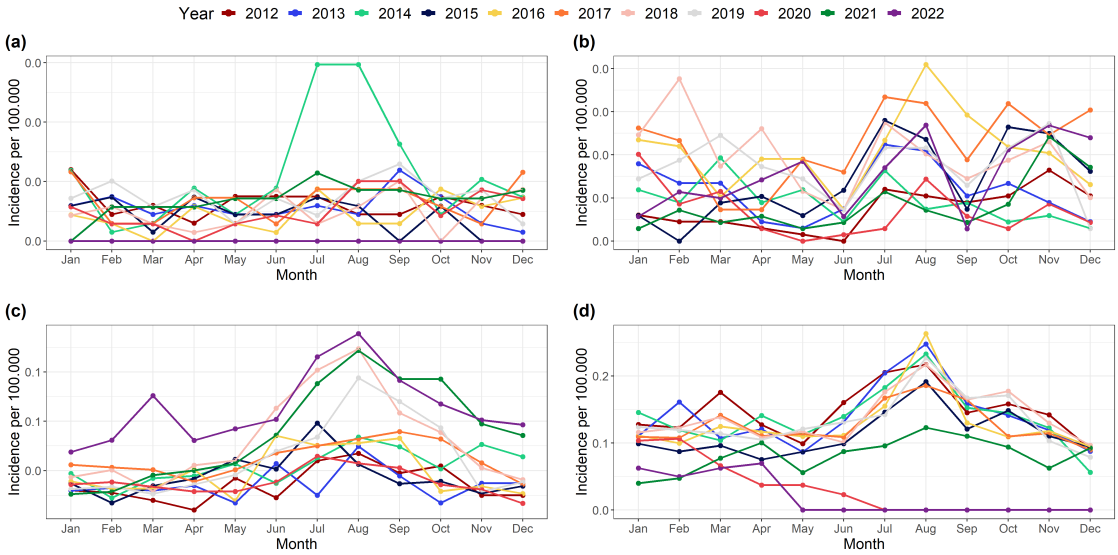

Figure 3.1: Epidemic curve showing the incidence per 100,000 in Denmark, 2012-2022, for the subset of diseases considered in this master thesis. **(a)** Listeriosis, **(b)** Shigellosis, **(c)** STEC, and **(d)** Salmonellosis.

In Figure 3.1, it is evident that all of the diseases display different levels of seasonal patterns on an annual basis.

In Figure 3.1a…

In Figure 3.1b…

Furthermore, in Figure 3.1c, a significant increase in the amplitude of the seasonal variation can be observed starting from 2018, with incidences doubling compared to the preceding years. At a first glance, this increase in the incidences might be recognized as a serious, reoccurring outbreak of the disease, but a more reasonable explanation can be found. Up to 2018, most departments of clinical microbiology used culture-based methods as a diagnostic test for bacterial pathogens and the process of changing the test method to polymerase chain reaction (PCR) methods was ongoing (Svendsen et al. 2023). In general, PCR resulted in higher incidences compared to other test methods, which is to no surprise as higher sensitivity is well documented for PCR (Buss et al. 2015; Knabl, Grutsch, and Orth-Höller 2016).

Figure 3.1d shows a significant decrease in the number of cases during the period from July to December 2020, with no cases identified during that time. This observation can

be attributed to the strict lockdown measures implemented in Denmark in response to the Covid-19 pandemic. The lockdown measures, which included restrictions on movement and social interactions, likely contributed to a reduced transmission of infectious diseases, including the one under investigation.

It is interesting to note that, unlike the specific disease being studied, the other diseases did not experience similarly severe decreases in the number of cases during the lockdown period. This could be due to various factors, including differences in the mode of transmission and susceptibility of different diseases, variations in testing and reporting practices, and the impact of preventive measures specific to each disease.

The Covid-19 pandemic and the associated lockdown measures had unique and multifaceted effects on public health and disease dynamics. While the decrease in cases observed for the studied disease during the lockdown period may be a direct result of reduced transmission, further analysis and investigation are necessary to understand the specific factors contributing to this pattern and to draw conclusive interpretations.

Some summary statistics for each of the disease considered in this master's thesis is gathered in Table 3.1.

Table 3.1: Summary statistics of the monthly count observations for the subset of diseases considered in this master's thesis. Boxplot: median (red line), IQR (grey box), whiskers (1.5 IQR), outliers (points). Time series: normalized observations (0-1), first time points minimum and maximum count (red)

Case definition	Min	Max	Mean	Median	Std. Deviation	Boxplot	Time series
LIST	0	20	4.150	4.0	3.044		
SHIG	0	28	9.078	8.0	5.652		
STEC	2	80	21.933	16.5	14.902		
SALM	0	583	101.767	83.5	83.438		

In Table 3.1, it is readily seen that the diseases exhibit different properties. WRITE SOME MORE!

All the diseases within the selected subset pose a significant risk of infection and can vary in terms of severity for affected individuals. Therefore, early identification of disease outbreaks is of utmost importance in order to promptly implement necessary interventions. Timely detection allows for swift and targeted actions to control the spread of these diseases and mitigate their impact on public health.

3.2.1 *Listeriosis*

Listeriosis is a foodborne illness that is caused by consuming food contaminated with *Listeria monocytogenes*. This disease primarily affects pregnant women, unborn or newborn babies, the elderly, and individuals with weakened immune systems. *Listeriosis* is associated with high mortality (Goulet et al. 2012) and manifests in three ways: sepsis, meningitis, and mother-to-child transmission. Pregnancy-associated *Listeriosis* can can have severe consequences for the fetus or newborn, including miscarriage, stillbirth, neonatal sepsis, and meningitis (Awofisayo et al. 2015). *Listeriosis* is uncommon among individuals in other demographic groups. The bacteria is ubiquitous in the environment, found in moist environments, soil, water, decaying vegetation, and animals. Furthermore, it can survive and even grow under refrigeration and other food preservation measures.

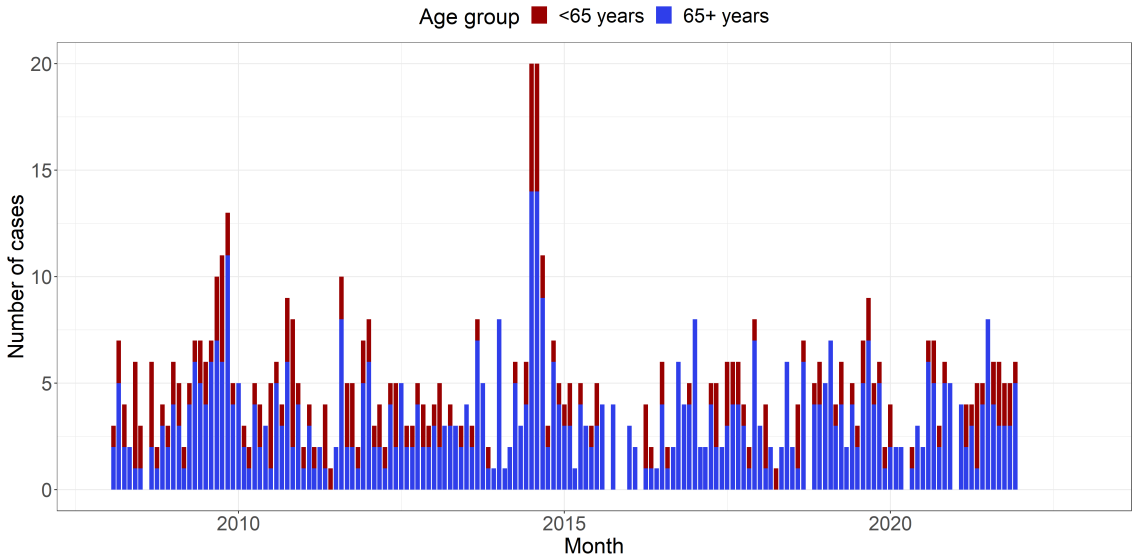

Figure 3.2: Placeholder caption

In general, SSI employs whole-genome sequencing (WGS) as the state-of-the-art method to detect disease outbreaks caused by *Listeria monocytogenes*. This method involves mapping the entire DNA of the bacteria and enables SSI to identify cases where patients are infected with identical Listeria bacteria. However, it is important to note that for this master's thesis, the DNA typing data is unavailable for use.

One notable outbreak investigated by SSI occurred between September 2013 and October 2014. This *Listeriosis* outbreak involved a total of 41 cases, resulting in 17 deaths. Deli meat products from a specific company were identified as the source of the outbreak. The high mortality rate may be attributed to the consumption of these products in nursing homes and hospitals, where patients are more vulnerable. Following the discovery of Listeria at the facility, the Danish Veterinary and Food Administration recalled all products from the company.

In another *Listeriosis* outbreak investigated by SSI, the source was traced back to cold-smoked and cured salmon products. A total of 5 related cases were identified, with 4 of them occurring in August 2017, and the fifth case in May 2017.

In some cases, despite extensive investigations, the source of contamination in an outbreak cannot always be identified. Such was the case in an unresolved outbreak that took place between May 13 and June 6, 2022. During this period, a total of nine cases were infected with the same type of Listeria, with the majority of affected patients located in the Capital Region of Denmark. Despite thorough efforts, the specific source of contamination remained unknown.

Early identification of outbreaks caused by *Listeria monocytogenes* is crucial to implement timely interventions and mitigate the impact of the disease. Otherwise, these outbreaks can persist over an extended period. SSI has successfully resolved several long-spanned outbreaks in the last decade. For example, one investigation revealed that a single outbreak was actually two simultaneous outbreaks caused by the consumption of smoked fish. Each outbreak consisted of ten cases and spanned from May 2013 to July 2015 (Gillesberg Lassen et al. 2016).

Other documented long-spanned outbreaks investigated by SSI include:

- A cold-smoked fish outbreak with 9 cases spanning from December 2016 to February 2019.
- A prolonged outbreak with 6 cases from 2016 to 2019, traced back to a local greengrocer.
- An outbreak with 8 cases from October 2021 to June 2022, caused by a deli meat product.
- Two unresolved outbreaks with 9 cases and 12 cases from the end of 2018 to November 2021 and October 2020 to May 2022, respectively.

The use of WGS in these outbreaks provided the ability to link cases that occurred over a period of years and revealed that they were, in fact, continuous-source outbreaks.

In a recent outbreak investigated by SSI, the Danish Veterinary and Food Administration, and the National Food Institute at the Technical University of Denmark, fish patties were identified as the source of contamination. This outbreak occurred from August 2022 to December 2022 and affected a total of 11 cases.

3.2.2 *Shiggellosis*

Shiggellosis is a diarrheal illness that is caused by a group of bacteria called *Shigella*. The bacteria are highly contagious and can be transmitted through direct person-to-person contact, consumption of contaminated food, or ingestion of water contaminated with human feces. *Shigella* infections are most commonly observed in children under the age of 5, individuals traveling to regions with poor sanitation and unsafe water and food practices, as well as gay, bisexual, and other men who have sex with men.

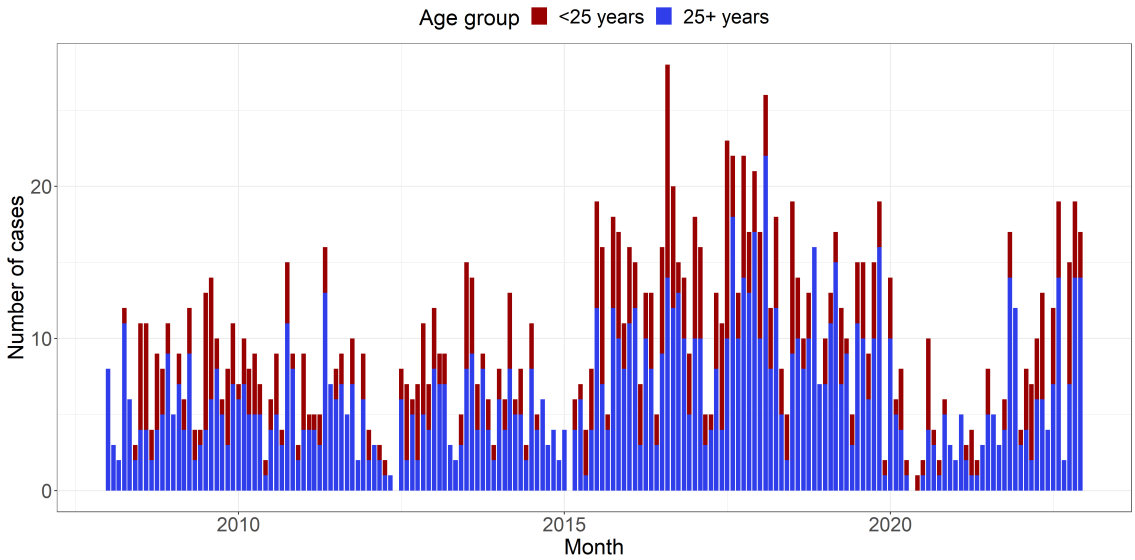

Figure 3.3: Placeholder caption

In Denmark, another significant cause of *Shigellosis* outbreaks is the importation of contaminated vegetables. This was evident in several incidents, including a 2007 outbreak where 215 individuals fell ill after consuming imported contaminated baby corn, a smaller outbreak in 2009 linked to sugar snap peas from Kenya, and a 2020 outbreak associated with fresh mint as the source of infection.

The 2020 outbreak is indeed a significant focus of this study, as it serves as a benchmark for evaluating the effectiveness of outbreak detection algorithms. It took place from August 25th to September 15th and was investigated by SSI in collaboration with the Danish Veterinary and Food Administration and the National Food Institute at the Technical University of Denmark. The outbreak affected 44 patients, mainly concentrated in the Capital Region of Denmark. During the investigation, at least five events were identified where individuals subsequently developed *Shigellosis*.

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

STEC primarily spreads through contaminated food. Less common sources of infection include contaminated drinking and bathing water, as well as direct or indirect contact with infected animals. Cattle and other ruminants are primary reservoirs for STEC serotypes that are frequently associated with human disease (Menge 2020). Therefore, in Denmark, the source of infection is often products derived from beef, non-heat-treated dairy products, or other foods such as ready-to-eat vegetables, leafy greens, vegetable sprouts, and berries contaminated with feces from cows. *Hemolytic uremic syndrome* (HUS) is a severe complication that, in some cases, particularly in children, can develop following an infection with STEC.

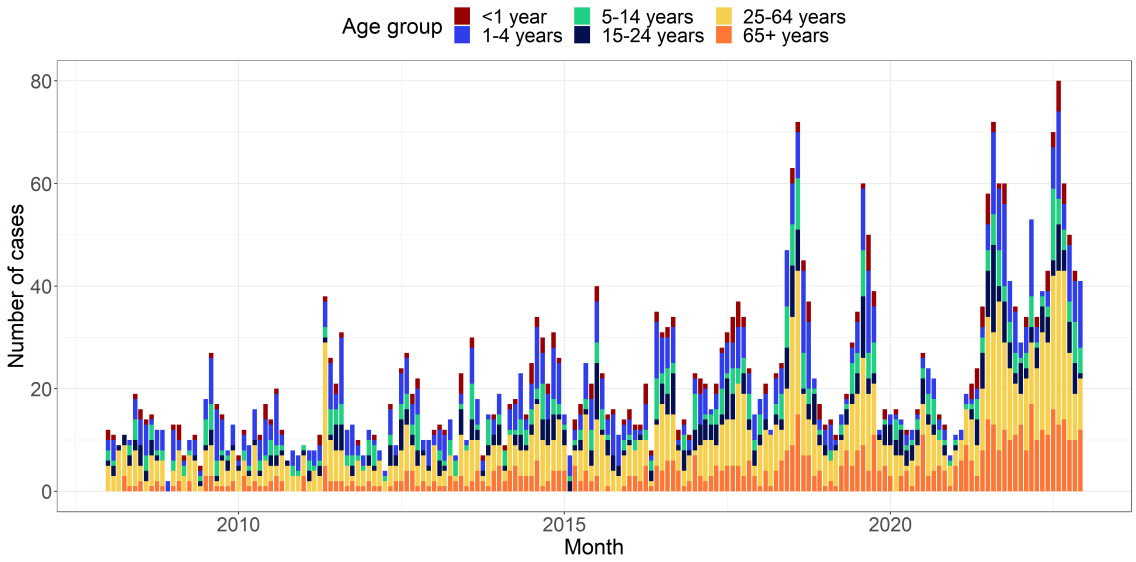

Figure 3.4: Placeholder caption

In general, stool samples are commonly used for diagnostic purposes in cases of STEC infections. Until 2018, most clinical microbiology departments relied on culture-based methods to detect and identify STEC bacteria in stool samples. However, in recent years, PCR methods have been increasingly adopted as a replacement for culture-based methods in the diagnosis of STEC infections(Svendsen et al. 2023). PCR methods offer advantages such as increased sensitivity and faster turnaround time, contributing to their growing popularity in clinical laboratories.

It is important to note that not all patients are routinely tested for STEC, and therefore, physicians need to specifically request STEC testing when submitting stool samples.

One of the earliest identified STEC outbreaks occurred in 2007, involving 18 laboratory-confirmed cases over a six-week period. The outbreak primarily affected children in day-care settings, and most patients experienced mild symptoms without bloody diarrhea.

Investigations indicated a specific brand of organic beef sausage as the likely source of infection.

In September to October 2012, a STEC outbreak with a high risk of HUS was observed. Thirteen cases were diagnosed, with eight individuals developing HUS. Epidemiological investigations suggested that ground beef was the vehicle of the outbreak (Soborg et al. 2013).

More recent outbreaks include a 38-case outbreak from September to November 2018, with a suspected association with beef sausage as the source of infection. Additionally, there were two unresolved outbreaks with 11 and 14 cases occurring from May to July 2019 and from December 2021 to January 2022, respectively. The latter outbreak included three cases of HUS.

3.2.4 *Salmonellosis*

Salmonellosis is a bacterial disease that primarily affects the intestinal tracts of humans. The *Salmonella* bacteria are commonly found in the intestines of animals and humans and are excreted in feces. Human infection typically occurs through the consumption of contaminated food or water. Salmonella infections are often associated with the consumption of raw or undercooked meat, poultry, eggs or egg products, as well as unpasteurized milk.

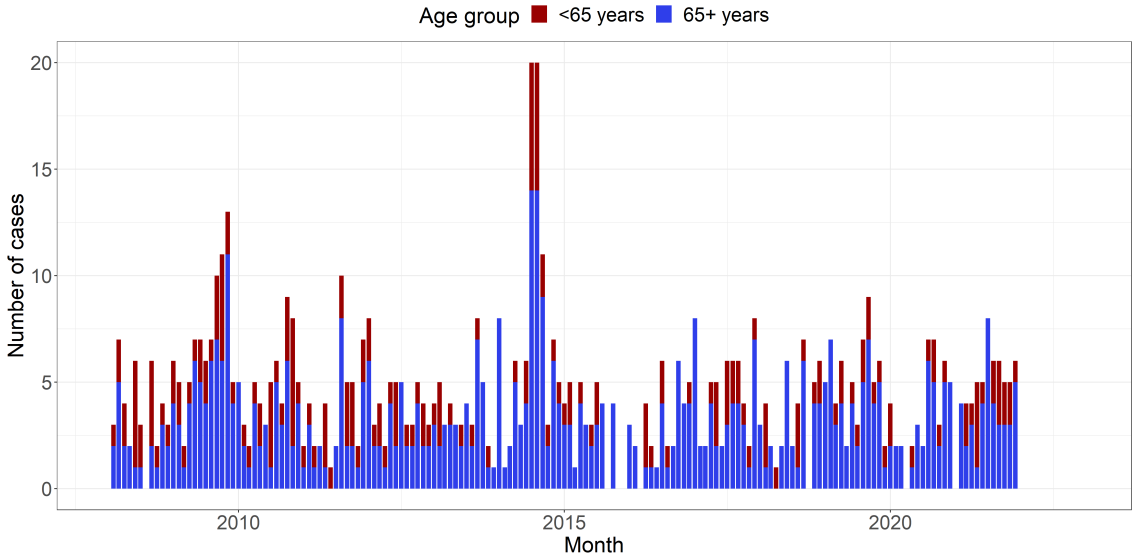

Figure 3.5: Placeholder caption

It is worth noting that an increasing proportion of infections in Denmark are now observed in connection with international travel, particularly since *Salmonella* has been eliminated from commercial chicken flocks in Denmark, making Danish eggs and poultry meat free from the bacteria. However, imported meat products can still pose a risk of contamination.

4 Methods

In this chapter, the current state-of-the-art methods for disease outbreak detection will be outlined. Furthermore, the novel outbreak detection algorithm will be introduced, along with the theory related to generalized mixed effects models and hierarchical generalized linear models. These models are utilized in this master's thesis to analyze the count observations denoted as $\boldsymbol{y}$, but more importantly they play a crucial role in assessing the unobserved random variables or random effects represented by $\boldsymbol{u}$, which are directly employed in the detection algorithm for characterizing outbreaks.

Due to the complexity of generalized mixed effects models, obtaining closed-form solutions is generally not feasible. Therefore, an overview of the Laplace approximation technique will be provided in this chapter, which allows for approximating the likelihood function in these models. Additionally, the implementation of these models in the R programming language will be presented. The presentation of this chapter is mostly inspired by Madsen and Thyregod (2011).

4.1 State-of-the-art outbreak detection algorithm

In this section, the Farrington method introduced by Farrington et al. (1996) and the subsequent improvements proposed by Noufaily et al. (2013) will be outlined. These methods are recognized as the current state-of-the-art for disease outbreak detection and will be used as benchmarks to evaluate the performance of the novel outbreak detection algorithm proposed in this master's thesis. The presentation of these methods are strongly inspired by Salmon, Schumacher, and Höhle (2016).

Both methods follow the same steps in the algorithm. The first step involves fitting an over-dispersed Poisson generalized linear model (GLM) with a log link to the reference data $y_{t_0} \subseteq \{y_t; t \leq t_0\}$. In this model, the baseline count y_t corresponding to the baseline time point t is assumed to have an expected value λ_t and a variance $\phi\lambda_t$, where $\phi \geq 1$ is ensured to account for over-dispersion. The systematic component of the model includes only a linear time trend in the frequency of reports. Therefore, the systematic component can be expressed as

$$\log(\lambda_t) = \alpha + \beta t \tag{4.1}$$

The original method incorporated seasonal effects by considering counts from comparable periods in past years for threshold calculation. This approach is similar to the one used by Stroup et al. (1989). The baseline weeks, which are used as reference, are determined by two integers: b represents the number of years back, and w represents the window half-width. For a given current week x of year y, only data from weeks $x - w$ to $x + w$ of years $y - b$ to $y - 1$ are considered, resulting in a total of $n = b(2w + 1)$ baseline weeks. The default values are $b = 5$ and $w = 3$, resulting in a total of $n = 35$ baseline values.

However, Noufaily et al. (2013) demonstrated that the algorithm performs better when utilizing more historical data without disregarding seasonality. To achieve this, the author introduced a 10-level factor with a 7-week reference period and nine additional 5-week periods in each year. As a result, the systematic component of the model is modified as follows

$$\log(\lambda_t) = \alpha + \beta t + \delta_{j(t)} \tag{4.2}$$

In this equation, $j(t)$ represents the seasonal factor level corresponding to time point t. The reference week t_0 is always associated with the reference seasonal level, denoted by $j(t_0) = 0$ and $\delta_0 = 0$.

The idea of incorporating more data while preserving seasonality has been further expanded in the implementation of the method in the **surveillance** package by Salmon, Schumacher, and Höhle (2016). The package allows the user to choose an arbitrary number of periods in each year. Consequently, the systematic component is adjusted as follows

$$\log(\lambda_t) = \alpha + \beta t + \delta_{c(t)} \tag{4.3}$$

In this equation, $c(t)$ represents the coefficients of a zero-order spline with noPeriods $+$ 1 knots. It can be conveniently represented as a noPeriods-level factor that captures seasonality. The function $c(t)$ indicates which season or period of the year t belongs to.

Furthermore, Noufaily et al. (2013) demonstrated that it is beneficial to exclude the last 26 weeks before t_0 from the baseline calculation. This exclusion helps prevent a reduction in sensitivity when an outbreak has recently started before t_0.

In the second step, the algorithm predicts the expected number of counts λ_{t_0} for the current time point t_0 using the fitted generalized linear model. Both methods differ in their assumptions for calculating the upper bound U_{t_0}.

The original method assumes that a transformation of the prediction error, denoted as $g(y_{t_0} - \hat{\lambda}_{t_0})$, follows a normal distribution. For example, when using the identity transformation $g(x) = x$, the assumption becomes

$$y_{t_0} - \hat{\lambda}_{t_0} \sim \mathsf{N}(0, \mathsf{V}(yt_0 - \hat{\lambda}_{t_0})) \tag{4.4}$$

The upper bound of the prediction interval is then calculated based on this distribution. The variance of the prediction error is given by

$$\mathsf{V}(y_{t_0} - \hat{\lambda}_{t_0}) = \mathsf{V}(\hat{y}_{t_0}) + \mathsf{V}(\hat{\lambda}_{t_0}) = \phi\lambda_{t_0} \tag{4.5}$$

Here, $\mathsf{V}(\hat{y}_{t_0})$ represents the variance of an observation, and $V(\hat{\lambda}_{t_0})$ represents the variance of the estimate. The threshold, defined as the upper bound of a one-sided $(1-\alpha)\cdot 100\%$ prediction interval, is calculated as

$$U_{t_0} = \hat{\lambda}_0 + z_{1-\alpha}\hat{V}(y_{t_0} - \hat{\lambda}_{t_0}) \tag{4.6}$$

However, this method's weakness lies in the assumption of normality itself. Therefore, an alternative assumption was presented in Noufaily et al. (2013). This approach assumes that y_{t_0} follows a negative binomial distribution, denoted as $\mathsf{NB}(\lambda_{t_0}, \nu)$, where λ_{t_0} represents the mean and $\nu = \frac{\lambda_{t_0}}{\phi - 1}$ represents the over-dispersion parameter. In this parameterization, the expected value of y_t remains λ_t, and the variance of y_t is $\phi\lambda_t$, with $\phi > 1$. If $\phi \leq 1$, a Poisson distribution is assumed for the observed count. The threshold

is defined as a quantile of the negative binomial distribution using the plug-in estimates $\hat{\lambda}_{t_0}$ and $\hat{\phi}$.

In the final step, the observed count y_{t_0} is compared to the upper bound U_{t_0}, and an alarm is raised if $y_{t_0} > U_{t_0}$. The fitting of the GLM in both methods involves three important steps.

First, the algorithm optionally performs a power transformation to correct for skewness and stabilize the variance of the data.

Next, the significance of the time trend is checked. The time trend is included in the model only if it is statistically significant at a chosen significance level, there are more than three years of reference data, and there is no over-extrapolation due to the time trend.

Finally, past outbreaks are reweighted based on their Anscombe residuals. If the Anscombe residual of a count exceeds a certain weight threshold, it is reweighted in a second fitting of the GLM. In the original method by Farrington et al. (1996), a reweighting threshold of 1 was used. However, Noufaily et al. (2013) suggests using a value of 2.56 for the weight threshold to make the reweighting procedure less drastic, as it also reduces the variance of the observations.

4.2 Novel outbreak detection algorithm

In this section, the novel algorithm for the prospective detection of disease outbreaks proposed in this master's thesis is outlined. The algorithm utilizes a generalized mixed effects model or a hierarchical generalized linear model to model the count observations $\boldsymbol{y}$ and assess the unobserved random effects $\boldsymbol{u}$. These random effects are used directly in the detection algorithm to characterize an outbreak. The theoretical foundations of these models will be further discussed in Section 4.3 and Section 4.4.

The first step involves fitting either a hierarchical Poisson Normal model or a hierarchical Poisson Gamma model with a log link to the reference data $y_{t_0} \subseteq \{y_t; t \leq t_0\}$. Here, it is possible for the user to include an arbitrary number of covariates by supplying a model formula. In order to account for structural changes in the time series, e.g. an improved and more sensitive diagnostic method or a new screening strategy at hospitals, a rolling window with width k is used to estimate the model parameters. Also, it is assumed that the count is proportional to the population size $\boldsymbol{n}$. Hence, in terms of the canonical link the model for the fixed effects is

$$\log(\lambda_{it}) = x_{it}\beta + \log(n_{it}), \quad i = 1, \ldots, m, \quad t = T, \ldots, 0 \tag{4.7}$$

Here x_{it} and β are p-dimensional vectors of covariates and fixed effects parameters respectively, where p denotes the number of covariates or fixed effects parameters, m denotes the number of groups, and T denotes the length of the period, i.e. $T = t_0 - k$.

In the second step, the algorithm infers the one-step ahead random effect $\boldsymbol{u}_{t_1}$ using the fitted model. The threshold for detecting outbreaks is defined as a quantile of the distribution of the random effects in the second stage model. This can be either a Gaussian distribution using the plug-in estimate $\hat{\sigma}$ or a Gamma distribution using the plug-in estimate $\hat{\phi}$.

In the final step, the inferred random effect u_{t_1} is compared to the upper bound U_{t_0}, and an alarm is raised if $u_{t_1} > U_{t_0}$. If an outbreak is detected, the related observation is omitted from the parameter estimation in the future. Thus, resulting in a smaller sample size for the rolling window until that specific observation is thrown away.

4.3 General mixed effects models

In this section selected theory related to generalized mixed effects models is presented. The general mixed effects model can be represented by its likelihood function

$$L_M(\boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathbb{R}^q} L(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y}) d\boldsymbol{u} \tag{4.8}$$

where $\boldsymbol{y}$ is the observed random variable, $\boldsymbol{\theta}$ is the model parameters to be estimated and $\boldsymbol{U}$ is the q unobserved random variables. The likelihood function L is the joint likelihood of both the observed and the unobserved random variables. The likelihood function for estimating $\boldsymbol{\theta}$ is the marginal likelihood L_M obtained by integrating out the unobserved random variables. In general it is difficult to solve the integral in (4.8) if the number of unobserved random variables is more than a few and hence numerical methods must be used.

4.3.1 Hierarchical models

It is useful to formulate the model as a hierarchical model containing a *first stage model*

$$f_{Y|u}(\boldsymbol{y}; \boldsymbol{u}, \boldsymbol{\beta}) \tag{4.9}$$

which is a model for the observed random variables given the unobserved random variables, and a *second stage model*

$$f_U(\boldsymbol{u}; \boldsymbol{\Psi}) \tag{4.10}$$

which is a model for the unobserved random variables. Here $\boldsymbol{\beta}$ represent the fixed effects parameters and $\boldsymbol{\Psi}$ is a model parameter. The total set of parameters is $\boldsymbol{\theta} = (\beta, \Psi)$. Hence the joint likelihood is given as

$$L(\boldsymbol{\beta}, \boldsymbol{\Psi}; \boldsymbol{u}, \boldsymbol{y}) = f_{Y|u}(\boldsymbol{y}; \boldsymbol{u}, \boldsymbol{\beta}) f_U(\boldsymbol{u}; \boldsymbol{\Psi}) \tag{4.11}$$

To obtain the likelihood for the model parameters $(\boldsymbol{\beta}, \boldsymbol{\Psi})$ the unobserved random variables are integrated out. The likelihood function for estimating $(\boldsymbol{\beta}, \boldsymbol{\Psi})$ is as in (4.8) the marginal likelihood

$$L_M(\boldsymbol{\beta}, \boldsymbol{\Psi}; \boldsymbol{y}) = \int_{\mathbb{R}^q} L(\boldsymbol{\beta}, \boldsymbol{\Psi}; \boldsymbol{u}, \boldsymbol{y}) d\boldsymbol{u} \tag{4.12}$$

where q is the number of unobserved random variables, and $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ are the parameters to be estimated.

4.3.2 Laplace Approximation

The Laplace approximation will be outlined in the following. A thorough description of the Laplace approximation in nonlinear mixed effects models is found in Wolfinger and Lin (1997).

For a given set of model parameters $\boldsymbol{\theta}$ the joint log-likelihood $\ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) = \log\left(L(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})\right)$ is approximated using a second order Taylor approximation around the optimum $\tilde{\boldsymbol{u}} = \hat{\boldsymbol{u}}_{\boldsymbol{\theta}}$ of the log-likelihood function w.r.t. the unobserved random variables $\boldsymbol{u}$, i.e.,

$$\ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) \approx \ell(\boldsymbol{\theta}, \tilde{\boldsymbol{u}}, \boldsymbol{y}) - \frac{1}{2}(\boldsymbol{u} - \tilde{\boldsymbol{u}})^T \boldsymbol{H}(\tilde{\boldsymbol{u}})(\boldsymbol{u} - \tilde{\boldsymbol{u}}) \tag{4.13}$$

where the first-order term of the Taylor expansion disappears since the expansion is done around the optimum $\tilde{\boldsymbol{u}}$ and $\boldsymbol{H}(\tilde{\boldsymbol{u}}) = -\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})|_{\boldsymbol{u}=\tilde{\boldsymbol{u}}}$ is the negative Hessian of the joint log-likelihood evaluated at $\tilde{\boldsymbol{u}}$.

It is readily seen that the joint log-likelihood for the hierarchical model specified in Section 4.3.1 is

$$\ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) = \ell(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{u}, \boldsymbol{y}) = \log f_{Y|u}(\boldsymbol{y}; \boldsymbol{u}, \boldsymbol{\beta}) + \log f_U(\boldsymbol{u}; \boldsymbol{\Psi}) \tag{4.14}$$

which implies that the Laplace approximation becomes

$$\ell_{M,LA}(\boldsymbol{\theta}, \boldsymbol{y}) = \log f_{Y|u}(\boldsymbol{y}; \tilde{\boldsymbol{u}}, \boldsymbol{\beta}) + \log f_U(\tilde{\boldsymbol{u}}, \boldsymbol{\Psi}) - \frac{1}{2} \log \left| \frac{\boldsymbol{H}(\tilde{\boldsymbol{u}})}{2\pi} \right| \tag{4.15}$$

4.3.3 Formulation of the hierarchical Poisson Normal model

In order to simplify the notation, the probability density functions are presented for a specific observation and hence the subscripts indicating the group and time are omitted. The conditional distribution of the count observations is assumed to be a Poisson distribution with intensities $\boldsymbol{\lambda}$

$$f_{Y|u}(y; u, \boldsymbol{\beta}) = \frac{\lambda \exp(u)^y}{y!} \exp\big(- \lambda \exp(u)\big) \tag{4.16}$$

Also, it is assumed that the count is proportional to the population size $\boldsymbol{x}$. Hence, in terms of the canonical link for the Poisson distribution the model for the fixed effects is

$$\log(\lambda_{it}) = \boldsymbol{X}_t^T \boldsymbol{\beta} + \log(x_{it}), \quad i = 1, \ldots, m, \quad t = 1, \ldots, T \tag{4.17}$$

The probability density function for the distribution of the random effects is assumed to follow a Gaussian distribution, $\boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, sigma^2)$, i.e.

$$f_U(u; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{u^2}{2\sigma^2} \right) \tag{4.18}$$

where σ is a model parameter.

Henceforth, the total set of parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ and the model can be formulated as a two-level hierarchical model

$$\boldsymbol{Y}|\boldsymbol{u} \sim \mathrm{Pois}\left(\boldsymbol{\lambda} \exp(\boldsymbol{u})\right) \tag{4.19a}$$

$$\boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, I\sigma^2) \tag{4.19b}$$

The joint likelihood for the count observations $\boldsymbol{y}$ and the random effects $\boldsymbol{u}$ becomes

$$L(\boldsymbol{\beta}, \sigma; u_{it}, y_{it}) = \prod_{t=1}^{T}\prod_{i=1}^{m} \frac{(\lambda_{it}\exp(u_{it}))^{y_{it}}}{y_{it}!}\exp\big(-\lambda_{it}\exp(u_{it})\big) \prod_{t=1}^{T}\prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{u_{it}^2}{2\sigma^2}\right) \tag{4.20}$$

4.4 Hierarchical generalized linear models

In this section selected theory related to hierarchical generalized linear models is presented. The model class was initially formulated by Lee and Nelder (1996) as a natural generalization of the generalized linear models to also incorporate random effects. A starting point in hierarchical modelling is an assumption that the distribution of random effects may be modeled by an exponential dispersion family. This family of models were first introduced by Fisher and Russell (1922), and has proven to play an important role in mathematical statistics because of their simple inferential properties. The exponential dispersion family considers a family of distributions, which can be written on the form

$$f_Y(y;\theta) = c(y,\lambda)\exp\big(\lambda\{\theta y - \kappa(\theta)\}\big) \tag{4.21}$$

Here the parameter $\lambda > 0$ is called the *precision parameter*, which in some cases represents a known shape parameter as for the Gamma distribution. In other cases the precision parameter represents an over-dispersion that is not related to the mean. These distributions combine with the so-called *standard conjugate distributions* in a simple way, and lead to marginal distributions that may be expressed in a closed form suited for likelihood calculations.

4.4.1 Standard conjugate distribution

Now the general notion of a *standard conjugate distribution* for an exponential dispersion family is introduced.

Consider an exponential dispersion family ED$(\mu, \mathrm{V}(\mu)/\lambda)$ with density (4.21) for $\theta \in \Omega$. Let $\mathcal{M} = \tau(\Omega)$ denote the mean value space for this family. Let $m \in \mathcal{M}$ and consider

$$g_\theta(\theta; m, \gamma) = \frac{1}{C(m,\gamma)}\exp\Big(\frac{\theta m - \kappa(\theta)}{\gamma}\Big) \tag{4.22}$$

with

$$C(m,\gamma) = \int_\Omega \exp\Big(\frac{\theta m - \kappa(\theta)}{\gamma}\Big)d\theta \tag{4.23}$$

for $\gamma \in \mathbb{R}_+$ for which the integral converges. Then (4.22) defines the density function of a probability distribution for θ. This distribution is called the *standard conjugate distribution* for θ corresponding to (4.21).

4.4.2 Definition of the hierarchical generalized linear model

Consider a set of observations $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)^T$ such that for a given value of a parameter θ the distribution of Y_i is given by an exponential dispersion model with density (4.21) and with canonical parameter space Ω (for θ), mean value $\mu = \kappa'(\theta)$, mean value space $\mathcal{M}$ (for μ) and canonical link $\theta = g(\mu)$.

Let the *conjugate distribution* of θ be given by (4.22), and the corresponding conjugate distribution of μ, (e.g., $f_Y(y)$ Poisson distribution; $g_\mu(\mu)$ Gamma distribution; link $g(\mu) = \log(\mu)$).

The variables in a hierarchical generalized linear model are

i) the observed responses $y_1, y_2, \ldots, y_k \quad (\in \mathcal{M})$

ii) the unobserved random effects $u_1, u_2, \ldots, u_q \quad (\in \mathcal{M})$

iii) and the corresponding unobserved canonical variables $v_i = g(u_i) \quad (\in \Omega)$

The *linear predictor* is of the form

$$\boldsymbol{\theta} = g(\boldsymbol{\mu}|\boldsymbol{v}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} \tag{4.24}$$

The distribution of $\boldsymbol{V} \in \Omega$ is a conjugated distribution to the canonical parameter θ. The derived distribution of $\boldsymbol{U} \in \mathcal{M}$ is the corresponding conjugated distribution to the mean value parameter μ such that $\mathrm{E}[U] = \psi$. When the conditional distribution of $Y|\mu$ is a Poisson distribution, and the distribution of V is constructed in such a way that the distribution of $U = \log(V)$ is a Gamma distribution with mean value $\mathrm{E}[U] = \psi = 1$, then it follows that the distribution of Y is a negative binomial distribution with parameters determined by $\boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{Z}\boldsymbol{v}$.

4.4.3 Formulation of the hierarchical Poisson Gamma model

In the compound Poisson Gamma model the conditional distribution of the count observations are assumed to be a Poisson distribution with intensities $\boldsymbol{\lambda}$

$$f_{Y|u}(y; u, \boldsymbol{\beta}) = \frac{(\lambda u)^y}{y!} \exp(-\lambda u) \tag{4.25}$$

The probability density function for the random effects $\boldsymbol{u}$ are assumed to follow a reparametrized Gamma distribution with mean $\mathbf{1}$, $\boldsymbol{u} \sim \mathrm{G}(\mathbf{1}/\phi, \phi)$ that is

$$f_u(u; \phi) = \frac{1}{\phi\Gamma(1/\phi)} \left(\frac{u}{\phi}\right)^{1/\phi - 1} \exp(-u/\phi) \tag{4.26}$$

Subsequently, the model can be formulated as a two-level hierarchical model

$$\boldsymbol{Y}|\boldsymbol{u} \sim \mathrm{Pois}(\boldsymbol{\lambda}\boldsymbol{u}) \tag{4.27a}$$

$$\boldsymbol{u} \sim \mathrm{G}(\mathbf{1}/\phi, \phi) \tag{4.27b}$$

Given (4.25) and (4.26), the probability function for the marginal distribution of $\boldsymbol{Y}$ is determined from

$$\begin{aligned} g_Y(y; \boldsymbol{\beta}, \phi) &= \int_{u=0}^{\infty} f_{Y|u}(y; ,u, \beta) f_u(u; \phi)\, du \\ &= \int_{u=0}^{\infty} \frac{(\lambda u)^y}{y!} \exp(-\lambda u) \frac{1}{\phi\Gamma(1/\phi)} \left(\frac{u}{\phi}\right)^{1/\phi - 1} \exp(-u/\phi)\, du \\ &= \frac{\lambda^y}{y!\Gamma(1/\phi)\phi^{1/\phi}} \int_{u=0}^{\infty} u^{y+1/\phi-1} \exp\big(-u(\lambda\phi + 1)/\phi\big)\, du \end{aligned} \tag{4.28}$$

In (4.28) it is noted that the integrand is the *kernel* in the probability density function for a Gamma distribution, G $\big(y+1/\phi, \phi/(\lambda\phi+1)\big)$. As the integral of the density shall equal one, we find by adjusting the norming constant that

$$\int_{u=0}^{\infty} u^{y+1/\phi-1} \exp\Big(-u/\Big(\phi/(\lambda\phi+1)\Big)\Big)\, du = \frac{\phi^{y+1/\phi}\Gamma(y+\mathbf{1}/\phi)}{(\lambda\phi+1)^{y+1/\phi}} \tag{4.29}$$

Therefore, it can be shown that the marginal distribution of Y is a negative binomial distribution, $Y \sim$ NB $\big(1/\phi, 1/(\lambda\phi+1)\big)$. The probability function for Y is

$$\begin{aligned} P[Y=y] &= g_Y(y;\boldsymbol{\beta},\phi) \\ &= \frac{\lambda^y}{y!\Gamma(1/\phi)\phi^{1/\phi}} \frac{\phi^{y+1/\phi}\Gamma(y+1/\phi)}{(\lambda\phi+1)^{y+1/\phi}} \\ &= \frac{\Gamma(y+1/\phi)}{\Gamma(1/\phi)y!} \frac{1}{(\lambda\phi+1)^{1/\phi}} \left(\frac{\lambda\phi}{\lambda\phi+1}\right)^y \\ &= \binom{y+1/\phi-1}{y} \frac{1}{(\lambda\phi+1)^{1/\phi}} \left(\frac{\lambda\phi}{\lambda\phi+1}\right)^y, \quad \text{for } y=0,1,2,\ldots \end{aligned} \tag{4.30}$$

where we have used the convention

$$\binom{z}{y} = \frac{\Gamma(z+1)}{\Gamma(z+1-y)y!} \tag{4.31}$$

for z real and y integer values. Consequently, the mean and variance of Y are given by

$$\mathsf{E}[Y] = \lambda \qquad \mathsf{V}[Y] = \lambda(\lambda\phi+1) \tag{4.32}$$

The likelihood function for estimating $(\boldsymbol{\beta}, \phi)$ is

$$L(\boldsymbol{\beta},\phi;y_{it}) = \prod_{t=1}^{T}\prod_{i=1}^{m} \binom{y_{it}+1/\phi-1}{y_{it}} \frac{1}{(\lambda_{it}\phi+1)^{1/\phi}} \left(\frac{\lambda_{it}\phi}{\lambda_{it}\phi+1}\right)^{y_{it}} \tag{4.33}$$

Inference on individual groups
In order to simplify the notation, the subscript indicating the group and time are omitted. Consider the compound Poisson Gamma model in (4.27), and assume that a value $Y=y$ has been observed.

Then the conditional distribution of u for given $Y=y$ is found using Bayes Theorem

$$\begin{aligned} g_u(u|Y=y) &= \frac{f_{y,u}(y,u)}{g_Y(y;\lambda,\phi)} \\ &= \frac{f_{y|u}(y;u)g_u(u)}{g_Y(y;\lambda,\phi)} \\ &= \frac{1}{g_Y(y;\lambda,\phi)} \left(\frac{(\lambda u)^y}{y!}\exp(-\lambda u) \frac{1}{\phi\Gamma(1/\phi)} \left(\frac{u}{\phi}\right)^{1/\phi-1} \exp(-u/\phi)\right) \\ &\propto u^{y+1/\phi-1} \exp\big(-u(\lambda\phi+1)/\phi\big) \end{aligned} \tag{4.34}$$

We identify the *kernel* of the probability density function

$$u^{y+1/\phi-1} \exp(-u(\lambda\phi+1)/\phi) \tag{4.35}$$

as the kernel of a Gamma distribution, $\mathsf{G}(y+1/\phi, \phi/(\lambda\phi+1))$, i.e. the conditional distribution of u for given $Y=y$ can be written as

$$u|Y=y \sim \mathsf{G}\left(y+1/\phi, \phi/(\lambda\phi+1)\right) \tag{4.36}$$

The mean of the conditional distribution is given by:

$$\mathsf{E}[u|Y=y] = \frac{y\phi+1}{\lambda\phi+1} \tag{4.37}$$

And the variance of the conditional distribution is:

$$\mathsf{V}[u|Y=y] = \frac{(\phi^2+\phi)}{(\lambda\phi+1)^2} \tag{4.38}$$

These formulas provide the mean and variance of the conditional distribution of u given the observed value $Y=y$.

Why do we choose the Gamma distribution to represent the variation between days?
The choice of the Gamma distribution for modeling the random effects has been motivated by several reasons. Firstly, the support of the Gamma distribution, which ranges from 0 to infinity, aligns with the mean-value space, denoted as $\mathcal{M}$, for the Poisson distribution. This ensures that the random effects are constrained within a meaningful range for the underlying Poisson process.

Secondly, the two-parameter family of Gamma distributions offers considerable flexibility, encompassing a wide range of shapes and distributions that can span from exponential-like distributions to fairly symmetrical distributions on the positive real line. This flexibility allows the model to capture various patterns and characteristics observed in the data.

Additionally, the choice of the Gamma distribution has benefits in terms of the derivation of the marginal distribution of the response variable Y. The kernel $u^{\alpha-1}\exp(-u/\beta)$ of the Gamma distribution used for modeling the random effects exhibits a similar structure to the kernel $u^y \exp(-u)$ of the likelihood function corresponding to the sampling distribution of Y. This similarity facilitates the analytical computation of the integral involved in deriving the marginal distribution, as it can be expressed in terms of known functions.

Overall, the Gamma distribution is selected due to its alignment with the mean-value space of the Poisson distribution, its flexibility in capturing diverse distributions, and its analytical convenience in computing the marginal distribution of the response variable.

4.5 Parameter estimation

In this section, the parameter estimation and implementation of the novel outbreak detection algorithm in R using the **TMB** (Template model Builder) package is presented. **TMB** is an open-source R package developed by Kristensen et al. (2016). This package facilitates efficient maximum likelihood estimation and uncertainty calculations for the parameter set $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Psi})$ and random effects $\boldsymbol{u}$. The presentation of the parameter estimation

conducted in **TMB** is strongly inspired by Chapter 2 in Kristensen et al. (2016) and Section 5.10 in Madsen and Thyregod (2011).

The **TMB** package maximizes a user-provided objective function in the form of a C++ template to estimate the maximum likelihood for the parameter set $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Psi})$. See Appendix B to access the C++ template files used in this master's thesis. The objective function maximizes the marginal log-likelihood function, which integrates out the random effects $\boldsymbol{u}$

$$\ell_M(\boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathbb{R}^q} \ell(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y}) d\boldsymbol{u} \tag{4.39}$$

where $\ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})$ is the joint log-likelihood function of the data given the parameters and random effects. We use $\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}$ to denote the maximizer of the joint log-likelihood $\ell(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y})$ w.r.t. $\boldsymbol{u}$; i.e.,

$$\hat{\boldsymbol{u}}_{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{u}} \ell(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y}) \tag{4.40}$$

Using $H(\hat{\boldsymbol{u}}_{\boldsymbol{\theta}})$ to denote the negative Hessian of the joint log-likelihood evaluated at $\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}$; i.e,

$$H(\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}) = -\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}} \tag{4.41}$$

The Laplace approximation for the marginal log-likelihood $\ell_M(\boldsymbol{\theta})$ is

$$\ell_{M,LA}(\boldsymbol{\theta}, \boldsymbol{y}) = \ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) - \frac{1}{2} \log \left| \frac{H(\hat{\boldsymbol{u}}_{\boldsymbol{\theta}})}{2\pi} \right| \tag{4.42}$$

Our estimate of θ minimizes the negative log of the Laplace approximation, i.e.,

$$-\ell_{M,LA}(\boldsymbol{\theta}, \boldsymbol{y}) = -\ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) + \frac{1}{2} \log \left| \frac{H(\hat{\boldsymbol{u}}_{\boldsymbol{\theta}})}{2\pi} \right| \tag{4.43}$$

The maximization of the Laplace approximation for the marginal likelihood is then performed using conventional R optimization routines (e.g., BFGS) to optimize the objective and obtain our estimate $\hat{\boldsymbol{\theta}}$. Uncertainty of the estimate $\hat{\boldsymbol{\theta}}$, or any differentiable function of the estimate $\phi(\hat{\boldsymbol{\theta}})$, is obtained by the δ-method:

$$\mathrm{V}\left(\phi(\hat{\boldsymbol{\theta}})\right) = -\phi'_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \Big(\Delta^2 \ell_{M,LA}(\hat{\boldsymbol{\theta}}, \boldsymbol{y}) \Big)^{-1} \phi'_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})^T \tag{4.44}$$

WRITE SOMETHING ABOUT THE LIKELIHOOD ESTIMATION IN THE HIERARCHICAL POISSON GAMMA MODEL!

Additionally, **TMB** utilized Automatic Differentiation (AD) techniques (Griewank and Walther 2008) to evaluate first, second, and potentially third-order derivatives. For a comprehensive introduction to the concept of AD, it is recommended to read Section 2.1 and Section 2.2 of Fournier et al. (2012). This approach enhances the computational efficiency and accuracy of the parameter estimation process in the implemented models.

4.6 Scoring rule

In this section, the scoring rule used to evaluate the overall score of the models is outlined. The approach is inspired by Bjerregård, Møller, and Madsen (2021). For a given time series $y_t = y_1, \ldots, y_N$, each forecast and its corresponding realized observation pair (G_t, y_t) is evaluated. The overall score of the model is then reported as the average score:

$$\bar{S}(G, y) = \frac{1}{N} \sum_{t=1}^{N} S(G_t, y_t) \tag{4.45}$$

One commonly used scoring rule is the *logarithmic score* derived from likelihood theory, which is defined as $S(G, y) = -\log\big(f(y)\big)$ (Good 1992). This scoring rule is based on the probability density function and is equivalent to the log-likelihood of the forecast model. It has desirable properties as it captures all possible information about the observed data in relation to the model. However, it has a potential drawback in that it heavily penalizes unlikely observations. Consequently, even small changes in the tails of a density forecast can lead to significant changes in the *logarithmic score*, even when the overall shape of the density remains unchanged.

The calculation of the logarithmic score is shown in Example 4.1, which is adapted from Bjerregård, Møller, and Madsen (2021).

Example 4.1 (Calculation of the logarithmic score)**.** The Gamma distribution is used to represent the probabilistic forecast in this example. The Gamma distribution is parametrized by two parameters, shape (α) and rate (β), and its probability density function (PDF) is given by:

$$f(y) = \frac{1}{\Gamma(\alpha)\beta} \left(\frac{y}{\beta}\right)^{\alpha-1} \exp\left(-\frac{y}{\beta}\right) \tag{4.46}$$

In this example, the parameters of the true model, denoted as f, is chosen to be $(\alpha, \beta) = (3, 3)$. We simulate 10 observations, denoted as $y_1, y_2, \ldots, y_{10}$, which are shown in Table 4.1. The true model f is compared to a competing model, denoted as g, which is a Gamma distribution with parameters $(\alpha, \beta) = (3, 8)$. The true model f, the competing model g, and the observations $\boldsymbol{y}$ are illustrated in Figure 4.1.

Table 4.1: 10 simulated observations following a G(3,3)-distribution.

i	1	2	3	4	5	6	7	8	9	10
y_i	1.278	2.233	0.657	0.831	1.281	0.287	1.363	1.612	0.790	0.161

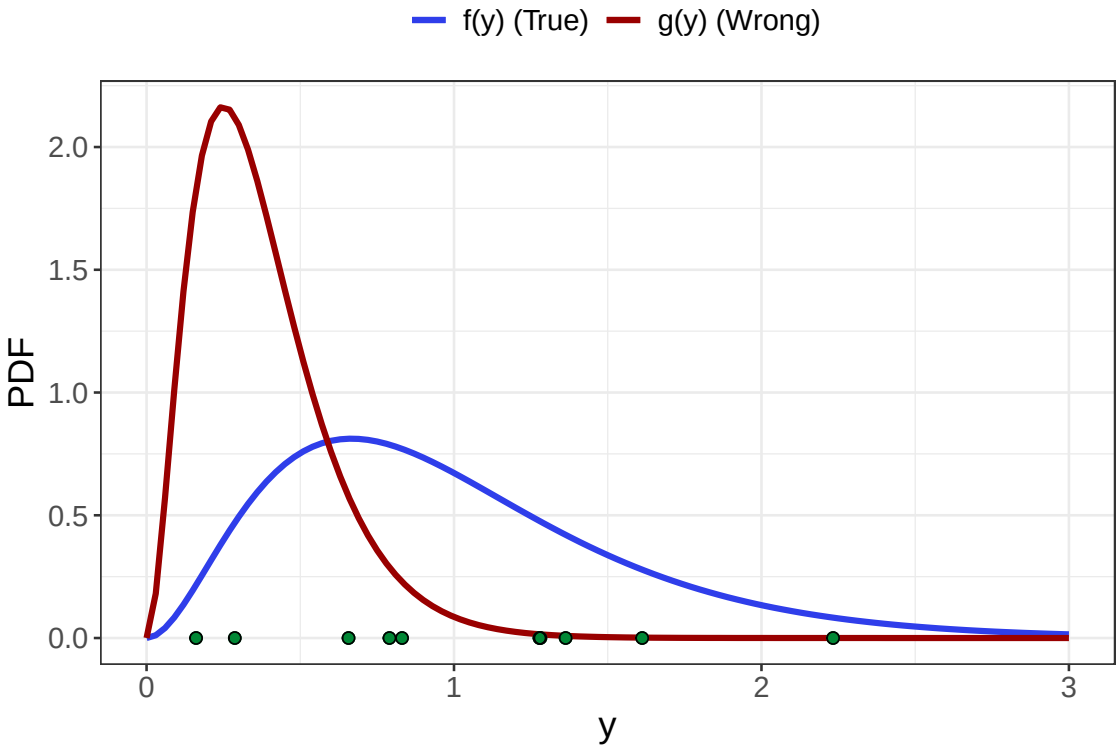

Figure 4.1: Observations (green dots) were simulated from a G$(3, 3)$-distribution. The true model f (blue) is shown, along with the PDF for the competing model g, which follows a $G(3, 8)$-distribution.

The logarithmic score of the true model f for the first observation, $y_1 = 1.278$, is calculated as follows

$$\begin{aligned} -\log\big(f(y_1)\big) &= -\log\big(f(1.278)\big) \\ &= -\log\left[\frac{1}{\Gamma(3)4}\Big(\frac{1.278}{4}\Big)^{3-1}\exp(-1.278/4)\right] \\ &= -1.156 \end{aligned} \tag{4.47}$$

Similarly, the logarithmic scores for the other observations can be calculated using the same formula. The individual logarithmic scores for all 10 observations are presented in Table 4.2. Among the 10 observations, 8 of them are more likely to occur under the true model f compared to the competing model g. The final key quantity, the average logarithmic score, $\bar{S}(G, \boldsymbol{y})$, can be calculated using Equation (4.45).

$$\begin{aligned} \bar{S}(f, \boldsymbol{y}) &= 1.1 \\ \bar{S}(g, \boldsymbol{y}) &= 3.22 \end{aligned} \tag{4.48}$$

Table 4.2: Logarithmic scores of the two different Gamma models w.r.t. the 10 individual observations.

i	1	2	3	4	5	6	7	8	9	10
$\bar{S}(f, y_i)$	1.16	3.86	0.00	0.23	1.16	0.18	1.37	2.03	0.17	0.83
$\bar{S}(g, y_i)$	4.19	10.71	0.55	1.47	4.21	-0.75	4.74	6.40	1.25	-0.60

5 Results

In this chapter…

5.1 Case studies

In this section, the results of applying the novel outbreak detection algorithm to the subset of diseases will be presented. The performance of the algorithm in detecting outbreaks in the selected diseases will be thoroughly discussed and analyzed. Additionally, a comparative analysis will be conducted, evaluating the performance of the novel outbreak detection algorithm against the current state-of-the-art methods, namely the Farrington method (Farrington et al. 1996) and the improved Noufaily method (Noufaily et al. 2013). This comparative analysis will provide valuable insights into the strengths and limitations of the novel algorithm, as well as its potential contributions to the field of outbreak detection.

5.1.1 *Listeriosis*

Consider the models given in (4.19) and (4.27)

5.1.2 *Shiggellosis*

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

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

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \log(n_{it}) \tag{5.1}$$

$$\log(\lambda_{it}) = \beta(ageGroup_i) + \sin\left(\frac{\pi \cdot \tau_t}{6}\right)\beta_{\sin} + \cos\left(\frac{\pi \cdot \tau_t}{6}\right)\beta_{\cos} + \log(n_{it}) \tag{5.2}$$

Here τ_t represents the month in year as a decimal number (01-12)-

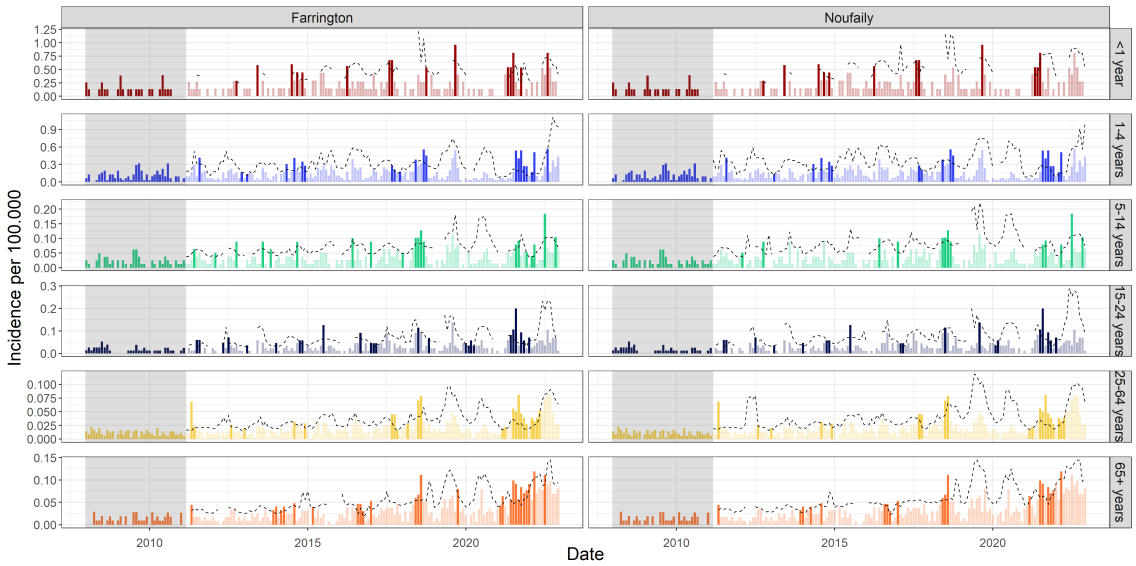

Figure 5.1: Placeholder caption

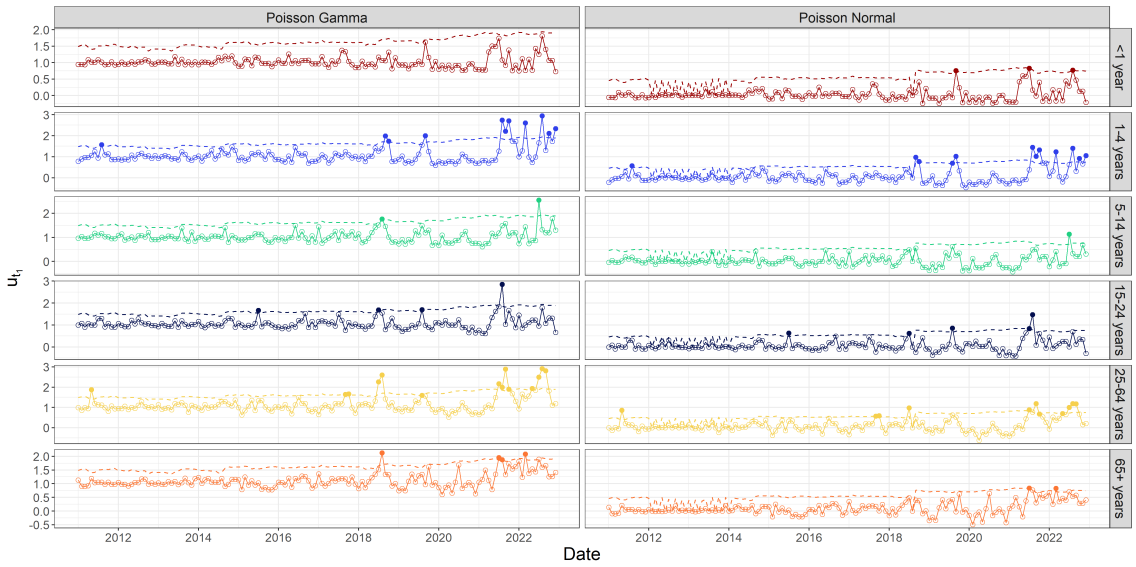

Figure 5.2: Placeholder caption

Compare_alarms.png

5.1.4 *Salmonellosis*

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5.2 Simulation study

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5.3 Performance comparison of novel methods

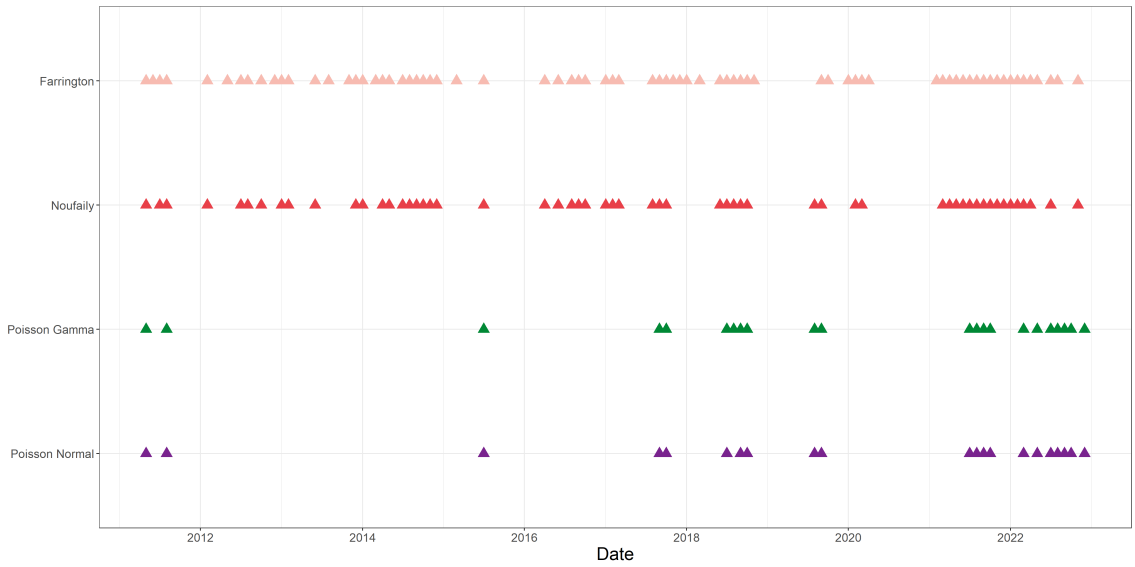

Figure 5.3: Placeholder caption

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

6 Discussion

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

In the future, the utilization of MiBa-based surveillance has immense potential for disease surveillance. It has already demonstrated its value in various surveillance systems, such as the Healthcare-Associated Infections Database (HAIBA) for monitoring hospital-acquired infections and the COVID-19 surveillance system.

HAIBA, launched in 2015, was the first fully automated surveillance system built on MiBa data. It provides monitoring capabilities for hospital-acquired infections, enabling healthcare professionals to track and manage these infections more effectively. Similarly, the COVID-19 surveillance system, developed during 2020 and 2021, utilizes MiBa data to monitor and respond to the COVID-19 pandemic.

In addition to these systems, MiBa-based surveillance includes monitoring respiratory infections (such as influenza, pertussis, Mycoplasma pneumonia, and respiratory syncytial virus) and sexually transmitted diseases like chlamydia. While these surveillance systems currently have partial automation in data processing, there are plans to fully automate them in the near future.

Expanding on the field of automated disease outbreak detection is crucial to fully harness the potential of MiBa. By developing advanced algorithms and methodologies, it becomes possible to automatically analyze MiBa data and detect disease outbreaks in a timely manner. This can lead to early identification of outbreaks, allowing for prompt interventions and preventive measures.

Further research and development in automated disease outbreak detection, specifically tailored to leverage MiBa data, can significantly enhance our ability to detect and respond to infectious disease outbreaks more proactively and efficiently. By maximizing the potential of MiBa-based surveillance and continuously improving automated detection methods, we can strengthen our overall disease surveillance efforts and better protect public health.

7 Conclusion

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Bibliography

Madsen, Henrik and Poul Thyregod (2011). *Introduction to general and generalized linear models*. English. Texts in statistical science. CRC Press. ISBN: 9781420091557.

Buckeridge, David L. (2007). "Outbreak detection through automated surveillance: A review of the determinants of detection". In: *Journal of Biomedical Informatics* 40.4. Public Health Informatics, pp. 370–379. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2006.09.003. URL: https://www.sciencedirect.com/science/article/pii/S1532046406000980.

Unkel, Steffen et al. (2012). "Statistical methods for the prospective detection of infectious disease outbreaks: a review". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 175.1, pp. 49–82. ISSN: 09641998, 1467985X. URL: http://www.jstor.org/stable/41409708 (visited on 02/15/2023).

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.

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

Hulth, A. et al. (2010). "Practical usage of computer-supported outbreak detection in five european countries". In: *Eurosurveillance* 15.36. Cited by: 31, pp. 1–6. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77957607700&partnerID=40&md5=ff916d06bcf38218b2388d0874eed9f8.

Svendsen, Anna Tølbøll et al. (2023). "The incidence of laboratory-confirmed cases of enteric pathogens in Denmark 2018: a national observational study". In: *Infectious Diseases* 55.5. PMID: 36868794, pp. 340–350. DOI: 10.1080/23744235.2023.2183253. eprint: https://doi.org/10.1080/23744235.2023.2183253. URL: https://doi.org/10.1080/23744235.2023.2183253.

Buss, Sarah N. et al. (2015). "Multicenter evaluation of the BioFire FilmArray gastrointestinal panel for etiologic diagnosis of infectious gastroenteritis". In: *Journal of Clinical Microbiology* 53.3. Cited by: 327; All Open Access, Bronze Open Access, Green Open Access, pp. 915–925. DOI: 10.1128/JCM.02674-14. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84923376418&doi=10.1128%2fJCM.02674-14&partnerID=40&md5=a31098bbc8437c1849360924f660dacd.

Knabl, L., I. Grutsch, and D. Orth-Höller (2016). "Comparison of the BD MAX® Enteric Bacterial Panel assay with conventional diagnostic procedures in diarrheal stool samples". In: *European Journal of Clinical Microbiology and Infectious Diseases* 35.1. Cited by: 22, pp. 131–136. DOI: 10.1007/s10096-015-2517-4. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84954361539&doi=10.1007%2fs10096-015-2517-4&partnerID=40&md5=2800622c88a129b3ed2c8dbf533d2a39.

Goulet, Véronique et al. (2012). "Incidence of listeriosis and related mortality among groups at risk of acquiring listeriosis". In: *Clinical infectious diseases* 54.5, pp. 652–660.

Awofisayo, A. et al. (Jan. 2015). “Pregnancy-associated listeriosis in England and Wales”. In: *EPIDEMIOLOGY AND INFECTION* 143.2, pp. 249–256. ISSN: 0950-2688. DOI: 10.1017/S0950268814000594.

Gillesberg Lassen, S. et al. (2016). “Two listeria outbreaks caused by smoked fish consumption—using whole-genome sequencing for outbreak investigations”. In: *Clinical Microbiology and Infection* 22.7, pp. 620–624. ISSN: 1198-743X. DOI: https://doi.org/10.1016/j.cmi.2016.04.017. URL: https://www.sciencedirect.com/science/article/pii/S1198743X16301148.

Menge, Christian (2020). “The Role of Escherichia coli Shiga Toxins in STEC Colonization of Cattle”. In: *Toxins* 12.9. ISSN: 2072-6651. DOI: 10.3390/toxins12090607. URL: https://www.mdpi.com/2072-6651/12/9/607.

Soborg, B et al. (2013). “A verocytotoxin-producing E. coli outbreak with a surprisingly high risk of haemolytic uraemic syndrome, Denmark, September-October 2012”. In: *Eurosurveillance* 18.2, 20350. DOI: https://doi.org/10.2807/ese.18.02.20350-en. URL: https://www.eurosurveillance.org/content/10.2807/ese.18.02.20350-en.

Stroup, D. F. et al. (Mar. 1989). “Detection of aberrations in the occurrence of notifiable diseases surveillance data”. In: *Statistics in medicine* 8.3, pp. 323–329. ISSN: 0277-6715. DOI: 10.1002/sim.4780080312.

Wolfinger, Russell D. and Xihong Lin (1997). “Two Taylor-series approximation methods for nonlinear mixed models”. In: *Computational Statistics and Data Analysis* 25.4. Cited by: 87, pp. 465–490. DOI: 10.1016/S0167-9473(97)00012-1. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0031222254&doi=10.1016%2fS0167-9473%2897%2900012-1&partnerID=40&md5=dbf7ca1e47d836029af4044ff9880463.

Lee, Y. and J. A. Nelder (1996). “Hierarchical Generalized Linear Models”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.4, pp. 619–678. ISSN: 00359246. URL: http://www.jstor.org/stable/2346105 (visited on 05/16/2023).

Fisher, R. A. and Edward John Russell (1922). “On the mathematical foundations of theoretical statistics”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222.594-604, pp. 309–368. DOI: 10.1098/rsta.1922.0009. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.1922.0009. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1922.0009.

Kristensen, Kasper et al. (2016). “TMB: Automatic Differentiation and Laplace Approximation”. In: *Journal of Statistical Software* 70.5, pp. 1–21. DOI: 10.18637/jss.v070.i05.

Griewank, Andreas and Andrea Walther (2008). *Evaluating Derivatives*. Second. Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9780898717761. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9780898717761. URL: https://epubs.siam.org/doi/abs/10.1137/1.9780898717761.

Fournier, David A. et al. (2012). “AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models”. In: *Optimization Methods and Software* 27.2, pp. 233–249. DOI: 10.1080/10556788.2011.597854. eprint: https://doi.org/10.1080/10556788.2011.597854. URL: https://doi.org/10.1080/10556788.2011.597854.

Bjerregård, Mathias Blicher, Jan Kloppenborg Møller, and Henrik Madsen (2021). “An introduction to multivariate probabilistic forecast evaluation”. In: *Energy and AI* 4, p. 100058. ISSN: 2666-5468. DOI: https://doi.org/10.1016/j.egyai.2021.100058. URL: https://www.sciencedirect.com/science/article/pii/S2666546821000124.

Good, I. J. (1992). “Rational Decisions”. In: *Breakthroughs in Statistics: Foundations and Basic Theory*. Ed. by Samuel Kotz and Norman L. Johnson. New York, NY: Springer

New York, pp. 365–377. ISBN: 978-1-4612-0919-5. DOI: 10.1007/978-1-4612-0919-5_24. URL: https://doi.org/10.1007/978-1-4612-0919-5_24.

A Some probability functions

This chapter serves as a reference, specifying notation, properties, and moments related to the various distributions used in this master thesis.

Name	Support	Density	E$[Y]$	V$[Y]$
Poisson Pois(λ)	$0, 1, 2, \ldots$ $\lambda \in \mathbb{R}_+$	$\frac{\lambda^y}{y!}\exp(-\lambda)$	λ	λ
Gamma G(α, β)	$\mathbb{R}_+$ $\alpha \in \mathbb{R}_+, \beta \in \mathbb{R}_+$	$\frac{1}{\Gamma(\alpha)\beta}\left(\frac{y}{\beta}\right)^{\alpha-1}\exp(-y/\beta)$	$\alpha\beta$	$\alpha\beta^2$
Neg. Bin. NB(r, p)	$0, 1, 2, \ldots$ $r \in \mathbb{R}_+, p \in]0, 1]$	$\binom{r+y-1}{y}p^r(1-p)^y$	$\frac{r(1+p)}{p}$	$\frac{r(1-p)}{p^2}$
Normal N(μ, σ^2)	$\mathbb{R}$ $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	μ	σ^2

Table A.1: Density, support, mean value, and variance for a number of distributions used in this master thesis.

B C++ templates for the negative joint log-likelihood

This chapter presents the user templates for the hierarchical Poisson Normal model in (4.19) and the hierarachical Poisson Gamma model in (4.27).

The user template for the hierarchical Poisson Normal model specified in (4.19) is

```
#include <TMB.hpp>
template<class Type>
Type objective_function<Type>::operator() ()
{
  // R input data
  DATA_VECTOR(y);                                 // Count data
  DATA_VECTOR(x);                                 // Population size
  DATA_MATRIX(X);                                 // Design matrix
  PARAMETER_VECTOR(u);                            // Random effects
  // Parameters
  PARAMETER_VECTOR(beta);                         // Fixed effects parameters
  PARAMETER(log_sigma_u);                         // Model parameter
  vector<Type> lambda  = exp(X*beta-log(x)+u);    // Construct 'lambda'
  Type sigma_u = exp(log_sigma_u);                // And the model parameters
  Type mean_ran = Type(0);
  // Objective function
  Type f = 0;                                     // Declare the objective
  f -= sum(dnorm(u,mean_ran,sigma_u,true));       // Calculate the objective
  f -= sum(dpois(y,lambda,true));                 // Calculate the objective
  return f;
}
```

The user template for the hierarchical Poisson Gamma model specified in (4.27) is

```
#include <TMB.hpp>
template<class Type>
Type objective_function<Type>::operator() ()
{
  // Data
  DATA_VECTOR(y);                                 // Count data
  DATA_VECTOR(x);                                 // Population size
  DATA_MATRIX(X);                                 // Design matrix
  // Parameters
  PARAMETER_VECTOR(beta);                         // Fixed effects parameters
  PARAMETER(log_phi_u);                           // Model parameter
  vector<Type> lambda  = exp(X*beta-log(x));      // Construct 'lambda'
  Type phi_u = exp(log_phi_u);                    // And the model parameters
  Type r = 1/phi_u;                               // Construct the size
  vector<Type> p = 1/(lambda*phi_u+1);            // And the prob. parameter
  // Objective function
  Type f = -sum(dnbinom(y, r, p,true));           // Calculate the objective
```

```
  return f;
}
```

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Technical
University of
Denmark

Richard Petersens Plads, Building 324
2800 Kgs. Lyngby
Tlf. 4525 1700

www.compute.dtu.dk