

# 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

Copyright: Reproduction of this publication in whole or in part must include the customary bibliographic citation, including author attribution, report title, etc.

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](http://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 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

Preface . . . . .	ii
Abstract . . . . .	iii
Acknowledgements . . . . .	iv
Notation . . . . .	v
<b>1 Introduction</b>	<b>1</b>
<b>2 Setting the scene</b>	<b>3</b>
<b>3 Surveillance data in Denmark</b>	<b>5</b>
3.1 Data collection and data quality . . . . .	5
3.2 Introducing the case studies . . . . .	6
<b>4 Methods</b>	<b>17</b>
4.1 State-of-the-art outbreak detection algorithm . . . . .	17
4.2 Novel outbreak detection algorithm . . . . .	19
4.3 General mixed effects models . . . . .	20
4.4 Hierarchical generalized linear models . . . . .	22
4.5 Parameter estimation . . . . .	26
4.6 Scoring rule . . . . .	28
<b>5 Case studies</b>	<b>31</b>
5.1 Analysis of STEC outbreak detection . . . . .	31
5.2 Outbreak detection in other case studies . . . . .	40
5.3 Challenges in statistical outbreak detection . . . . .	46
5.4 Performance comparison of novel methods . . . . .	48
<b>6 Simulation study</b>	<b>51</b>
6.1 The simulated baseline data . . . . .	51
6.2 The simulated outbreaks . . . . .	53
6.3 Evaluation measures . . . . .	54
6.4 Results of the simulations . . . . .	55
<b>7 Discussion</b>	<b>59</b>
7.1 Harnessing the potential of MiBa . . . . .	59
7.2 Potential extensions of the novel algorithm . . . . .	59
<b>8 Conclusion</b>	<b>61</b>
<b>Bibliography</b>	<b>63</b>
<b>A Some probability functions</b>	<b>67</b>
<b>B C++ templates for the negative joint log-likelihood</b>	<b>68</b>
<b>C State-of-the-art detection algorithm</b>	<b>69</b>
C.1 Controls . . . . .	69

<b>D Figures and Tables related to the case studies</b>	<b>72</b>
D.1 <i>Listeriosis</i> . . . . .	72
D.2 <i>Shigellosis</i> . . . . .	72
D.3 Shiga toxin (verotoxin)-producing <i>Escherichia coli</i> . . . . .	73
D.4 <i>Salmonellosis</i> . . . . .	75

# 1 Introduction

The fight against infectious diseases not only requires proper treatment of patients and implementation of preventive measures but also demands early detection of emerging disease outbreaks. Timely identification and intervention can mean the difference between containing an outbreak or facing a devastating epidemic.

In Denmark, the surveillance of infectious diseases is mainly carried out by Statens Serum Institut (SSI), which plays a pivotal role in national and international disease preparedness. The high quality of the Danish surveillance registers, facilitated by a robust regulatory community framework, presents a significant opportunity for data analysis and enables timely detection of aberrant case counts. This empowers the prevention of additional disease cases through early interventions.

This master's thesis proposes a novel method for automated and early detection of disease outbreaks. The method utilizes hierarchical models in an innovative way, offering an alternative to state-of-the-art outbreak detection algorithms. The models are formulated and used to estimate the random effects in a subset of diseases in the mandatory notification system. Subsequently, novel automated procedures for detection of aberrant counts are implemented and compared against the methods currently employed at SSI and state-of-the-art method identified in the literature.

Primarily, the focus is on investigating the effectiveness of these methods in detecting disease outbreak in case studies, particularly in diseases where timely detection allows for feasible interventions. Emphasis is placed on diseases caused by foodborne illnesses, as they are often associated with specific food sources and provide an opportunity for intervention and preventive measures. However, it is important to note that the statistical algorithms discussed in this thesis are not limited to these specific diseases. They can be applied to detect abnormal case counts in any type of disease, enabling early detection and timely response.

Currently, several approaches are employed at SSI to detect outbreaks of foodborne illnesses. These include:

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 multiple cases can strongly indicate a common source of infection.

These approaches contribute to the early detection and investigation of foodborne disease outbreaks, enabling timely intervention and prevention measures. However, at the time of writing this thesis, there are no automated procedures in place at SSI, and the identification of outbreaks relies on the work of individual epidemiologists. Consequently,

there is a significant demand for an automated procedure for the early identification of disease outbreaks. Such a method could effectively guide the work of epidemiologists at SSI and enhance the overall surveillance of infectious diseases at national and international levels.

In addition to the case studies, a comprehensive simulation study is conducted to evaluate the performance of the novel outbreak detection algorithm compared to the state-of-the-art algorithms. The objective is to assess the ability of the automated procedures to timely detect disease outbreaks. By leveraging the potential of hierarchical models, this research seeks to improve the timeliness and accuracy of disease outbreak detection.

The novel outbreak detection algorithm proposed in this master's thesis is open-source and can be accessed at <https://github.com/telkamp7/AEDDO>

## 2 Setting the scene

Outbreak investigations have a long history, dating back to John Snow's iconic removal of the handle of London's Broad Street pump during the cholera epidemic in 1865 (Tulchin-sky 2018). Indeed, while John Snow's work was groundbreaking for his time, modern disease outbreak investigations require more advanced and sophisticated techniques. Today, epidemiologists and public health professionals utilize a range of tools and methodologies to effectively tackle disease outbreaks. Here, laboratory-based approaches play a crucial role in outbreak investigations and may involve technique such as molecular epidemiology (Honardoost, Rajabpour, and Vakil 2018; Struelens and Brisse 2013) and, more recently, Whole Genome Sequencing (WGS) (Koeser et al. 2012; Baldry 2010).

However, in recent years, there has been a growing interest in statistical methods for automated and early detection of disease outbreaks. These methodologies encompass various 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 studies by Buckeridge (2007) and Unkel et al. (2012).

To establish a golden standard, this master's thesis will focus on the Farrington method initially introduced by Farrington et al. (1996) and the subsequent improvement proposed by Noufaily et al. (2013). These methods offer advanced statistical tools for detecting and monitoring disease outbreaks and are currently *the* methods of choice at European public health institutes (Hulth et al. 2010) and can be accessed through the R package called **surveillance** developed by Salmon, Schumacher, and Höhle (2016).

It is a well known fact, that one limitation of these detection algorithms is an occasional lack of specificity, leading to false alarms that can overwhelm the epidemiologist with verification tasks (Bédubourg and Strat 2017). 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 already existing methods.

While generalized linear mixed effects models and hierarchical models have earned a reputation within ecology (Bolker et al. 2009; Zuur et al. 2009), urban energy modeling (Palmer Real et al. 2022; Real et al. 2021), and other fields, their application in the automatic detection of disease outbreaks is relatively unproven. However, there is a promising paper by Heisterkamp, Dekkers, and Heijne (2006) that applied a hierarchical time series model to detect infectious disease outbreaks in empirical data from *Rubella* and *Salmonella*.



# 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*, and *Salmonellosis*. These diseases will be referred to using either their full name or their related case definition:

- LIST: *Listeriosis*
- SHIL: *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/ita/2000/277>). This order specifies that several diseases<sup>1</sup> 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 (Voldstedlund et al. 2014). 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<sup>2</sup>. The national surveillance system focuses on diseases of a severe nature, those that are highly contagious, and the majority of vaccine-preventable diseases.

---

<sup>1</sup>For a full list of diseases see [https://www.ssi.dk/sygdomme-beredskab-og-forskning/anmeldelse-af-sygdomme/lovpligtige-meldesystemer/individ\\_anmeldelses\\_sygdomme](https://www.ssi.dk/sygdomme-beredskab-og-forskning/anmeldelse-af-sygdomme/lovpligtige-meldesystemer/individ_anmeldelses_sygdomme)

<sup>2</sup>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*, 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 investigated by SSI in the past decade, which adds to their relevance for the study.

The count observations are observed in the period from January 2008 to June 2023. An epidemic curve graph for an excerpt of the data for each of the diseases considered in this master's thesis is shown in Figure 3.1.

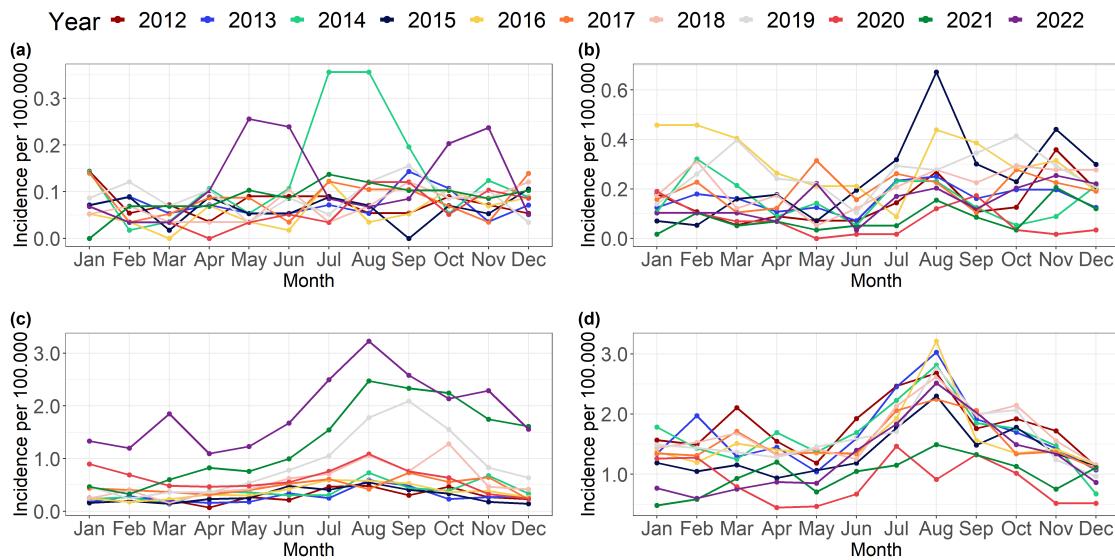


Figure 3.1: Epidemic curve showing the incidence per 100,000 in Denmark, 2012-2022, for the subset of diseases considered in this master thesis. (a) LIST, (b) SHIL, (c) STEC, and (d) SALM.

Evidently, the subset of diseases in Figure 3.1 exhibits varying incidences and different levels of seasonal patterns on an annual basis. It is interesting to note that the incidences peak in August, which can be attributed to several factors, including:

- **Increased travel activity:** Especially when individuals travel to countries with unsafe drinking water, poor sanitation, and insufficient hygiene practices.
- **Large social gatherings:** Events such as weddings, large festivals, or other gatherings where a significant number of people consume potentially contaminated food or drinking water.

Moreover, one could hypothesize that the warmer climate during the summer could potentially directly influence the proliferation of bacteria, which in turn may have an impact on the transmission and spread of these diseases.

In general, there is a noticeable decrease in the number of observed cases starting from March 2020 and continuing until January 2021. This decline can be attributed to the strict lockdown measures implemented in Denmark in response to the Covid-19 pandemic. These measures, which involved restrictions on movement and social interactions, likely

played a significant role in reducing the transmission of infectious diseases, including the ones being investigated.

From Figure 3.1a, it is evident that the incidence of LIST remains relatively low but stable over time. However, there are six specific months that stand out due to higher incidence rates. These months include July and August 2014, May and June 2022, and October and November 2022. These periods show a notable increase in the number of reported cases compared to the rest of the time series.

In Figure 3.1b, the incidence of the SHIL exhibits a consistent pattern over the years, with sporadic peaks occurring in the time series. Notable examples include a peak in August 2015, where the incidence reaches its highest level in the entire series. Additionally, January of 2016 stands out compared to other years. From the series alone, it is not evident whether there is a significant seasonality in the data. Another interesting observation is the substantial increase in incidence during August and September 2020 compared to the period leading up to these months.

Furthermore, in Figure 3.1c, a significant increase in the amplitude of the seasonal variation in STEC 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).

Overall, the highest incidences among the diseases considered in this master's thesis are observed for SALM in Figure 3.1d. SALM exhibits a notable pattern of high incidence throughout the observed period. However, there is a significant drop in incidences observed in 2020, which is consistent with the impact of the COVID-19 pandemic on disease surveillance and reporting. Additionally, 2021 shows generally lower incidences compared to previous years.

Some summary statistics for each of the diseases considered in this master's thesis are 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	4.909	5	3.145	*	
SHIL	0	38	10.941	11	6.253	*	
STEC	2	190	33.177	21	33.108	---	
SALM	17	588	107.677	86	78.708	---	

In Table 3.1, it is readily seen that the diseases exhibit different statistical properties, including variations in mean incidence, standard deviation, and range. These variations

indicate that each disease has its own unique epidemiological characteristics.

From an epidemiological perspective, 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

LIST is a foodborne illness that is caused by consuming food contaminated with *Listeria monocytogenes*. 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.

This disease primarily affects pregnant women, unborn or newborn babies, the elderly, and individuals with weakened immune systems. The disease is associated with high mortality (Goulet et al. 2012) and manifests in three ways: sepsis, meningitis, and mother-to-child transmission. Pregnancy-associated LIST can have severe consequences for the fetus or newborn, including miscarriage, stillbirth, neonatal sepsis, and meningitis (Awofisayo et al. 2015). LIST is uncommon among individuals in other demographic groups. T

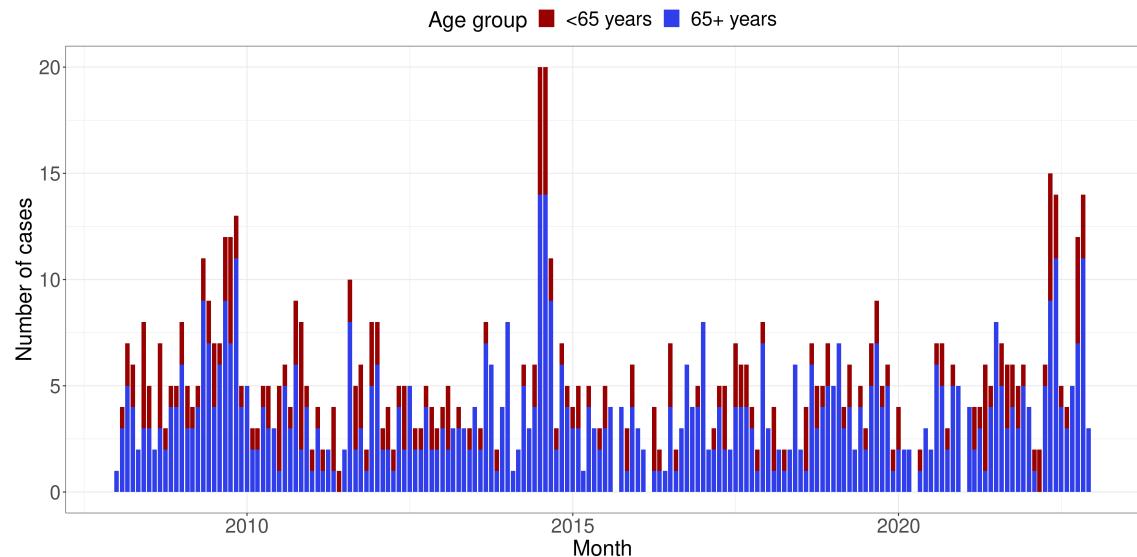


Figure 3.2: A stacked bar graph illustrating the number of LIST cases observed in the period from 2008 to 2022 for the age groups below and above 65 years.

The first notified outbreak in the 20th century occurred in 2009 (Smith et al. 2011). This outbreak affected 8 people, of which two died. The patient samples were isolated between the 6th and 11th of May and it was believed that the cause was infected beef meat from a meals-on-wheels delivery.

In 2014, Whole Genome Sequencing (WGS) was employed by SSI as the routine testing of *Listeria* isolates in Denmark (Wingstrand et al. 2015). 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.

A notable outbreak investigated by SSI occurred between September 2013 and October 2014 (Kvistholm Jensen et al. 2016). This LIST 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. Following the outbreak, a series of control activities and research project were initiated to improve the control and management of *Listeria* in the food production (Takeuchi-Storm et al. 2023).

In another LIST outbreak investigated by SSI, the source was traced back to cold-smoked and cured salmon products (Schjorring et al. 2017). 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 the 13th of May and the 6th of June, 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 (Helwigh and Müller 2018; Helwigh, Petersen, and Torpdahl 2020; Lassen et al. 2023) 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 green-grocer.
- 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 (Lassen et al. 2023).

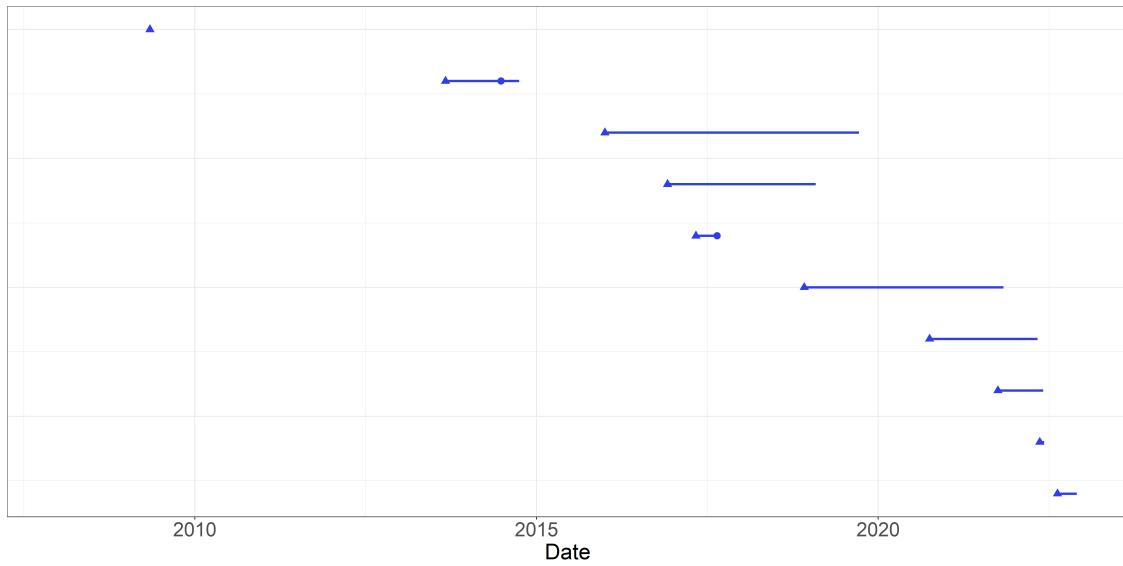


Figure 3.3: Timeline plot indicating the start (triangle) and duration (line) of documented outbreaks of LIST by SSI and possible collaborators. Time of detection (circle) is indicated when the information is available.

### 3.2.2 *Shiggellosis*

SHIL is a diarrheal illness that is caused by a group of bacteria called *Shigella*. There are four types of *Shigella* bacteria, namely: *Shigella dysenteriae*, *Shigella boydii*, *Shigella flexneri*, and *Shigella sonnei*. The latter is the most common species in Denmark. For surveillance purposes, it is generally preferred to consider data from the laboratory notification systems, as it includes information about the specific type of *Shigella* bacteria. However, for the purpose of this master's thesis, all four types of *Shigella* bacteria are analyzed collectively, regardless of the availability of type-specific information in the data.

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. SHIL 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.

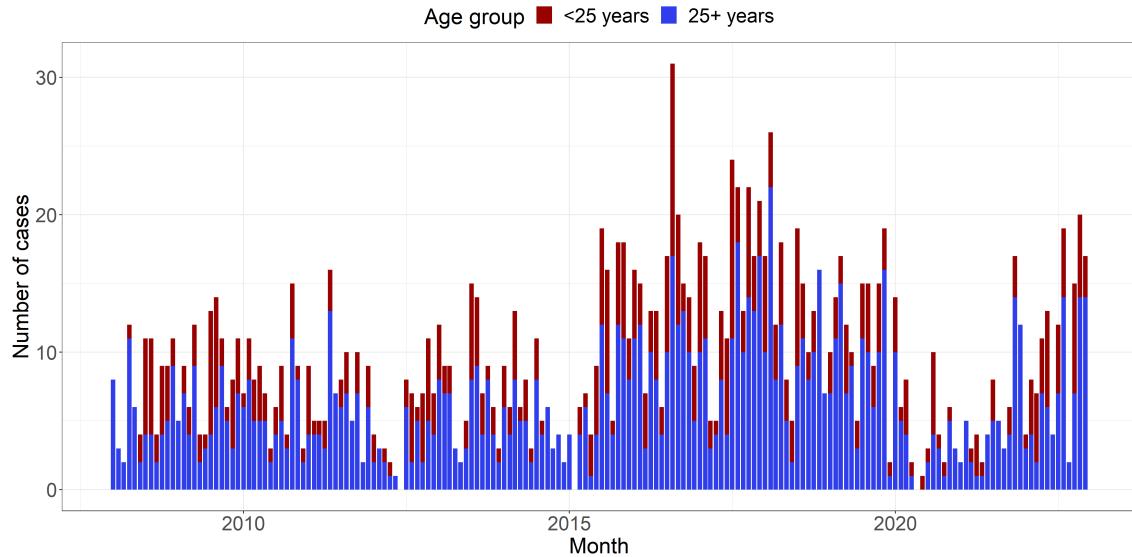


Figure 3.4: A stacked bar graph illustrating the number of SHIL cases observed in the period from 2008 to 2022 for the age groups below and above 25 years.

In Denmark, another significant cause of SHIL 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 (Lewis et al. 2009), a smaller outbreak in 2009 linked to sugar snap peas from Kenya (Muller et al. 2009), 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 the 22th of August to the 9th of September 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*.

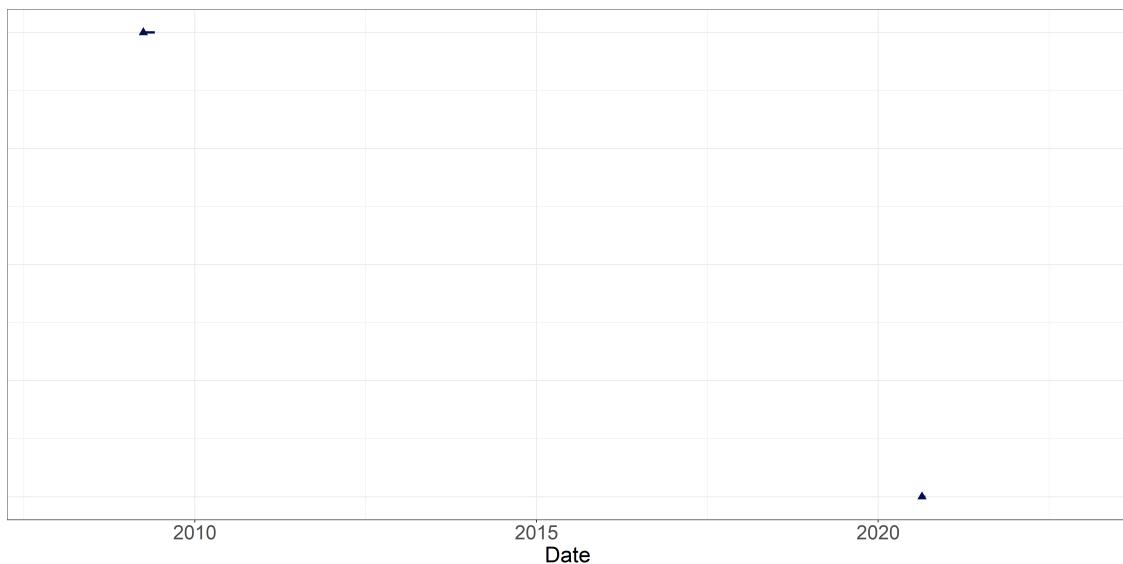


Figure 3.5: Timeline plot indicating the start (triangle) and duration (line) of documented outbreaks of SHIL by SSI and possible collaborators.

### 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.

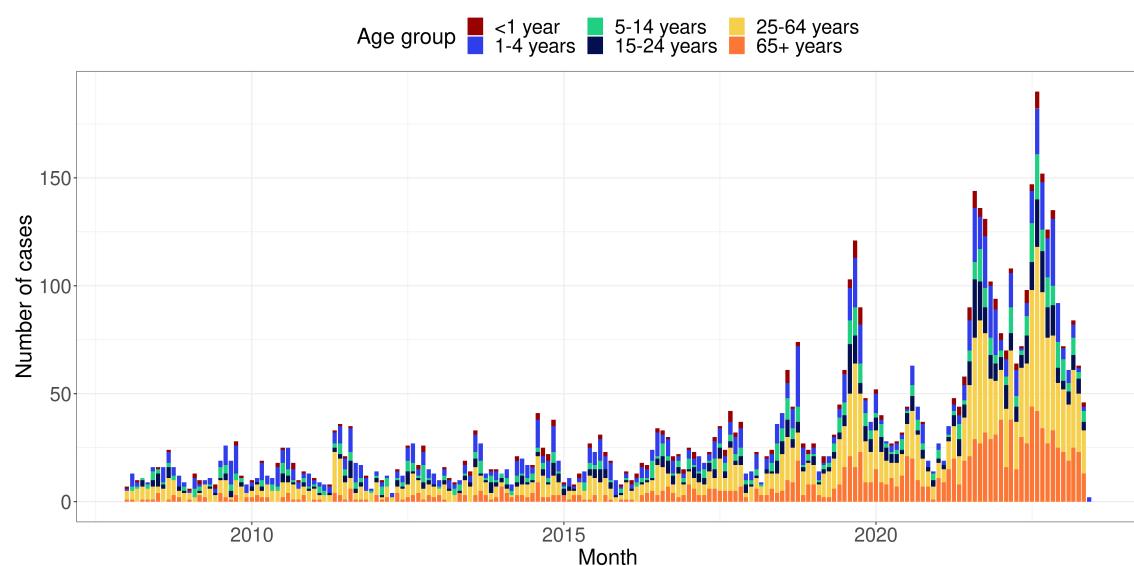


Figure 3.6: A stacked bar graph illustrating the number of STEC cases observed in the period from 2008 to 2022 for the six age groups.

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 documented 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.

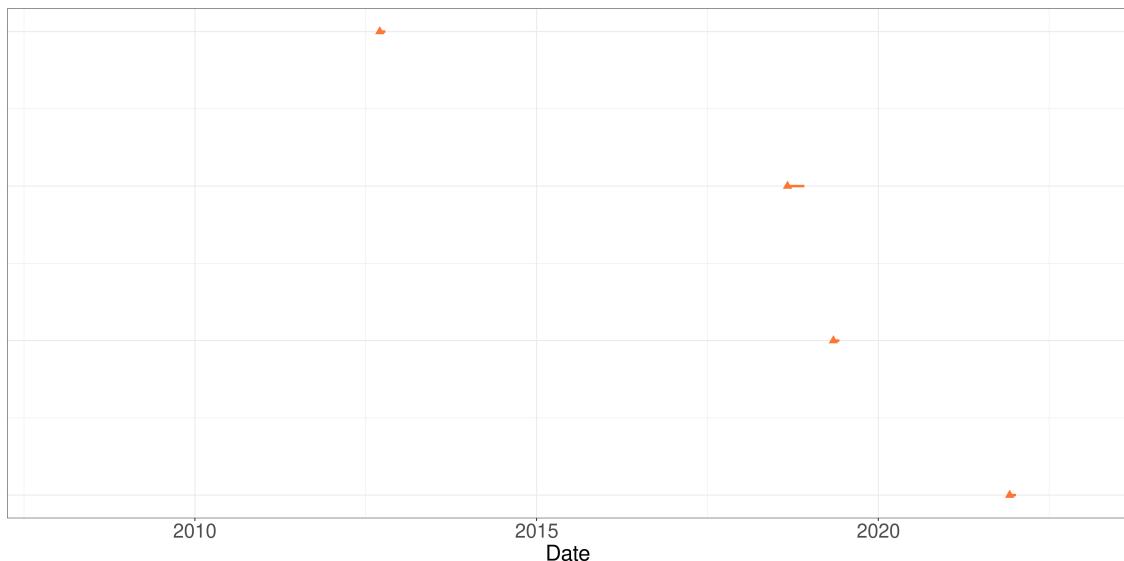


Figure 3.7: Timeline plot indicating the start (triangle) and duration (line) of documented outbreaks of STEC by SSI and possible collaborators.

### 3.2.4 *Salmonellosis*

SALM 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 con-

taminated food or water. *Salmonella* infections are often associated with the consumption of raw or under cooked meat, poultry, eggs or egg products, as well as unpasteurized milk.

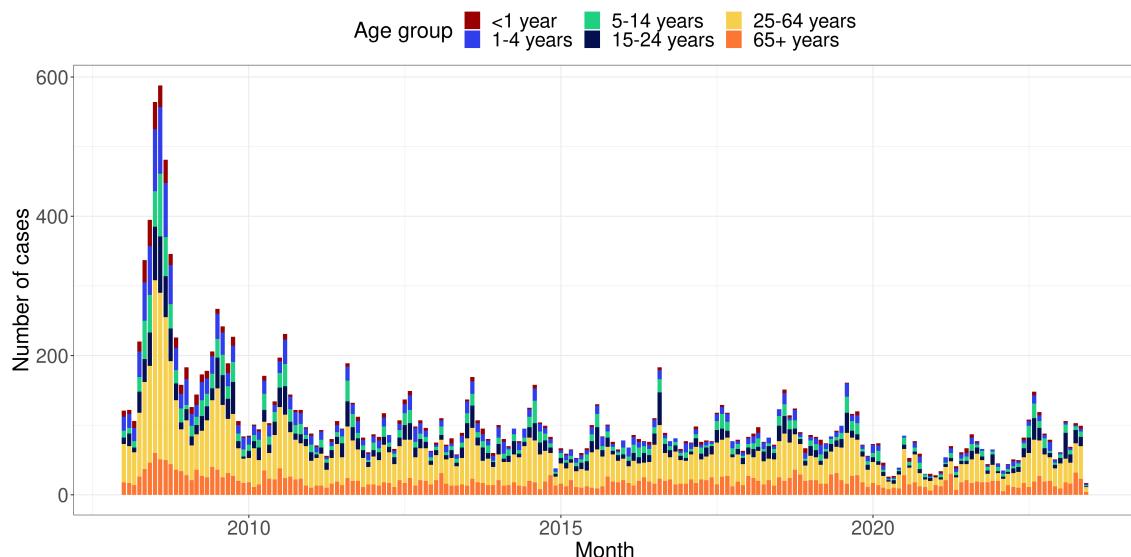


Figure 3.8: A stacked bar graph illustrating the number of SALM cases observed in the period from 2008 to 2022 for the six age groups.

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.

In 2015, three outbreaks of SALM with patients in two or more regions were investigated (Helwigh, Christensen, and Müller 2016):

- An outbreak caused by *S. Newport* affecting 6 people in the period from March to April 2015.
- A long-lasting outbreak caused by *S. Oranienburg* with 14 genetically linked cases from July 2015 to January 2016.
- An outbreak with 6 patients from November 2015 to January 2016.

Other resolved and well-documented outbreaks include:

- An outbreak with 49 cases from October 2018 to January 2019, where Mediterranean sausage was identified as a possible source of contamination.
- An outbreak with 45 cases from November 2020 to April 2021, linked to the consumption of the natural remedy HUSK Psyllium.
- An international outbreak from March to July 2021, with more than 300 cases in Europe, including 39 cases in Denmark. Imported melons were suspected as the source of infection.
- An outbreak caused by eggs from a Danish producer, resulting in 24 cases registered from September to November 2021.

- An international outbreak with a total of 392 cases across 12 countries in the EU-/EEA and the UK, including 4 cases in Denmark. Kinder chocolate products were identified as the source of infection.

There were also outbreaks where it was not possible to identify the source of contamination, including:

- An outbreak with 26 cases in the period from May to August 2019.
- An outbreak with 11 cases in the period from June to July 2020.
- An outbreak with 24 cases in the period from March to September 2022.
- An outbreak with 15 cases in the period from August to September 2022.

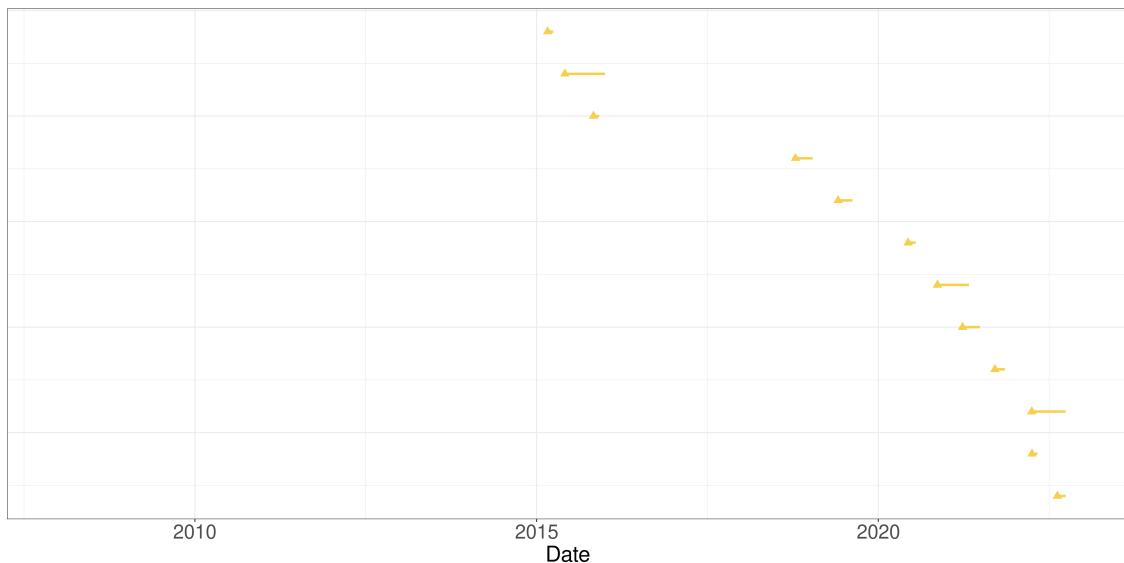


Figure 3.9: Timeline plot indicating the start (triangle) and duration (line) of documented outbreaks of SALM by SSI and possible collaborators.



## 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 modeling frameworks are utilized in this master's thesis to analyze the count observations denoted as  $y$ , but more importantly they play a crucial role in assessing the unobserved random variables or random effects represented by  $u$ , 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.

### 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. Both of the aforementioned methods have been implemented in the R package called **surveillance** by Salmon, Schumacher, and Höhle (2016), which can be accessed from the Comprehensive R Archive Network (CRAN) at <https://cran.r-project.org/web/packages/surveillance/index.html>. 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. In this model, the baseline count  $y_t$  corresponding to the baseline time point  $t$  is assumed to have an expected value  $\lambda_t$  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

$$\log(\lambda_t) = \alpha + \beta t \quad (4.1)$$

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  $h$ , only data from weeks  $x - w$  to  $x + w$  of years  $h - b$  to  $h - 1$  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

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

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** R package. The package allows the user to choose an arbitrary number of periods in each year. Consequently, the systematic component is adjusted as follows

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

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  $\lambda_{t_0}$  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

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

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

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

Here,  $V(\hat{y}_{t_0})$  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

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

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  $NB(\lambda_{t_0}, \nu)$ , where  $\lambda_{t_0}$  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  $\lambda_t$ , 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 re-weighted in a second fitting of the GLM. In the original method by Farrington et al. (1996), a re-weighting threshold of 1 was used. However, Noufaily et al. (2013) suggests using a value of 2.56 for the weight threshold to make the re-weighting procedure less drastic, as it also reduces the variance of the observations. For a more detailed description of the re-weighting scheme employed in the Farrington and Noufaily method, refer to Section 3.6 in Farrington et al. (1996) and Section 2.1 in Noufaily et al. (2013), respectively.

## 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 as a modeling framework to model the count observations  $y$  and assess the unobserved random effects  $u$ . 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 or Poisson Gamma model with a log link to the reference data. 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 time-varying model parameters. Also, it is assumed that the count is proportional to the population size  $n$ . Hence, in terms of the canonical link the model for the fixed effects is

$$\log(\lambda_{it}) = \mathbf{x}_{it}\boldsymbol{\beta} + \log(n_{it}), \quad i = 1, \dots, m, \quad t = 1, \dots, T \quad (4.7)$$

Here  $\mathbf{x}_{it}$  and  $\boldsymbol{\beta}$  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.

In the second step of the algorithm, as a new observation becomes available, the algorithm infers the one-step ahead random effect  $u_{it_1}$  for each group using the obtained model estimates  $\hat{\theta}_{t_0}$ . Here,  $t_0$  represents the current time point, and  $t_1$  represents the one-step ahead time point. The threshold  $U_{t_0}$  for detecting outbreaks is defined as a quantile of the distribution of the random effects in the second stage model. This threshold can be calculated based on either a Gaussian distribution using the plug-in estimate  $\hat{\sigma}$  or a Gamma distribution using the plug-in estimate  $\hat{\phi}$ . The choice of distribution depends on the specific modeling framework and assumptions used in the analysis.

In the final step, the inferred random effect  $\hat{u}_{it_1}$  is compared to the upper bound  $U_{t_0}$ , and an alarm is raised if  $\hat{u}_{it_1} > U_{t_0}$ . If an outbreak is detected, the related observation  $y_{i1}$  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 discarded.

### 4.3 General mixed effects models

In this section selected theory related to generalized mixed effects models is presented. The presentation of Section 4.3 and Section 4.4 is mostly inspired by Madsen and Thyregod (2011).

The general mixed effects model can be represented by its likelihood function

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

where  $\mathbf{y}$  is the observed random variable,  $\theta$  is the model parameters to be estimated and  $\mathbf{U}$  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  $\theta$  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. Thus, an outline of the Laplace approximation is included in this section.

#### 4.3.1 Hierarchical models

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

$$f_{Y|u}(\mathbf{y}; \mathbf{u}, \beta) \quad (4.9)$$

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

$$f_U(\mathbf{u}; \Psi) \quad (4.10)$$

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

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

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

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

where  $q$  is the number of unobserved random variables, and  $\beta$  and  $\Psi$  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  $\theta$  the joint log-likelihood  $\ell(\theta, u, y) = \log(L(\theta, u, y))$  is approximated using a second order Taylor approximation around the optimum  $\tilde{u} = \hat{u}_\theta$  of the log-likelihood function w.r.t. the unobserved random variables  $u$ , i.e.,

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

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

It is readily seen that the joint log-likelihood for the hierarchical model specified in (4.9) and (4.10) is

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

which implies that the Laplace approximation becomes

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

### 4.3.3 Formulation of the generalized mixed effects model

The generalized mixed effects model utilized in this master's thesis to model the count observations  $y$  and assess the random effects  $u$  is presented in Theorem 4.1, along with the joint likelihood function for the first and second stage models.

**Theorem 4.1** (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  $\lambda$*

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

*Also, it is assumed that the count is proportional to the population size  $n$ . Hence, in terms of the canonical link for the Poisson distribution the model for the fixed effects is*

$$\log(\lambda_{it}) = \mathbf{x}_{it}\beta + \log(n_{it}), \quad i = 1, \dots, m, \quad t = 1, \dots, T \quad (4.17)$$

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

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

where  $\sigma$  is a model parameter.

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

$$Y|u \sim \text{Pois}(\lambda \exp(u)) \quad (4.19a)$$

$$u \sim N(0, I\sigma^2) \quad (4.19b)$$

The joint likelihood for the count observations  $y$  and the random effects  $u$  becomes

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

## 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

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

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. For an introduction to the concept of *standard conjugate distributions* and the definition of a hierarchical generalized linear model, refer to Section 6.3 and Section 6.5 of Madsen and Thyregod (2011), respectively.

In general, when the conditional distribution of  $Y|u$  is a Poisson distribution, and the the conjugate distribution is assumed to be a Gamma distribution with mean value 1, it follows that the distribution of  $Y$  is a negative binomial distribution.

To further motivate this choice of distribution for the second stage model, an illustrative example is presented.

**Example 4.1** (Variation between observed cases of LIST). Usually, it is reasonable to assume a Poisson distribution for count data, where the expected value and variance are both equal to  $\lambda$ . However, it should be noted that this assumption may not always hold true when modeling count data.

Table 4.1: The distribution of cases with 1, 2, ⋯, 10+ cases of LIST.

Number of LIST cases	Number of times observed	Poisson expected	Negative binomial expected
0	4	6.26	14.44
1	5	15.52	21.41
2	28	25.66	24.75
3	25	31.83	24.58
4	26	31.58	22.01
5	34	26.11	18.27
6	19	18.51	14.31
7	14	11.48	10.72
8	10	6.33	7.73
9	3	3.14	5.41
10+	12	2.34	10.46

Take, for example, the monthly cases of LIST analyzed in this master's thesis. Table 4.1 presents the distribution of these cases from 2008 to 2022, along with the expected numbers from both the Poisson and Negative binomial distributions. The distribution of cases over time is visualized in Figure 4.1.

As observed, the actual distribution has significantly heavier tails compared to the Poisson distribution. Additionally, the mean of the Poisson distribution for the monthly LIST cases is  $\hat{\lambda} = \bar{y} = 4.96$ , while the empirical variance is  $s^2 = 10$ , which is considerably larger than the mean. These findings indicate overdispersion, suggesting that a Negative binomial model would be more appropriate for modeling the data.

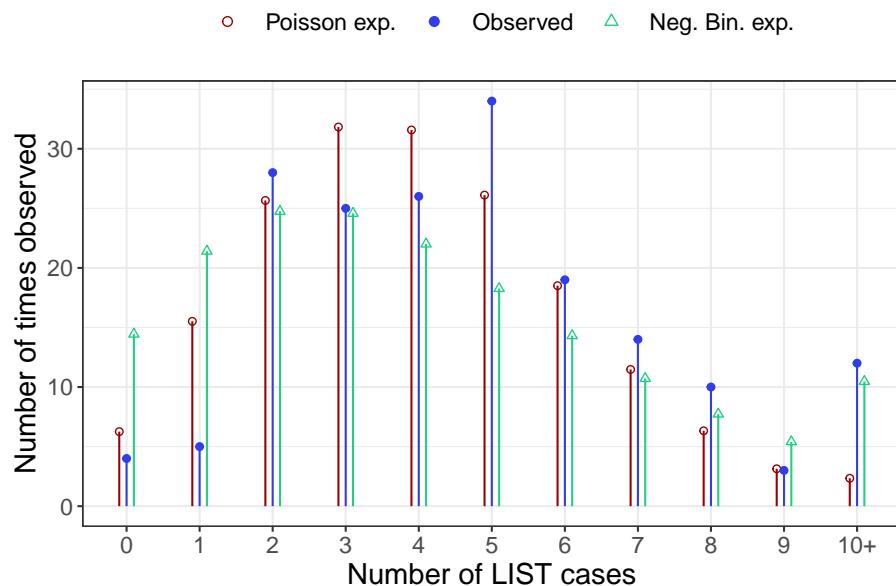


Figure 4.1: Distribution of monthly cases of LIST together with the expected numbers from both the Poisson and Negative binomial distributions

#### 4.4.1 Formulation of the hierarchical model

The hierarchical model used in this master's thesis to model the count observations  $y$  and assess the random effects  $u$  is presented in Theorem 4.2, along with the derivation of the marginal distribution of  $Y$  and the joint likelihood function.

**Theorem 4.2** (Compound 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  $\lambda$*

$$f_{Y|u}(y; u, \beta) = \frac{(\lambda u)^y}{y!} \exp(-\lambda u) \quad (4.22)$$

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

$$f_u(u; \phi) = \frac{1}{\phi \Gamma(1/\phi)} \left(\frac{u}{\phi}\right)^{1/\phi-1} \exp(-u/\phi) \quad (4.23)$$

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

$$Y|u \sim \text{Pois}(\lambda u) \quad (4.24a)$$

$$u \sim G(1/\phi, \phi) \quad (4.24b)$$

*Given (4.22) and (4.23), the probability function for the marginal distribution of  $Y$  is determined from*

$$\begin{aligned} g_Y(y; \beta, \phi) &= \int_{u=0}^{\infty} f_{Y|u}(y; u, \beta) f_u(u; \phi) du \\ &= \int_{u=0}^{\infty} \frac{(\lambda u)^y}{y!} \exp(-\lambda u) \frac{1}{\phi \Gamma(1/\phi)} \left(\frac{u}{\phi}\right)^{1/\phi-1} \exp(-u/\phi) du \\ &= \frac{\lambda^y}{y! \Gamma(1/\phi) \phi^{1/\phi}} \int_{u=0}^{\infty} u^{y+1/\phi-1} \exp(-u(\lambda\phi + 1)/\phi) du \end{aligned} \quad (4.25)$$

*In (4.25) it is noted that the integrand is the kernel in the probability density function for a Gamma distribution,  $G(y + 1/\phi, \phi/(\lambda\phi + 1))$ . As the integral of the density shall equal one, we find by adjusting the norming constant that*

$$\int_{u=0}^{\infty} u^{y+1/\phi-1} \exp\left(-u/\left(\phi/(\lambda\phi + 1)\right)\right) du = \frac{\phi^{y+1/\phi} \Gamma(y + 1/\phi)}{(\lambda\phi + 1)^{y+1/\phi}} \quad (4.26)$$

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

$$\begin{aligned}
P[Y = y] &= g_Y(y; \beta, \phi) \\
&= \frac{\lambda^y}{y! \Gamma(1/\phi) \phi^{1/\phi}} \frac{\phi^{y+1/\phi} \Gamma(y+1/\phi)}{(\lambda\phi + 1)^{y+1/\phi}} \\
&= \frac{\Gamma(y+1/\phi)}{\Gamma(1/\phi)y!} \frac{1}{(\lambda\phi + 1)^{1/\phi}} \left( \frac{\lambda\phi}{\lambda\phi + 1} \right)^y \\
&= \binom{y+1/\phi-1}{y} \frac{1}{(\lambda\phi + 1)^{1/\phi}} \left( \frac{\lambda\phi}{\lambda\phi + 1} \right)^y, \quad \text{for } y = 0, 1, 2, \dots
\end{aligned} \tag{4.27}$$

where we have used the convention

$$\binom{z}{y} = \frac{\Gamma(z+1)}{\Gamma(z+1-y)y!} \tag{4.28}$$

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

$$\mathbb{E}[Y] = \lambda \quad \mathbb{V}[Y] = \lambda(\lambda\phi + 1) \tag{4.29}$$

The joint likelihood function for estimating  $(\beta, \phi)$  is

$$L(\beta, \phi; y_{it}) = \prod_{t=1}^T \prod_{i=1}^m \binom{y_{it} + 1/\phi - 1}{y_{it}} \frac{1}{(\lambda_{it}\phi + 1)^{1/\phi}} \left( \frac{\lambda_{it}\phi}{\lambda_{it}\phi + 1} \right)^{y_{it}} \tag{4.30}$$

### Inference on individual groups

Consider the compound Poisson Gamma model in (4.24), 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. In order to simplify the notation, the subscript indicating the group and time are omitted.

$$\begin{aligned}
g_u(u|Y = y) &= \frac{f_{y,u}(y, u)}{g_Y(y; \lambda, \phi)} \\
&= \frac{f_{y|u}(y; u)g_u(u)}{g_Y(y; \lambda, \phi)} \\
&= \frac{1}{g_Y(y; \lambda, \phi)} \left( \frac{(\lambda u)^y}{y!} \exp(-\lambda u) \frac{1}{\phi \Gamma(1/\phi)} \left( \frac{u}{\phi} \right)^{1/\phi-1} \exp(-u/\phi) \right) \\
&\propto u^{y+1/\phi-1} \exp(-u(\lambda\phi + 1)/\phi)
\end{aligned} \tag{4.31}$$

We identify the *kernel* of the probability density function

$$u^{y+1/\phi-1} \exp(-u(\lambda\phi + 1)/\phi) \tag{4.32}$$

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

$$u|Y = y \sim G(y + 1/\phi, \phi/(\lambda\phi + 1)) \tag{4.33}$$

The mean of the conditional distribution is given by:

$$\mathbb{E}[u|Y = y] = \frac{y\phi + 1}{\lambda\phi + 1} \quad (4.34)$$

And the variance of the conditional distribution is:

$$\text{V}[u|Y = y] = \frac{(\phi^2 + \phi)}{(\lambda\phi + 1)^2} \quad (4.35)$$

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

### Why is the Gamma distribution chosen to represent the variation between months?

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 models used in the novel outbreak detection algorithm are presented. These models are implemented in R using the open-source R package **TMB** (Template Model Builder) developed by Kristensen et al. (2016). This package facilitates efficient maximum likelihood estimation and uncertainty calculations for the parameter set  $\theta = (\beta, \Psi)$  and random effects  $u$ . 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).

**TMB** maximizes a user-provided objective function in the form of a C++ template, to estimate the maximum likelihood for the parameter set  $\theta = (\beta, \Psi)$ . In the following code chunk, the C++ template for the hierarchical Poisson normal model specified in (4.19), is shown:

```
#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;
}

```

Refer to B to access the C++ template for the hierarchical Poisson Gamma model. The objective function maximizes the marginal log-likelihood function, which integrates out the random effects  $u$

$$\ell_M(\boldsymbol{\theta}; \mathbf{y}) = \int_{\mathbb{R}^q} \ell(\boldsymbol{\theta}; \mathbf{u}, \mathbf{y}) d\mathbf{u} \quad (4.36)$$

where  $\ell(\boldsymbol{\theta}, \mathbf{u}, \mathbf{y})$  is the joint log-likelihood function of the data given the parameters and random effects. The maximizer  $\hat{\mathbf{u}}_\theta$  of the joint log-likelihood  $\ell(\boldsymbol{\theta}; \mathbf{u}, \mathbf{y})$  with respect to the random effects  $\mathbf{u}$  is defined as:

$$\hat{\mathbf{u}}_\theta = \arg \max_{\mathbf{u}} \ell(\boldsymbol{\theta}; \mathbf{u}, \mathbf{y}) \quad (4.37)$$

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

$$H(\hat{\mathbf{u}}_\theta) = -\ell''_{uu}(\boldsymbol{\theta}, \mathbf{u}, \mathbf{y})|_{\mathbf{u}=\hat{\mathbf{u}}_\theta} \quad (4.38)$$

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

$$\ell_{M,LA}(\boldsymbol{\theta}, \mathbf{y}) = \ell(\boldsymbol{\theta}, \mathbf{u}, \mathbf{y}) - \frac{1}{2} \log \left| \frac{H(\hat{\mathbf{u}}_\theta)}{2\pi} \right| \quad (4.39)$$

Our estimate of  $\boldsymbol{\theta}$  minimizes the negative log of the Laplace approximation, i.e.,

$$-\ell_{M,LA}(\boldsymbol{\theta}, \mathbf{y}) = -\ell(\boldsymbol{\theta}, \mathbf{u}, \mathbf{y}) + \frac{1}{2} \log \left| \frac{H(\hat{\mathbf{u}}_\theta)}{2\pi} \right| \quad (4.40)$$

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  $\hat{\theta}$ . Uncertainty of the estimate  $\hat{\theta}$ , or any differentiable function of the estimate  $\phi(\hat{\theta})$ , is obtained by the  $\delta$ -method:

$$V(\phi(\hat{\theta})) = -\phi'_{\theta}(\hat{\theta}) \left( \Delta^2 \ell_{M,LA}(\hat{\theta}, y) \right)^{-1} \phi'_{\theta}(\hat{\theta})^T \quad (4.41)$$

Additionally, **TMB** utilizes Automatic Differentiation (AD) techniques (Griewank and Walther 2008) to evaluate first, second, and potentially third-order derivatives. This approach enhances the computational efficiency and accuracy of the parameter estimation process in the implemented models. Therefore, even though the random effects are analytically integrated out in (4.30), and the Laplace approximation is not needed, implementing the joint log-likelihood function in **TMB** can still result in more efficient computations.

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).

## 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  $y_t = y_1, \dots, y_N$ , 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:

$$\bar{S}(G, y) = \frac{1}{N} \sum_{t=1}^N S(G_t, y_t) \quad (4.42)$$

One commonly used scoring rule is the *logarithmic score* derived from likelihood theory, which is defined as  $S(G, y) = -\log(f(y))$  (Good 1992; Gneiting and Raftery 2007). 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 4.2, which is adapted from Bjerregård, Møller, and Madsen (2021).

**Example 4.2** (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 ( $\alpha$ ) and rate ( $\beta$ ), and its probability density function (PDF) is given by:

$$f(y) = \frac{1}{\Gamma(\alpha)\beta} \left( \frac{y}{\beta} \right)^{\alpha-1} \exp\left(-\frac{y}{\beta}\right) \quad (4.43)$$

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  $y_1, y_2, \dots, y_{10}$ , which are shown in Table 4.2. 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  $y$  are illustrated in Figure 4.2.

Table 4.2: 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

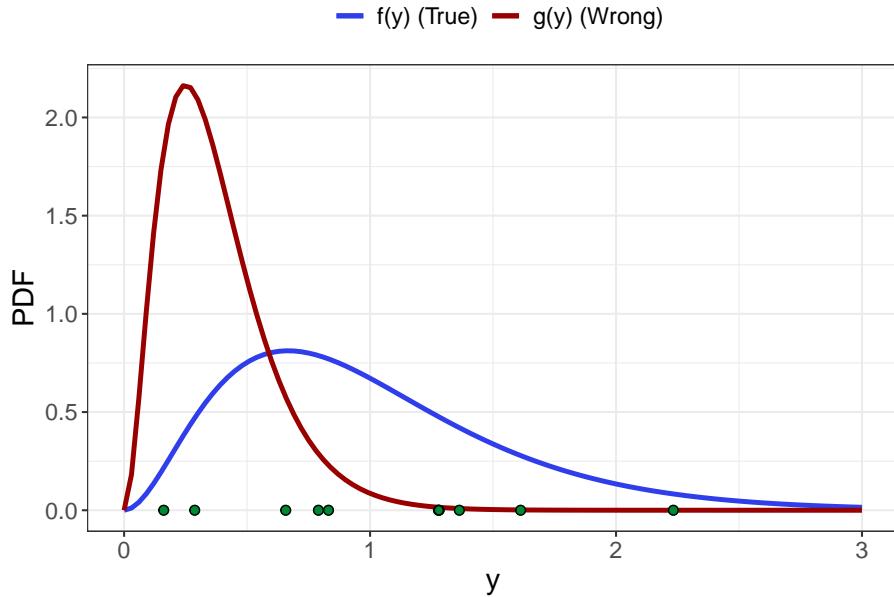


Figure 4.2: Observations (green dots) were simulated from a G(3,3)-distribution. The true model  $f$  (blue) is shown, along with the PDF for the competing model  $g$ , which follows a  $G(3,8)$ -distribution.

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

$$\begin{aligned}
 -\log(f(y_1)) &= -\log(f(1.278)) \\
 &= -\log\left[\frac{1}{\Gamma(3)4}\left(\frac{1.278}{4}\right)^{3-1} \exp(-1.278/4)\right] \\
 &= -1.156
 \end{aligned} \tag{4.44}$$

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 4.3. 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,  $\bar{S}(G, y)$ , can be calculated using Equation (4.42).

$$\begin{aligned}
 \bar{S}(f, y) &= 1.1 \\
 \bar{S}(g, y) &= 3.22
 \end{aligned} \tag{4.45}$$

Table 4.3: 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
$S(f, y_i)$	1.16	3.86	0.00	0.23	1.16	0.18	1.37	2.03	0.17	0.83
$S(g, y_i)$	4.19	10.71	0.55	1.47	4.21	-0.75	4.74	6.40	1.25	-0.60

# 5 Case studies

This chapter presents the findings obtained from applying both the state-of-the-art and novel outbreak detection algorithms to the subset of diseases examined in this master's thesis. To demonstrate the practical application of these algorithms, a comprehensive case study focused on STEC will be presented. The results of the remaining case studies will be presented in a more concise manner to maintain reader engagement. For a complete collection of figures and tables related to the case studies, please refer to Appendix D.

It is widely recognized that effective monitoring of a surveillance time series necessitates accurate modeling of the time series prior to assessing aberrations. Therefore, the performance of these algorithms in identifying outbreaks within the selected diseases will be thoroughly discussed and analyzed. Both the state-of-the-art and the novel algorithms take into account trends and seasonality, which will be addressed in the analysis. Additionally, a comparison of the two modeling frameworks used in the novel algorithm will be presented, providing valuable insights into the strengths and limitations of generalized mixed effects models and hierarchical generalized linear models.

Moreover, this chapter outlines some of the challenges encountered during the process of conducting this master's thesis, specifically related to statistical outbreak detection. These challenges are showcased using real case studies of diseases in the mandatory notification system, including LIST, SHIL and SALM. Among other things, the role of overdispersion in statistical outbreak detection, the impact of context and observational bias, and how diseases with many alarms are handled will be discussed. At last, a comprehensive comparative analysis will be conducted.

For a detailed presentation of the data used to generate these results, please refer to Chapter 3.

## 5.1 Analysis of STEC outbreak detection

The initial analysis focuses on STEC. The data set comprises monthly counts of Danish STEC cases, denoted as  $y_{it}$ , where  $i = 1, \dots, 6$  represents the six age groups and  $t = 1, \dots, T$  represents the time period of  $T = 180$  months starting in 2008. The first step involves applying the state-of-the-art outbreak detection algorithms to identify potential outbreaks in the time series.

Following that, the novel algorithm is utilized, and different models for the fixed effects are proposed. Both the hierarchical Poisson Normal model and the hierarchical Poisson Gamma model is considered for the modeling framework. To determine the most appropriate models for further analysis, the performance of the different fixed effects models is compared using the average logarithmic score, denoted as  $\bar{S}(G, y)$ , within each of the modeling frameworks. The average logarithmic score provides a measure of how well the models fit the observed data and allows for objective comparison between different models.

### 5.1.1 Applying the state-of-the-art outbreak detection algorithm to STEC

To investigate outbreak detection, the Farrington method and the subsequently improved Noufaily method are initially explored. These methods can be implemented using the `farringtonFlexible` function, which is available in the R package called `surveillance`.

The methods are controlled using specific control arguments. The specific arguments used in this master's thesis can be found in Appendix C.1.

For the analysis of STEC, the reference values are based on data collected from January 2008 to February 2011. Subsequently, surveillance is performed using the data spanning from March 2011 to December 2022. The resulting time series and alarms are visualized in Figure 5.1.

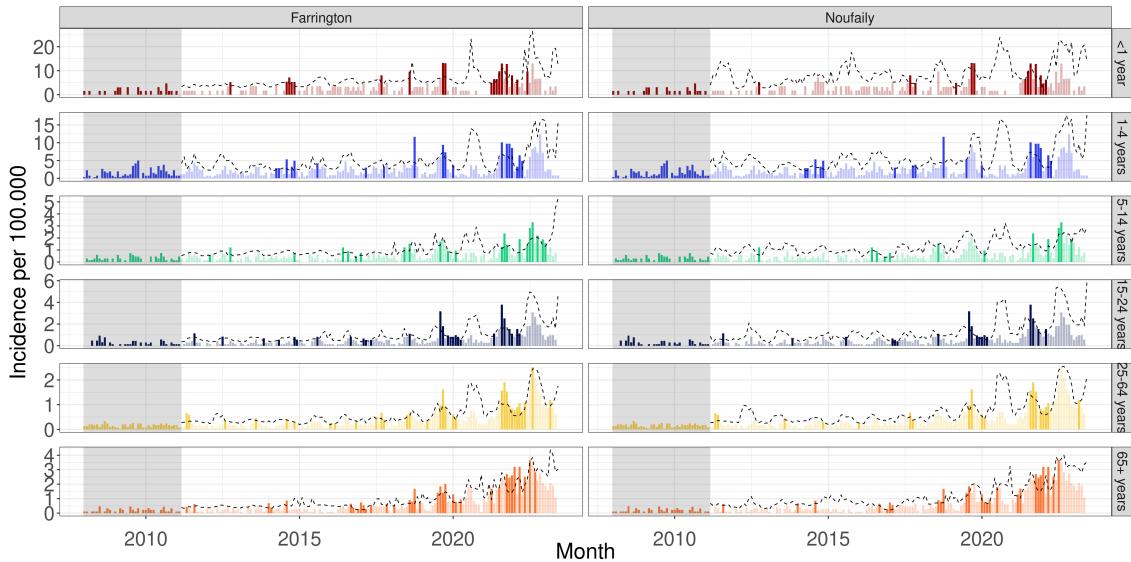


Figure 5.1: Monthly STEC incidence per 100,000 in Denmark, 2008-2022. Monitored by (left) Farrington and (right) Noufaily method. Reference data for the estimation of model parameters from January 2008 to March 2011 (grey area). Threshold (dashed line) is computed for observations time points outside reference data. Alarm triggered (dark color) if observations exceeds threshold.

Multiple alarms are displayed in Figure 5.1 for both the Farrington and Noufaily method. The Farrington method generates a total of 151 alarms, while the Noufaily method produces a significantly lower count of 104 alarms. It is not surprising that the Farrington method generates more alarms, as it is known for its high sensitivity. This means that the method is able to detect very small aberrations in the time series, but it also leads to a significant number of “false positive” alarms.

The substantial number of alarms generated by both methods raises valid concerns regarding their practical utility. The sheer volume of alarms can overwhelm epidemiologists and pose challenges in effectively prioritizing and investigating potential outbreaks. It is essential to consider the overall disease picture and avoid overly burdening epidemiologists with verification tasks. Striking a balance between sensitivity and specificity is crucial to ensure that the generated alarms are meaningful and actionable, leading to efficient outbreak detection and response.

As a general rule, alarms that are triggered simultaneously in multiple age groups or during concurrent time points are of particular importance. Such patterns indicate potential outbreaks that warrant further attention and investigation by epidemiologist. By analyzing and interpreting the alarms in context, epidemiologist can gain a more comprehensive understanding of the outbreak situation and make informed decision about resource allocation and intervention strategies.

When comparing the Farrington and Noufaily methods, it is interesting to note that the thresholds are inflated towards the end of the series. Specifically, the Noufaily method exhibits a higher threshold compared to the Farrington method. This behavior is expected since the Noufaily method employs a less drastic approach in weighting prior outbreaks, allowing more information to be transmitted from previous outbreaks. As a result, the Noufaily method allows for a greater variation in the series and may generate fewer alarms. This characteristic highlights the different sensitivities of the two methods.

### 5.1.2 Applying the novel outbreak detection algorithm to STEC

The subsequent investigation focuses on the application of the novel outbreak detection algorithm to monitor STEC for potential outbreaks. In this analysis, a rolling window with a width of  $k = 36$  months is selected for the parameter estimation. Thus, the reference values are established using monthly data collected from January 2008 to December 2010. Hereafter, the time series is recursively monitored for outbreaks by moving the rolling window one month at a time using data from January 2011 to December 2022.

To identify outbreaks, an observation is classified as such if the one-step ahead random effect  $u_{it_1}$  for the specific age group exceeds the upper bound  $U_{t_0}$ . In this case study, the upper bound is determined as the 90% quantile of the distribution of random effects obtained from the second stage model. This threshold helps identify significant deviations in the outbreak intensity.

#### Introducing the fixed effects models

To ensure effective monitoring, a series of models for the fixed effects are proposed, with each model building upon the previous one. These models aim to capture different aspects of the disease dynamics and improve the accuracy of outbreak detection.

In the initial model, the intensity  $\lambda_{it}$  is assumed to be solely dependent on the age group and the population size  $n_{it}$ . The model is formulated as:

$$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \log(n_{it}) \quad (5.1)$$

Here,  $\beta(\text{ageGroup}_i)$  represents the fixed effect specific to the age group  $i$ , capturing the age group's influence on the outbreak intensity. The term  $\log(n_{it})$  acts as an offset, accounting for the population size at time  $t$  for age group  $i$ . This model will be referred to as the *constant* model.

In the first extension of the constant model, the trend over time across all age groups is incorporated. The extended model is termed the *trend* model and is formulated as:

$$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\text{trend}} t + \log(n_{it}) \quad (5.2)$$

In this model,  $\beta_{\text{trend}}$  quantifies the rate of change in the outbreak intensity over time. By including this parameter, the model accounts for the overall trend in the outbreak intensity across all age groups, allowing for a more comprehensive analysis of the outbreak dynamics.

In addition to the models described earlier, another model is proposed, which incorporates seasonality into the analysis. This model builds upon the constant model and assumes an annual seasonality pattern. This model is termed the *seasonality* model. The intensity  $\lambda_{it}$  is modeled as follows:

$$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \sin\left(\frac{2\pi \cdot \tau_t}{12}\right)\beta_{\sin} + \cos\left(2\frac{\pi \cdot \tau_t}{12}\right)\beta_{\cos} + \log(n_{it}) \quad (5.3)$$

In this model,  $\tau_t$  represents the time period  $t$  within a year, ranging from 1 to 12 (corresponding to the months of January to December). The parameters  $\beta_{\sin}$  and  $\beta_{\cos}$  capture the effect of the seasonal pattern on the outbreak intensity. By including sine and cosine functions of  $\tau_t$ , the model accounts for the periodic fluctuations in the outbreak intensity observed throughout the year. This allows for the detection and analysis of seasonality patterns in the outbreak data, providing insights into the seasonal variations in disease occurrence.

In addition to the models described earlier, a final model is proposed that combines both trend and seasonality components. This model builds upon the previous models and includes the effects of both the overall trend over time and the seasonal patterns. Thus, the model is labeled as the *combined* model. The intensity  $\lambda_{it}$  is modeled as follows:

$$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{trend}t + \sin\left(\frac{2\pi \cdot \tau_t}{12}\right)\beta_{\sin} + \cos\left(\frac{2\pi \cdot \tau_t}{12}\right)\beta_{\cos} + \log(n_{it}) \quad (5.4)$$

### Performance evaluation and parameter estimates of hierarchical models

The proposed models above are subsequently implemented and estimated in two different modeling frameworks: the hierarchical Poisson Normal model and the hierarchical Poisson Gamma model. The performance of these models is evaluated using the average logarithmic score,  $\bar{S}(G, y)$ .

In Table 5.1, an excerpt of these results, namely for the models that minimizes the logarithmic score,  $\bar{S}(G, y)$ , for both modeling framework, are summarized. For the full table of results, refer to Table D.3.

Table 5.1: An excerpt of the STEC modeling results. The average logarithmic score,  $\bar{S}(G, y)$ , is computed for all the fixed effects models. The parameters are estimated in a rolling window with width  $k = 36$  and the estimates at time  $t_0$ , i.e. the last time point, are presented in this table for both modeling frameworks. Confidence intervals for the parameter estimates are calculated using 95% profile likelihood confidence intervals.

	Parameter	Estimate	95% CI
<b>Poisson Normal</b>			
$\bar{S}(G, y) = 14.18$	$\hat{\beta}_{trend}$	0.0442	[0.04, 0.05]
	$\hat{\beta}_{<1year}$	12.5572	[12.3, 12.81]
	$\hat{\beta}_{1-4years}$	15.2251	[15.03, 15.41]
	$\hat{\beta}_{5-14years}$	15.6763	[15.46, 15.88]
	$\hat{\beta}_{15-24years}$	16.1862	[15.99, 16.38]
	$\hat{\beta}_{25-64years}$	18.7354	[18.56, 18.9]
	$\hat{\beta}_{65+years}$	17.5977	[17.43, 17.77]
	$\hat{\beta}_{\sin}$	-0.3402	[-0.44, -0.25]
	$\hat{\beta}_{\cos}$	-0.0556	[-0.15, 0.04]
	$\log(\hat{\sigma})$	-1.1158	[-1.33, -0.92]
<b>Poisson Gamma</b>			

Table 5.1: (continued)

Model	Parameter	Estimate	95% CI
$\bar{S}(G, y) = 14.15$	$\hat{\beta}_{trend}$	0.0438	[0.04, 0.05]
	$\hat{\beta}_{<1year}$	12.6075	[12.34, 12.85]
	$\hat{\beta}_{1-4years}$	15.2839	[15.09, 15.47]
	$\hat{\beta}_{5-14years}$	15.7232	[15.51, 15.92]
	$\hat{\beta}_{15-24years}$	16.2339	[16.03, 16.42]
	$\hat{\beta}_{25-64years}$	18.7678	[18.6, 18.94]
	$\hat{\beta}_{65+years}$	17.6451	[17.47, 17.82]
	$\hat{\beta}_{\sin}$	-0.3350	[-0.43, -0.24]
	$\hat{\beta}_{\cos}$	-0.0415	[-0.14, 0.05]
	$\log(\hat{\phi})$	-2.2059	[-2.63, -1.82]

It can be shown, that the combined models for the fixed effects minimizes the logarithmic score for both modeling frameworks, indicating a better fit to the STEC data. As a result, these models have been selected for further investigation regarding their ability to detect outbreaks.

Additionally, it is worth noting that the parameter estimates and confidence intervals for the fixed effects are consistent across the two modeling frameworks, despite the differences in assumptions about the distribution of random effects. This suggests that the choice of modeling framework does not have a substantial impact on the estimation of the fixed effects in this analysis.

It is noteworthy that all parameter estimates, except one, are statistically significant at  $t_0$ . The only parameter estimate that is not significant is  $\hat{\beta}_{\cos}$ , as its 95% confidence interval overlaps with zero in both modeling frameworks.

A slight, yet consistent, increase in the contribution from the age group to the overall intensity of STEC cases is observed in Figure 5.2 over the analyzed period, with only one instance of a decrease in intensity immediately after the 2020 Covid-19 lockdown.

It is worth mentioning that in the hierarchical Poisson Gamma model, the maximum likelihood estimate for  $\hat{\beta}(\text{ageGroup}_i)$  directly represents the contribution of the age group to the disease intensity. This is possible due to the properties of the expected value described in Equation (4.29). However, this direct interpretation does not hold for the parameter estimates obtained in the hierarchical Poisson Normal model. Nevertheless, it can be observed that both sets of parameter estimates are similar throughout the entire time period, indicating that this discrepancy is not significant.

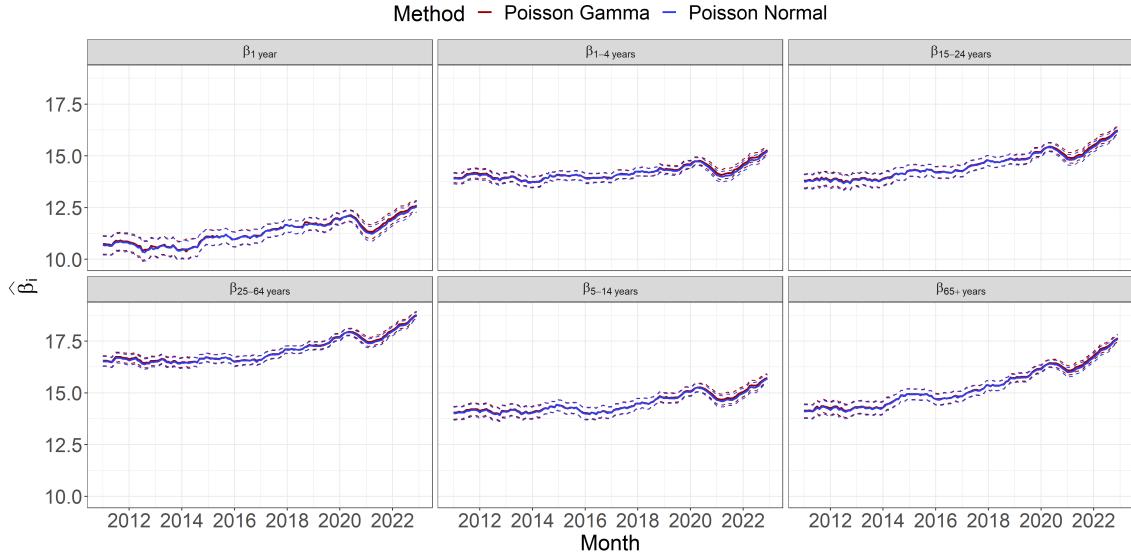


Figure 5.2: Considering STEC in the time period between January 2011 and December 2022. Estimate (solid line) and 95% profile likelihood confidence interval (dashed line) of  $\hat{\beta}(ageGroup_i)$  for the models given in Equation (5.4) estimated using the rolling window with a width of  $k = 36$  months.

For the most, the maximum likelihood estimate for  $\hat{\beta}_{trend}$  indicates a positive trend during the observed period (See Figure 5.3). However, it is important to note that the trend is not consistently statistically significant. The first significant positive trend is observed in the mid-2014, lasting until 2016. Another significant positive trend is observed in 2018, coinciding with the transition from a culture-based diagnostic method to a PCR method in most departments of clinical microbiology (Svendsen et al. 2023).

Notably, a significant drop in the trend is observed in the mid-2020, immediately after the Covid-19 lockdown, lasting until the mid-2021. Since then, the trend has been steadily increasing and is currently at an all-time high.

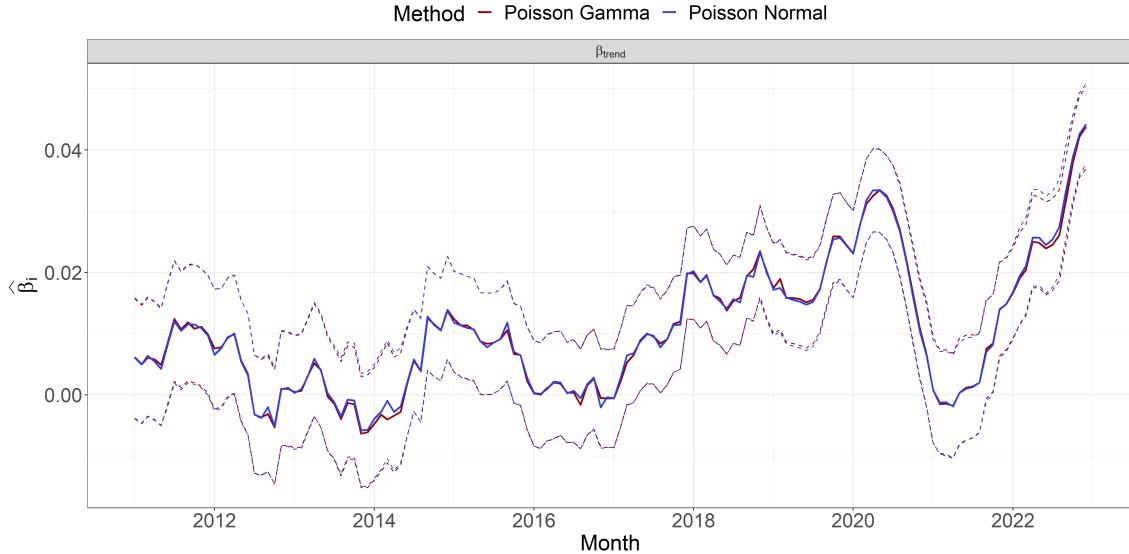


Figure 5.3: Disease trend in STEC in the time period between January 2011 and December 2022. Recursive estimates (solid line) and 95% profile likelihood confidence intervals (dashed line) of  $\hat{\beta}_{trend}$  for the combined model. Estimates are obtained using a rolling window with a width of  $k = 36$  months and discarding observations that are characterized as alarms.

From Figure 5.4, it is evident that there is a pronounced and statistically significant seasonality in the observed period. Additionally, it is noteworthy that the amplitude of the seasonal component varies over time. Specifically, there is an increase in the amplitude of the seasonal pattern after 2018, coinciding with the introduction of a PCR method for diagnostics. This change in diagnostic practices may have influenced the detection and reporting of STEC cases, leading to a seemingly intensified seasonal pattern in the data. However, it is most likely due to a change in the observational bias, as the PCR method is more sensitive than previous methods used. The varying amplitude emphasizes the importance of considering temporal changes and external factors when modeling and interpreting disease outbreak data.

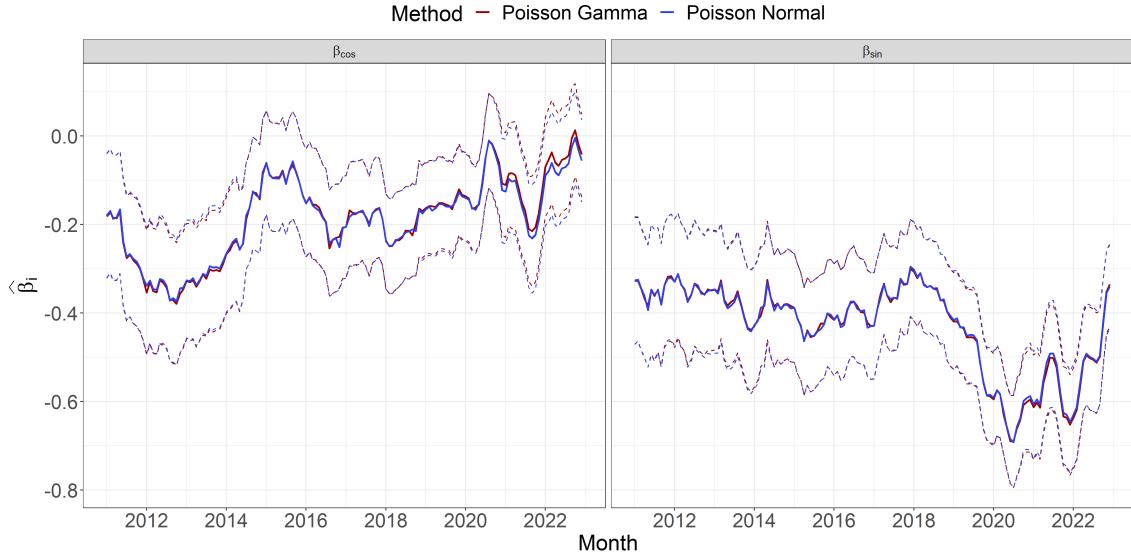


Figure 5.4: Disease seasonality in STEC in the time period between January 2011 and December 2022. Recursive estimates (solid line) and 95% profile likelihood confidence intervals (dashed line) of  $\hat{\beta}_{\cos}$  and  $\hat{\beta}_{\sin}$  for the combined model. Estimates are obtained using a rolling window with a width of  $k = 36$  months and discarding observations that are characterized as alarms.

The estimated variance  $\hat{\sigma}^2$  for the hierarchical Poisson Normal model and the estimated dispersion  $\hat{\phi}$  for the hierarchical Poisson Gamma model are visualized in Figure 5.5. These parameters provide insights into the variability of the random effects and play a crucial role in determining the threshold for detecting outbreaks. The values of  $\hat{\sigma}^2$  and  $\hat{\phi}$  help assess the magnitude of deviations from the expected disease intensity and aid in distinguishing significant aberrations from normal fluctuations in the data.

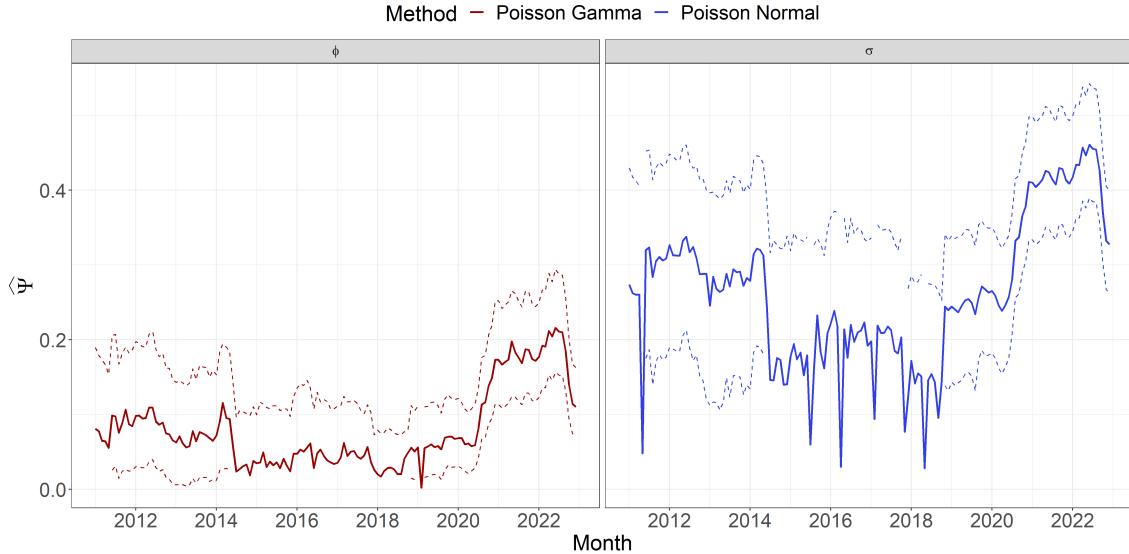


Figure 5.5: Variability and dispersion in STEC in the time period between January 2011 and December 2022. Recursive estimates (solid line) and 95% profile likelihood confidence intervals (dashed line) of  $\hat{\phi}$  and  $\hat{\sigma}$ . Estimates are obtained using a rolling window with a width of  $k = 36$  months and discarding observations that are characterized as alarms.

It should be noted that the inability of the second stage model parameters ( $\sigma$  or  $\phi$ ) to converge during certain time periods does not affect the performance of the novel outbreak detection algorithm. These periods typically coincide with time periods with relatively few reported cases, i.e. periods with no outbreaks. Generally, these periods occur when the data show no overdispersion. In such cases, the time series can be adequately modeled by a Poisson distribution, and therefore the second stage model is not informed. Specifically, this becomes relevant in the case study of LIST, where the example of missing overdispersion is very explicit. The role of overdispersion in the disease data will be discussed in Section 5.3.

#### Visualization of estimated one-step ahead random effects and upper bounds

Figure 5.6 illustrates the estimated one-step ahead random effects  $\hat{u}_{it_1}$  for each age group, along with their corresponding upper bounds  $U_{t_0}$ , in both modeling frameworks.

The random effects  $u_{it_1}$  represent the deviations from the expected outbreak intensity in the subsequent time period for each age group. These random effects provide information on whether there is an unusual or unexpected increase or decrease in the outbreak intensity compared to the expected values.

The upper bounds  $U_{t_0}$  are calculated based on the 90% quantile of the distribution of the random effects obtained from the second stage model. They serve as threshold values to identify potential outbreaks. If the one-step ahead random effects exceed the upper bounds, it suggests a significant deviation from normal variation and indicates a potential outbreak.

By visualizing the one-step ahead random effects and the upper bounds together, epidemiologists can easily identify periods or age groups where the random effects surpass the upper bounds, highlighting potential outbreaks that require further investigation and monitoring.

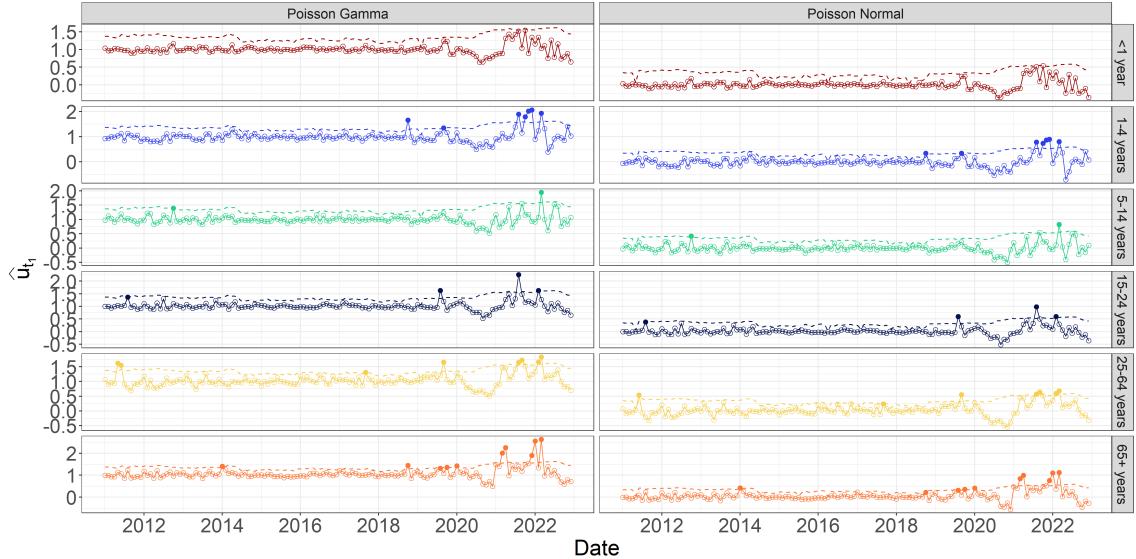


Figure 5.6: Estimated one-step ahead random effects  $\hat{u}_{it_1}$  (circles) given the model described in (5.4) for STEC in Denmark, 2011-2022. Poisson Normal model (right) and Poisson Gamma model (left) monitor the disease. Alarm raised (solid circle) if  $\hat{u}_{it_1}$  exceeds the threshold  $U_{t_0}$  (dashed line).

From Figure 5.6, it is evident that the estimated one-step ahead random effects  $u_{it_0}$  surpass the threshold  $U_{t_0}$  on multiple occasions. Specifically, a total of 30 alarms are generated using the hierarchical Poisson Normal framework, while 31 alarms are generated in the hierarchical Poisson Gamma framework. Notably, during the period from March 2021 to March 2022, a significant number of alarms are triggered in multiple age groups in both modeling frameworks.

## 5.2 Outbreak detection in other case studies

In this section, the results of applying the state-of-the-art and novel outbreak detection algorithms to the other case studies, including LIST, SHIL, and SALM, are presented.

For these investigations, the reference data used in the state-of-the-art outbreak detection algorithms is based on monthly data from January 2008 to February 2011. Subsequent surveillance is conducted using data from March 2011 to December 2022. In the novel outbreak detection algorithm, a rolling window of width  $k = 36$  is used. Thus, the reference data is established using monthly data collected from January 2008 to December 2010. The time series is then monitored for outbreaks using data from January 2011 to December 2022. Additionally, it is important to note that different age groupings are employed in each of the case studies.

### 5.2.1 Analysis of outbreak detection in LIST

Intuitively, it is not common to employ statistical methods for monitoring and detecting outbreaks of LIST. The current state-of-the-art approach for detecting outbreaks of this disease relies heavily on laboratory-based methods, specifically WGS. WGS enables epidemiologists to link cases that have occurred over an extended period, even months or years apart, and identify them as part of a continuous-source outbreak. Statistical method may face challenges in detecting such outbreaks due to their complex and prolonged nature. However, it is crucial to include this case study in the master's thesis to highlight both the strengths and limitations of the novel and state-of-the-art methods.

The data used in this case study consist of Danish LIST case counts from the laboratory notification system. These counts are denoted as  $y_{it}$ . Here, the subscript  $i$  distinguishes between two age groups, with  $i = 1$  representing the age group below 65 years and  $i = 2$  representing the age group above 65 years. The subscript  $t$  represents the time period, ranging from  $t = 1, \dots, T$ , where  $T = 180$  corresponds to the total number of months starting in 2008.

### Applying the state-of-the-art outbreak detection algorithms to LIST

The Farrington and Noufaily methods are both investigated on their ability to detect outbreaks of LIST. The resulting series is visualized in Figure 5.7 together with the alarms triggered by each of the methods.

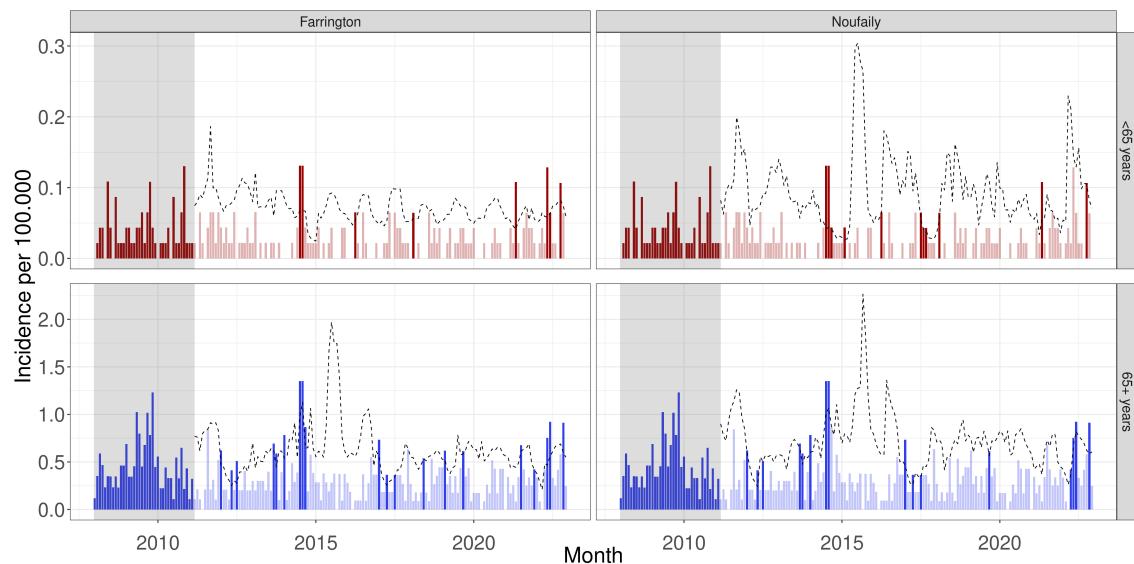


Figure 5.7: Monthly LIST incidence per 100,000 in Denmark, 2008-2022. Monitored by (left) Farrington and (right) Noufaily method. Reference data for the estimation of model parameters from January 2008 to March 2011 (grey area). Threshold (dashed line) is computed for observations time points outside reference data. Alarm triggered (dark color) if observations exceeds threshold.

In Figure 5.7, multiple alarms can be observed for both the Farrington Noufaily methods. The Farrington method triggers a total of 28 alarms, while the Noufaily method produces slightly fewer alarms, specifically 27 alarms.

Interestingly, both methods demonstrate the ability to correctly raise alarms during ongoing outbreak investigations by SSI. One notable example is the outbreak investigated by SSI, which involved 41 cases and resulted in 17 deaths. The Farrington method successfully flags this event in September 2013, with alarms occurring sporadically in the subsequent period. The Noufaily method also identifies this outbreak, although it does so one month later in October 2013.

It is worth noting that the other observations flagged by the methods may be related to some of the numerous other long-spanned outbreaks that occurred concurrently in the period from 2016 to 2021. However, it is out of scope for this master's thesis, to directly link individual alarms to specific outbreaks. For a full list over the alarms raised by the methods see Table C.1.

### Applying the novel outbreak detection algorithm to LIST

For the LIST data, it is found that only the constant age group model for the fixed effects is able to converge. However, even the constant fixed effects models struggle to accurately capture the underlying process of the disease. Figure 5.8 provides a visual representation of this, showing the estimated one-step ahead random effects  $\hat{u}_{it_1}$  and the calculated thresholds  $U_{t_0}$ . It can be observed that the estimated random effects and the calculated thresholds collapse for prolonged periods, indicating that there is insufficient information in the data to inform the second stage model. This is particularly evident in the hierarchical Poisson Normal model, where the estimate of the variance parameter  $\hat{\sigma}$  approaches zero. The role of overdispersion in the disease data will be further discussed in Section 5.3.

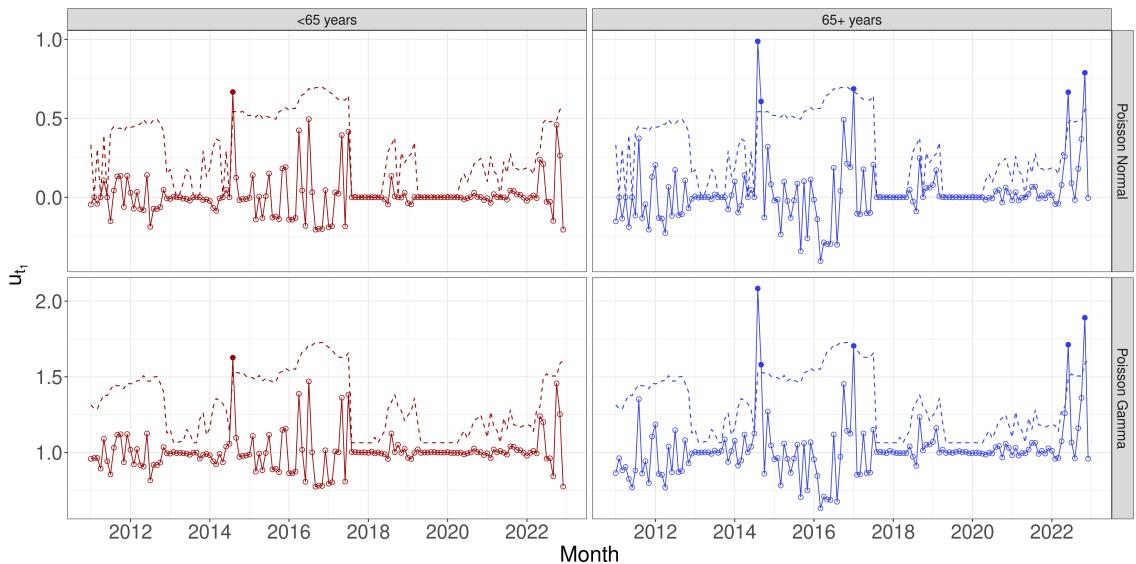


Figure 5.8: Estimate one-step ahead random effects  $\hat{u}_{it_1}$  (circles) given the model described in (5.1) of LIST in Denmark, 2011-2022, as circles. Poisson Normal model (top) and Poisson Gamma model (bottom) monitor the disease. Alarm raised (solid circle) if  $\hat{u}_{it_1}$  exceeds the threshold  $U_{t_0}$  (dashed line).

Nevertheless, it is worth noting that both modeling framework identifies the same observations as outbreaks, all of which coincide with one or more concurrent outbreaks investigated by SSI. The first two observations characterized as outbreaks occur in August and September 2014. The third outbreak is in January 2017, and the last two outbreaks are in June and November 2022.

### 5.2.2 Analysis of outbreak detection in SHIL

The SHIL data set consists of Danish SHIL case counts categorized into two age groups. These counts are represented as  $y_{it}$ , where the subscript  $i = 1$  corresponds to the age group below 25 years, and  $i = 2$  corresponds to the age group above 25 years. It should be noted that the selection of these age groups was not based on clear epidemiological reasoning. In retrospect, it would have been more informative to differentiate between sexes rather than age groups, as men who have sex with men have a higher risk of contracting and spreading the disease. However, this information was received towards the end of writing the master's thesis and was therefore not given priority.

### Applying the state-of-the-art outbreak detection algorithm to *Shigellosis*

In Figure 5.9, a significant number of alarms are observed sporadically in the time series for both of the state-of-the-art methods. For the Farrington method, the number of alarms

is 54, while for the Noufaily method, the number of alarms is 50.

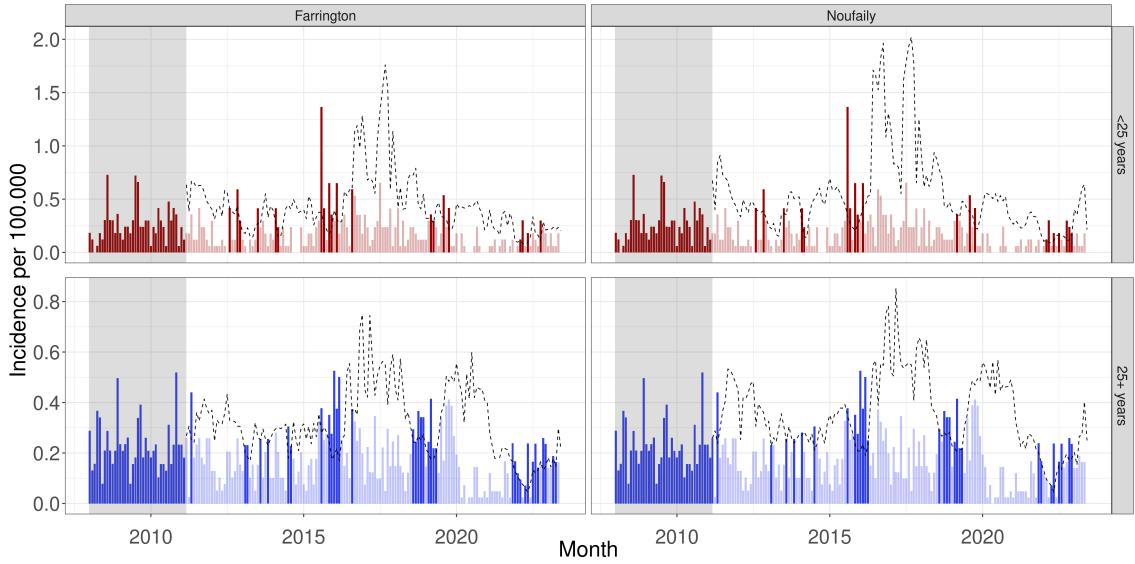


Figure 5.9: Monthly SHIL incidence per 100,000 in Denmark, 2008-2022. Monitored by (left) Farrington and (right) Noufaily method. Reference data for the estimation of model parameters from January 2008 to March 2011 (grey area). Threshold (dashed line) is computed for observations time points outside reference data. Alarm triggered (full opacity) if observations exceeds threshold.

It is noteworthy that some of the alarms occur in clusters over specific time periods and are raised simultaneously in both age groups. These clustering patterns are particularly evident during periods with multiple alarms, such as from July 2015 to April 2016, from July 2018 to October 2019, and from November 2021 to December 2022. Surprisingly, none of these periods coincide with outbreaks investigated by SSI, and all methods fail to identify the outbreak investigated in August and September 2020.

However, it is important to recognize the significant challenge faced by statistical algorithms in detecting the outbreak during the summer of 2020. This difficulty arises due to the concurrent Covid-19-related lockdown, which results in different disease dynamics and a relatively low number of SHIL cases compared to similar periods in other years. This highlights the importance of considering the context in which the statistical outbreak detection algorithm is applied. Further discussion on this topic can be found in Section 5.3.

### Applying the novel outbreak detection algorithm to *Shigellosis*

For SHIL it was found that the trend model for the fixed effects minimizes the average logarithmic score,  $\bar{S}(G, y)$ , in both modeling frameworks. Thus, the estimated one-step ahead random effects  $\hat{u}_{it_1}$  and thresholds  $U_{t_0}$  visualized in Figure 5.10 is calculated using this model.

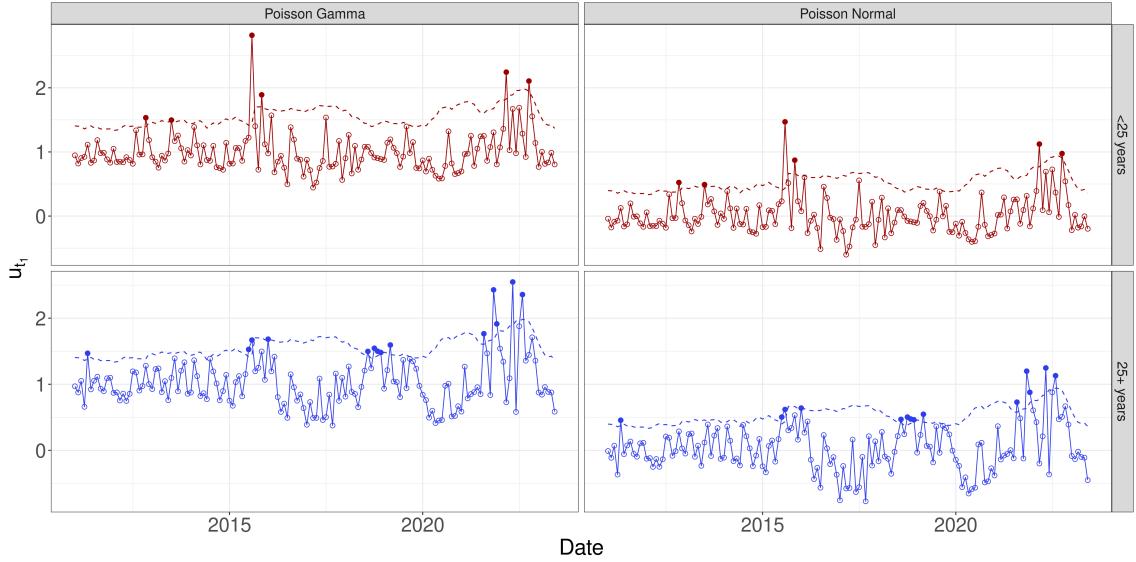


Figure 5.10: Estimated one-step ahead random effects  $\hat{u}_{t_1}$  of SHIL in Denmark, 2011–2022, as circles. Poisson Normal model (top) and Poisson Gamma model (bottom) monitor the disease. Alarm raised (solid circle) if  $\hat{u}_{t_1}$  exceeds the threshold  $U_{t_0}$  (dashed line).

It is noteworthy that the 20 alarms generated by both modeling frameworks coincide, indicating consistency in their detection. Additionally, it is observed that the novel outbreak detection algorithms identify significantly fewer alarms compared to the state-of-the-art algorithms. However, it is important to highlight that both algorithms detect outbreaks during certain overlapping time periods, demonstrating their ability to capture similar patterns of disease occurrence.

### 5.2.3 Analysis of outbreak detection in SALM

Historically, SALM is associated with frequent disease outbreaks. This is no different during the time period investigated here. The data set consist of monthly counts of Danish SALM cases, denoted as  $y_{it}$ . Here, the subscript  $i$  differentiates between the six age groups  $i = 1, \dots, 6$ . The subscript  $t$  represents time.

#### Applying the state-of-the-art outbreak detection algorithm to *Salmonellosis*

From Figure 5.11, it is evident that both of the state-of-the-art methods generate an overwhelming number of alarms across all six age groups. The Farrington method produces 131 alarms, while the Noufaily method produces 98 alarms. Although these methods exhibit high sensitivity, the excessive number of alarms undermines their usefulness and practicality.

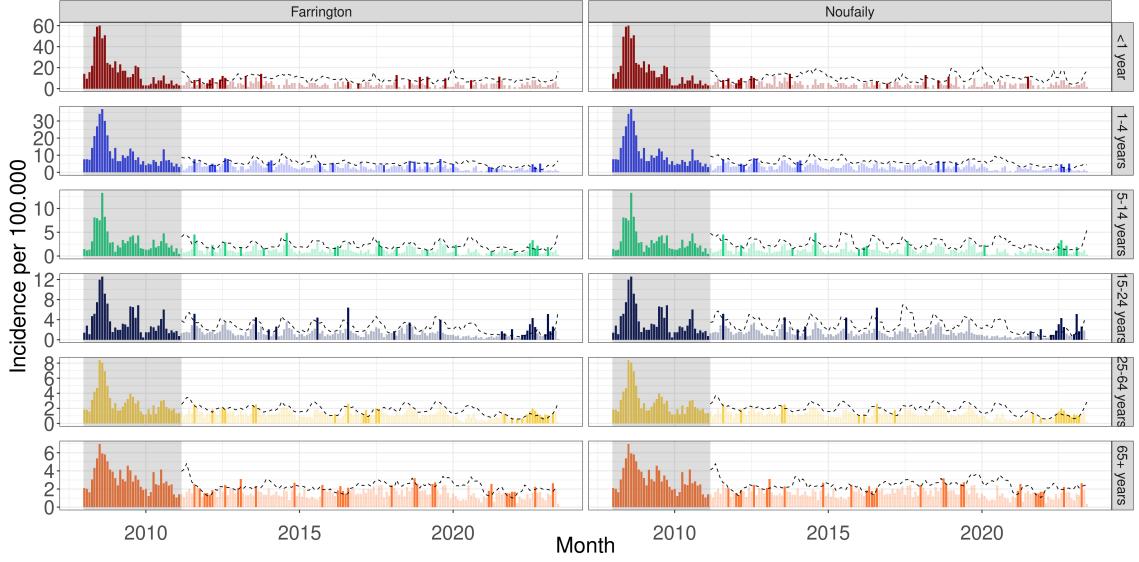


Figure 5.11: Monthly SALM incidence per 100.000 in Denmark, 2008-2022. Monitored by (left) Farrington and (right) Noufaily method. Reference data for the estimation of model parameters from January 2008 to March 2011 (grey area). Threshold (dashed line) is computed for observations time points outside reference data. Alarm triggered (full opacity) if observations exceeds threshold.

### Applying the novel outbreak detection algorithm to *Salmonellosis*

In Figure 5.12 the estimated one-step ahead random effects  $\hat{u}_{it_1}$  are visualized together with the calculated threshold  $U_{t_0}$ .

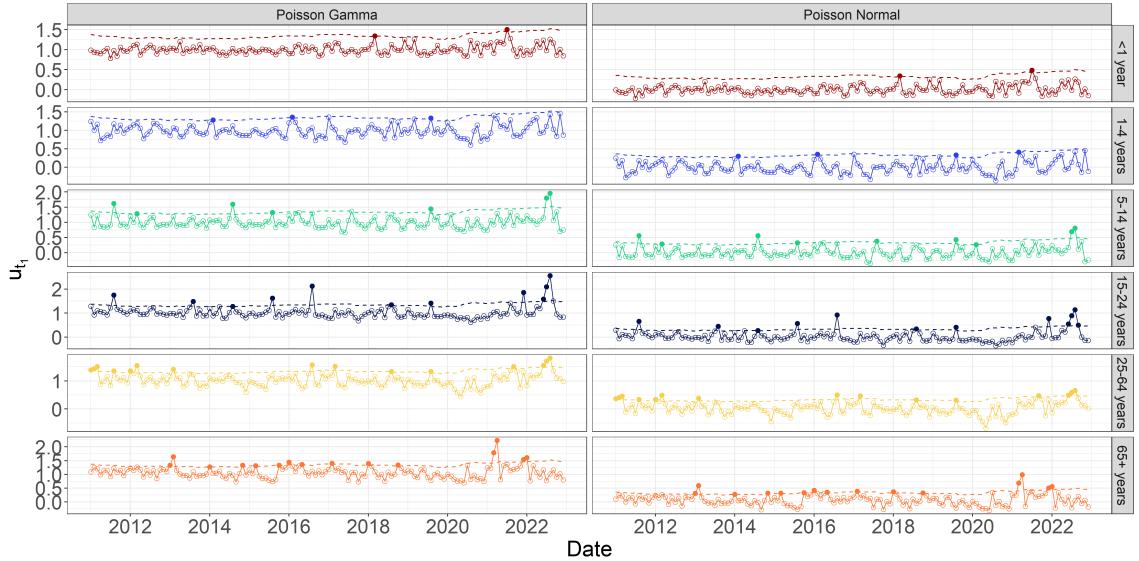


Figure 5.12: Estimated one-step ahead random effects  $\hat{u}_{it_1}$  of SALM in Denmark, 2011-2022, as circles. Poisson Normal model (top) and Poisson Gamma model (bottom) monitor the disease. Alarm raised (solid circle) if  $\hat{u}_{it_1}$  exceeds the threshold  $U_{t_0}$  (dashed line).

In total, the hierarchical Poisson Normal framework generates 57 alarms, while the hierarchical Poisson Gamma framework generates 54 alarms. This demonstrates that the novel

outbreak detection algorithm has significantly lower sensitivity compared to the state-of-the-art algorithms.

However, the large number of alarms raises valid concerns about how to handle them effectively. In the novel algorithm, observations that trigger an alarm are identified and excluded from future parameter estimation to prevent inflating the parameter estimates. This approach aims to mitigate the impact of false alarms and maintain the accuracy of the modeling process. However, as the alarms accumulate, the effective sample window for parameter estimation diminishes, reducing the amount of available data for estimating reliable parameters. This challenge is expanded upon in Section 5.3.

### 5.3 Challenges in statistical outbreak detection

In this section, the challenges encountered in this master's thesis regarding statistical outbreak detection are outlined. These challenges encompass specific issues related to the use of hierarchical modeling as well as more generalized challenges that are relevant to statistical modeling in general.

The discussion will focus on the role of overdispersion in statistical outbreak detection, the impact of context and bias, and the handling of a large number of alarms.

#### 5.3.1 Investigating the role of overdispersion in statistical outbreak detection for LIST

Convergence issues were observed in the hierarchical models during the case studies, particularly in the LIST case study. Only the constant fixed effects models were able to converge, indicating challenges with the other models. The low incidence rate in the LIST data and the absence of overdispersion are likely factors contributing to these convergence difficulties.

To further investigate the lack of convergence, the empirical mean and variance of the LIST data series in the two age groups are visualized in Figure 5.13 using a rolling window approach. The rolling window width was set to match the parameter estimation window, which was 36 months. This visualization provides a closer examination of the patterns in the data and helps to identify potential factors affecting convergence.

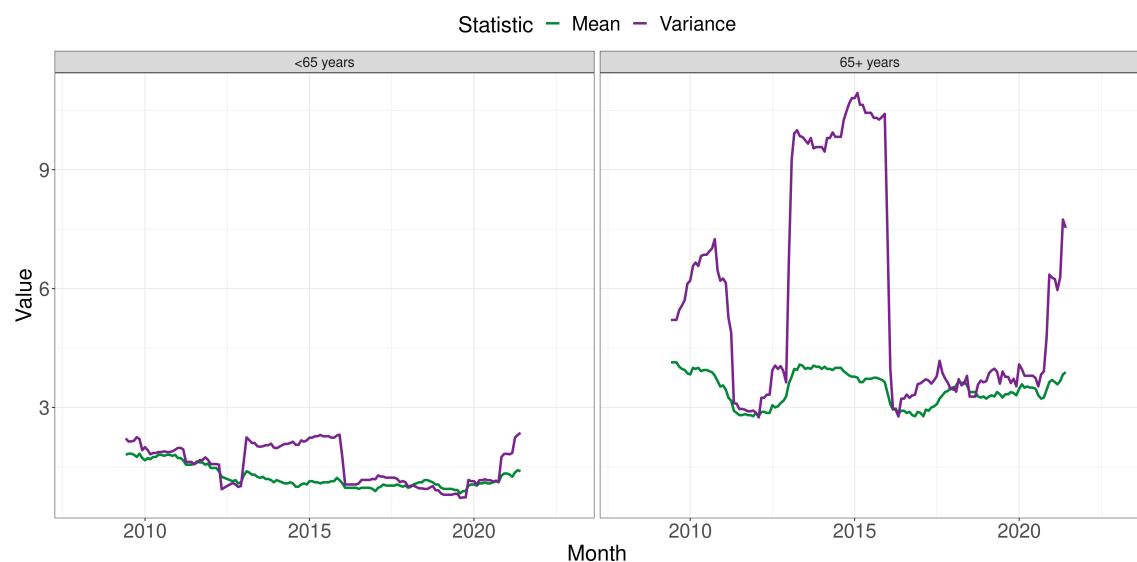


Figure 5.13: Empirical mean and variance of the LIST data in two age groups.

The figure clearly shows prolonged periods in the sample window where there is no overdispersion, as indicated by the similarity between the empirical mean and variance. The duration of the window with overdispersion is directly influenced by the chosen window size, and even a small number of observations can introduce significant overdispersion.

It is worth noting that the periods where the estimated second stage model parameters ( $\hat{\sigma}$  or  $\hat{\phi}$ ) fail to converge align with periods where statistical evidence for an ongoing outbreak is not feasible. These periods are characterized by a low number of cases and, more importantly for the hierarchical models, the absence of overdispersion in the sample window.

As the number of cases are very low, it is also unlikely that an outbreak is occurring. Regardless, it is not possible for statistical models to detect these outbreaks.

In the case of the novel outbreak detection algorithm, particular attention is given to periods with overdispersion. This is because the modeling frameworks assume that the distribution of random “noise” can be modeled by a Gaussian distribution, viz. hierarchical Poisson Normal model, or an exponential dispersion family, viz. hierarchical Poisson Gamma model. In the absence of overdispersion, count data can be adequately modeled by a Poisson distribution alone, making an approach based on generalized linear models more suitable.

Further investigation into modeling periods without overdispersion using hierarchical models are required but out of scope of this master’s thesis.

### 5.3.2 The impact of context and observational bias

Deploying an outbreak detection algorithm blindly, without considering the context and potential changes in observational bias, can lead to unreliable results. It is essential to closely monitor the algorithms and critically assess the model inputs and the specific context in which it operates. This is evident in the case study of SHIL, where an outbreak investigated by SSI in August and September 2020 was challenging to detect using statistical methods due to the concurrent Covid-19-related lockdown and the resulting decrease in reported cases.

Considering the specific circumstances and potential changes in observational bias is crucial when interpreting outbreak detection results. It is important to determine whether the number of cases during an outbreak is considered aberrant compared to similar periods in other years. However, in situations where there is a sudden decrease in reported cases, such as during a lockdown, detecting outbreaks using statistical methods becomes more challenging.

An important potential extension of automated statistical outbreak detection algorithms is the incorporation of data drift monitoring. This involves detecting changes in the underlying data distribution and adapting the model accordingly to maintain its predictive performance. Specifically, detecting sudden decreases in case counts can be considered a form of data drift.

Indeed, sudden decreases in case counts can be treated similarly to sudden increases in case counts, and detecting such data drift can be incorporated into hierarchical models using the unobserved one-step ahead random effects  $u_{it_0}$ . One approach is to define a lower threshold by selecting an appropriate quantile of the random effects distribution obtained from the second stage model, such as the 5% or 10% quantile.

However, the implementation of automated data drift monitoring requires further investigation and is beyond the scope of this master's thesis. It is an area that calls for future research and development to enhance the robustness and adaptability of statistical outbreak detection algorithms. The inclusion of data drift monitoring can improve the algorithms' ability to adapt to changing outbreak patterns and contextual factors, leading to more reliable and accurate detection outcomes.

### 5.3.3 How do we handle many alarms?

Dealing with diseases that generate frequent alarms presents a challenge in statistical outbreak detection. The novel outbreak detection algorithm addresses this challenge by excluding flagged observations from future parameter estimation. This exclusion reduces the sample window for parameter estimation and results in a loss of information and reduced variation in the data. The purpose of this exclusion is to prevent inflated parameter estimates.

On the other hand, the state-of-the-art detection algorithm employs a re-weighting scheme using Anscombe residuals. It has been observed that the Farrington method's correction for past outbreaks is too drastic, and a higher re-weighting threshold of 2.58 rather than 1 is recommended in the Noufaily method to improve specificity (Noufaily et al. 2013).

Effectively handling previous outbreaks in statistical outbreak detection requires domain-specific knowledge from epidemiologists to determine the most appropriate approach. While the novel method completely excludes past outbreaks, this may not be the optimal practice as it significantly reduces the sample window. If the method is employed in an automated setting, it is essential to report the size of the sampling window to the epidemiologists so they can interpret the results appropriately. Further research in this area is necessary, but it is beyond the scope of this master's thesis.

## 5.4 Performance comparison of novel methods

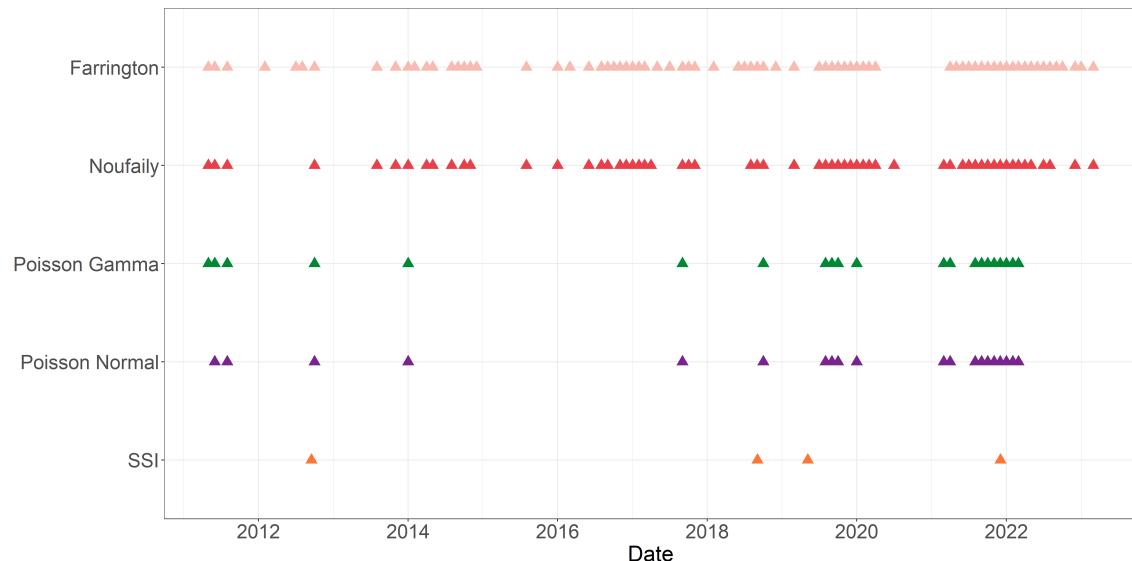


Figure 5.14: Alarm plot displaying alarms for Shiga toxin (verotoxin)-producing *Escherichia coli* time series using four different algorithms, along with outbreaks investigated by SSI.

WRITE HERE

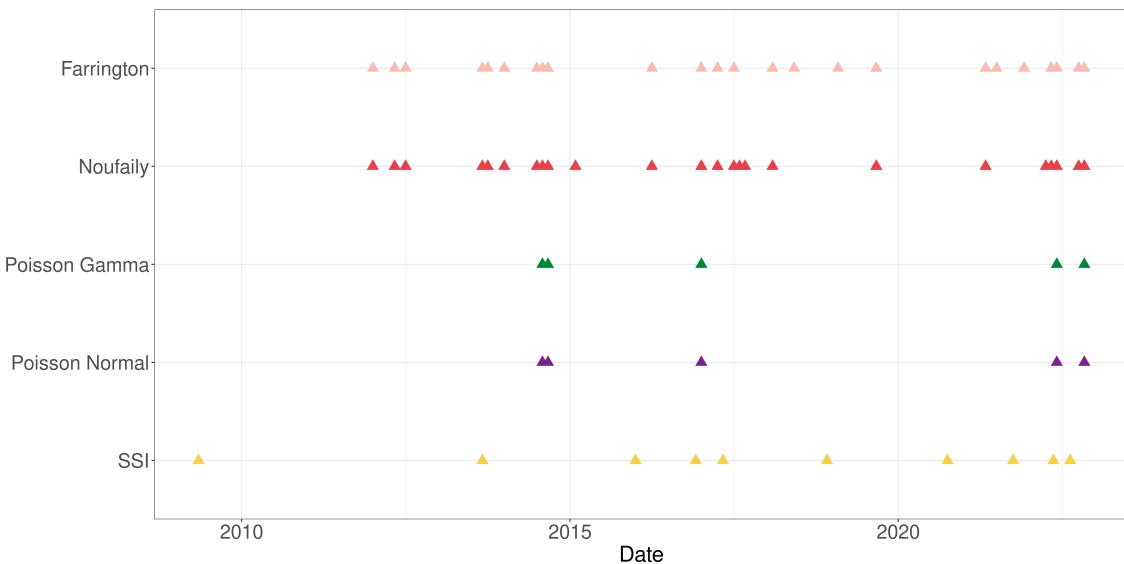


Figure 5.15: Alarm plot displaying alarms for LIST time series using four different algorithms, along with outbreaks investigated by SSI.

WRITE HERE

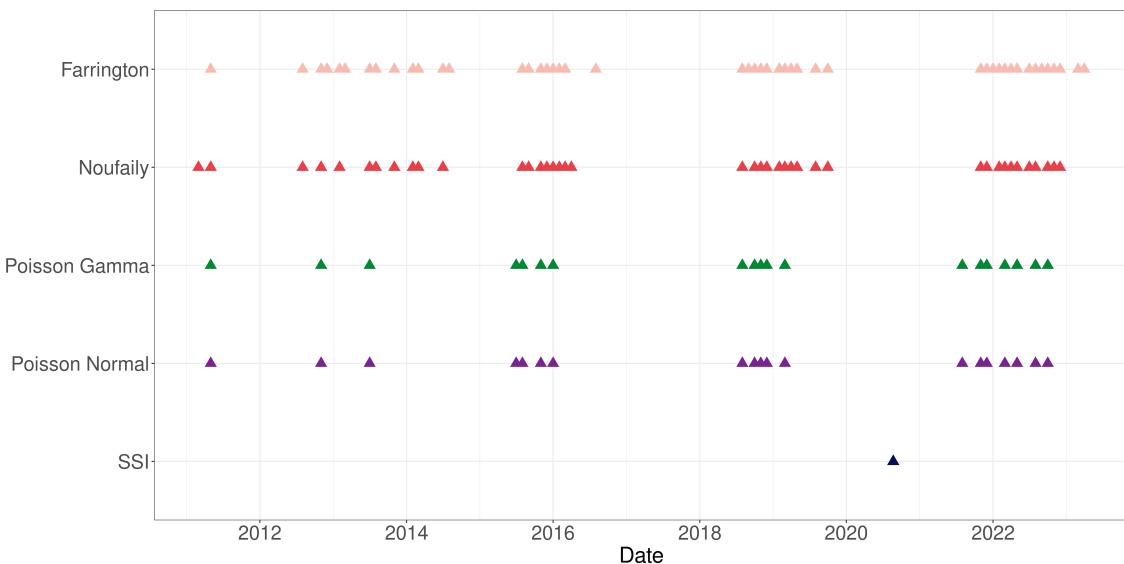


Figure 5.16: Alarm plot displaying alarms for SHIL time series using four different algorithms, along with outbreaks investigated by SSI.

WRITE HERE



Figure 5.17: Alarm plot displaying alarms for SALM time series using four different algorithms, along with outbreaks investigated by SSI.

INCLUDE A TABLE THAT SUMMARIZES THESE RESULTS. AN IDEA IS TO CALCULATE A CONFUSION MATRIX!

# 6 Simulation study

In this chapter, a comprehensive simulation study is conducted to evaluate the performance of the novel outbreak detection algorithm compared to state-of-the-art algorithms. The simulations cover various scenarios, adapted from the study by Noufaily et al. (2013). However, for the purpose of this master's thesis, the focus is not on diseases with bi-annual seasonality. Therefore, the scenarios involving bi-annual seasonality are excluded from the simulation study.

The chapter begins by describing the method used to simulate the baseline data. These data are generated using a negative binomial model with a time-dependent mean  $\mu(t)$ . Next, the assumptions regarding the simulated outbreaks are outlined, including the outbreak size and distribution in time.

The evaluation measures used to assess the performance of the outbreak detection algorithms are then presented. These measures are designed to capture relevant quantities in the context of outbreak detection.

Finally, the chapter presents the results of applying both the state-of-the-art outbreak detection algorithms and the novel outbreak detection algorithm to the simulated data. The performance of each algorithm is evaluated based on the simulation results.

## 6.1 The simulated baseline data

The simulated baseline data is generated using a negative binomial model with a mean parameter  $\mu$  and a variance parameter  $\phi\mu$ . The dispersion parameter  $\phi$  is assumed to be greater than or equal to 1. The mean  $\mu(t)$  is defined by a linear predictor that includes a trend component and a seasonality component represented by Fourier terms.

The equation for  $\mu(t)$  is given as:

$$\mu(t) = \exp \left( \theta + \beta_t + \sum_{j=1}^m \left( \gamma_1 \cos \left( \frac{2\pi jt}{52} \right) + \gamma_2 \sin \left( \frac{2\pi jt}{52} \right) \right) \right) \quad (6.1)$$

In this equation,  $m$  represents the number of Fourier terms used to model seasonality. When  $m = 0$ , it indicates the absence of seasonality, while  $m = 1$  corresponds to annual seasonality.

To cover a wide range of data sets encountered in practical applications, 28 different parameter combinations are generated. These combinations vary in terms of trends (represented by different values of  $\beta$ ), seasonalities (represented by different values of  $\gamma_1$  and  $\gamma_2$ ), baseline frequencies of reports (represented by different values of  $\theta$ ), and dispersion (represented by different values of  $\phi$ ). The specific parameter values for the 28 scenarios are provided in Table 6.1.

Table 6.1: Parameters and criteria utilized to generate the 28 scenarios.

Scenario	$\theta$	$\phi$	$\beta$	$\gamma_1$	$\gamma_2$	$m$	Trend
1	0.10	1.5	0.0000	0.00	0.00	0	0
2	0.10	1.5	0.0000	0.60	0.60	1	0

Table 6.1: (*continued*)

Scenario	$\theta$	$\phi$	$\beta$	$\gamma_1$	$\gamma_2$	$m$	Trend
3	0.10	1.5	0.0025	0.00	0.00	0	1
4	0.10	1.5	0.0025	0.60	0.60	1	1
5	-2.00	2.0	0.0000	0.00	0.00	0	0
6	-2.00	2.0	0.0000	0.10	0.30	1	0
7	-2.00	2.0	0.0050	0.00	0.00	0	1
8	-2.00	2.0	0.0050	0.10	0.30	1	1
9	1.50	1.0	0.0000	0.00	0.00	0	0
10	1.50	1.0	0.0000	0.20	-0.40	1	0
11	1.50	1.0	0.0030	0.00	0.00	0	1
12	1.50	1.0	0.0030	0.20	-0.40	1	1
13	0.50	5.0	0.0000	0.00	0.00	0	0
14	0.50	5.0	0.0000	0.50	0.50	1	0
15	0.50	5.0	0.0020	0.00	0.00	0	1
16	0.50	5.0	0.0020	0.50	0.50	1	1
17	2.50	3.0	0.0000	0.00	0.00	0	0
18	2.50	3.0	0.0000	1.00	0.10	1	0
19	2.50	3.0	0.0010	0.00	0.00	0	1
20	2.50	3.0	0.0010	1.00	0.10	1	1
21	3.75	1.1	0.0000	0.00	0.00	0	0
22	3.75	1.1	0.0000	0.10	-0.10	1	0
23	3.75	1.1	0.0010	0.00	0.00	0	1
24	3.75	1.1	0.0010	0.10	-0.10	1	1
25	5.00	1.2	0.0000	0.00	0.00	0	0
26	5.00	1.2	0.0000	0.05	0.01	1	0
27	5.00	1.2	0.0001	0.00	0.00	0	1
28	5.00	1.2	0.0001	0.05	0.01	1	1

To simulate the baseline data without outbreaks, 100 replicates are generated for each of the 28 parameter scenarios. Each replicate consist of a time series of size  $T = 624$  weeks.

The 624 weeks are divided into three periods: 1-313 are used as training weeks for the adaptive re-weighting schemes, weeks 313-575 are considered as baseline weeks, and weeks 576-624 are designated as the “current” weeks for evaluation.

The simulation results are based on the “current” weeks of all the replicates, totaling  $100 \times 49 = 4900$  replicates, for each of the 28 data scenarios and each method investigated.

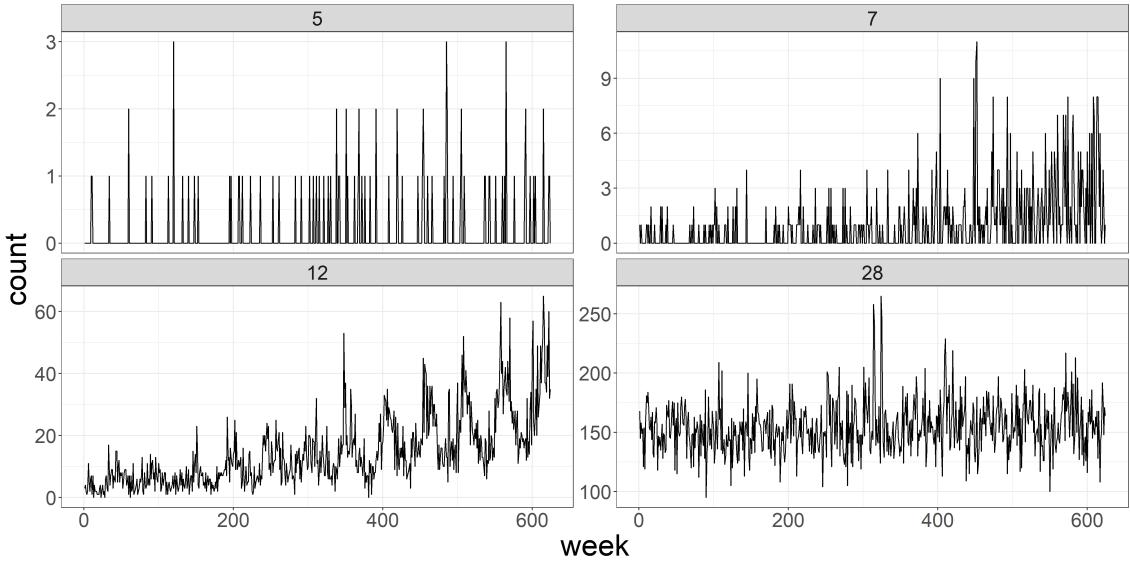


Figure 6.1: Plots of an excerpt of the realizations for scenario 5, scenario 7, scenario 12, and scenario 28 (see Table 6.1).

## 6.2 The simulated outbreaks

The outbreaks starting in week  $t$  are simulated using the following procedure. First, a constant value  $k$  is chosen. The size of the outbreak, denoted as  $n$ , is then generated randomly from a Poisson distribution with a mean equal to  $k$  times the standard deviation of the baseline count at time  $t$ .

Next, the outbreak is distributed randomly in time according to a log-normal distribution with a mean of 0 and a standard deviation of 0.5, represented as  $Z \sim \text{LN}(0, 0.5^2)$ . This is achieved by drawing  $n$  random numbers from the log-normal distribution, which correspond to the outbreak size, and then rounding down these numbers to the nearest integer, denoted as  $\lfloor z \rfloor$ .

The probability density function for the log-normal distribution is visualized in Figure 6.2.

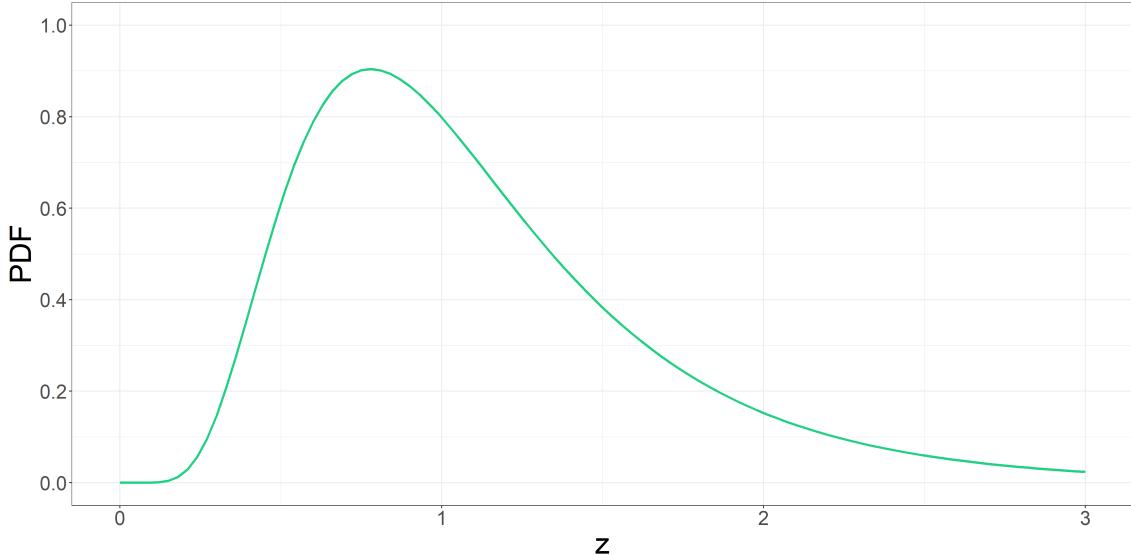


Figure 6.2: Probaility density function for the log-normal distribution with a mean of 0 and a standard deviation of 0.5, i.e.  $Z \sim \text{LN}(0, 0.5^2)$ .

Typically, outbreak durations of 2-8 weeks are observed when values of  $k$  are in the range of 2-10. To simulate the outbreaks, the outbreak cases are added to the baseline count in week  $t + z_i$ , where  $t$  represents the start time of the outbreak and  $z_i$  represents the number of weeks after the start of the outbreak. The start and end times of the outbreaks are recorded for evaluating the performance of the methods.

To simulate outbreaks, the following procedure is followed:

- *Outbreaks in baseline weeks.* For each data series, four outbreaks are generated. The start time of each outbreak is randomly selected from the baseline weeks (weeks 313-575). The value of  $k$  is sampled randomly with replacement from the set  $\{2, 3, 5, 10\}$ . It should be noted that different outbreaks are generated for each of the 2800 runs.
- *Outbreaks in current weeks.* For each data series, one outbreak is generated. The start time of the outbreak is randomly chosen from the last 49 weeks (weeks 576-624). The value of  $k$  is sampled randomly in the range of 1 to 10. Similar to the previous case, different outbreaks are generated for each of the 2800 runs.

### 6.3 Evaluation measures

To evaluate the performance of the outbreak detection system, several measures are employed to assess its effectiveness in detecting current outbreaks. These measures are specifically designed to capture relevant quantities in the given context.

One of the primary measures used is the False Positive Rate (FPR), which is calculated for each of the 28 scenarios and methods. The FPR is determined by calculating the proportion of the 49 weeks and 100 replicates in which the observed value exceeds the threshold, even though there are no actual outbreaks during those weeks.

Another measure is the probability of detecting an outbreak (POD), also known as power. Likewise, this is calculated for each of the 28 scenarios. The algorithm is applied iteratively for the 49 current weeks, and an outbreak is considered detected if the observed value

exceeds the threshold at least one within the start and end times of the outbreak. The POD is then determined as the proportion of outbreaks detected out of the 100 runs.

It is important to note that the FPR is a rate per week, while the POD is a rate per outbreak. These evaluation measures are chosen because they provide insights into the performance of the detection system on individual time series.

## 6.4 Results of the simulations

In this section, the results regarding the simulation study is presented. Both the state-of-the-art outbreak detection algorithm and the novel outbreak detection algorithm are applied to the simulated data.

The control arguments for the Farrington and Noufaily method respectively can be found in Appendix C.1.

For both of the state-of-the-art outbreak detection algorithms

In all of the methods, a

The model for the fixed effects in each of the modeling frameworks respectively, is defined as:

$$\log(\lambda_t) = \beta_{intercept} + \beta_{trend}t + \sin\left(\frac{2\pi \cdot \tau_t}{52}\right)\beta_{\sin} + \cos\left(\frac{2\pi \cdot \tau_t}{52}\right)\beta_{\cos} \quad (6.2)$$

in each of the modeling frameworks

In general, one obtains relatively higher FPRs when using the Farrington method compared to the other methods. This is to no surprise, as this method is known for being overly sensitive, and thus producing more false alarms.

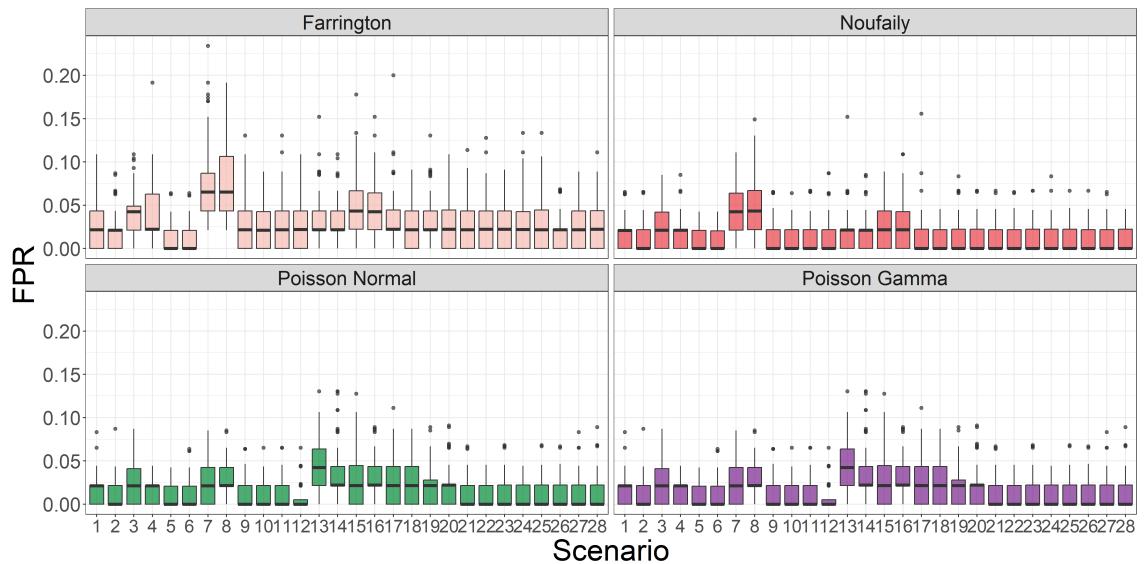


Figure 6.3: FPRs obtained in each of the 28 scenarios for each of the methods applied. Thresholds are based on the 0.95 quantile

The PODs for the 28 scenarios and eac

To no surprise, the PODs increase with size of  $k$ . Intuitively, the larger the outbreak size is, the more likely it is that the outbreak is detected.

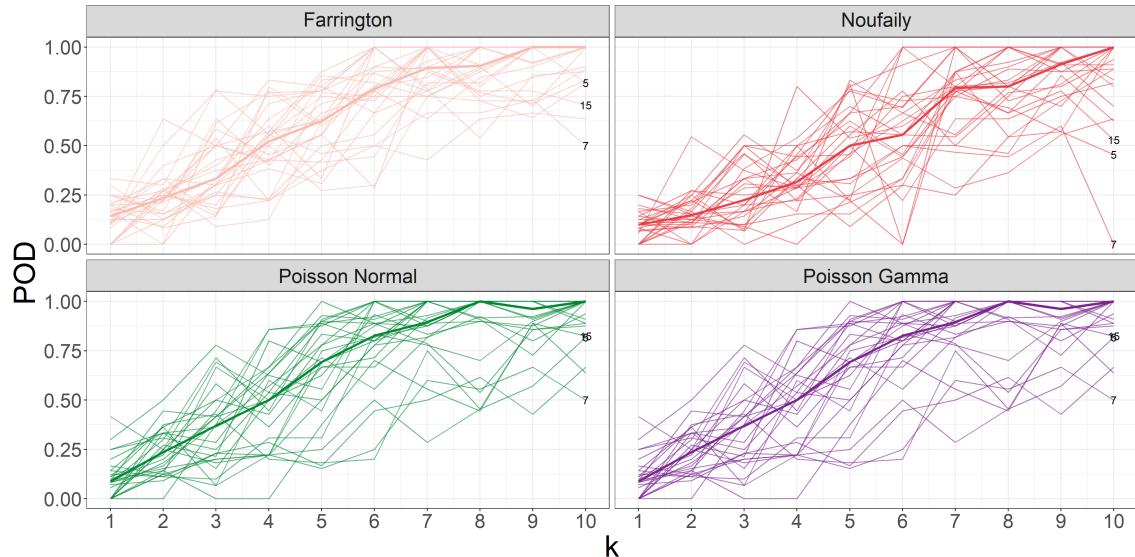


Figure 6.4: PODs of  $k$  standard deviations, along with the median curves (bold), corresponding to the 28 scenarios for each of the methods applied.

Interestingly, some scenarios produce unusual low power curves. Specifically, this is seen in the Noufaily method for scenario 5, 7, and 15. Here, these scenarios leads to less than 60% of the outbreaks detected with this method, when the outbreak size is  $k = 10$  times the standard deviation. In general, scenario 7 leads to the lowest power curve across all methods, with none of the outbreaks being detected by the Noufaily method.

It is important to remember, that the outbreak of size  $n$  is distributed randomly in time according to a log-normal distribution with mean 0 and a standard deviation of 0.5. Thus, the simulated outbreaks is not observed in a single time point, but rather in several concurrent time points.

In general, this means that an outbreak generated from a Poisson distribution with mean equal to  $k$  times the standard deviation of the baseline series is observed to be relatively smaller, than what it initially perceived from the simulation setup, i.e. an outbreak of size  $k = 4$  times the standard deviation may only be perceived as an outbreak of size 2 times the standard deviation in an individual time point.

Table 6.2: Placeholder

Method	min(FPR)	max(FPR)	mean(FPR)	median(FPR)	sd(FPR)
Farrington	0	0.2340	0.0306	0.0217	0.0314
Noufaily	0	0.1556	0.0163	0.0000	0.0215
Poisson Normal	0	0.1304	0.0179	0.0208	0.0215
Poisson Gamma	0	0.1304	0.0179	0.0208	0.0215

Table 6.3: Placeholder

Method	k	min(POD)	max(POD)	mean(POD)	median(POD)	sd(POD)
Farrington	2	0.0000	0.6364	0.2334	0.2361	0.1361
	4	0.1250	0.8333	0.5071	0.5227	0.1925
	6	0.2857	1.0000	0.7657	0.7889	0.2113
	8	0.5385	1.0000	0.8797	0.9045	0.1371
	10	0.5000	1.0000	0.9277	1.0000	0.1297
Noufaily	2	0.0000	0.5455	0.1603	0.1484	0.1072
	4	0.0000	0.8000	0.3398	0.3167	0.1567
	6	0.0000	1.0000	0.5534	0.5556	0.2665
	8	0.3636	1.0000	0.7797	0.8000	0.1872
	10	0.0000	1.0000	0.8579	1.0000	0.2321
'Poisson Normal'	2	0.0000	0.5000	0.2348	0.2361	0.1191
	4	0.0000	0.8571	0.4691	0.5000	0.2178
	6	0.2000	1.0000	0.7838	0.8258	0.2180
	8	0.4444	1.0000	0.8683	1.0000	0.2025
	10	0.5000	1.0000	0.9266	1.0000	0.1304
'Poisson Gamma'	2	0.0000	0.5000	0.2348	0.2361	0.1191
	4	0.0000	0.8571	0.4691	0.5000	0.2178
	6	0.2000	1.0000	0.7838	0.8258	0.2180
	8	0.4444	1.0000	0.8683	1.0000	0.2025
	10	0.5000	1.0000	0.9266	1.0000	0.1304



# 7 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.

## 7.1 Harnessing the potential of MiBa

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 health-care 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.

## 7.2 Potential extensions of the novel algorithm

One potential extension of the novel outbreak detection method proposed in this master's thesis is the incorporation of correlation, either in time or between groups.

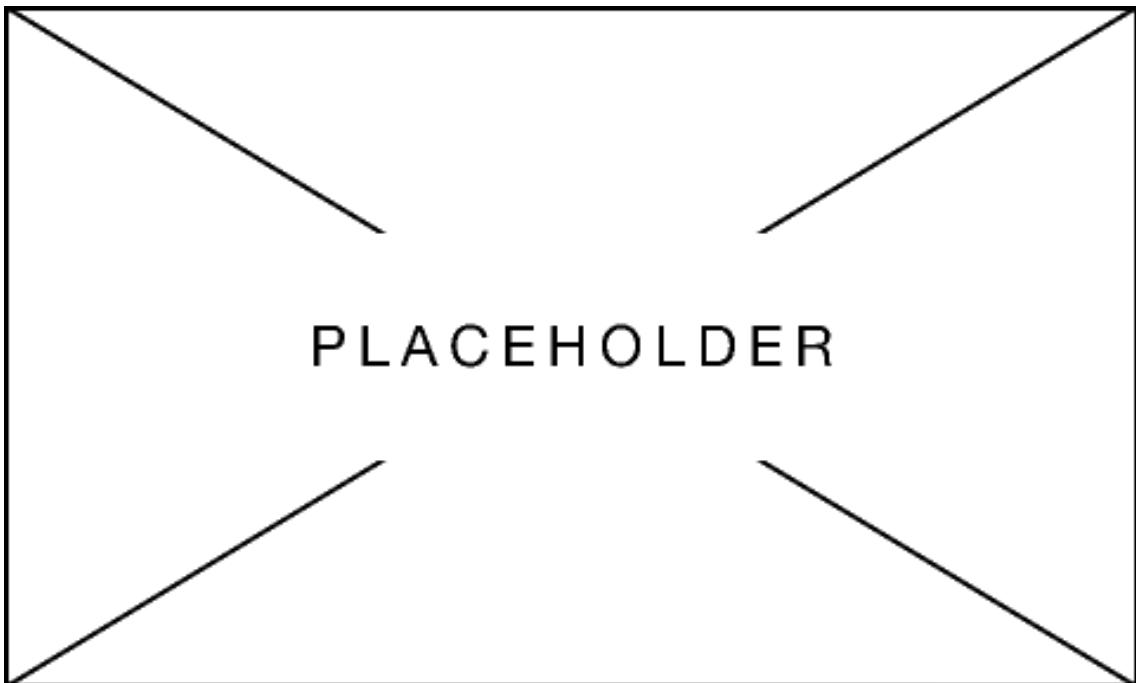


Figure 7.1: Maybe include an ACF plot of one of the disease investigated in this masters thesis, to highlight the correlation in time.

This extension can be readily implemented within the framework of the hierarchical Poisson Normal model. However, the implementation in the hierarchical Poisson Gamma model is more complex and requires further investigation.

## 8 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.
- Tulchinsky, Theodore H. (2018). "Chapter 5 - John Snow, Cholera, the Broad Street Pump; Waterborne Diseases Then and Now". In: *Case Studies in Public Health*. Ed. by Theodore H. Tulchinsky. Academic Press, pp. 77–99. ISBN: 978-0-12-804571-8. DOI: <https://doi.org/10.1016/B978-0-12-804571-8.00017-2>. URL: <https://www.sciencedirect.com/science/article/pii/B9780128045718000172>.
- Honardoost, Maryam, Azam Rajabpour, and Ladan Vakil (2018). "Molecular epidemiology; New but impressive". In: *Medical Journal of the Islamic Republic of Iran* 32.1. Cited by: 4; All Open Access, Gold Open Access, Green Open Access. DOI: 10.14196/mjiri.32.53. URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065203514&doi=10.14196%2fmjiri.32.53&partnerID=40&md5=31ba202c2698dce55f73f5bfb6efb1a7>.
- Struelens, M J and S Brisse (2013). "From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases". In: *Eurosurveillance* 18.4, 20386. DOI: <https://doi.org/10.2807/ese.18.04.20386-en>. URL: <https://www.eurosurveillance.org/content/10.2807/ese.18.04.20386-en>.
- Koeser, Claudio U. et al. (June 2012). "Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak". In: *NEW ENGLAND JOURNAL OF MEDICINE* 366.24, pp. 2267–2275. ISSN: 0028-4793.
- Baldry, Susannah (2010). "Attack of the clones". In: *Nature Reviews Microbiology* 8.6. Cited by: 9; All Open Access, Bronze Open Access, p. 390. DOI: 10.1038/nrmicro2369. URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-77952857458&doi=10.1038%2fnrmicro2369&partnerID=40&md5=228603d5ee11cdcdf04bea3f5de7c9a8>.
- 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).
- 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>.
- 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>.

- Bédubourg, Gabriel and Yann Le Strat (2017). "Evaluation and comparison of statistical methods for early temporal detection of outbreaks: A simulation-based study". In: *PLoS ONE* 12.7. Cited by: 27; All Open Access, Gold Open Access, Green Open Access. DOI: 10.1371/journal.pone.0181227. URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85024904807&doi=10.1371%2fjournal.pone.0181227&partnerID=40&md5=0a18530ff75e9efd8643ca4a743097f7>.
- Bolker, Benjamin M. et al. (2009). "Generalized linear mixed models: a practical guide for ecology and evolution". In: *Trends in Ecology & Evolution* 24.3, pp. 127–135. ISSN: 0169-5347. doi: <https://doi.org/10.1016/j.tree.2008.10.008>. URL: <https://www.sciencedirect.com/science/article/pii/S0169534709000196>.
- Zuur, Alain F. et al. (2009). *Mixed Effects Models and Extensions in Ecology with R*. English. Statistics for Biology and Health. Springer New York, NY. ISBN: 9780387874586.
- Palmer Real, Jaume et al. (2022). "A data-driven framework for characterising building archetypes: A mixed effects modelling approach". English. In: *Energy* 254. ISSN: 0360-5442. doi: 10.1016/j.energy.2022.124278.
- Real, J. Palmer et al. (2021). "Simulating heat load profiles in buildings using mixed effects models". English. In: vol. 2069. 1. IOP Publishing. doi: 10.1088/1742-6596/2069/1/012138.
- Heisterkamp, Simon H., Arnold L. M. Dekkers, and Janneke C. M. Heijne (2006). "Automated detection of infectious disease outbreaks: hierarchical time series models". In: *Statistics in Medicine* 25.24, pp. 4179–4196. doi: <https://doi.org/10.1002/sim.2674>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2674>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2674>.
- Voldstedlund, M et al. (2014). "The Danish Microbiology Database (MiBa) 2010 to 2013". In: *Eurosurveillance* 19.1, 20667. DOI: <https://doi.org/10.2807/1560-7917.ES2014.19.1.20667>. URL: <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES2014.19.1.20667>.
- 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.

- Smith, B. et al. (2011). "Outbreak of listeriosis caused by infected beef meat from a meals-on-wheels delivery in Denmark 2009". In: *Clinical Microbiology and Infection* 17.1, pp. 50–52. ISSN: 1198-743X. DOI: <https://doi.org/10.1111/j.1469-0691.2010.03200.x>. URL: <https://www.sciencedirect.com/science/article/pii/S1198743X14609125>.
- Wingstrand, Anne et al., eds. (2015). *Annual Report on Zoonoses in Denmark 2014*. English. Annual Report on Zoonoses in Denmark. DTU Food.
- Kvistholm Jensen, Anne et al. (Mar. 2016). "Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014". In: *Clinical Infectious Diseases* 63.1, pp. 64–70. ISSN: 1058-4838. DOI: 10.1093/cid/ciw192. eprint: <https://academic.oup.com/cid/article-pdf/63/1/64/6711273/ciw192.pdf>. URL: <https://doi.org/10.1093/cid/ciw192>.
- Takeuchi-Storm, Nao et al. (2023). "Presence and Persistence of *Listeria monocytogenes* in the Danish Ready-to-Eat Food Production Environment". In: *Hygiene* 3.1, pp. 18–32. ISSN: 2673-947X. DOI: 10.3390/hygiene3010004. URL: <https://www.mdpi.com/2673-947X/3/1/4>.
- Schjorring, Susanne et al. (Dec. 2017). "Cross-border outbreak of listeriosis caused by cold-smoked salmon, revealed by integrated surveillance and whole genome sequencing (WGS), Denmark and France, 2015 to 2017". In: *EUROSURVEILLANCE* 22.50, pp. 8–12. ISSN: 1025-496X. DOI: 10.2807/1560-7917.ES.2017.22.50.17-00762.
- 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>.
- Helwigh, Birgitte and Luise Müller, eds. (2018). *Annual Report on Zoonoses in Denmark 2017*. English. Annual Report on Zoonoses in Denmark. Technical University of Denmark.
- Helwigh, Birgitte, Channie Kahl Petersen, and Mia Torpdahl, eds. (2020). *Annual Report on Zoonoses in Denmark 2019*. English. Annual Report on Zoonoses in Denmark. Technical University of Denmark.
- Lassen, Brian et al., eds. (2023). *Annual Report on Zoonoses in Denmark 2022*. English. Annual Report on Zoonoses in Denmark. DTU Food.
- Lewis, H. C. et al. (2009). "Outbreaks of *Shigella sonnei* infections in Denmark and Australia linked to consumption of imported raw baby corn". In: *Epidemiology and Infection* 137.3, pp. 326–334. DOI: 10.1017/S0950268808001829.
- Muller, L. et al. (June 2009). "Imported fresh sugar peas as suspected source of an outbreak of *Shigella sonnei* in Denmark, April – May 2009". In: *Eurosurveillance* 14.24. ISSN: 1560-7917.
- 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>.
- Helwigh, Birgitte, Julia Christensen, and Luise Müller, eds. (2016). *Annual report on zoonoses in Denmark 2015*. English. Annual Report on Zoonoses in Denmark. National Food Institute, Technical University of Denmark.

- 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](https://doi.org/10.1007/978-1-4612-0919-5_24).
- Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378. doi: 10.1198/016214506000001437. eprint: <https://doi.org/10.1198/016214506000001437>. URL: <https://doi.org/10.1198/016214506000001437>.

# 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	$0, 1, 2, \dots$			
$\text{Pois}(\lambda)$	$\lambda \in \mathbb{R}_+$	$\frac{\lambda^y}{y!} \exp(-\lambda)$	$\lambda$	$\lambda$
Gamma	$\mathbb{R}_+$			
$G(\alpha, \beta)$	$\alpha \in \mathbb{R}_+, \beta \in \mathbb{R}_+$	$\frac{1}{\Gamma(\alpha)\beta} \left(\frac{y}{\beta}\right)^{\alpha-1} \exp(-y/\beta)$	$\alpha\beta$	$\alpha\beta^2$
Neg. Bin.	$0, 1, 2, \dots$			
$\text{NB}(r, p)$	$r \in \mathbb{R}_+, p \in ]0, 1]$	$\binom{r+y-1}{y} p^r (1-p)^y$	$\frac{r(1+p)}{p}$	$\frac{r(1-p)}{p^2}$
Normal	$\mathbb{R}$			
$N(\mu, \sigma^2)$	$\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$\mu$	$\sigma^2$

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 present the user template for the hierarchical Poisson Gamma model specified in (4.24).

```
#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;
}
```

# C State-of-the-art detection algorithm

## C.1 Controls

In the function `farringtonFlexible`, users can select either the original Farrington method or the improved method by Noufaily by specifying the appropriate `control` arguments. The choice of algorithm variant is determined by the contents of the `control` slot. In the example provided, `con.farrington` indicates the use of the original method, while `con.noufaily` represents the options for the improved method.

```
con.farrington <- list(
  range = NULL, b = 3, w = 2,
  reweight = TRUE, weightsThreshold = 1,
  verbose = TRUE, glmWarnings = TRUE,
  alpha = 0.05, trend = TRUE, pThresholdTrend = 0.05,
  limit54 = c(0,4), powertrans = "2/3",
  fitFun = "algo.farrington.fitGLM.flexible",
  populationOffset = TRUE,
  noPeriods = 1, pastWooksNotIncluded = NULL,
  thershodMethod = "delta"
)

con.noufaily <- list(
  range = NULL, b = 3, w = 2,
  reweight = TRUE, weightsThreshold = 2.58,
  verbose = TRUE, glmWarnings = TRUE,
  alpha = 0.05, trend = TRUE, pThresholdTrend = 0.05,
  limit54 = c(0,4), powertrans = "2/3",
  fitFun = "algo.farrington.fitGLM.flexible",
  populationOffset = TRUE,
  noPeriods = 1, pastWooksNotIncluded = NULL,
  thershodMethod = "Noufaily"
)
```

In this chapter an excerpt of the results

Table C.1: Longtable

Date	Age group	$y_t$	Farrington		Noufaily	
			Threshold	Alarm	Threshold	Alarm
2012-01-01	<65 years	2	3.407	FALSE	5.504	FALSE
	65+ years	6	4.345	TRUE	5.646	TRUE
2012-05-01	<65 years	1	3.594	FALSE	3.274	FALSE
	65+ years	4	2.444	TRUE	2.327	TRUE
2012-07-01	<65 years	0	4.424	FALSE	4.186	FALSE
	65+ years	5	2.715	TRUE	4.976	TRUE

Table C.1: (*continued*)

Date	Age group	$y_t$	Threshold	Alarm	Threshold	Alarm
<b>2013-09-01</b>	<65 years	1	2.749	FALSE	2.604	FALSE
	65+ years	7	6.321	TRUE	6.556	TRUE
<b>2013-10-01</b>	<65 years	0	4.384	FALSE	3.420	FALSE
	65+ years	6	4.551	TRUE	4.828	TRUE
<b>2014-01-01</b>	<65 years	0	3.920	FALSE	3.539	FALSE
	65+ years	8	5.530	TRUE	6.166	TRUE
<b>2014-07-01</b>	<65 years	6	3.093	TRUE	1.949	TRUE
	65+ years	14	10.223	TRUE	10.398	TRUE
<b>2014-08-01</b>	<65 years	6	3.093	TRUE	1.982	TRUE
	65+ years	14	11.690	TRUE	11.084	TRUE
<b>2014-09-01</b>	<65 years	2	3.061	FALSE	1.936	TRUE
	65+ years	9	6.705	TRUE	9.057	FALSE
<b>2015-02-01</b>	<65 years	2	2.918	FALSE	1.260	TRUE
	65+ years	3	6.405	FALSE	8.324	FALSE
<b>2016-04-01</b>	<65 years	3	2.474	TRUE	2.862	TRUE
	65+ years	1	5.990	FALSE	6.170	FALSE
<b>2017-01-01</b>	<65 years	0	2.228	FALSE	4.696	FALSE
	65+ years	8	6.005	TRUE	5.049	TRUE
<b>2017-04-01</b>	<65 years	1	2.440	FALSE	4.417	FALSE
	65+ years	4	2.789	TRUE	2.856	TRUE
<b>2017-07-01</b>	<65 years	3	4.541	FALSE	2.558	TRUE
	65+ years	4	3.736	TRUE	3.688	TRUE
<b>2017-08-01</b>	<65 years	2	4.542	FALSE	1.884	TRUE
	65+ years	4	4.599	FALSE	5.132	FALSE
<b>2017-09-01</b>	<65 years	2	4.541	FALSE	1.837	TRUE
	65+ years	4	4.572	FALSE	4.957	FALSE
<b>2018-02-01</b>	<65 years	3	2.618	TRUE	2.577	TRUE
	65+ years	1	6.040	FALSE	8.134	FALSE
<b>2018-06-01</b>	<65 years	0	3.569	FALSE	6.739	FALSE
	65+ years	6	5.935	TRUE	6.069	FALSE
<b>2019-02-01</b>	<65 years	0	2.622	FALSE	4.037	FALSE
	65+ years	7	5.962	TRUE	8.418	FALSE
<b>2019-09-01</b>	<65 years	2	3.446	FALSE	5.072	FALSE
	65+ years	7	6.686	TRUE	6.468	TRUE

Table C.1: (continued)

Date	Age group	$y_t$	Threshold	Alarm	Threshold	Alarm
<b>2021-05-01</b>	<65 years	5	1.905	TRUE	2.581	TRUE
	65+ years	1	5.209	FALSE	4.320	FALSE
<b>2021-07-01</b>	<65 years	0	2.802	FALSE	4.204	FALSE
	65+ years	8	7.258	TRUE	8.618	FALSE
<b>2021-12-01</b>	<65 years	1	2.595	FALSE	1.692	FALSE
	65+ years	5	4.979	TRUE	5.359	FALSE
<b>2022-04-01</b>	<65 years	1	2.759	FALSE	9.852	FALSE
	65+ years	5	5.776	FALSE	4.461	TRUE
<b>2022-05-01</b>	<65 years	6	2.714	TRUE	6.674	FALSE
	65+ years	9	5.562	TRUE	8.050	TRUE
<b>2022-06-01</b>	<65 years	3	2.957	TRUE	6.030	FALSE
	65+ years	11	6.700	TRUE	9.800	TRUE
<b>2022-10-01</b>	<65 years	5	3.864	TRUE	4.930	TRUE
	65+ years	7	8.298	FALSE	7.491	FALSE
<b>2022-11-01</b>	<65 years	3	3.481	FALSE	3.839	FALSE
	65+ years	11	7.213	TRUE	7.046	TRUE

# D Figures and Tables related to the case studies

## D.1 Listeriosis

Table D.1: The average logarithmic score,  $\bar{S}(G, y)$ , along with the parameter estimates at  $t_0$  for a suite of models modelling *Listeriosis*, assuming either the hierarchical Poisson Normal model or the hierarchical Poisson Gamma model. The confidence intervals for the estimates are calculated using profile likelihood confidence intervals.

$\bar{S}(G, y)$	Parameter	Estimate (95% CI)
<b>Poisson Normal</b>		
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \log(n_{it})$		
3.75	$\beta_{<65\text{years}}$	15.6 (15.24, 15.91)
	$\beta_{65+\text{years}}$	15.11 (14.83, 15.35)
	$\log(\sigma)$	-0.81 (-1.77, -0.35)
<b>Poisson Gamma</b>		
	$\beta_{<65\text{years}}$	15.7 (15.38, 16.01)
	$\beta_{65+\text{years}}$	15.21 (14.96, 15.45)
	$\log(\phi)$	-1.58 (-3.57, -0.67)

## D.2 Shigellosis

Table D.2: The average logarithmic score,  $\bar{S}(G, y)$ , along with the parameter estimates at  $t_0$  for a suite of models modelling *Shigellosis*, assuming either the hierarchical Poisson Normal model or the hierarchical Poisson Gamma model. The confidence intervals for the estimates are calculated using profile likelihood confidence intervals.

$\bar{S}(G, y)$	Parameter	Estimate (95% CI)
<b>Poisson Normal</b>		
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \log(n_{it})$		
5.297	$\beta_{<25\text{years}}$	14.5 (14.11, 14.84)
	$\beta_{25+\text{years}}$	16.5 (16.2, 16.76)
	$\log(\sigma)$	-0.54 (-0.98, -0.19)
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\text{trend}} t + \log(n_{it})$		
5.234	$\beta_{\text{trend}}$	0.03 (0.01, 0.05)
	$\beta_{<25\text{years}}$	15.01 (14.55, 15.44)
	$\beta_{25+\text{years}}$	17.01 (16.59, 17.4)
	$\log(\sigma)$	-0.6 (-1.06, -0.24)
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$		
5.278	$\beta_{<25\text{years}}$	14.51 (14.14, 14.83)
	$\beta_{25+\text{years}}$	16.5 (16.22, 16.73)

Table D.2: (continued)

Model	Formula	$\bar{S}(G, y)$	Parameter	Estimate (95% CI)
		NA		-0.34 (-0.62, -0.07)
		NA		0.32 (0.05, 0.6)
		$\log(\sigma)$		-0.78 (-1.45, -0.35)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{trend} + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$			
5.246		$\beta_{trend}$		0.03 (0.01, 0.05)
		$\beta_{<25\text{years}}$		15.01 (14.55, 15.43)
		$\beta_{25+\text{years}}$		16.99 (16.58, 17.37)
		NA		-0.13 (-0.41, 0.15)
		NA		0.28 (0.01, 0.56)
		$\log(\sigma)$		-0.71 (-1.27, -0.32)
<b>Poisson Gamma</b>				
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \log(n_{it})$			
5.302		$\beta_{<25\text{years}}$		14.66 (14.31, 15)
		$\beta_{25+\text{years}}$		16.67 (16.41, 16.94)
		$\log(\phi)$		-1.04 (-1.87, -0.39)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{trend} + \log(n_{it})$			
5.242		$\beta_{trend}$		0.02 (0.01, 0.04)
		$\beta_{<25\text{years}}$		15.11 (14.67, 15.53)
		$\beta_{25+\text{years}}$		17.11 (16.73, 17.5)
		$\log(\phi)$		-1.18 (-2.08, -0.5)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$			
5.266		$\beta_{<25\text{years}}$		14.61 (14.27, 14.93)
		$\beta_{25+\text{years}}$		16.6 (16.36, 16.84)
		NA		-0.35 (-0.62, -0.08)
		NA		0.33 (0.05, 0.61)
		$\log(\phi)$		-1.46 (-2.68, -0.67)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{trend} + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$			
5.253		$\beta_{trend}$		0.02 (0.01, 0.04)
		$\beta_{<25\text{years}}$		15.09 (14.67, 15.51)
		$\beta_{25+\text{years}}$		17.08 (16.7, 17.47)
		NA		-0.13 (-0.4, 0.15)
		NA		0.29 (0.01, 0.56)
		$\log(\phi)$		-1.38 (-2.49, -0.64)

### D.3 Shiga toxin (verotoxin)-producing *Escherichia coli*

Table D.3: The average logarithmic score,  $\bar{S}(G, y)$ , along with the parameter estimates at  $t_0$  for a suite of models modelling Shiga toxin (verotoxin)-producing *Escherichia coli*, assuming either the hierarchical Poisson Normal model or the hierarchical Poisson Gamma model. The confidence intervals for the estimates are calculated using profile likelihood confidence intervals.

$\bar{S}(G, y)$	Parameter	Estimate (95% CI)
<b>Poisson Normal</b>		
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \log(n_{it})$		
17.89	$\beta_{<1\text{year}}$	11.7 (11.41, 11.97)
	$\beta_{1-4\text{years}}$	13.88 (13.61, 14.13)
	$\beta_{5-14\text{years}}$	14.67 (14.42, 14.9)
	$\beta_{15-24\text{years}}$	15.01 (14.75, 15.26)
	$\beta_{25-64\text{years}}$	17.39 (17.15, 17.63)
	$\beta_{65+\text{years}}$	16.29 (16.03, 16.54)
	$\log(\sigma)$	-0.86 (-1.12, -0.63)
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\text{trend}} t + \log(n_{it})$		
14.87	$\beta_{\text{trend}}$	0.05 (0.04, 0.06)
	$\beta_{<1\text{year}}$	12.66 (12.38, 12.93)
	$\beta_{1-4\text{years}}$	15.33 (15.11, 15.54)
	$\beta_{5-14\text{years}}$	15.78 (15.55, 16.01)
	$\beta_{15-24\text{years}}$	16.3 (16.08, 16.52)
	$\beta_{25-64\text{years}}$	18.85 (18.65, 19.05)
	$\beta_{65+\text{years}}$	17.71 (17.51, 17.91)
	$\log(\sigma)$	-0.87 (-1.05, -0.71)
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$		
14.58	$\beta_{<1\text{year}}$	11.78 (11.48, 12.06)
	$\beta_{1-4\text{years}}$	14.41 (14.18, 14.64)
	$\beta_{5-14\text{years}}$	14.89 (14.64, 15.13)
	$\beta_{15-24\text{years}}$	15.42 (15.19, 15.64)
	$\beta_{25-64\text{years}}$	17.96 (17.76, 18.16)
	$\beta_{65+\text{years}}$	16.83 (16.62, 17.03)
	NA	-0.51 (-0.64, -0.38)
	NA	-0.1 (-0.23, 0.03)
	$\log(\sigma)$	-0.57 (-0.71, -0.42)
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\text{trend}} t + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$		
14.18	$\beta_{\text{trend}}$	0.04 (0.04, 0.05)
	$\beta_{<1\text{year}}$	12.56 (12.3, 12.81)
	$\beta_{1-4\text{years}}$	15.23 (15.03, 15.41)
	$\beta_{5-14\text{years}}$	15.68 (15.46, 15.88)
	$\beta_{15-24\text{years}}$	16.19 (15.99, 16.38)
	$\beta_{25-64\text{years}}$	18.74 (18.56, 18.9)
	$\beta_{65+\text{years}}$	17.6 (17.43, 17.77)
	NA	-0.34 (-0.44, -0.25)
	NA	-0.06 (-0.15, 0.04)
	$\log(\sigma)$	-1.12 (-1.33, -0.92)

Table D.3: (continued)

Model	Formula	$\bar{S}(G, y)$	Parameter	Estimate (95% CI)
<b>Poisson Gamma</b>				
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \log(n_{it})$			
	19.79		$\beta_{<1\text{year}}$	11.8 (11.53, 12.04)
			$\beta_{1-4\text{years}}$	13.84 (13.61, 14.08)
			$\beta_{5-14\text{years}}$	14.8 (14.6, 15)
			$\beta_{15-24\text{years}}$	15.12 (14.91, 15.33)
			$\beta_{25-64\text{years}}$	17.29 (17.07, 17.51)
			$\beta_{65+\text{years}}$	16.15 (15.91, 16.39)
			$\log(\phi)$	-2.46 (-3.89, -1.7)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\text{trend}} t + \log(n_{it})$			
	14.94		$\beta_{\text{trend}}$	0.05 (0.04, 0.06)
			$\beta_{<1\text{year}}$	12.74 (12.46, 13)
			$\beta_{1-4\text{years}}$	15.43 (15.22, 15.64)
			$\beta_{5-14\text{years}}$	15.86 (15.64, 16.09)
			$\beta_{15-24\text{years}}$	16.37 (16.16, 16.59)
			$\beta_{25-64\text{years}}$	18.91 (18.72, 19.11)
			$\beta_{65+\text{years}}$	17.76 (17.56, 17.96)
			$\log(\phi)$	-1.73 (-2.07, -1.41)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$			
	14.63		$\beta_{<1\text{year}}$	11.94 (11.66, 12.21)
			$\beta_{1-4\text{years}}$	14.61 (14.4, 14.83)
			$\beta_{5-14\text{years}}$	15.05 (14.82, 15.28)
			$\beta_{15-24\text{years}}$	15.54 (15.33, 15.77)
			$\beta_{25-64\text{years}}$	18.09 (17.9, 18.3)
			$\beta_{65+\text{years}}$	16.98 (16.79, 17.19)
			NA	-0.51 (-0.63, -0.38)
			NA	-0.07 (-0.21, 0.05)
			$\log(\phi)$	-1.15 (-1.42, -0.89)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\text{trend}} t + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$			
	14.15		$\beta_{\text{trend}}$	0.04 (0.04, 0.05)
			$\beta_{<1\text{year}}$	12.61 (12.34, 12.85)
			$\beta_{1-4\text{years}}$	15.28 (15.09, 15.47)
			$\beta_{5-14\text{years}}$	15.72 (15.51, 15.92)
			$\beta_{15-24\text{years}}$	16.23 (16.03, 16.42)
			$\beta_{25-64\text{years}}$	18.77 (18.6, 18.94)
			$\beta_{65+\text{years}}$	17.65 (17.47, 17.82)
			NA	-0.33 (-0.43, -0.24)
			NA	-0.04 (-0.14, 0.05)
			$\log(\phi)$	-2.21 (-2.63, -1.82)

## D.4 Salmonellosis

Table D.4: The average logarithmic score,  $\bar{S}(G, y)$ , along with the parameter estimates at  $t_0$  for a suite of models modelling *Listeriosis*, assuming either the hierarchical Poisson Normal model or the hierarchical Poisson Gamma model. The confidence intervals for the estimates are calculated using profile likelihood confidence intervals.

$\bar{S}(G, y)$	Parameter	Estimate (95% CI)
<b>Poisson Normal</b>		
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \log(n_{it})$		
21.43	$\beta_{<1\text{year}}$	11.78 (11.52, 12.02)
	$\beta_{1-4\text{years}}$	13.94 (13.75, 14.13)
	$\beta_{5-14\text{years}}$	14.94 (14.74, 15.12)
	$\beta_{15-24\text{years}}$	15.07 (14.87, 15.26)
	$\beta_{25-64\text{years}}$	17.95 (17.81, 18.08)
	$\beta_{65+\text{years}}$	16.65 (16.51, 16.79)
	$\log(\sigma)$	-1.12 (-1.35, -0.91)
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\text{trend}}t + \log(n_{it})$		
18.48	$\beta_{\text{trend}}$	0.02 (0.01, 0.03)
	$\beta_{<1\text{year}}$	12.12 (11.83, 12.39)
	$\beta_{1-4\text{years}}$	14.31 (14.07, 14.54)
	$\beta_{5-14\text{years}}$	15.38 (15.14, 15.6)
	$\beta_{15-24\text{years}}$	15.62 (15.39, 15.84)
	$\beta_{25-64\text{years}}$	18.36 (18.16, 18.56)
	$\beta_{65+\text{years}}$	16.97 (16.77, 17.18)
	$\log(\sigma)$	-0.88 (-1.05, -0.71)
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$		
18.40	$\beta_{<1\text{year}}$	11.83 (11.59, 12.06)
	$\beta_{1-4\text{years}}$	13.99 (13.81, 14.16)
	$\beta_{5-14\text{years}}$	14.96 (14.77, 15.13)
	$\beta_{15-24\text{years}}$	15.08 (14.89, 15.27)
	$\beta_{25-64\text{years}}$	17.99 (17.87, 18.11)
	$\beta_{65+\text{years}}$	16.64 (16.51, 16.76)
	NA	-0.26 (-0.35, -0.16)
	NA	-0.07 (-0.16, 0.02)
	$\log(\sigma)$	-1.31 (-1.61, -1.06)
$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\text{trend}}t + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$		
17.89	$\beta_{\text{trend}}$	0.01 (0.01, 0.02)
	$\beta_{<1\text{year}}$	12.01 (11.74, 12.28)
	$\beta_{1-4\text{years}}$	14.22 (13.99, 14.44)
	$\beta_{5-14\text{years}}$	15.28 (15.06, 15.49)
	$\beta_{15-24\text{years}}$	15.53 (15.31, 15.74)
	$\beta_{25-64\text{years}}$	18.26 (18.08, 18.44)
	$\beta_{65+\text{years}}$	16.87 (16.68, 17.06)
	NA	-0.28 (-0.38, -0.17)
	NA	-0.11 (-0.21, -0.01)
	$\log(\sigma)$	-1.03 (-1.23, -0.85)

### Poisson Gamma

Table D.4: (continued)

Model	Formula	$\bar{S}(G, y)$	Parameter	Estimate (95% CI)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \log(n_{it})$			
	20.53		$\beta_{<1\text{year}}$	11.83 (11.57, 12.08)
			$\beta_{1-4\text{years}}$	14.04 (13.84, 14.23)
			$\beta_{5-14\text{years}}$	14.99 (14.79, 15.19)
			$\beta_{15-24\text{years}}$	15.21 (15.02, 15.41)
			$\beta_{25-64\text{years}}$	18.14 (18, 18.29)
			$\beta_{65+\text{years}}$	16.72 (16.57, 16.87)
			$\log(\phi)$	0 (NA, 0.03)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{trend} t + \log(n_{it})$			
	18.60		$\beta_{trend}$	0.02 (0.01, 0.03)
			$\beta_{<1\text{year}}$	12.2 (11.92, 12.47)
			$\beta_{1-4\text{years}}$	14.4 (14.17, 14.63)
			$\beta_{5-14\text{years}}$	15.48 (15.26, 15.7)
			$\beta_{15-24\text{years}}$	15.73 (15.52, 15.95)
			$\beta_{25-64\text{years}}$	18.45 (18.25, 18.64)
			$\beta_{65+\text{years}}$	17.04 (16.84, 17.24)
			$\log(\phi)$	-1.76 (-2.1, -1.44)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$			
	18.41		$\beta_{<1\text{year}}$	11.87 (11.63, 12.09)
			$\beta_{1-4\text{years}}$	14.03 (13.85, 14.2)
			$\beta_{5-14\text{years}}$	15 (14.82, 15.17)
			$\beta_{15-24\text{years}}$	15.12 (14.93, 15.3)
			$\beta_{25-64\text{years}}$	18.03 (17.91, 18.15)
			$\beta_{65+\text{years}}$	16.67 (16.54, 16.8)
			NA	-0.25 (-0.34, -0.16)
			NA	-0.07 (-0.16, 0.02)
			$\log(\phi)$	-2.61 (-3.2, -2.13)
	$\log(\lambda_{it}) = \beta(\text{ageGroup}_i) + \beta_{trend} t + \beta_{\sin} \sin\left(\frac{\pi \cdot \tau_t}{6}\right) + \beta_{\cos} \cos\left(\frac{\pi \cdot \tau_t}{6}\right) + \log(n_{it})$			
	17.95		$\beta_{trend}$	0.01 (0.01, 0.02)
			$\beta_{<1\text{year}}$	12.07 (11.8, 12.33)
			$\beta_{1-4\text{years}}$	14.28 (14.06, 14.5)
			$\beta_{5-14\text{years}}$	15.34 (15.13, 15.55)
			$\beta_{15-24\text{years}}$	15.6 (15.4, 15.8)
			$\beta_{25-64\text{years}}$	18.3 (18.13, 18.49)
			$\beta_{65+\text{years}}$	16.92 (16.74, 17.11)
			NA	-0.27 (-0.37, -0.17)
			NA	-0.1 (-0.2, -0.01)
			$\log(\phi)$	-2.07 (-2.48, -1.7)

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](http://www.compute.dtu.dk)