LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Juliane Manitz

Automated Detection of Infectious Disease Outbreaks

Diploma Thesis (Corrected and revised version)

Department of Statistics, Ludwig-Maximilians-Universität München

Supervisor: PD Michael Höhle, Ph.D.
Closing Date: 9^{th} August 2010

Automated Detection of

Infectious Disease Outbreaks

Diploma Thesis
by
Juliane Manitz

Corrected and revised version (22. November 2010)

Department of Statistics, Ludwig-Maximilians-Universität München

Supervisor: PD Michael Hohle, Ph.D.
Closing Date: 9^{th} August 2010

Hiermit erkläre ich, dass ich die Diplomarbeit selbstständig angefertigt und nur die angegebenen Quellen verwendet habe.

Juliane Manitz

München, 6. August 2010

Acknowledgement

Today, the day has come, I will complete my Diploma thesis. I am deeply grateful for the support of many people who helped me during these last months while I worked on this project. A special thank you to....

... Michael Höhle for his excellent supervision, his accurate proofreading, his support, his commitment and his patience. I also want to thank him for his R package developement `surveillance` which is a great framework to automated outbreak detection methods.

... the Robert Koch-Institute, especially Dr. Hermann Claus and Doris Altmann, for providing the Campylobacter data and their great support regarding content-related issues.

... Professor Håvard Rue and Birgit Schrödle for several helpful emails on the implementation with INLA.

... Andreas Mayr, Andrea Wiencierz, and Lucie Wink for proofreading and their fair comments.

... Gunna Höwing who helped me to cope with the thesis and printed it

... my friends, especially Marcus Scherl, and my family for their constant support, encouragement, and bearing my daily statement 'Today, I will complete my Diploma thesis!'.

Juliane Manitz
Munich, August 6, 2010

Contents

Chapter 1

Introduction

It's always an uncertainty. We're always at the infectious disease roulette table .

Dr. William Schaffner (Anonymous, 2010)

William Schaffner is consultant for, e.g. the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO). The main task of a statistician and this thesis is the handling of the uncertainty, he speaks about, and find pattern to forecast how things will happen.

1.1 Presentation of Problem and Interest of Research

The most well known pandemic in history is the plague which appeared in different waves and caused high proportions of deaths in Europe. Between the middle of the 14^{th} until the beginning of the 18^{th} century the plague has been the infection par excellence. The Western World understood the infection as a punishment of God (see Bulst, 1989). Because religious and secular departments worked out defensive measures in close cooperation, they advised people to attend church services as "first aid'. Other popular misconceptions appeared such as the plague being transmitted by bad breathing air. In the middle of 16^{th} century a variety of defensive measures originated like controlling the access to towns, quarantine of infected, destroying belonging of the dead, but the displacement of marginalized groups as well. Finally, with John Snow (1813–1858) and his investigation on the 1854 cholera outbreak in London, modern epidemiology was developed. He explored the spread of the disease systematically and detected quality of water source as its root. This study came to be a major break-through in the history of public health.

This thesis will investigate automated outbreak detection of infectious diseases. Thereby, outbreaks, which means an unusual high incidence, should be identified early to control and prevent their spread.

Today, infectious diseases are studied in a more advanced manner compared to the time of the plague. Usually, it is known how infections are transmitted. In general, societies have a basic infection protection by hygienically standards, and in case of an outbreak, efficient ways to prevent the spread of an infectious disease are known.

Institutions for public health are established to monitor infectious diseases. In general, they combine risk research with political advice. In Germany, through the Law for the Prevention of Infection (Infektionsschutzgesetz, IfSG), the Robert-Koch-Institute (RKI) was given the responsibilities of a federal epidemiological centre for infectious diseases. The RKI cooperates with various public health institutions in Germany and around the world, including the World Health Organization (WHO). Guided by the United Nations, its main task is the combat of diseases with focus on infectious kind.

Thereby, the public health organizations have to work hand in glove with various scientists. Physicians, who need to be alert and detect the disease are important. Afterwards, statisticians have to detect a significantly increased number of incidences as quickly as possible and define time points of alarms. And finally, public health scientists need to investigate these alarms, while public health politicians have to initiate appropriate ways to prevent the spread of the disease.

1.2 Research Question

In cooperation with the RKI, automated outbreak detection of infectious diseases based on routine surveillance data will be studied. The particular interest lies on the hierarchical time series models approach of Heisterkamp et al. (2006). Its description, implementation in the R package `surveillance` (Höhle, 2007), evaluation, and application on RKI's Campylobacter data are the main tasks of this work. Furthermore, covariate progresses will be integrated into the outbreak detection. The potential of hierarchical time series models in comparison with the established Farrington algorithm (Farrington et al., 1996) or other public health surveillance algorithms will be figured out.

1.3 Introducing Issues to Automated Outbreak Detection

Detection of infectious disease outbreaks belongs to the area of surveillance, where time series of disease counts are monitored and change points or outbreaks should be detected. Thereby, the surveillance methods are not only applied in the context of public health. The statistical methods are applied, e.g. in other medical areas, in economics, environmental control, as well. Thus, different terminology for statistical surveillance developed such as optimal stopping, change point stopping, early warning system, statistical process control, quality control, etc. Each terminology emphasize different issues corresponding to the area of application (see Frisén, 2003).

1.3.1 What is an Outbreak?

An outbreak does not have an unique definition. As an technical epidemiological description an outbreak is defined by linked cases of a certain definition. From the public health's point of view an outbreak occurs if more cases than expected are recognized (Farrington, 2010). There are other definitions as well, but in the following I will use the later definition.

It is distinguished between different types of outbreaks, which will be explained especially in the context of food borne diseases. A traditional food borne outbreak is highly local with usually a high attack rate in the group which is exposed to the source. For example, in 2005, there has been an outbreak of Campylobacter infections on a school excursion due to raw milk in Landsberg am Lech, Germany (Anonymous, 2006). This kind of outbreak is usually identified early by the local public health authority (Gesundheitsamt).

Another type of outbreak is emerging as the world is getting smaller. Since persons and food can travel more easily and faster around the world, linked cases might be widespread and involve many countries. An example is the Salmonella outbreak, caused by a kosher snack, detected in the United Kingdom in February 2005 (Farrington et al., 1996). A non food-borne example is the Swine influenza in 2009.

Moreover, there can be an outbreak by the introduction of a new pathogen into a new geographic area, such as tropical disease outbreaks are expected in European countries due to the climate change. For example, a Chikungunya outbreak, a disease which is transmitted by mosquitos, has been recognized in Ravenna, Italy, in summer 2007 (Stark, 2009).

1.3.2 Which Statistical Methods are Used?

As the range of application areas, the variety of statistical methods for surveillance is wide. They can be summarized in two groups: Methods based on reference values, and the framework of statistical process control. In the context of this thesis, it will be focused on methods which construct a reference set with values of comparable time points, e.g. the observed number of infected in the corresponding 18 weeks in the last two years.

Special attention will be given to a simple system used at the RKI, the Farrington algorithm which is based on a generalized linear model, the Bayes algorithm using the predictive posterior, and the hierarchical time series algorithm which includes all past values as an enhancement of the Farrington algorithm. Furthermore, a full Bayesian approach will be developed on basis of this algorithm.

1.3.3 Which Software can be Used?

For practical application different statistical software such as R, SAS, or Stata can be used. In this thesis, all analyses are carried out using the statistical programming environment R (R Developer Core Team, 2009) in its version 2.9.2. This software and every add-on package used in this thesis is available at `http://www.cran.r-project.org`. Especially, the add-on package `surveillance` and its application will be introduced which provides a wide range of statistical methods for automated outbreak detection of infectious diseases. Its environment for developers of new algorithms will be used for implementation of the originated Bayesian hierarchical time series algorithm.

1.4 Surveillance Data

To control and prevent diseases, large amounts of data about various diseases and health events are collected, not only in various nationwide public health programmes such as the RKI in Germany, but in international organisations such as the European Center for Disease Prevention and Control (ECDC) in the European Union or the WHO in the United Nations (UNO) as well.

This thesis comes into being in cooperation with the RKI, Germany, who provided reports on Campylobacter infections between 2001 and 2009 in Germany (Robert Koch-Institute, 2010). Its time series is introduced in Figure 1.1. The on-going data collection is determined by the IfSG. Campylobacter is the most common cause of diarrhoeal diseases in Germany and many other industrial states. Food borne diseases concern governments and food industry today more than decades ago due to the increasing number of reported outbreaks and the impact on children, ageing population and immune compromised. Since the ease of worldwide shipment of fresh and frozen food, migrant population demanding their traditional food in the country of settlement, and the development of new food industries including aquaculture, more world-wide associated and therefore difficult to detect outbreaks appear (World Health Organization, 2007).

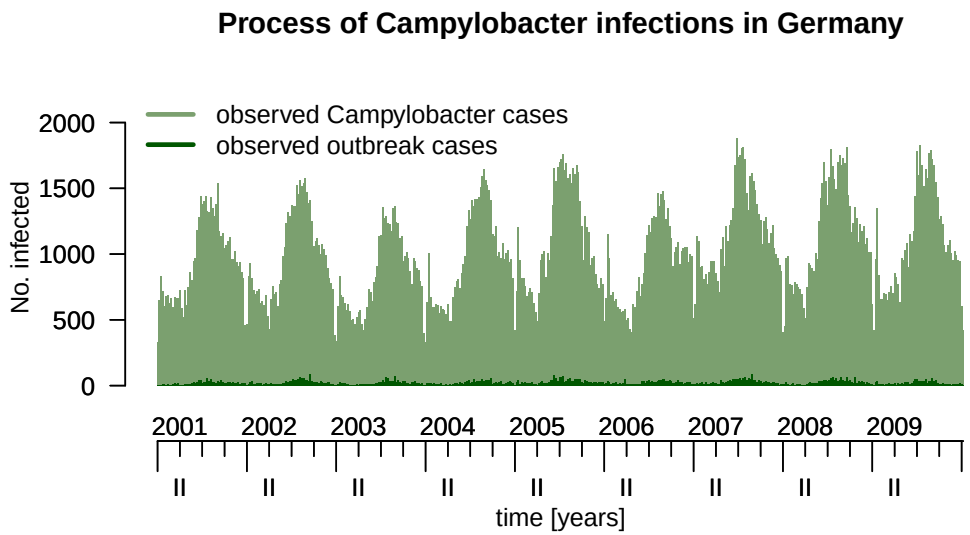

Figure 1.1: Time series of weekly Campylobacter data and of outbreak cases counts

Using the reports of Campylobacter infections automated outbreak detection algorithms under influence of weather parameters will be evaluated. This relation is taken into consideration by different theoretical and practical investigations. In the RKI bulletin reasons for the high incidence of Campylobacter infection in 2007 was discussed (Jansen et al., 2007). It is believed that the comparative warm spring resulted in different leisure activities such as picnics and barbecues. This fact, in combination with a higher proportion of contamination in chicken meat are assumed

to be reasonable. A greater incidence in chicken farms by higher outdoor temperature is already proofed by different studies (see Jansen et al., 2007). The direct influence of specific weather on Campylobacter infection of chicken could not be investigated.

In an internship during summer 2008 at the RKI, the time series of Campylobacter incidence by weather parameters was modelled (an der Heiden et al., 2010). It was revealed that temperature and humidity of the past weeks are able to explain the time series of infection cases. For the evaluation of the described algorithms the time period between 2002 and 2006 will be used for modelling and defining a rule. The years 2007 until 2009 are monitored. Therefore, it is necessary to flag known outbreaks. As usually in real-world data not all outbreaks are known. Using a variable of the data frame, which identifies cases belonging to an outbreak, an indicator for the state of outbreak will be defined and investigated.

1.5 Overview of Strategy and Outline

This thesis is organized as follows. The subject of Campylobacter disease and reporting system will be introduced, followed by a descriptive analysis of the RKI's Campylobacter count time series and description of the weather parameters in the second chapter. It will proceed with the basic statistical theory for surveillance methods by representing general statements on surveillance and describing the Farrington algorithm in chapter 3. Chapter 4 contains description, enhancement and implementation of the Heisterkamp algorithm. To explore the potential and for evaluation of the Heisterkamp algorithm there will be chapter 5 including simulation studies, and chapter 6 with an application of the presented algorithms on the Campylobacter data. Finally, the findings will be discussed in chapter 7.

For structuring this thesis, special environments for *Example* and *Excursus* are applied. Using them ancillary information which is not of main importance for the statistical chain of reasoning is provided for the interested reader. The end of every section is indicated by the diamond symbol, i.e. $\diamond$. Thereby, an excursus gives further information of subject-related issues. Throughout, a very short summary in given in the surrounding text, so that the reader could easily skip the part. In each chapter, a worked through example is used to illustrate the statistical theory. In this context the special environment is reopened.

A common notation is used. Bold symbols denote vectors or matrices, e.g. $\boldsymbol{x} = (x_1, \ldots, x_n)'$. Furthermore, different typefaces are used. `Variables` in a data set and `programming code` are emphasized by typescript. *Quotations* are printed in cursive characters.

Chapter 2

Introduction to Campylobacter and the Data

In this chapter, the reporting data of Campylobacter infections is described. Beginning with an introduction to Campylobacter with its medical aspects, a basic understanding of the disease is given. Afterwards, the German system of data collection based on the Law for the Prevention of Infection (Infektionsschutzgesetz, IfSG), is described. This is important to understand the structure and irregularities in the data. Furthermore, a descriptive analysis of the Campylobacter data is given, and the relevant weather data are shown. After reading this chapter, one should have received an impression of public health surveillance data and its specialities in general, and particular for Campylobacter infection in Germany.

2.1 Medical Issues to Campylobacter

In this section, the subject of Campylobacteriosis disease, due to infection with the Campylobacter bacterium, and its prevention is introduced.

History

The bacterium Campylobacter was identified first in 1913 by Mc Faydean and Stockman in fetal tissue of aborted sheep. As a human diarrhoeal pathogen Campylobacter was identified much later, in 1972 (see Altekruse et al., 1999). Hence, more and more laboratories tested faecal specimen for Campylobacter and soon it became one of the most common causes of diarrhoeal diseases. Today Campylobacter is considered to be the leading cause of enteritis illness in Germany and other industrial countries (see Levin, 2007; Robert Koch-Institut, 2005). In general, only a fraction is reported. US-American studies estimated about 1% of the population in the United States and Europe is affected with Campylobacter each year (see Louis et al., 2005).

In 2005, the World Health Organization recognized that about 1.8 million people died from diarrhoeal diseases. Furthermore, the consequences of food contamination create an enormous social and economic burden on communities and their health systems (see World Health Organization, 2007). Several studies estimated the annual

costs of food borne diseases due to medical treatment and lost of productivity (see Buzby and Roberts, 2009).

Microbiological Bacterium

The Campylobacter bacterium is a small, spiral-shaped rod (see Figure 2.1) exhibiting corkscrew motility (Levin, 2007). It has high demands for reproduction. Thus, it is highly sensitive to a lot of environmental influences, e.g. it needs humidity and a special atmosphere with specific concentrations of Carbon dioxide and oxygen. The Campylobacter bacterium has optimum growth conditions at temperatures of 42–43°C. It is unable to grow below 30°C and does not survive temperatures over 55°C (see Levin, 2007).

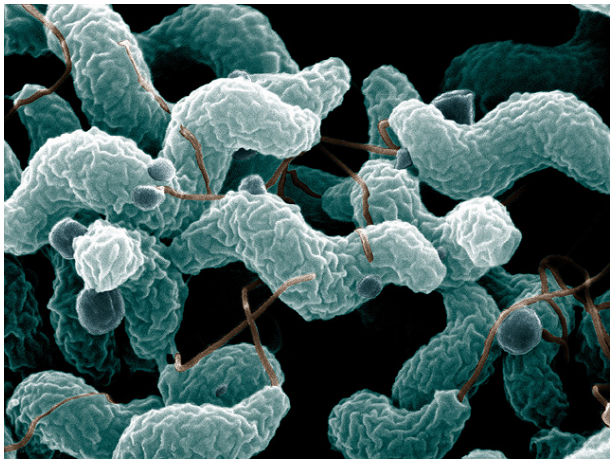

Figure 2.1: Scanning electron microscope image of Campylobacter jejuni (Source: Wood and Pooley, 2007)

Twenty species of Campylobacter bacterium can be identified where Campylobacter jejuni, coli, and lari are the most common. The organism is omnipresent and its initial infectious reservoir is not surely explored. It is carried in the intestine of many wild and domestic animals and is present in untreated surface waters. Many authors assume that wild animals as suppositious infectious reservoir. Others consider the water borne route of infection as the common factor linking infections in humans and animals (Levin, 2007). Campylobacter jejuni is mostly found in poultry while Campylobacter coli is mostly found in pig meat.

In 99% of human Campylobacteriosis the associated bacteria are jejuni or coli (Louis et al., 2005). Identification of Campylobacter species is based on biochemical tests, antibiotic resistance patterns, and growth temperatures (Levin, 2007).

Campylobacteriosis

The described bacterium causes the world-wide appearing of the infectious disease Campylobacteriosis. In industrial countries, especially children under the age of five

years are exposed which is usual for enteritis diseases. Furthermore, young adults between 20 and 29 years have higher risk. The infection is caused by the described Campylobacter bacteria of a low dose greater than 500 germs. The incubation time varies between two and five days, in special cases it is one up to ten days (Robert Koch-Institut, 2005).

Many infections are asymptomatic. If an infection is apparent, the symptoms are acute enteritis, fever, headache, myalgia, anthralia, diarrhoea, abdominal ache and cramps, and tiredness. In general, the disease lasts no longer than one week and is self-limiting. Therefore, a symptomatic therapy is usually adequate. In a severe case a therapy with antibiotics is applied. Complications are rare, but a chronic progress in cases of immune weakened persons is possible.

Transmission

Campylobacter bacteria are transmitted from animals to humans, while direct transmission by close contact with infected animals is rare. Animals usually do not get ill. Nevertheless, there is an especially high contamination in their excrements. While butchering or milking the animals food can get contaminated. Already a low number of bacteria can cause an infection. Campylobacter does not spoil the food, so one cannot recognize the bacteria by appearance or smell (Bundesinstitut für Risikobewertung, 2009). Infection is possible by direct transmission or indirect by other food, people's hands, kitchen utensils or at the work space (see Table 2.1).

exposure	odds ratio	95% confidence interval
contact with raw meat	9.37	[2.03 ; 43.3]
having a pet with diarrhoea	2.39	[1.09 ; 5.25]
ingesting untreated water from rivers etc.	4.16	[1.45 ; 11.9]
consumption/handling of chicken with giblets eaten at home	0.44	[0.24 ; 0.79]
contact with animal faeces	0.44	[0.21, 0.92]

Table 2.1: Significant risk factors for Campylobacteriosis determined by conditional logistic regression analysis (Source: Adak et al., 1995)

Campylobacter can be transmitted as well by drinking surface water during water sport activities. A direct person-to-person transmission is possible with children. An infected person releases between two and four weeks excretions of germs, which are potentially infectious.

Prevention

As learned before, the bacteria are in general transmitted by contaminated food. The initial prevention is the reduction of contamination in poultry butcheries, as well as strict obedient butchering hygiene.

The consumer prevents Campylobacteriosis by consequent hygiene in the kitchen. The bacteria can be destroyed by heating the food at least two minutes at 70°C

(Robert Koch-Institut, 2005), but indirect transmission is common as well. Therefore, high attention when preparing raw poultry meat is necessary. Freezing decreases contamination with Campylobacter bacteria, but does not destroy it sufficiently.

2.2 The Reporting System in Germany

In this section, the background of Campylobacter infection data collection in Germany is presented. The data gathered by the Robert-Koch-Institute in Germany are part of the system for compulsory notifiable infectious diseases. With the Law for the Prevention of Infection of 2001 by the German government, an expert-based registration system for infectious diseases was implemented and with this new tools for data generation, prevention, surveillance and research were created.

Law for the Prevention of Infection

The Law for the Prevention of Infection (Infektionsschutzgesetz, IfSG) replaced the Federal Pandemic Disease Law (Bundes-Seuchengesetz, BSeuchG) on the 1^{st} of January 2001. It provided a new basis for public health surveillance data collection. It was adopted to create a statutory basis for cooperation of all parties concerned, and has the aim of developing an overview of infections and disease dynamics reflecting reality as closely as possible. On this basis, the target of preventing and keeping infectious diseases at bay should be attained (Forßbohm, 2000). The Law regulates exact content, persons of responsibility and flow of a report.

A report includes the disease and its characteristics. The list of notifiable diseases was selected in consideration of the illness dangerousness, the need of an immediate reaction by the public health department, and the importance of the disease as indicator to reveal hygiene faults. The resulting list with 18 infectious diseases is briefer than the pre-existing one due to the BSeuchG. In general, even suspected cases need to be reported. Gastroenteritis, which includes Campylobacteriosis, constitute an exception where suspicions are only reported if the case belongs to an outbreak or the patient is working in the food industry. The report of only already confirmed suspicions should reduce the large number of gastroenteritis cases to a manageable level.

The Flow of Reporting

The Law for the Prevention of Infection describes the flow and sets deadlines to accelerate the reporting flow. The reporting system itself is well-structured. It is a sophisticated system with four levels of reporting, mapping the federal structure of Germany: A doctor or person obliged to report, the local health department (Gesundheitsamt), the state health department (Landesgesundheitsamt), and the Robert Koch-Institute. The flow involves exchanges, feedbacks and inquiries between the institutions (see Figure 2.2). The SurvNet software of the RKI (Robert Koch-Institute, 2010) organizes the electronic transmission of case-based datasets between the different departments.

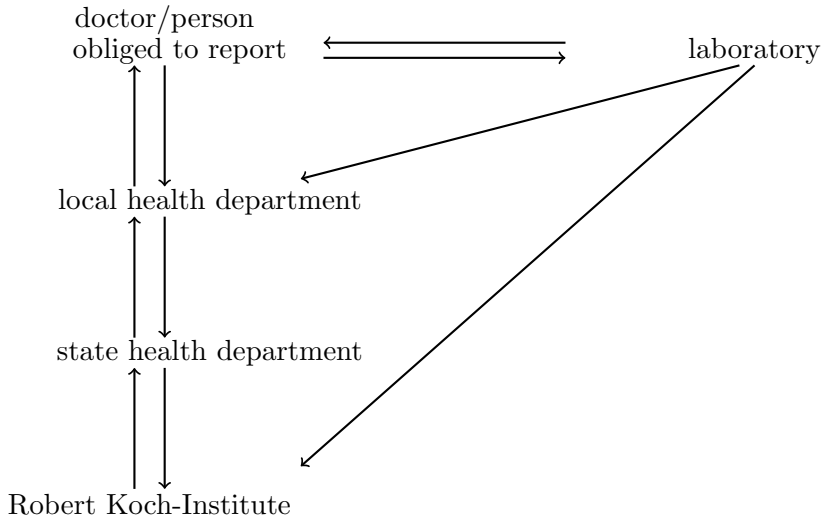

Figure 2.2: Systematic structure of reporting proceedings (Robert Koch-Institut (2000))

If a doctor or laboratory is having suspicion of a notifiable infectious disease it has to be reported within 24 hours. This first report includes the basic information about the patient, infection chain, possibilities of spreading, consulted laboratories for diagnosis, and hospital of treatment.

This information is given to the responsible local health department. In Germany, there are 412 of those departments, one in every district. They are responsible to investigate the cases and to complete further characteristics such as type, cause, infection source, and spreading of the disease. For this research, the local health department may consult the doctor, the laboratory of diagnosis, and the patient. If necessary, the institution may already take protective measures. Finally, until the third working day of the following week, a data set of 20 characteristics is formed and reported to the responsible state health department.

The responsible state health department merges the reported cases. At this stage, double reports are avoided by using the knowledge about names and birth dates. Afterwards, these personal data are removed to fulfil privacy regulations. Within one week the reorganized and standardized cases should be reported to the Robert Koch-Institute.

In many cases, an infected person consults a doctor in another district than he or she is living, such as the nearest largest town, the job location or in holidays. In this instance, the first report is done to the local health department of the patient's whereabouts. Later on, the case is reported to the local health department of the patient's residence where further combination is possible and the procedure continues as explained above.

Beside this system there are some infectious diseases such as HIV (human immunodeficiency virus), which are handled more anonymously. A case of these diseases are reported without any personal data directly to the RKI (Robert Koch-Institut, 2000).

Resulting Data Sets

Finally, in the system of the Robert-Koch-Institute there is a standardized data set for each reported case. The report involves not only the diagnosis of disease, but as well the bacteria type. Beside these, there are information about the infection such as the most probable way and risk, patient's whereabouts, admission to a hospital, and belonging to an outbreak. Personal data are removed, only age and sex redescribe the patient (Robert Koch-Institut, 2000). The date of report is considered to be the date of report to the first stage, which means to the local health department. Moreover, other dates such as a day of diagnosis or a period for illness beginning are stored.

2.3 Descriptive Analysis of Campylobacter Data

In this section, the reported cases of Campylobacter are analysed descriptively. After introducing the data, the variables and their irregularities, and the incidence is displayed. Then, the reporting delay is investigated. Finally, the infection cases are describes and the patients characterized.

2.3.1 Introduction to Data

The data provided by the Robert Koch-Institute (Robert Koch-Institute, 2010) include a data set for every case recorded from 1^{st} January 2001 to 3^{rd} January 2010 in Germany, which corresponds to the calendar weeks of the years 2001 until 2009.

Variables in the Data Set

The Table B.1 in Appendix B gives a detailed overview of the initial variables of the data set and their meaning. Information about several characteristics of infection and patient are given.

The moment and reporting of infection is detailed by several variables indicating the day of reporting, the date of arriving in the system of RKI, the time of diagnostic and report of the laboratories as well as the time period for the start of illness. Further on, the case of infection is qualified by the diagnosed subtype of Campylobacter bacterium and a variable indicating outbreaks. Beside this, the location is given with the resolution of districts. The patient's age and sex are recorded as well.

Missing Data

In general, the first check of a given data set is the proportion of missing values in the data frame (see Table B.1 in Appendix B). The most important variables for surveillance analysis reporting date, sex, age, location and bacteria type are almost

complete (less than 1% missings). The variable `outbreak` indicates the outbreak ID, if the case belongs to one. The large number of missings (97.1%) is reasonable, because non linked cases are indicated by `NA`. The second value of bacterium type is used only if a second type of bacterium is diagnosed by the laboratory (98.5% missings). There are some missings for the time period of illness beginning (11.5% and 69.2% missings respectively). If the beginning of the illness is known exactly, the value for `start2` is usually not given, which explains 69.2% of missing data in this variable. Of less relevance for the analysis in this thesis seem to be variables `locStart` and `locEnd` which are dates indicating time period the patient spend at the location of the infection (93.4% and 94.2% missings respectively). The variables `labReport` and `labDiag`, the dates at which the laboratories report the diagnosed Campylobacter infection, have a remarkable high rate of missings (43.9% and 30.8% respectively). The variables `lastUpd`, for inserting the full version at the local health department, and `arriveRKI`, indicating the date inserting the first version at the RKI, are recorded electronically and therefore are almost complete (0.1% and 0.2% missings respectively).

Quality of the Data

At the beginning of 2001, when the new reporting system was introduced, irregularities in the data can be observed. The users needed to find their way around in the system, which caused abnormal reporting behaviour. In this period, many cases were not reported and input errors appeared. Therefore, it is preferred to exclude these data for the later on investigations.

Furthermore, the RKI knows about misunderstandings in the meaning of variables and typing errors. With identification of this matter the RKI is working on an automatic quality control. Especially the variables representing a date consist of a large amount of impossible values and a very wide range. Here, only a superficial quality check is made. Dates outside the time period of interest are set as missings. Furthermore, for descriptive analysis robust measures such as median, 5%- and 95%-quantiles, instead of mean, minimum, and maximum, are used to describe the variables. Figures are plotted in a truncated manner. This strategy is confirmed by the RKI methods.

2.3.2 Incidence

In this part, the number of Campylobacter infections is examined in aggregated, spatial and local level.

Aggregated Incidence in Germany

As illustrated in Figure 2.3 the frequency of weekly cases ranges between 327 and 1880 in Germany. Yearly totals range between 43716 in 2003 and 61680 cases in 2007. A consistent seasonality in the data is observed. Using the weekly means of the reported cases over the years the Campylobacter incidence peaks between the 28^{th} and 35^{th} calendar week. It is assumed, that this is related to different leisure activities and to the fact that German summer holidays are in this time period.

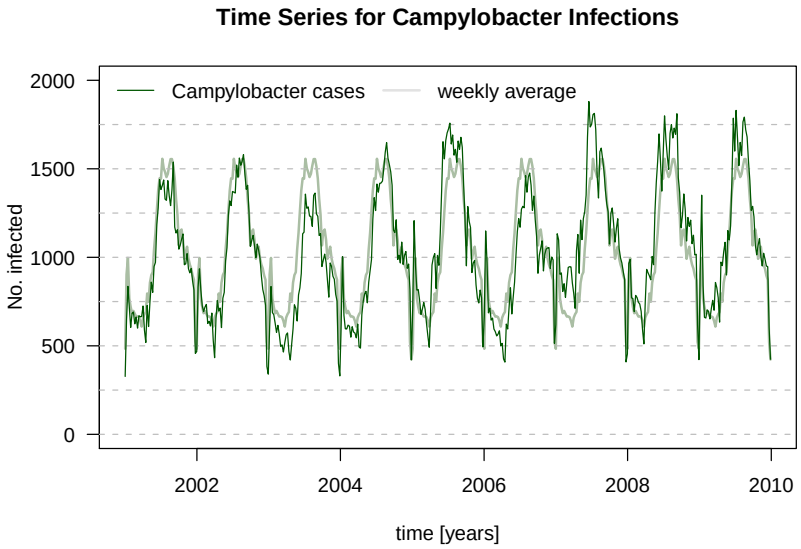

Figure 2.3: Progress of Campylobacter incidence between 2001 and 2009 with superimposed weekly average computed from the corresponding nine values.

Using additional plotting of the weekly average, computed over all nine years, can identify just small changes in seasonality every year. Therefore, the peak in 2003 is later than usual. Additionally, the 'intensive' years are in 2007 and 2008, in which there have been much more cases than on average.

Spatial Distribution of Incidence

Even though the spatial distribution of incidence is not of major interest here, a map of Germany's Campylobacter incidence 2009 is displayed in Figure 2.4. The data is stratified by the level of districts in 2009. In Germany, there were 412 districts at this moment. The number varies, because there have been several local government reorganizations. The living condition, especially the population size, might be very different in the various districts. For that reason, the measure of incidence is used which is the number of new cases of a disease per unit of time in a given population. In general, this means the number of cases by 100.000 inhabitants within one year. Figure 2.4 shows the yearly incidence in Germany for every district in 2009.

In the map, there are more cases in coast regions in the north of Germany. Furthermore, one can find higher incidences in Saxony and the north of North Rhine-Westphalia. Another cluster seems to be in Saarland and Rhineland-Palatinate, which is located in the south-west of Germany.

Local Incidence

In order to examine the local incidence, ten districts (see Table 2.2) are randomly selected. The district of Landsberg am Lech is added, because there was a described outbreak in 2005. Four of the district's yearly incidences are introduced in Figure 2.5. The overall incidence in Germany is plotted with a dark line to recognize local

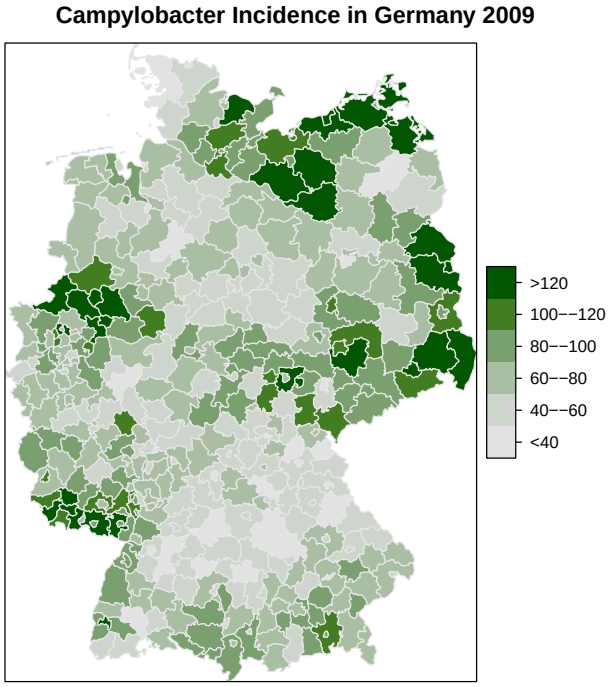

Figure 2.4: Map of the raw Campylobacter incidence rates per 100.000 inhabitants for each of 412 administrative districts in Germany, 2009. (Map source: Bundesamt für Kartographie und Geodäsie, 2010)

differences. Moreover, the grid distance in the plot is kept fixed to emphasize the different scales. Corresponding key parameters are given in Table 2.2.

It can be recognized that the incidence time series vary more than the general one which includes cases of all districts. Therefore, the absolute maximum of local incidence is usually much larger than the incidence for whole Germany (see Table 2.2). In all districts occur weeks without reported Campylobacter cases. Furthermore, there are district specific differences between the general incidence levels.

The district Schmalkalden-Meiningen is located in the Thuringian Forest, in the center of Germany. Its yearly incidence has on average a similar time series as the overall incidence in Germany, but in some time period, for example in 2005, there are less cases.

The district of region Hanover, with 1.130.039 inhabitants, is the most populated district in Germany when excluding the big cities. Here, the incidences are in general larger than the overall progress.

Kempten is a town in the Bavarian Allgau where a remarkable low incidence was examined. There are some time periods without any reported Campylobacter cases. The mean seasonal peak is very late in the 38^{th} week of the year.

	population	maximal incidence	max year	week of mean peak
Germany	82314906	2.28	2007	28
Region.Hannover	1130039	59.96	2008	30
LK.Offenbach	336671	17.83	2007	35
SK.Trier	103888	11.34	2007	34
LK.Freising	164692	8.1	2009	28
LK.Landsberg.a.Lech	113311	13.36	2005	28
SK.Nürnberg	503110	16.21	2007	28
SK.Kempten	61703	4.86	2008	35
SK.Schwerin	95855	9.72	2004	34
LK.Görlitz	288735	45.38	2008	32
LK.Schmalkalden.Meiningen	134262	11.34	2009	35
SK.Berlin	3416255	153.96	2001	27

Table 2.2: Key parameters and population for selected districts

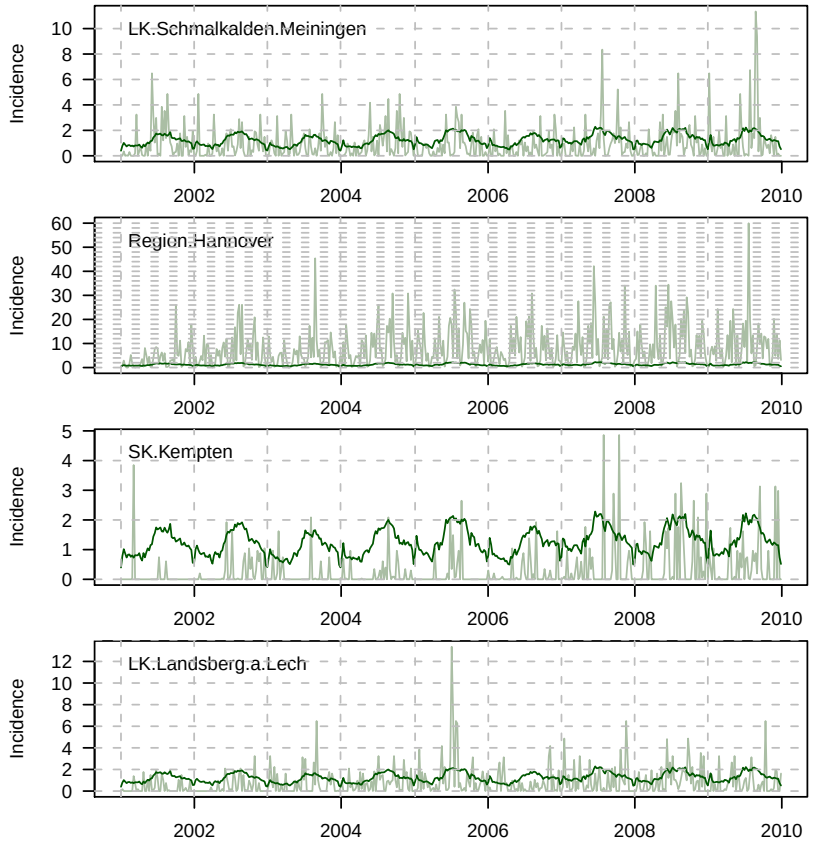

Figure 2.5: Progresses of yearly Campylobacter incidences in chosen districts between 2001 and 2010 in comparison to the overall incidence (darker line) in Germany

Landsberg am Lech, in Bavaria, has usually a lower incidence than Germany in general. In 2005, a high peak is, however, recognizable. In the RKI bulletin an outbreak in June 2005 due to raw milk was reported. At a school excursion to a farm 18 children got infected by Campylobacteriosis (Anonymous, 2006).

2.3.3 Investigation of Durations and Delay

Many authors mention the importance to investigate reporting delays and their bias (see Heisterkamp et al., 2006). Most of the algorithms for automated outbreak detection of infectious diseases are not able to include delays properly. In the last section, some irregularities due to delayed reporting were mentioned which are now examined in more detail.

Variables of Delay

The variables given in the data frame representing a date, have a logical chronological order which is caused by the regulations of the reporting system (see section 2.2). First, dates corresponding the period of illness beginning are defined by the variables `start1` and `start2`. In general, it follows the diagnose and report date of the laboratory. At the date of report, the case is reported first to the local health department. According to the time period required for the investigations of the department the date of last update is defined. Finally, there is the report to the RKI in its first standardized version. In the special event, where a patient working in the food industry or belongs to an outbreak this logical sequence can vary due to a different reporting system (see section 2.2).

	0.05-quantile	Median	Mean	0.95-quantile
Illness Start Time Period	0.00	0.00	3.10	12.00
Delay Laboratory Diagnose	1.00	6.00	7.80	17.50
Delay Laboratory Report	3.00	8.00	10.10	20.00
Reporting Delay	0.00	7.00	9.20	21.00
Delay last Update	5.40	12.00	15.50	30.50
Delay RKI Arrival	8.50	17.40	21.00	38.50
Laboratory Delay	0.00	1.00	2.30	6.00
Investigation Delay	0.50	4.00	6.40	15.50

Table 2.3: Descriptive measures for durations and delays in days

First, the time period of illness starting is investigated. In the data, it is indicated by the two variables `start1` and `start2`. The time period is on average 3 days, but the median is 0 (see Table 2.3). If the beginning of the illness is known exactly, the value for `start2` is usually not given, which explains 69.2% of missing data in this variable (see Table B.1 in Appendix B). Hereafter, the value of `start1` will be used for the illness beginning, if no value is given in `start2`. Otherwise the midpoint of their interval will be used. The computed measure will be the reference value for the delays worked out in the following.

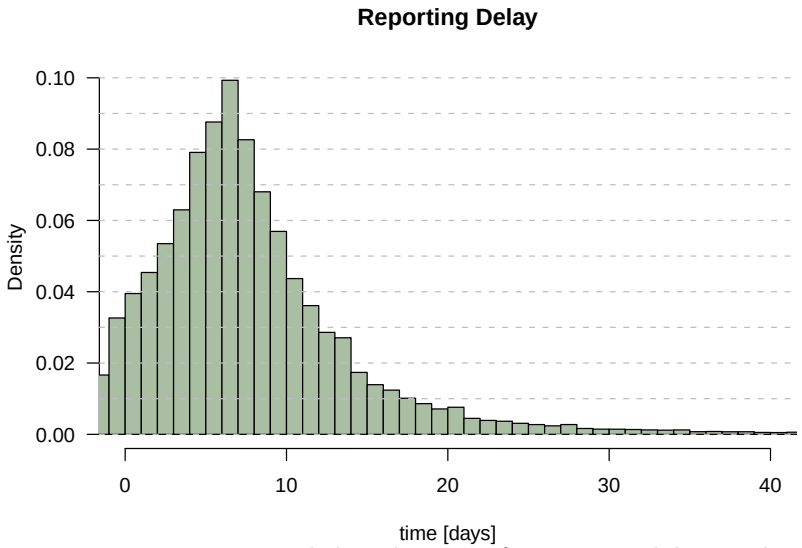

Figure 2.6: Truncated distribution of reporting delay in days

Distribution of Delay

For the investigations, the reporting delay is defined as the difference between start of illness and date of report to the local health department in days. Referring to Figure 2.6, it has a slightly left skewed distribution with an expected delay of 9.2 days. The histogram is truncated at 0 and 40 and values range mostly between 0 and 21 days (see Table 2.3). Only 50% of the cases are reported within 7 days after the start of illness, which means 10 days after date of infection implying an average incubation time of three days.

Composition of Delay

The variables giving different dates and related durations will be viewed in more detail. Thereby, the date at which the illness started will be the reference. As before, robust measures are chosen for description (see Table 2.3).

Figure 2.7 presents boxplots, truncated between -5 and 30 days, for duration corresponding to each stage of delay. Due to missings in the variables, different number of data sets are basis for the calculations which refers to the heights of the boxplots. The chronological order, given by reporting flow regulations, can be identified. Moreover, with increasing delay is examined increasing variation.

Campylobacter is diagnosed with a median of six days delay. The report takes place a bit later, in general after two days, which means eight days after start of illness. The already investigated reporting delay has a bit lower median of seven days, which is assumed to be caused different regulations for reporting. The following investigations by the local health department occupy mostly between a half and four days so that the last update in general is made 12 days after start of illness. Finally, the data set arrives at the system of the RKI after 17 days. After three weeks 95% of the Campylobacter cases arrived at the RKI.

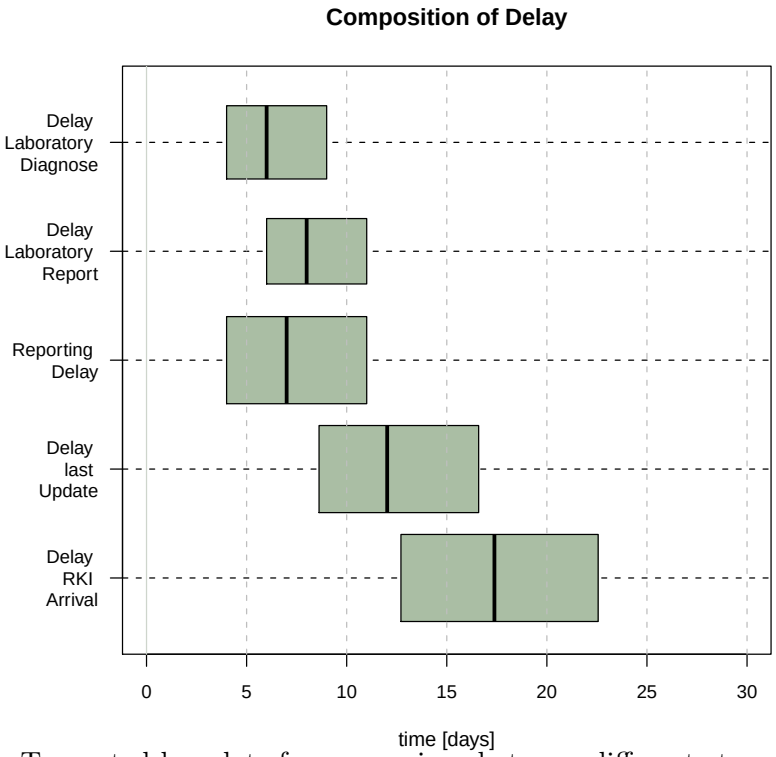

Figure 2.7: Truncated boxplots for comparison between different stages of delay in days

Artefact of Reporting at the Turn of the Year

A special type of delay occurs during holidays, which attract most attention on Christmas and New Years Eve (see Figure 2.8). A lot of physicians and laboratories are closed during these days. Therefore, less reports are issued during this time. Usually, after these holidays, there is a large increase of reports. It is assumed that a large proportion of these cases actually originate from previous weeks.

2.3.4 Case Characteristics

In this section, the characteristics of the reported cases, regarding the type of Campylobacter bacteria and belonging to an outbreak, are investigated. The characteristics of infection are examined in relation to the demographical characteristics as well.

Characteristics of Infection

The different bacteria types are summarized into four groups of interest. Most cases are caused by Campylobacter jejuni bacteria (60%) and 20% by Campylobacter spp. Thereby, spp. means species pluralis, therefore the occurrence of various not declared species. Beside this, 12% of Campylobacter species could not be differentiated and 7% are other Campylobacter types such as coli or lari. For future analysis, the last class of cases will be excluded to obtain a data set as homogeneous as possible.

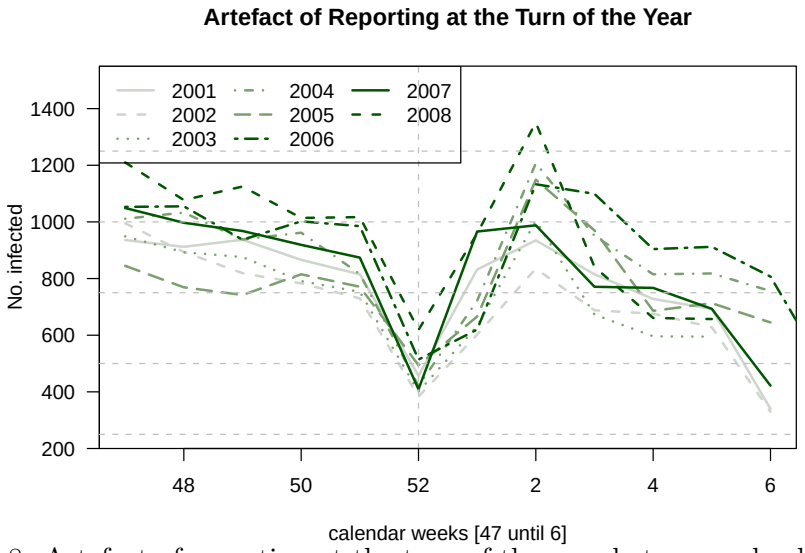

Figure 2.8: Artefact of reporting at the turn of the year between calendar weeks 47 and 6

With every case of Campylobacter infection, it is reported whether it belongs to recognized outbreak or not. At the same time, not every outbreak is detected by the public health system. In the analysed data set the size of an outbreak ranges between one and 45 cases. Outbreaks of size two are most common (64%). Outbreaks with more than ten cases are very rare which reflects the difficulties to link outbreak cases.

Another measure for the true and undetected size of outbreaks can be derived by the number of weekly reported outbreaks displayed in Figure 2.9, which ranges between zero and 88 outbreaks with its average at 25.

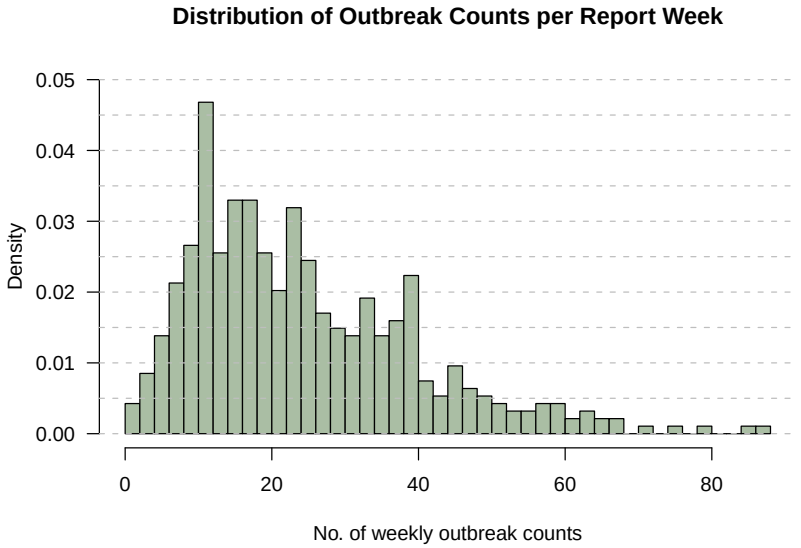

Figure 2.9: Truncated distribution of weekly number of reported outbreaks

Demographical Characteristics

In this section, demographical characteristics age and sex are investigated. There are 53% male and 47% female cases. Missings are rare (0.1%). This finding corresponds with the results of other studies where a higher proportions of males was found as well (Louis et al., 2005).

In the following Figure 2.10, the age distributions differentiated by gender are displayed in an age pyramid. Furthermore, the age distribution of the general population in Germany is plotted to figure out age groups which get infected more frequently than others, thus are specially exposed.

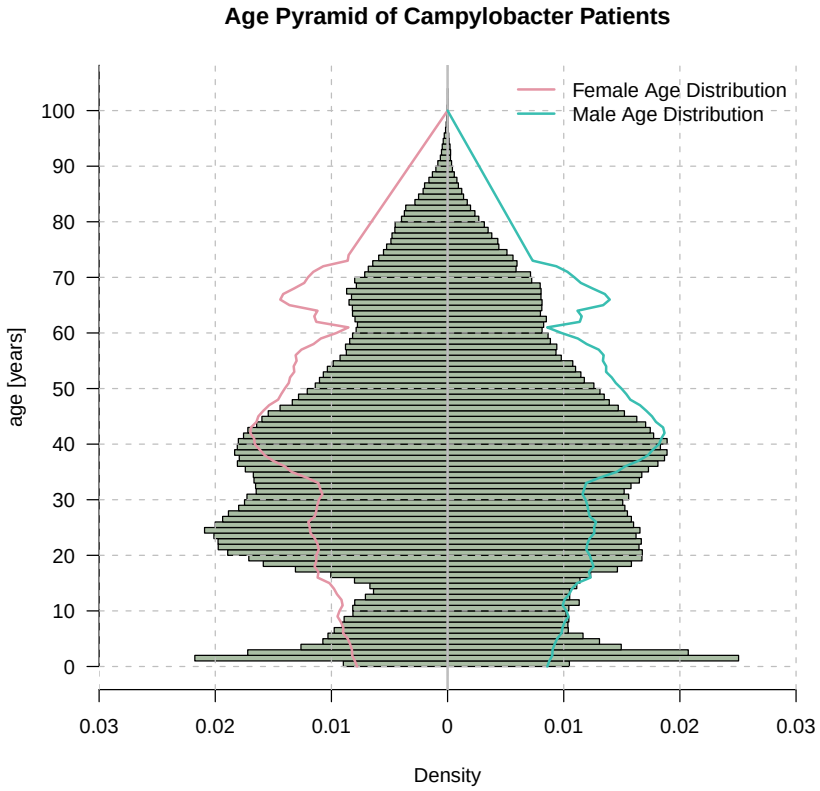

Figure 2.10: Age distribution of the Campylobacter cases given the gender compared to the overall age distribution in Germany.

First, the general age distribution of the Campylobacter cases is investigated. It is multi-modal with modes in children's first year, 25 years, and 40 years. Children until the age of five years seem to have special exposition to fall sick. This might be caused by generally high parents' and doctor's caution with diseases of small children. It is hypothesized that children are more likely to be diagnosed and reported and assumed to be a group almost without underreporting. Furthermore, there are more cases in the age span between 15 and 35 years, while after the age of 50 years less cases appear.

Moreover, in Figure 2.10 one can recognize that boys in the age between 10 and 15 years seem to be more likely to get infected with Campylobacter than girls. On the other hand, young women between 20 and 30 years are more exposed to get infected than young men. For higher ages no difference can be recognized.

Factors of Belonging to an Outbreak

In this part, it is investigated, if there are any associations for belonging to an outbreak on other case characteristics. The factors sex, age, and bacteria type are examined.

Table 2.4 shows that females are a bit more likely to belong to an outbreak. A χ^2-test confirms this assumption (p-value < 0.001).

	male	female
no	0.511	0.461
yes	0.015	0.014

Table 2.4: Association between outbreak belonging and sex

For the age distribution for cases belonging to an outbreak some associations were recognized as well (see Figure 2.11). The distribution is bimodal while the mode at 25 years which was recognized in Figure 2.10, disappeared. Compared to the overall age distribution in the healthy population, it can easily be examined that children and teenagers until the age of 20 years are much more likely to belong to an outbreak. Furthermore, in the ages between 30 and 40 years more Campylobacter infections occur than it is expected.

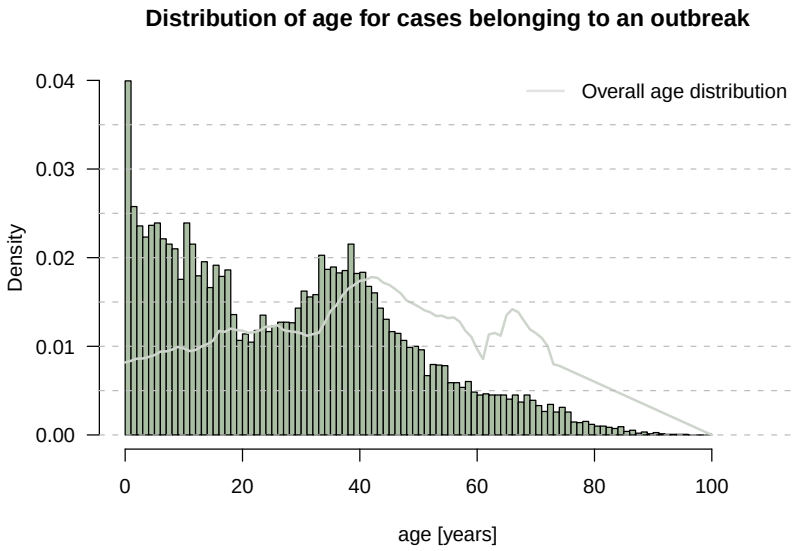

Figure 2.11: Distribution of age for cases belonging to an outbreak

Moreover, bacteria types Campylobacter jejuni and spp. are less often diagnosed if the case belongs to an outbreak (see Table 2.5). The bacteria type defined as 'others' appears more frequently in outbreaks. An association is confirmed by a χ^2-test (p-value < 0.001) as well.

	jejuni	spp.	not diff.	others
no	0.588	0.194	0.122	0.070
yes	0.015	0.004	0.003	0.004

Table 2.5: Association between outbreak belonging and bacteria type

2.4 Weather Data as extern Process

In this section, a general discussion about the influence of weather parameters on the incidence of Campylobacter based on a literature review is given. It follows a description of available climate data which will be used as a covariate process in a developed Bayesian hierarchical time series algorithm (see chapter 4.3) in section 6.

2.4.1 Weather Influence on Campylobacteriosis: A Literature Review

Weather, as the condition of atmosphere at a particular place and time, is assumed to influence the incidence of several diseases. It is a dynamic, seasonal, but irregular process. In this section the literature is reviewed regarding the influence of weather parameters on Campylobacter incidence. The section is organized as follows. First, the investigations which try to explain the seasonal pattern in the incidence by weather parameters are presented, and afterwards the hypotheses of weather influence on the possible reservoir of Campylobacter are resumed. Finally, consequences for modelling the weather influence in the surveillance analysis are derived.

Investigations on direct Influence of Weather Parameters to the Campylobacter Incidence

The Robert Koch-Institute detected an increased Campylobacter incidence in Germany in 2007 (Jansen et al., 2007). The increased number of infections was caused by bacteria Campylobacter jejuni. No abnormalities were found for any group of age, gender, or regions. Only the federal states Berlin and Bremen did not show increased incidence. An association to the special warm weather in the spring of 2007 is assumed, since the change of leisure time behaviour, such as picnics and barbeques, definitely changes the exposition for Campylobacteriosis.

In an internship project during summer 2008, the supposed association between weather and the Campylobacter infection cases was investigated by using a negative binomial regression model. The most convenient model included lagged weekly mean temperature and relative humidity. The parameters were aggregated to absolute humidity due to content-related argumentation and the strong association between

variable	estimate	95% C.I.	p-value
(Intercept)	0.040	(0.037, 0.044)	<0.001
time	1.000	(0.990, 1.010)	0.964
l1.hum	1.067	(1.005,1.080)	<0.001
l2.hum	1.038	(1.026,1.050)	<0.001
age[<10 years]	0.650	(0.593,0.713)	<0.001
sex[male]	1.289	(1.214,1.370)	<0.001
time:age[<10 years]	1.049	(1.038, 1.060)	<0.001
time:sex[male]	0.989	(0.979,1.000)	0.048
l1.hum:age[<10 years]	0.984	(0.969,0.999)	0.037
l2.hum:age[<10 years]	0.996	(0.981,1.012)	0.631
age[<10 years]:sex[male]	0.921	(0.880, 0.964)	<0.001

Table 2.6: Parameter estimates of final negative Binomial regression model (Source: an der Heiden et al., 2010)

temperature and relative humidity (an der Heiden et al., 2010). The resulting model estimates are displayed in Table 2.6 and the corresponding fit is shown in Figure 2.12.

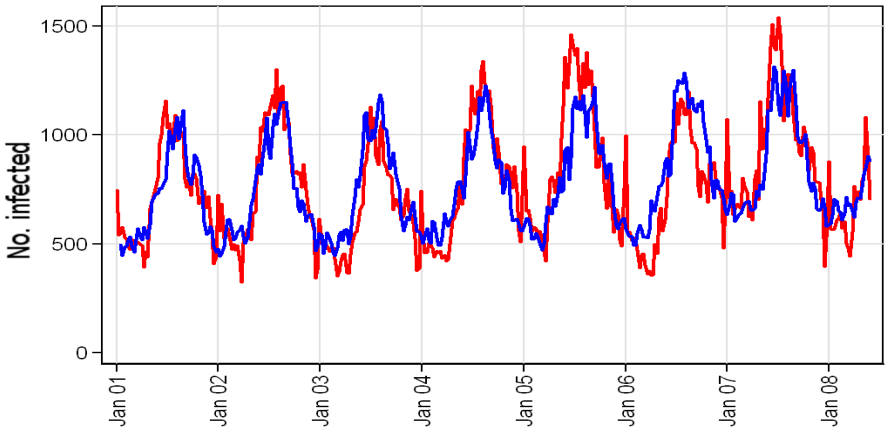

Figure 2.12: Negative Binomial regression model for reported number of Campylobacter infections while covariate absolute humidity enters in the 15mb truncated version. The red line represents observed, and the blue line modelled counts. (Source: an der Heiden et al., 2010)

Louis et al. (2005) investigated the relationship between seasonal variation in human Campylobacter infection in England and Wales. They used the laboratory-confirmed reported cases between 1990 and 1999 and weather data averaged per week. They figured out that the date and the shape of the Campylobacter time series varied with geographical location. The incidence of Campylobacter was modelled by a regression with autocorrelated errors. The disease rates were transformed with the Freeman-Tukey square root and averaged weekly temperature, precipitation, and hours of sunshine were included as covariates. Furthermore, an overall trend as long term correction was included, and weather lags or intervals of weather influences were

tested for better performance. Nevertheless, the three-variable-model remained the best one.

Patrick et al. (2004) fitted a locally weighted linear model with cross-validated optimized number of weather lags between zero and six weeks. The effects of temperature, precipitation, relative humidity, and hours of sunlight on Campylobacter incidence in humans in Denmark were examined. Moreover, they could reveal these effects in broiler flocks while the factors explained considerably less than in the model for humans.

Nylen et al. (2002) investigated the seasonality of kernel smoothed Campylobacter incidence in several European countries and New Zealand and extracted seasonal patterns with different peaks that were consistent within each country. Several possible explanations are discussed, and unexplored reservoirs linked to the differences in weather should be considered. A literature review summarizes them in the following.

Investigations on Weather Influence to Campylobacter Reservoirs

There is a remarkable pronounced and consistent seasonal pattern of Campylobacter infections whose cause is unclear. Several authors tried to find the environmental factors explaining the observed seasonality.

The most frequently discussed reservoir of Campylobacter, especially of subtype jejuni, is poultry. National monitoring of Campylobacter in broiler chicken in Germany in 2004 and 2005 observed a seasonality in animal incidences. There was a significantly higher prevalence in broiler in summer months even though the indoor temperature was consistent (Peters et al., 2006). The strongest increase was observed between 13–20°C degrees which is levelling off at higher temperatures.

Nichols (2005) attempts another explanation and hypothesizes flies as reservoir for Campylobacter in both, humans and animals. The seasonal distribution of Campylobacter infections around the world could be associated with conditions of fly reproduction which need rainy days in warm summers, therefore high relative humidity. This hypothesis is justified by several argumentations, but remains statistically not verified.

Hudson et al. (2001) studied serotypes of human cases, veterinary cases, raw chicken, milk, and water. The study could conclude that the type of serotype which was dominant in February was absent in August, and that the common type of serotype in August was absent in February. Therefore, it is hypothesized that different pathogens cause the observed seasonality.

Louis et al. (2005) figured out a qualitative association between Campylobacteriosis incidence and the level of agriculture, represented by the percentage of rural ward in a district. Furthermore, the source of drinking water (surface or not) was revealed to be associated, but quantitative analysis did not yield significant correlations.

Consequences for Modelling

The literature review showed, that the influence of weather to the Campylobacter incidence is confirmed by several independent studies and statistical methods, while the explicit causality in a reservoir remains unclear. The investigations agree more or less in their selection of mean temperature and precipitation or relative humidity as covariates.

In the following investigations, these weather parameters are replaced by the absolute humidity, as introduced by an der Heiden et al. (2010), to avoid possible model collinearity. An approach of monitoring, which is able to include covariate processes, will be introduced in chapter 4. Its application to the Campylobacter time series will be shown in chapter 6. In the following, the preprocessing of the weather data is described.

2.4.2 Available Weather Data

The data of the 44 freely available weather stations of the German Climate Service (Deutscher Wetterdienst, DWD, see Table B.2 in appendix B) are used. The stations are evenly spread in the country and each of them is typical for a natural region in Germany.

The basis for the covariate weather process are the daily climatological values. The DWD provides information for several weather parameters such as minimum, mean and maximum temperature, mean relative humidity, mean wind-force and strongest gust, total sunshine duration, mean degree of cloud coverage, and the total precipitation (Deutscher Wetterdienst, 2010).

Weather Data Preprocessing

The data is available starting from 1991, but only information starting from the late 2000 is needed for the present investigations.

Eight stations are not considered for the investigations, because they represent extreme weather conditions. These are five stations situated on the mountains, namely Brocken, Fichtelberg, Hohenpeissenberg, Kahler Asten, Nuerburg-Barweiler, and Zugspitze. Furthermore, the three stations on the islands Fehmann, Helgoland, and Sylt are also excluded. Additionally, the data of the station in Fritzlar is excluded due to the poor data quality.

As worked out above, the relevant variables are mean temperature and mean relative humidity. Their composition is absolute humidity which is computed according to formula given in the following excursus. According to an der Heiden et al. (2010),

large and infinity values in the resulting time series of absolute humidity are limited from above by 15 g/m^3.

Since the IfSG incidence data of the Robert Koch-Institute are available on weekly level, the daily weather values are averaged on a weekly basis as well. Therefore, the resulting data sets contain the mean absolute humidity of each calendar week and weather stations.

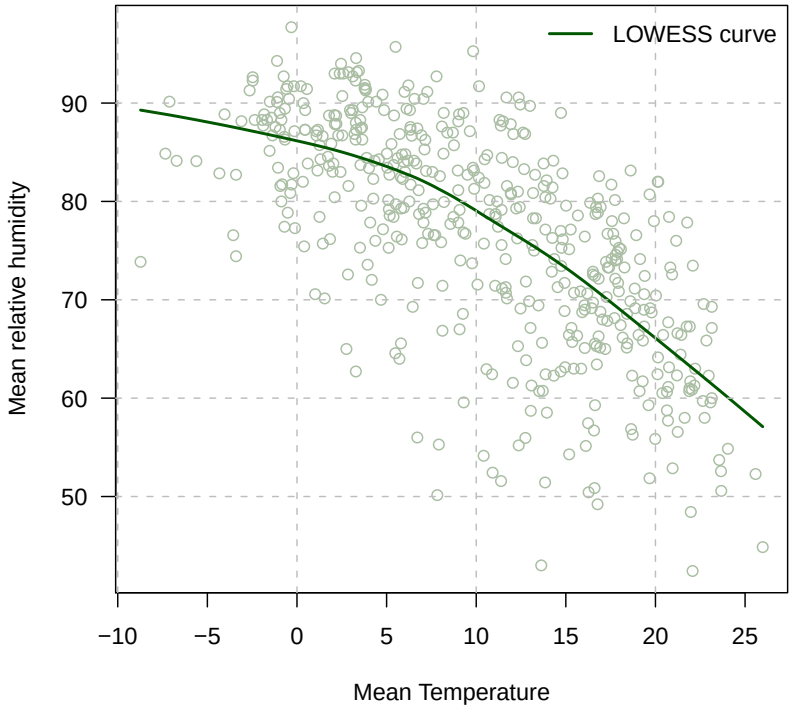

Figure 2.13: Association between mean temperature and mean relative humidity in Berlin, Germany

Excursus. Absolute Humidity

Water appears in three states of matter in the atmosphere: solid such as in snow, liquid such as in fog clouds, and in form of vapour, which exist always in the air, as well. Thereby, absolute humidity measures the actual amount of vapour in the air.

With increasing temperature, the air's ability to absorb water in form of vapour increases as well. Since relative humidity is the ratio of absolute humidity to the maximum of saturation, relative humidity depends on the temperature. Therefore, absolute humidity combines the knowledge about temperature and relative humidity.

The absolute humidity at temperature T is computed by using a unique way to do it in the physical literature (an der Heiden et al., 2010)

$$h_{abs} = h_{sat}(T) \ \cdot \ \frac{h_{rel}}{100},$$

where h_{rel} is the relative humidity and $h_{sat}(T)$ is the ability of the air to absorb water vapour at temperature T which is computed by

$$h_{sat}(T) = h_{sat}(T_0) \ \cdot \ \exp\left(\frac{L}{R_\nu}\left(\frac{1}{T_0} - \frac{1}{T+T_0}\right)\right)$$

while $L = 2.270.000\frac{J}{kg}$ is a constant latent heat for vaporization of water, $R_\nu = 461,5\frac{J}{kg \cdot K}$ a gas constant for water vapour, $T_0 = 273.15K$ the reference temperature measured in Kelvin with $h_{sat}(T_0) = 6.11mb$, measured in millibar, the reference water absorbing ability of the air (an der Heiden et al., 2010). The formula describes an exponential association between temperature and absolute humidity. Further details are described e.g. in Dengler (1997).

$\diamond$

Ordinary Kriging

A continuous absolute humidity map is created for each time point by using ordinary Kriging (see following excursus). In the present application, the weather stations have the location $s_{it} = (s_{itx}, s_{ity})$ represented by latitude and longitude coordinates, which vary continuously in $D \in \mathbb{R}^2$. For each of these reference points an absolute humidity for a given day $y(s_{it})$ is observed.

Excursus. Ordinary Kriging

Kriging is an interpolation of spatial phenomenons (Fahrmeir et al., 2009). The basis are geostatistical point-referenced data $y(s_1), \ldots, y(s_n)$ of n locations $s_1, \ldots, s_n$. Thereby, $y(s)$ can be seen as realization of a spatial stochastic process $\{Y(s), s \in D\}$ with $D \in \mathbb{R}^d$, a spatial domain. If s is varying continuously in D, a Gaussian random field $Y(s)$ is used, which means that $Y(s_1), \ldots, Y(s_n)$ are assumed to be multivariate normal distributed for all n and $s_1, \ldots, s_n$.

The classical geostatistical model is followed with this assumption and is the basis of estimation in ordinary kriging.

$$\begin{aligned} y(s) &= \mu(s) + \gamma(s) + \epsilon(s) \\ &= \mu + \gamma(s) + \epsilon(s), \end{aligned}$$

where $\mu(s)$ is the spatial trend, $\{\gamma(s), s \in \mathbb{R}^2\}$ is a stationary random field and $\epsilon(s) \overset{\text{iid}}{\sim} N(0,1)$ the model errors. In case of ordinary Kriging no covariates are included and the trend $\mu(s) = \mu$ is constant. Due to the assumption of stationariness, the random field $\gamma(s)$ is expected to be $\nu = 0$ with variance $\tau^2(s) = \tau^2$, and correlation $\rho(s,t) = \rho(s-t) = \rho(h)$ which depends only on the distance and

not on the exact location. The specification of correlation defines a spatial association structure while potence, exponential, spherical, or Matérn family could be used.

The resulting estimation equation can be represented as a linear combination of basis kernel functions. Therefore, the univariate simplification equals to a density estimation. The spatial effect can be treated as random effect, so that inference can be represented in a linear mixed models context.

◇

Example. Figure 2.14 shows one specific Kriging result for the 28^{th} calendar week in 2009. Higher absolute humidity in the north eastern and the south western corner of Germany can be spotted. The lowest humidity was recognized in the west of Germany.

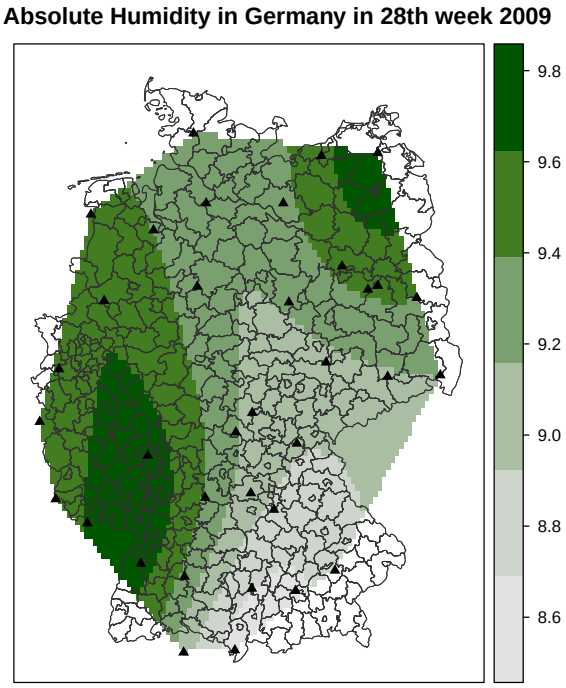

Figure 2.14: Kriging result for absolute humidity in mb in Germany in 18^{th} calendar week 2009, with the triangles indicating the positions of the included weather stations.

◇

Hereafter, the estimated constant trend $\hat{\mu}$, referring to the geostatistical model, is used as overall absolute humidity in Germany. Regarding possible future analyses, one is free to compute local humidity, e.g. for specific districts, as well.

2.4.3 Time Series of Absolute Humidity

In Figure 2.15, the process of absolute humidity is visualized. The values are generally very low, with exception in autumn and winter when the values are high, where

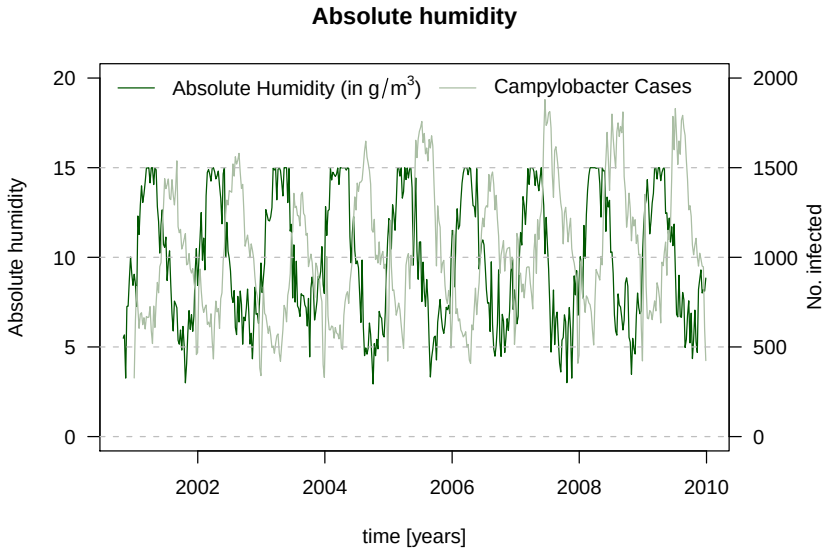

Figure 2.15: Absolute humidity with superimposed weekly Campylobacter cases

they range between 2.9mb and 15mb. Note the adjustment for large values, which are limited from above by 15mb. The median is 9.6mb while the mean is 9.9mb which marks a strong left-skewed distribution. To investigate the association with the Campylobacter incidence the disease cases are superimposed in the graphics. When the values for absolute humidity are large the season of Campylobacteriosis begins.

2.5 Wrap-up Notes

This chapter aims to set the subject of monitoring infectious outbreaks in its medical, structural, and environmental context. It was introduced exemplary to several surveillance issues such as a reporting system, characteristics of data, and explanation attempts by prospective modelling.

The reporting system of the Robert Koch-Institute was introduced. While the public health systems in other countries, such as the Centers for Disease Control and Prevention (CDC) in the United Kingdom or the Infectious disease Surveillance and Information System (ISIS) in the Netherlands, have variations according to their politically system, but they all work quite similar.

The surveillance data these systems provide, are introduced by the Campylobacter data of the RKI. A lot of specific and general characteristics of the surveillance data have been shown. Especially reporting delays, inaccuracies, and associations with outbreak belonging, were examined.

Furthermore, investigation was brought into line with the state of the disease research. In the literature review, potential influences to Campylobacteriosis including their modelling approaches were inspected. The time series of absolute humidity is extracted to have a substantial influence to the Campylobacter incidence.

In the following chapter, the exemplary findings are generalized and statistical methods of surveillance are introduced. Later on, in chapter 6, the Campylobacter data are monitored by using different surveillance algorithms. During the application of a newly developed Bayesian hierarchical time series approach absolute humidity as a covariate process will be included.

Chapter 3

Overview on Surveillance Methods

'Public health surveillance is defined as the on-going systematic collection, analysis and interpretation of outcome-specific data that are essential to the planning, implementation, and evaluation of public health programmes, closely integrated with the timely dissemination of these data to those who are responsible for prevention and control' (Thacker and Berkelman, 1988, quoted in Sonesson and Bock, 2003).

As the surveillance data and relevant issues were introduced in the last chapter, the present chapter gives a general introduction to statistical concepts for surveillance methods. First, the specific characteristics of surveillance data are described. An outline of the general strategy for detection of aberrations and an overview on the variety of surveillance methods will follow. Afterwards, the focus is on the Farrington algorithm as this method is in routine use in many public health institutions. Additionally, techniques for evaluation of algorithm performance and comparison of different methods are discussed.

3.1 Introduction to Surveillance Methods

Surveillance usually includes passive case detection due to an established reporting system and active surveillance in which the recipient takes some action to identify the cases. Passive surveillance simplifies reporting and is less costly while active surveillance is the best approach in outbreak investigations to elicit all cases (Straif-Bourgeois and Ratard, 2005).

This work is focussed on the analysis of these public health data with respect to collection and interpretation. Therefore, in the following sections the aim, purposes, and peculiarities of the data are summarized.

3.1.1 Aim and Purposes

"Public Health surveillance of emerging infectious diseases is an essential instrument in the attempt to control and prevent their spread' (Höhle, 2007). Beside this, epidemiology is the basic science to understand the disease and find appropriate

interventions to break the chain of transmission to prevent diseases. Due to global travellers and world-wide food distribution the spreading of diseases has become faster than before. Depending of the disease the arrangement, implementation and impact of the chosen intervention requires a sufficiently early outbreak alarm.

Furthermore, surveillance methods should be automated procedures, which are able to handle large quantities of data. On the one hand, they should be sufficiently robust and flexible to handle a wide range of diseases and specific characteristics of surveillance databases.

On the basis of a time series, models should detect aberrations, which means the detection of abnormalities, which usually are more recognized disease cases than expected. In case of the appearance of any aberration in the time series, the time point is flagged as alarm for further investigations.

3.1.2 Special Characteristics of Surveillance Data

The data in surveillance can be seen as a contrast to the perfectly planned data in clinical trials. In the following, general characteristics and specific problems in data from public health surveillance are summarized. The consideration as well as the appropriate handling of these issues is necessary for a successful statistical surveillance method.

Example. In Figure 3.1 are shown the counts of adult meningcoccal infections in France between 1985 and 1998. Meningococcal infection is rare, but the most common cause of bacterial meningitis, which is an inflammation of the membranes that cover the brain and spinal cord. On average six adult persons are infected each month, while seasonality is obvious.

The monthly counts are an example for the typical time series of surveillance data. In the following, this time series is used to illustrate important issues in surveillance methods.

◇

Electronic Reporting

In most countries, the surveillance system is based on computer programs and web based reporting which allows a collection of large amounts of data within a reasonably short period of time. But, persons who may not be very proficient with the software could enter data of poor quality (Straif-Bourgeois and Ratard, 2005). Because of the large number of persons involved in the reporting system it is difficult to provide adequate software training and achieve consistency in the system which leads to inaccuracies in the data.

Lack of Accuracy

These errors due to electronic reporting and other typing errors, i.e. by usage of short cuts, can lead to inaccuracies in the databases. The major issue is that of an

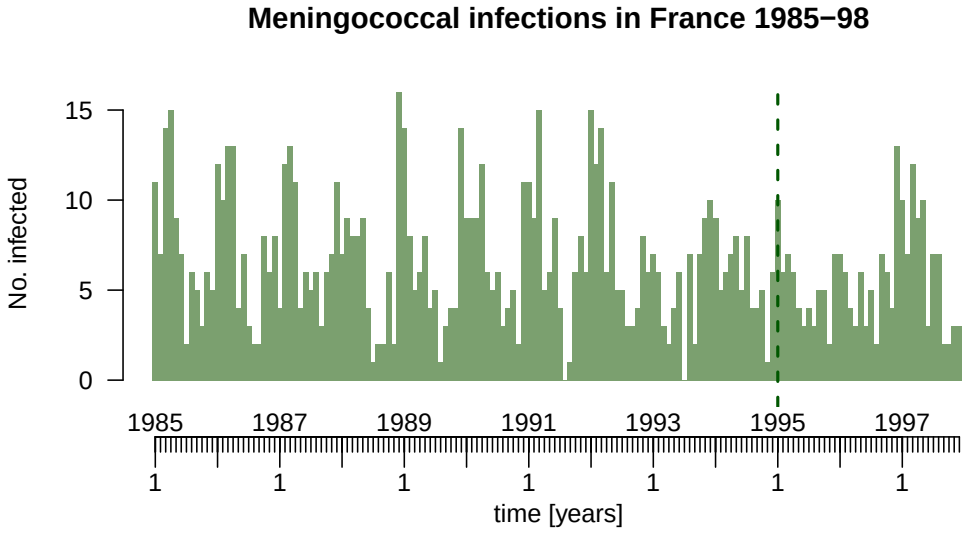

Figure 3.1: Monthly counts of Meningococcal infections in France 1985–1998 in the age group over 20 years (Data source: Höhle, 2007)

unclear case definition which results in a data-bias. This may depend on the one hand on the definition itself and on the other hand on the handling by the person involved in the reporting system.

Underreporting and Bias

Theoretically, public health surveillance data is assumed to be a case register which includes a complete list of all cases of a particular disease. In practical terms, the reports are incomplete and "cases reported are only the top of the iceberg' (Straif-Bourgeois and Ratard, 2005). Different stages of the disease may not be included in the surveillance data such as

- cases diagnosed, but not reported,
- cases which seek medical attention, but were not diagnosed or get misdiagnosed,
- cases which were symptomatic, but did not seek medical attention, and
- cases which were asymptomatic.

To which group a non-reported case belongs to is not random. In this context, Kleinman and Abrams (2008) speak about a syndromic surveillance system. Especially the cases of children as well as of weak and ill patients are more likely to be reported and examined in a more exact manner. This underreporting of other groups causes a bias within the data.

Reporting Delays

A reporting delay is the time between the time of infection and the report. This time period includes the incubation period, which differs by the type of disease, the

onset of the illness, the diagnose of a doctor, an additional diagnose in a laboratory and finally the report (see section 2.3.3).

A special type of delay results from district-specific holidays where most doctors, laboratories and public health departments are closed. The majority of the cases are reported after the holidays which could result in a bias in the data as well (see e.g. the artefact of reporting at the turn of the year in Figure 2.8 in section 2.3.3).

Thus the reporting delay, it's bias and it's effects need to be investigated formally. Unfortunately, hardly any statistical method for surveillance is able to consider the reporting delays properly.

Trends and Seasonality

Usually, there are structural changes in a time series over time. Therefore, time series are seen as a composition of the components: level, trend, seasonality, and errors. As a result, an aberration detection algorithm should be able to handle trend and seasonality.

Presence of Past Outbreaks

Case registers include cases of infections which belong to outbreaks as well. In order to model the natural appearance of a disease, it is desirable to exclude these cases. Even with a variable which is indicating outbreak, it is impossible to investigate all outbreaks with all associated cases. Farrington et al. (1996) suggested a reweighting procedure to correct the past for outbreaks.

Further Possible Influencing Variables

Morabia (1996) raised the question of what to monitor. He suggested to collect data not only in context of the disease, but also data of possible risk factors such as dietary data. Furthermore, other covariate processes such as climate parameters which explain the incidence by conditions of reproductions for the germs could be considered. With existing procedures this is not possible, but in chapter 4 an algorithm is introduced, which is able to consider covariates.

Repeated on-line Analysis

For prospective on-line analysis the data is collected sequentially and repeated analyses over time is done. Thereby, an appropriate handling of correlated observations and repeated decision problems is necessary.

Importance of Considering Data Characteristics

Therefore, surveillance data is characterized by trends, seasonality, and inaccuracies through underreporting, delays, and presence of past outbreaks. The data is on-going and electronically reported. In the following, the presented structure and

problems with surveillance data are considered in the selection of an adequate statistical method for surveillance. No golden rule is given, so that various approaches are studied.

3.1.3 Basic Setting of Surveillance Methods

This section introduces notation and general strategy of monitoring methods. Furthermore, an overview of existing methods is given. This list is not intended to be exhaustive. Moreover, an overview of a variety of possible solutions to the previously described problems with surveillance data is presented.

Notation and Point of Origin

General surveillance analysis is based on public health data, i.e. data sets collected by governmental institutions. These data sets are often aggregated by week or month of report. The result can be seen as an univariate time series of count data. The variable is of count nature: $y_t, t = 1, 2, \ldots, T$, where T is the current time point or date. In some methods, further covariates, i.e. the week, season, proportion of gender or region, can be considered. Such covariate processes are denoted by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$.

If one is interested in monitoring different strata simultaneously, the setting is based on a multivariate time series represented by $\boldsymbol{y}_t, t = 1, 2, \ldots, T$ with $\boldsymbol{y}_t = (y_{t1}, y_{t2}, \ldots, y_{tn})'$, while n being the number of strata, i.e. regions or age groups. During the analysis of these data it is important to recognize that correlations between the groups are taken into account.

This thesis focuses on univariate time series. Therefore, the methods are introduced only for the univariate case. Different univariate scenarios are simulated in chapter 5 and the aggregated time series of Campylobacter infections in Germany is monitored in chapter 6.

General Approach

Most of the surveillance methods follow the same general strategy. Outbreak detection means the detection of aberrations, therefore at a certain time occurs an important change in the time series. Thus, the process is assumed to be in-control until an unknown time point τ. Note the retrospective behaviour of the detection: at each time point one is only allowed to look back in time, never ahead in time.

To detect time points of aberration it is necessary to solve a binary decision problem: if everything is normal or a change in the structure of the process has occurred (Höhle and Mazick, 2010). Each time point t of the time series is in one of the two states. If the counts of a disease can be assumed as coming from a state of normality the process is termed to be in-control at time t when $t < \tau$. Otherwise, if the time point is later than the point of change, so $t > \tau$, the process is out-of-control. The true but unknown state of process is denoted by z_t, which takes the values 0 and 1, where 1 indicates an out-of-control time point.

Which kind of change should be detected depends on the type of application. The easiest type is the step change where a parameter changes from one fixed level to another one. Other types are gradual, linear change, or an exponential increase (Sonesson and Bock, 2003). Considering this, a threshold ξ is defined as a fixed threshold or varying threshold with respect to time, season, or covariate variables.

Finally, a detection method is a rule, which predicts the unknown state of z_t based on the observations $\boldsymbol{y} = (y_1, \ldots, y_t)'$. Consequently

$$\hat{z}_t = I(r(\boldsymbol{y}) > \xi),$$

where $I(\cdot)$ is the indicator function and $r(\cdot)$ a summary statistic (Höhle and Mazick, 2010). In order to find an appropriate summary statistic and threshold ξ, it is distinguished between two strategies: methods based on reference values and approaches inspired by statistical process control, which are introduced in the following.

3.1.4 Methods based on Reference Values

First of all, the time series is modelled with the purpose to determine the usual in-control properties of the time series. Thereby, trend, season, or/and influence of other variables describing the time series is derived. In general, this knowledge is used to create a prediction $\hat{y}_T$. The predicted value is then compared with the observed value of disease counts.

To create the prediction a set of reference values is used

$$R(w, w_0, b) = \left(\bigcup_{i=1}^{b} \bigcup_{j=-w}^{w} y_{T-ip+j} \right) \cup \left(\bigcup_{k=-w_0}^{-1} y_{T+k} \right), \tag{3.1}$$

where b is the number of years to go back in time, w the number of weeks around t to be included from these previous years, and w_0 the number of previous weeks in the current year, typically $w = w_0$ (Höhle, 2007). Parameter p represents the frequency of reports during a year, so that $p = 52$ refers to a weekly and $p = 12$ to a monthly data basis.

Example. The construction of the reference set is illustrated in Figure 3.2. The years since 1985 until 1994 are assumed to be the training data, while the data from 1995 until 1998 will be monitored later on. The reference set with $b = 6$ and $w = 2$ for the first point in the evaluation time period is highlighted. It is obvious, that more than half of the data is not considered for surveillance.

$\diamond$

The resulting prediction is compared with the observed value, whereby a decision is made, if an alarm is to be triggered. In the following, some specific methods to perform this prediction are summarized.

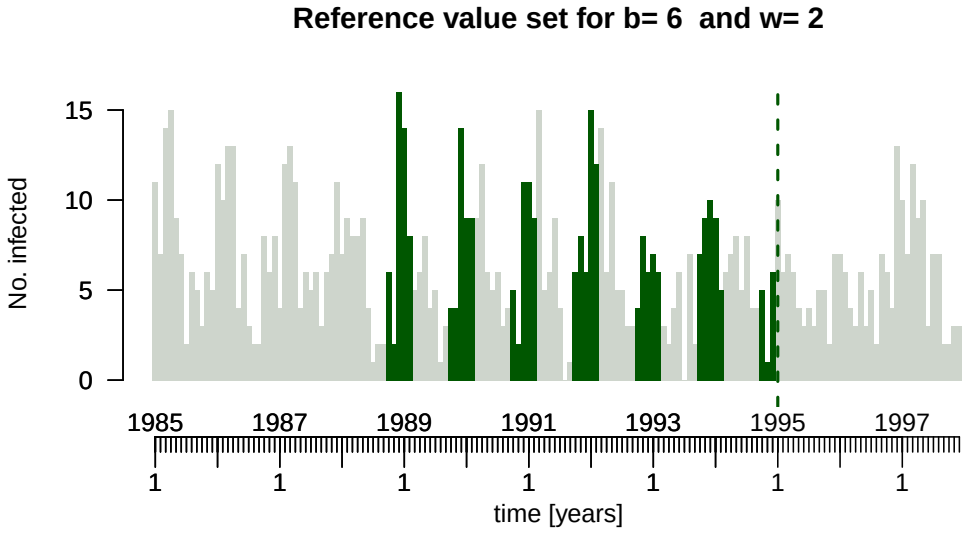

Figure 3.2: Illustration of reference values set generation in the example dataset of meningococcal infections in France

The System used at the Robert Koch-Institute, Germany

The method used at the RKI is inspired by Stroup et al. (1993). It observes the progression of the disease counts in a moving window of a fixed length in comparison with an expected number based on the previous years. The reference set is composed by the weekly incidence of previous w_0 weeks and parallel weeks in the previous years (Höhle, 2007). The window range depends on the disease, but typical values are $w = 3$ or $w = 4$. Note, that windowing implies a loss of information due to the reduction of the time series information. Furthermore, the selected window length has a strong influence on the outcome (Sonesson and Bock, 2003).

Farrington Algorithm

Farrington et al. (1996) introduced an algorithm finding a threshold based on the prediction intervals for expected values by a generalized linear model of the baseline rate. The automated algorithm should be applicable to various types of infections, thus should produce accurate results both for rare and common diseases. The method is described in detail in section 3.2.

Hierarchical Time Series Algorithm

Heisterkamp et al. (2006), ten years later than Farrington et al. (1996), proceeds essentially with the same strategy and compare the observed counts of each time point with the predicted count of that particular time point. By usage of different prior models for the intercept in a generalized linear model, an hierarchical model is obtained. Therefore, the Heisterkamp algorithm can be seen as an advancement of Farrington's algorithm. For a more detailed description of this algorithm see chapter 4.

A Bayesian Predictive Posterior Approach

According to Höhle (2007) a Bayesian approach is introduced. The reference values $R(w, w_0, b)$ are assumed to be Poisson distributed $y_i|\lambda \overset{\text{iid}}{\sim} \text{Po}(\lambda)$ with $\lambda \sim \Gamma(\kappa, \nu)$ being the conjugated prior distribution for λ. Accordingly, the predictive posterior of y_T can be simplified to a negative binomial distribution (Held, 2008).

$$y_{T+1}|y_1, \ldots, y_T \sim \text{NegBin}\left(\kappa + \sum_{t=1}^{T} y_i,\ \frac{\nu + T}{\nu + T + 1}\right).$$

Basically, a threshold is calculated using the α-quantile of the resulting predictive posterior distribution ξ_α. If the observed value y_{T+1} is greater than the threshold ξ_α an alarm is triggered at time $T + 1$.

The Assets and Drawbacks

The methods based on reference values handle seasonality automatically due to using corresponding values of the previous years. But, they use only a small subset of available data (see Figure 3.2), so that the methods based on reference values tend to be sub-optimal (Höhle, 2007). A critical assumption is that the base-line rate of the disease rate is known (Sonesson and Bock, 2003). Furthermore, no auto-correlation is taken into account.

3.1.5 Algorithms Inspired by Statistical Process Control

Other approaches are inspired by statistical process control techniques. The basic setting assumes that the observations during an in-control state has a specific distribution with mean μ_1, e.g. normal distribution $x_1, \ldots, x_T \overset{\text{iid}}{\sim} N(\mu_1, 1)$. A change of one level to another one is detected if the distribution changes to another mean μ_2, e.g. $N(\mu_2, 1)$ (see Hawkins and Olwell, 1998).

In surveillance models, cases are assumed to follow a Poisson process or a more complex time-varying process. Often likelihood ratios between in-control and out-of-control state are used to construct a decision rule. Assuming a Poisson process, an increased rate of incidence corresponds to an increased intensity of the process. The Poisson model can be replaced by a negative binomial model when over-dispersion occurs (Woodall, 2006).

Most common are methods based on cumulative sum statistics. The count data $y_1, \ldots, y_T \overset{\text{iid}}{\sim} \text{Po}(\lambda)$ is transformed into approximately normally distributed data. Furthermore, seasonality can be handled by letting λ vary over time by means of a periodical transformation (Höhle, 2007).

3.1.6 Detection using Search Engine Query Data

Even the method is not in common use and in the early stage of development, it presents an automated outbreak detection method which gives a basically different perspective to the problem. Thus, this part introduces an algorithm which is not

based on public health data and should outline the diversity of methods in the area of automated outbreak detection.
The search engine operator Google analysed their inquiries and modelled the influenza epidemic in several countries. The system assumes that patients first request the Internet with their problems before or even instead of consulting a physician. The usage of Internet data avoids the problem of reporting delays, because search queries can be proceeded quickly, so that outbreak could be detected on-time. (Grinsberg et al., 2009).

The model for prognosis is based on statistically related search queries to influenza during the influenza epidemics in the previous years. Therefore, a proper case definition is missing and only influenza-like illnesses can be prognosed. In Germany, so far, just the spatial view at the level for the federal states is possible. Furthermore, it is debatable whether the method is applicable to rare diseases. Therefore, the prognosis cannot replace public health surveillance, but may add information and help to solve some problems, especially reporting delay, with surveillance data.

Example. Figure 3.3 displays the Google Flu epidemic prognosis exemplary for Germany. Additional plottings of the acute respiratory data of Germany examine the fit, while not all data have been provided. In general, the prognosis is quite accurate, even though the increase during the year 2009 is overestimated. Note, that the case definition is not accurate.

Figure 3.3: Google Flu epidemic prognosis (blue line) compared to data of acute respiratory infection in Germany (yellow line). (Source: Google, 2010)

◇

3.2 Farrington Algorithm

Now, the algorithm of Farrington et al. (1996) is illustrated in detail. It is one of the mostly used methods in outbreak detection and will be evaluated later in comparison with the introduced hierarchical time series algorithm of Heisterkamp et al. (2006).

The algorithm was developed to have a robust and fast method applicable for the routine monitoring of weekly reports on infections at the former Communicable Disease Surveillance Centre (now Health Protection Agency), the national public health

department in the United Kingdom (Farrington et al., 1996). The method was designed for all organisms which includes diseases with sporadic cases, therefore low counts, but as well for diseases with large counts. The primary interest was to find aberrations of abnormally high number of observed cases.

As displayed in the last section, the strategy belongs to the group of methods which is based on reference values. First, an overdispersed Poisson log-linear model is fitted which results in a prediction for expected value. After that, thresholds are calculated to define an in-control area.

3.2.1 The Algorithm

The algorithm's strategy can be summarized in five steps. Here, an overview is given and in the following sections each step is explained in more detail.

1. Fit an initial generalized linear model and calculate estimates for expectation μ_t and dispersion ϕ.
2. Calculate past outbreak correcting weights with Anscombe residuals and refit the weighted model.
3. Calculate a revised estimate $\hat{\phi}$ and rescale the model.
4. Check if the trend component is significant. If not, repeat fitting procedure without time trend.
5. Calculate threshold.

Step 1: Fit Generalized Linear Model

Using the reference values $R(w, w_0, b) \subseteq \{y_1, \ldots, y_T\}$ a prediction for the current time point T is calculated. For prediction a generalized linear model is fitted, where the baseline count y_t corresponds to any time point $t < T$.

The linear predictor is defined by the assumptions of the influencing variables in the statistical model. Without loss of generality, only a linear time trend is included, so that the linear predictor is $\eta_t = \beta_0 + \beta_1 t$. Furthermore, a log-linear relation is assumed between the linear predictor η_t and the mean counts at one time point $\mu_t = \mathrm{E}(y_t)$. Therefore, the monotone link function of the generalized linear model is

$$\log(\mu_t) = \beta_0 + \beta_1 t, \tag{3.2}$$

The last component describing a generalized linear model uniquely is a probability distribution. Public health data are counts of infections with large variance. Therefore, without loss of generality, a quasi-Poisson model $y_t \sim \mathrm{Po}(\mu_t)$ with $\mathrm{Var}(y_t) = \phi\mu_t$ is assumed.

Following the theory of generalized linear models (Fahrmeir et al., 1996), the parameters are estimated by a quasi-likelihood method. In particular, the dispersion

parameter ϕ is estimated by

$$\hat{\phi} = \max\left\{\frac{1}{n-p}\sum_{t=1}^{T}\omega_t\frac{(y_t-\hat{\mu}_t)^2}{\hat{\mu}_t}, 1\right\}, \tag{3.3}$$

where ω_t is a weight, which is discussed later, and p is the number of predicted parameters in the linear predictor (3.2), e.g. $p = 2$ if the time trend is included and $p = 1$ without time trend. Assuming no over-dispersion, one would have $\hat{\phi} = 1$, the expected counts for the current week T is estimated by

$$\hat{\mu}_T = \exp(\hat{\beta}_0 + \hat{\beta}_1 T).$$

This model can be modified by removing the non-significant time trend or adding a seasonal component. If a seasonal component is present, the prediction bases only on counts from comparable periods in the past year, $R(w, w_0, b)$.

Example. Using the reference set with $w = 2$ and $b = 4$, as displayed in Figure 3.2, a generalized linear model is fitted to the time series of meningococcal infections in France. The estimation for the intercept is $\hat{\beta}_0 = 2.04$, while the time trend does not have a significant influence. Therefore, it is removed from the model. The dispersion parameter is estimated to be $\hat{\phi} = 1.81$. It follows the corresponding summary of the model in a R-Output.

```
Call:
glm(formula = ref$observed ~ 1, family = "quasipoisson")

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.04083    0.08453   24.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.814752)
```

Thus, the first time point of the evaluation period $T = 121$ is predicted, based on the defined reference set, one obtains $\mathrm{E}(y_{121}) = \exp(\eta_{121}) = \exp(2.04) = 7.70$. The prediction is shown in Figure 3.4. The reference points included in the model are displayed with blank dots, while the full progress is drawn by a grey line. The observed value is highlighted by a filled dot and is located a bit higher than its prediction.

◇

Step 2: Correction of Past Outbreaks

Past outbreaks cannot be identified completely. Therefore, Farrington et al. (1996) suggest a reweighting procedure to reduce the influence of high base-line counts. Corresponding weights in the generalized linear model (as mentioned in formula (3.3)) are defined by standardized Anscombe residuals

$$s_t = \frac{3}{2\hat{\phi}^{1/2}}\frac{y_t^{2/3} - \hat{\mu}^{2/3}}{\hat{\mu}^{1/6}(1-h_{tt})^{1/2}},$$

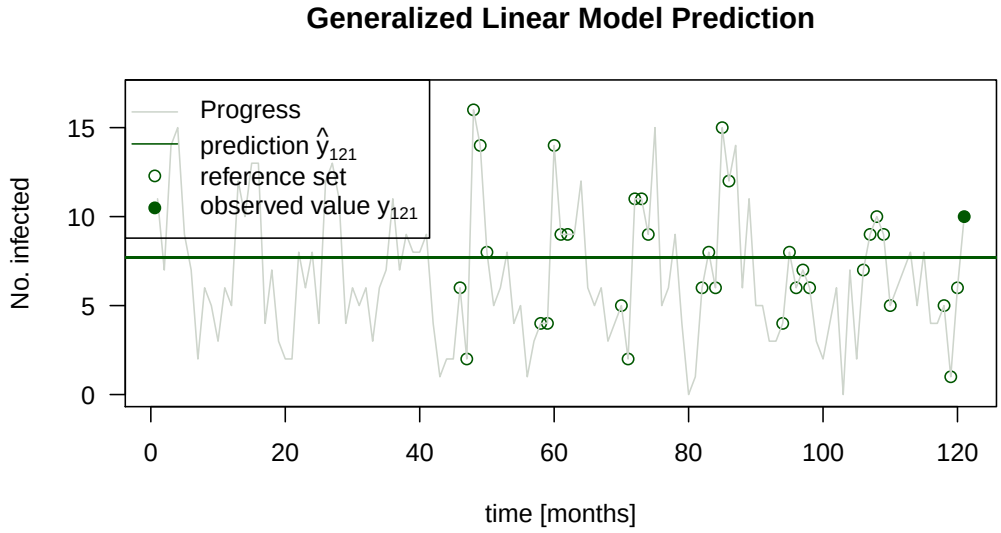

Figure 3.4: Illustration of generalized linear model fit in the example dataset of meningococcal infections in France

where h_{tt} are the diagonal elements of the hat matrix, so that the weights are

$$\omega_t = \begin{cases} \gamma s_t^{-2} & \text{if } s_t > 1, \\ \gamma & \text{otherwise,} \end{cases}$$

where γ is a normalization constant, so that $\sum \omega_i = n$. Hence, low weights to counts with large residuals are given. However, note that this shrinks not only the mean, but also the estimate of variance.

Example. In the time series of meningococcal infections in France between 1985 and 1998 no outbreaks are removed or marked. Therefore, a correction of past outbreaks is advised. Figure 3.5 shows the influence of adjusting the data by the described procedure. The threshold based on reweighted data due to the standardized Anscombe residuals is consistently lower.

◇

Steps 3 and 4: Rescaling and Checking the Model

With the calculated weights ω_t the generalized linear model is refitted and with a revised dispersion estimate $\hat{\phi}$ the model is rescaled. If the time trend is not significant, this component is omitted and the procedure is repeated.

Step 5: Calculation of Threshold

The threshold calculation is based on an assumed normal-distributed prediction of y_T. Farrington et al. (1996) introduced the algorithm with a two-sided prediction interval, but the general aim in monitoring of infectious diseases is the detection of

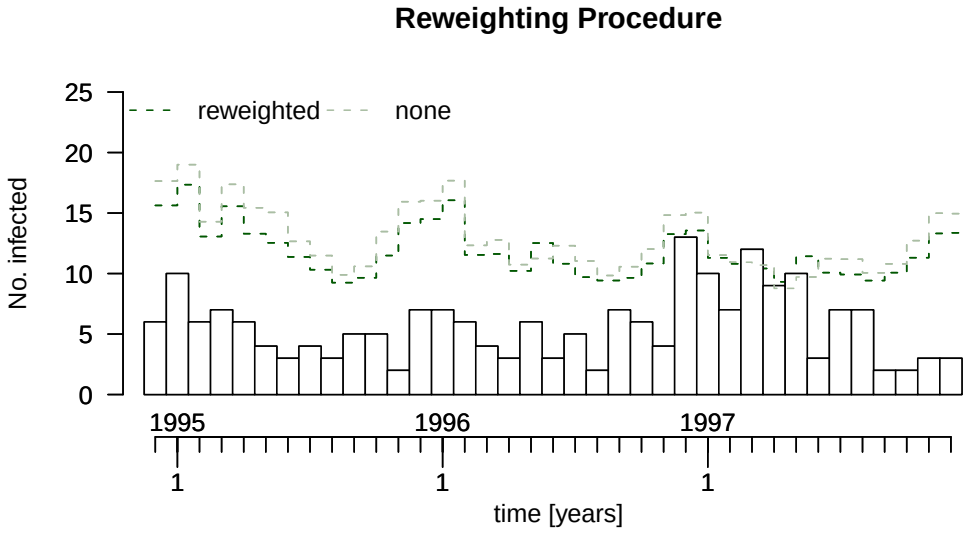

Figure 3.5: Illustration of influence of past outbreak correction in threshold calculation in the example dataset of meningococcal infections in France

aberrations of abnormally high counts. Therefore, an one-sided prediction interval for the current count y_T is constructed. Then, the prediction interval without any reweighting for y_T results in the general form

$$U_{\text{none}} = \text{E}(y_T) + z_{1-\alpha}\sqrt{\text{Var}(y_T - \hat{\mu}_T)},$$

where $z_{1-\alpha}$ is the $100\% \cdot (1-\alpha)$-quantile of the normal distribution, and $\text{Var}(y_T - \hat{\mu}_T)$ the prediction error variance. If no skewness transformation for y_T is necessary it follows with y_T and $\hat{\mu}_T$ independent, that

$$\begin{aligned} U_{\text{none}} &= \hat{\mu}_T + z_{1-\alpha}\sqrt{\text{Var}(y_T) + \text{Var}(\hat{\mu}_T)} \\ &= \hat{\mu}_T + z_{1-\alpha}\sqrt{\phi\hat{\mu}_T + \text{Var}(\hat{\mu}_T)} \\ &= \hat{\mu}_T \left\{ 1 + z_{1-\alpha}\sqrt{\frac{\phi\hat{\mu}_T + \text{Var}(\hat{\mu}_T)}{\hat{\mu}_T^2}} \right\} \\ &= \hat{\mu}_T \left\{ 1 + z_{1-\alpha}\sqrt{\frac{\hat{\tau}}{\hat{\mu}_T}} \right\} \qquad \text{with } \hat{\tau} = \phi + \frac{\text{Var}(\hat{\mu}_T)}{\hat{\mu}_T}. \end{aligned}$$

A common problem in diseases with low counts is a highly skewed distribution. Therefore, a $\frac{2}{3}$-power-transformation is studied. This skewness correction is based on the transformation function $g(x) = x^{2/3}$ with derivative $g'(x) = \frac{2}{3}x^{-1/3}$. Using the assumption of the quasi-Poisson model $y_T \sim F(\mu_T, \phi\mu_T)$ and by applying the

Δ-rule one obtains

$$\begin{aligned} \mathrm{E}(y_T^{2/3}) &= \mu_T^{2/3}, \\ \mathrm{Var}(y_T^{2/3}) &= \mathrm{Var}(g(y_T)) = [g'(E(y_T))]^2 \mathrm{Var}(y_T) \\ &= [\frac{2}{3}\mu_T^{-1/3}]^2 \phi \mu_T = \frac{4}{9}\phi \mu_T^{1/3}, \\ \mathrm{Var}(\hat{\mu}_T^{2/3}) &= \frac{4}{9}\mu_T^{-2/3}\mathrm{Var}(\hat{\mu}_T). \end{aligned}$$

The prediction error variance results as

$$\begin{aligned} \mathrm{Var}\left(y_T^{2/3} - \hat{\mu}_T^{2/3}\right) &= \mathrm{Var}(y_T^{2/3}) + \mathrm{Var}(\hat{\mu}_T^{2/3}) = \frac{4}{9}\ \mu_T^{1/3}\left[\phi + \frac{\mathrm{Var}(\hat{\mu}_T)}{\mu_T}\right] \\ &= \frac{4}{9}\ \tau \mu_T^{1/3}, \qquad\qquad \tau = \phi + \frac{\mathrm{Var}(\hat{\mu}_T)}{\mu_T}. \end{aligned}$$

With the transformed parameters, the prediction interval construction follows the same strategy like in the case without transformation.

$$\begin{aligned} U_{2/3}^{2/3} &= \hat{\mu}_T^{2/3} + z_{1-\alpha}\ \sqrt{\widehat{\mathrm{Var}}(y_T^{2/3} - \hat{\mu}_T^{2/3})} = \hat{\mu}_T^{2/3} + \frac{2}{3}\ z_{1-\alpha}\hat{\mu}_T^{1/6}\ \sqrt{\hat{\tau}} \\ \Rightarrow\ U_{2/3} &= \left\{\hat{\mu}_T^{2/3} + \frac{2}{3}\ z_{1-\alpha}\hat{\mu}_T^{1/6}\ \sqrt{\hat{\tau}}\right\}^{3/2} = \hat{\mu}_T\left\{1 + \frac{2}{3}\ z_{1-\alpha}\sqrt{\frac{\hat{\tau}}{\hat{\mu}_T}}\right\}^{3/2}, \end{aligned}$$

Those observations lying outside this interval are considered to be an aberration.

Another skewness correction by a square-root transformation results in the upper prediction interval border

$$U_{1/2} = \hat{\mu}_T\left\{1 + \frac{1}{2}z_{1-\alpha}\left(\frac{\hat{\tau}}{\hat{\mu}_T}\right)^{1/2}\right\}^2.$$

Example. In Figure 3.6 the different thresholds are shown. `'2/3'` displays the upper bound corrected by skewness correction in low count scenario, `'1/2'` by variance stabilizing square-root transformation, and `'none'` by no transformation.

$\diamond$

3.2.2 Enhancements and Limitations Discussion

The Farrington algorithm is very sensitive, and detects even small increases of rare infections, while in common disease processes only large excesses are detected. Therefore, small localized outbreaks of common diseases are unlikely to be identified (Farrington et al., 1996).

The method is able to handle trend and seasonality in a robust way. A trend component should be added only if the data basis covers more than three years in order to avoid unrealistic extrapolations.

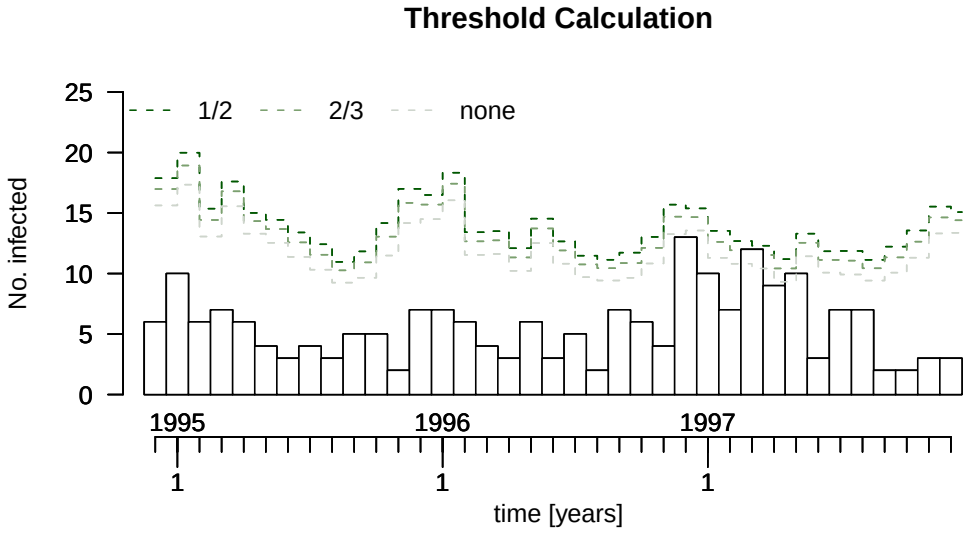

Figure 3.6: Illustration of different transformations in threshold calculation in the example dataset of meningococcal infections in France.

The distributional assumptions for the generalized linear model should be validated and replaced if necessary, like it is usually done in model-fitting.

The problem with the Farrington algorithm is the missing guidance on how to choose the values of α, b and w. The user has to have strong content-related evidence or large experience to choose these parameters. For significance level α, it is advised to determine it empirically to keep the number of detections to a manageable level (Farrington et al., 1996). A simulation study to the significance level is introduced in section 5.3.

An examination of Farrington on the behaviour of the past outbreak correction shows that it 'substantially alleviates the effect of past outbreak but does not eliminate it' (Farrington et al., 1996). But Höhle (2008) showed that the downweighting of large positive outliers not only shrinks the variance due to $\text{Var}(y_t) = \frac{\phi\mu_t}{\omega_t}$, but the prediction of mean as well.

Example. For the example, in Figure 3.7, it can be shown, that the predicted value $\hat{y}_{121} = 7.07$ is lower than before in the unadjusted fit, where it has been $\hat{y}_{121} = 7.70$. The observed value is much higher.

$\diamond$

The exceedance score defined as

$$\delta = \frac{y_T - \hat{\mu}_T}{U - \hat{\mu}_T}$$

is set to zero, if fewer than five reports were reported at the past four time points.

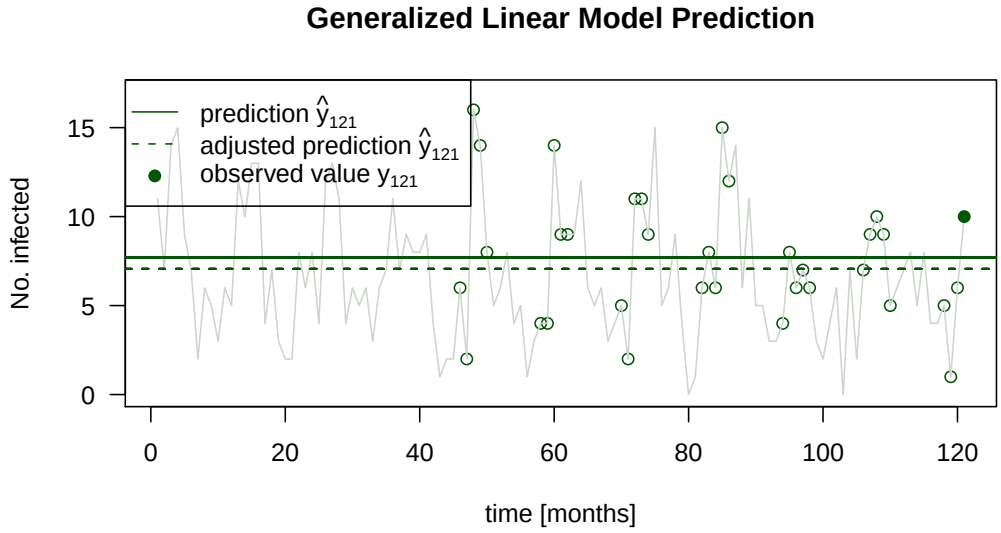

Figure 3.7: Illustration of influence of past outbreaks detection to the generalized linear model fit in the example dataset of meningococcal infections in France

The method is not able to consider the effect of delays, thus Farrington et al. (1996) refers to the importance of their prior investigation.

Each time point is evaluated separately, so that a continuous observation of the time series is not possible. Furthermore, serial correlations between base-line counts are ignored. Therefore, non-epidemic conditions are assumed which is given for infrequent diseases. Farrington et al. (1996) show that the bias ignoring the serial correlation is very small .

3.3 Evaluation of Performance

Work is not done once the system flags an outbreak. This section will examine the magnitude of evaluation, their criteria, and central key parameters.

3.3.1 Criteria for Evaluation of Surveillance Systems

Evaluation of surveillance programs is based on the criteria usefulness, cost, and quality, while the improvement of the one may impose or compromise the other (Thacker and Berkelman, 1988). In the following, relations are described and afterwards central criteria are defined.

Usefulness

The measurement of usefulness is inexact and it is the practicability and early detection of an outbreak which leads to effective intervention and provides information used for preventive purposes. For instance, the computing time of a monitoring algorithm is still relevant, because algorithms should be practical and not too time consuming.

Costs

It can be distinguished between fixed costs for running a surveillance system and variable costs. In case of an alarm, follow-up activities such as further diagnosis, case-management, or community interventions are necessary. The financial and public health costs of missing outbreaks (entirely or later) should not be ignored. To reduce the variable cost an optimal surveillance method regarding its quality is needed (Buehler et al., 2004).

Quality

There are eight attributes of quality: sensitivity, specificity, predictive value positives, representativeness, timeliness, simplicity, flexibility, and acceptability (Thacker and Berkelman, 1988).

By controlling the values of sensitivity, specificity, and predictive value positives the number of detections are kept in a manageable range while the probability of an aberration detection is kept high. Representativeness refers to the surveillance data quality, which is confronted with underreporting and bias. It ensures that the occurrence and distribution of cases represent the true situation in the population. Timeliness means to detect outbreaks as soon as possible to initiate interventions. A general claim to statistical methods is to be as simple as possible while still meeting their objectives. If a surveillance system is able to adapt changing operating conditions or random variability of trend is measured by flexibility. Acceptability reflects the willingness of participants and stakeholders to contribute to the data collection and analysis (Buehler et al., 2004).

Thereby, special importance for the evaluation of surveillance methods for infectious diseases have the criteria of sensitivity, specificity, predictive value positives, and timeliness, which are described in detail in section 3.3.3.

3.3.2 Choice of Evaluation Data

For most quality criteria a 'gold standard' from an alternative data source to confirm occurring outbreaks is necessary. Not all outbreaks are recognized in the public health system. With capture-recapture techniques estimations for outbreaks missed by the surveillance system can be achieved (Buehler et al., 2004).

Furthermore, it is possible to simulate historical data in different outbreak scenarios and to evaluate the performance of an algorithm. Höhle (2007) provides a method for simulation using hidden Markov models, which is described in detail in section 5.1.1. Beside this, Hutwagner et al. (2005) introduced simulated data of various scenarios which have been adjusted by typical irregularities and superimposed by different outbreak types. The data is provided by the CDC and section 5.4.1 puts a finer point on it. However, simulations are limited in their ability to imitate the diversity and unpredictability of real-life events (Buehler et al., 2004).

3.3.3 Key Parameters in Evaluation of Infectious Disease Outbreak Detection Methods

Validation of the surveillance system, i.e. data sources, case definitions, statistical methods, and timeliness of reporting, can provide indirect evidence of system performance (Buehler et al., 2004). Here, a focus is drawn on the statistical assessment of outbreak detection quality, especially the key parameters of sensitivity, specificity, predictive value, and timeliness.

It has to be clear in one's mind, that these measures are not known in real disease time series as long as the true outbreak state is unknown, and therefore can be used only for simulated data sets.

First, the general notation is introduced (see Frisén, 2003). The observed process is denoted by $y_t, t = 1, 2, \ldots, T$. The state of process is denoted by z_t. Assuming a process with step change, the in-control set $C(s) = \{\tau \leq s\}$ implies the acceptable value $z_t = z^0$ for $t = 1, 2, \ldots, \tau - 1$ and the out-of-control set $D(s) = \{\tau > s\}$ implies an unacceptable value $z_t = z^1$ for $t = \tau, \tau + 1, \ldots, T$. The time point of the first alarm generated by the surveillance method is a random variable defined as

$$t_A = \min\{s | r(s) \geq \xi\},$$

where $r(s)$ is an alarm statistic at time point s, and ξ is the selected threshold.

On this basis, the selected common quality criteria are introduced subsequently.

Sensitivity and Specificity

There are two kinds of errors in aberration detection: a false positive result, meaning an alarm is triggered at an in-control time point, and the false negative, meaning no alarm is triggered if an outbreak is present. Therefore, different values are defined (see Table 3.1): the number of correct found outbreaks TP $= |\{A(s)|s \in C(s)\}|$, of false found outbreaks FP $= |\{A(s)|s \in D(s)\}|$, of correct found non-outbreaks TN $= |\{\overline{A(s)}|s \in C(s)\}|$, and the number of not-detected outbreaks FN $= |\{\overline{A(s)}|s \in D(s)\}|$ (see Table 3.1).

Alarm \ Outbreak	yes	no	
TRUE	TP	FP	→ Positive predictive value
FALSE	FN	TN	→ Negative predictive value
	↓	↓	
	Sensitivity	Specificity	

Table 3.1: Illustration for detection rates

Aggregating these numbers to rates the criteria of sensitivity and specificity can be deduced. Sensitivity, as the ability to identify every single case of outbreak, is the true positive rate, i.e. the total number of correctly flagged outbreaks, divided by

the total number of outbreaks.

$$\text{se} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

Specificity is the true negative rate, i.e. the rate of correctly non-detection of non-outbreak time points.

$$\text{sp} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Plotting the $(1 - \text{sp})$ versus se across the values of ξ generates a receiver operating characteristic (ROC) curve. The Euclidean distance between this curve to $(0, 1)$ can be used as a measure for optimality.

Predictive Value of an Alarm

The positive predictive value is $\text{PV}(s) = P(C(s)|A(s))$ is the proportion of alarms that actually have been an outbreak event (see Table 3.1). Therefore, it is a general indicator for the uncertainty in a triggered alarm. A low predictive value means a large number of misclassification, erroneous conclusions and unnecessary interventions (Straif-Bourgeois and Ratard, 2005).

In passive surveillance the predictive value increases to one as time s increases. In active surveillance, which means the process is stopped if an alarm is triggered, the predictive value has a limiting value of less than one.

Average of the Run Length (ARL)

A generally used quality measure is the average of the run length (ARL). Hence, ARL^0 is defined as the expected value of in-control run length until the first alarm in a system of surveillance where no change occurs

$$\text{ARL}^0 = \text{E}(t_A|\tau = \infty),$$

and ARL^1 is the average out-of-control run length, where a true change occurred at the same time surveillance started,

$$\text{ARL}^1 = \text{E}(t_A|\tau = 1).$$

It can be understood as the expected time to detect an outbreak which already occurred. Optimality is found by minimizing ARL^1 holding ARL^0 fixed.

Expected Delay

A right decision could be useless if it is made too late. Therefore, the expected delay matters in the evaluation of timeliness. The expected delay is the time between the change τ and the time of alarm t_A summarized using the expectation with regard to the distribution of τ (Frisén, 2003).

$$\text{ED} = \text{E}[\text{ED}(\tau)] = \text{E}\left[\text{E}[\max(0, t_A - \tau)|\tau]\right].$$

Therefore, ARL^1 can be defined in terms of the expected delay $\mathrm{ARL}^1 = \mathrm{ED}(1) + 1$.

For prediction of the expected delay, the beginning of each outbreak chain is located. With this, the difference between the beginning of the outbreak and the first alarm is the specific delay. Furthermore, a penalty is defined as a maximum detection time range in case the outbreak is not detected. All delays of the monitored time series are averaged to obtain an estimate for the expected delay.

Example. The following Figure 3.8 explains how the expected delay is estimated. The specific observed lags are aggregated by its mean. Thus, $\mathrm{ED} = \frac{1}{n} \sum_{k=1}^{n} \mathrm{lag}_k$.

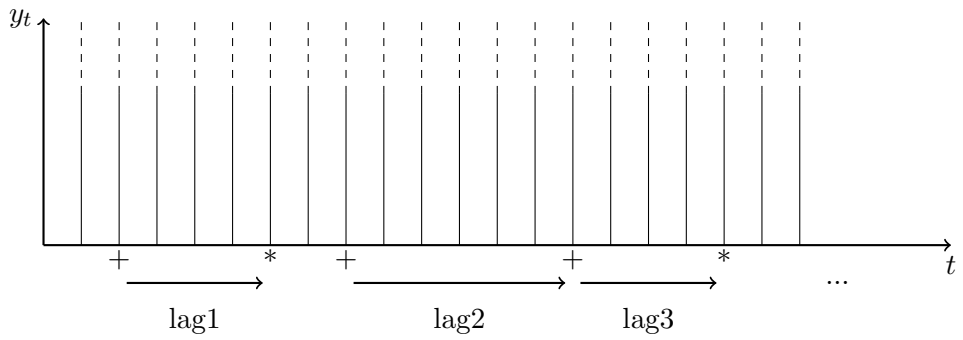

Figure 3.8: Illustration for estimation of expected delay where + indicates an outbreak and * the corresponding alarm

◇

Probabilities of a Successful Detection

Often there should be a limited time d from the change to the detection if a disaster is to be avoided (Frisén, 1992). Then, the probability of a successful detection is defined by

$$\mathrm{PSD}(\tau, d) = P(t_A - \tau \leq d | t_A \geq \tau),$$

which means the probability for an alarm delay smaller than a limit d (Frisén, 2003).

3.3.4 Comparisons between Algorithms

As presented before, there is a large variety of outbreak detection methods. Therefore, Kleinman and Abrams (2008) ask "how should system designers decide which detection method is the best for surveillance?".

The quality of outbreak detection may depend not only on the outbreak scenario, but on the specific characteristics of underlying data, as well. Therefore, it is necessary to define optimality referring to a given situation. In the last part a variety of criteria were introduced, while most important are specificity, sensitivity as well as expected delay for timeliness. Different surveillance methods can be compared using

these quality measures while the trade-off between them could be solved by content-related argumentation. Kleinman and Abrams (2008) suggested three-dimensional analogues of the ROC curve as well as including timeliness in a ROC curve by weighting the curve by the average proportion of time or lives saved by the detection. It is noted that for diseases with outbreaks affecting only a few cases the reweighting by the number of saved lives is not appropriate.

Chapter 4

Hierarchical Time Series Algorithm

After in the last chapter a selection of methods for monitoring surveillance time series were introduced, this chapter presents a hierarchical time series algorithm described by Heisterkamp et al. (2006). Firstly, the method and a newly developed full Bayesian version of it is described in detail. Secondly, technical details of implementation are given. Finally, a discussion of algorithm limitations is given. The algorithm's behaviour in application is investigated in the subsequent chapters 5 and 6, where simulation studies of different scenarios and the application to the Campylobacter time series shows its properties compared to the previously described RKI method, Farrington, and Bayes algorithm as established methods.

4.1 General View of the Algorithm

Heisterkamp et al. (2006) proposed an algorithm to improve the surveillance system in the Netherlands (Infectious disease Surveillance and Information System, ISIS). This system is based on daily laboratory reports of the test results of over 350 pathogens, which covered at that time laboratory results corresponding to about 20% of the Dutch population.

The algorithm can be seen as an extension of Farrington's algorithm (see section 3.2) and basically fits a generalized linear model with over-dispersion. The procedure can be summarized in three steps:

1. The model parameters are estimated using data from a predefined training period.
2. The expected values are updated for small time steps as new data arrive.
3. Thresholds are calculated conditionally on the expected value for the new time point.

An alarm is triggered when the threshold is exceeded.

Example. In the following, each step is described in detail and exemplified by the data displayed in Figure 4.1. The data are simulated from a hidden Markov model

(see section 5.1.1) such that the probability for switching into the outbreak state is 1%, while staying in the state is defined as 50%. If an outbreak appears the expectation of the counts is doubled. The training period is 2001–2003 while the years 2004 and 2005 will be monitored. The end of the training period is marked in the time series plot by a dashed line.

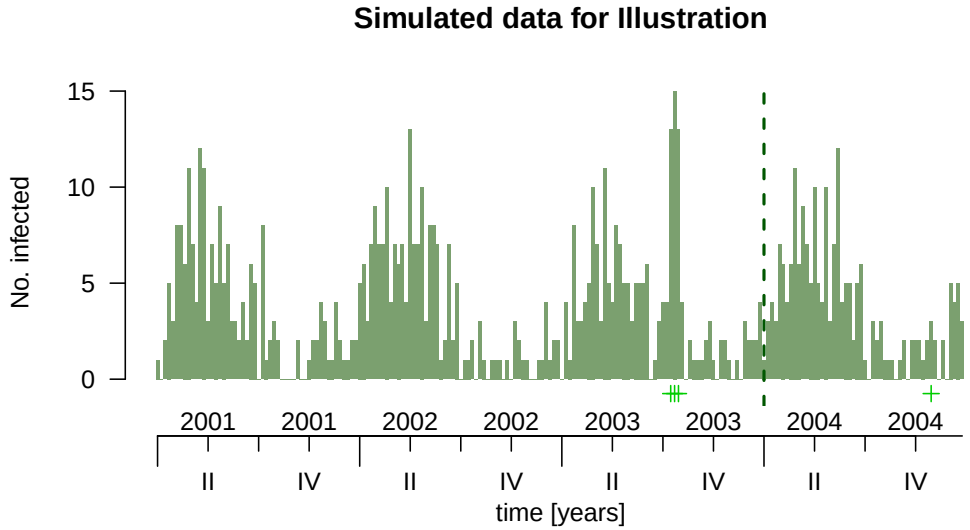

Figure 4.1: Illustrative data simulated by a hidden Markov model. The symbol + indicates an outbreak.

◇

4.1.1 Definition of the Hierarchical Time Series Model

Using a training period a hierarchical time series model is fitted, which is based on a generalized linear model and a stochastic model for the time varying parameter. Due to the count data nature of the time series, a Poisson distribution of the response is assumed.

$$y_t|\mu_t \overset{\text{iid}}{\sim} \text{Po}(\mu_t),\ t = 1, \ldots, T,$$

such that $\text{E}(y_t) = \mu_t$ and $\text{Var}(y_t) = \mu_t$. In case of over-dispersion a negative Binomial distribution $y_t|\mu_t \overset{\text{iid}}{\sim} \text{NegBin}(\mu_t, \alpha)$ with $\text{E}(y_t) = \mu_t$ and $\text{Var}(y_t) = \mu_t + \frac{\mu_t^2}{\alpha}$ can be used instead.

In the Poisson model, the expectation of the response is linked to the linear predictor by $\eta_t = \log(\mu_t)$, while the linear predictor is defined as

$$\eta_t = \beta_{0t} + \boldsymbol{x}_t'\boldsymbol{\beta}_x,\ t = 1, \ldots, T,$$

where $\boldsymbol{x}_t'$ is a row-vector of length p including time specific of covariates, and $\boldsymbol{\beta}_x = (\beta_1, \ldots, \beta_p)'$ the vector of regression coefficients. The intercept $\beta_0(t) = \beta_{0t}$ is

assumed to be time varying with dynamics governed by a stochastic process. Therefore, different models for this second level of the hierarchical model are investigated

$$\begin{aligned}
\text{stationary model:} \quad & \beta_{0t}|\beta_{0t-1},\ldots,\beta_{01} \sim \mathbb{N}(\beta_0,\lambda^{-1}), t=2,3,\ldots \\
\text{neighbour model:} \quad & \beta_{0t}|\beta_{0t-1},\ldots,\beta_{01} \sim \mathbb{N}(\beta_{0t-1},\lambda^{-1}), t=2,3,\ldots \\
\text{linear model:} \quad & \beta_{0t}|\beta_{0t-1},\ldots,\beta_{01} \sim \mathbb{N}(2\beta_{0t-1}-\beta_{0t-2},\lambda^{-1}), t=3,4,\ldots \\
\text{quadratic model:} \quad & \beta_{0t}|\beta_{0t-1},\ldots,\beta_{01} \sim \mathbb{N}(3\beta_{0t-1}-3\beta_{0t-2}-\beta_{0t-3},\lambda^{-1}), \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t=4,5,\ldots
\end{aligned}$$

Here, λ is the precision parameter of the normal distribution, i.e. the variance is equal to λ^{-1}. In summary, the models can be written by the usage of the d^{th} order difference operator Δ^d, i.e. $\Delta^d(\beta_{0t}) = \Delta^{d-1}(\beta_{0t}-\beta_{0t-1})$ for $d>1$ and $\Delta^0(\beta_{0t}) = \beta_{0t}$. In particular, for the stationary model is defined $\Delta^0(\beta_{0t}) = \beta_{0t} - \beta_0$.

$$\Delta^d(\beta_{0t})|\beta_{0t-1},\ldots,\beta_{01} \quad \sim \quad N(0,\lambda^{-1}), \quad \text{for } d=0,1,2,3. \tag{4.1}$$

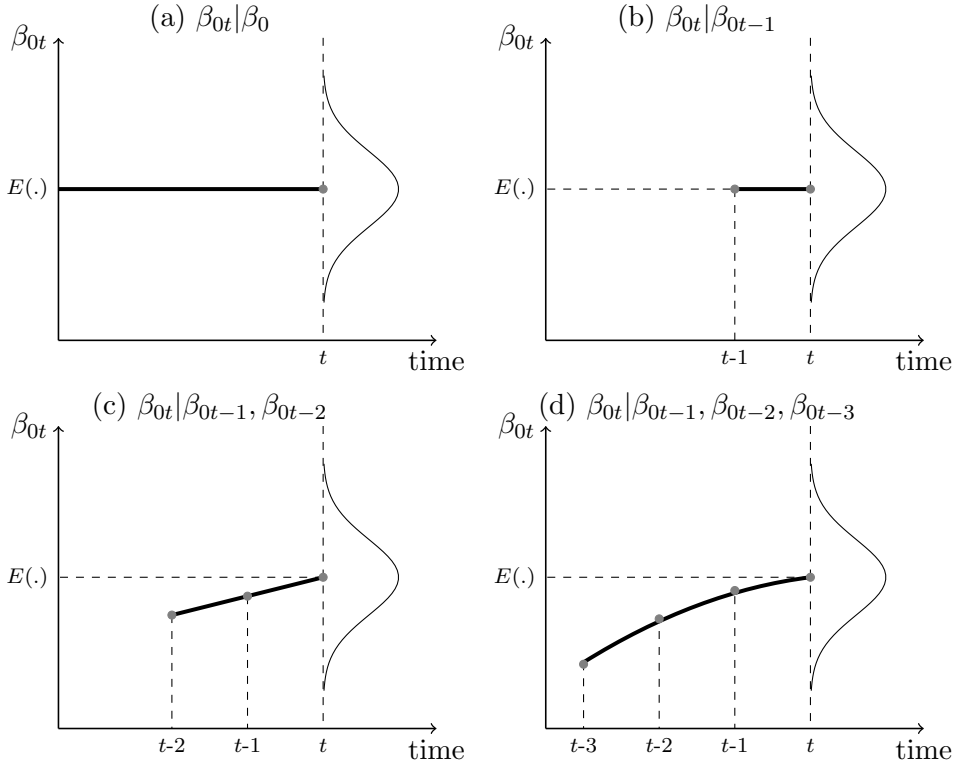

Figure 4.2: Illustration for different latent models for intercept: (a) stationary model, (b) neighbour model, (c) linear model, and (d) quadratic model while E(·) means the expectation for the current parameter $\beta_0 t$ given the corresponding previous values.

The latent models for β_{0t} can also be seen as random walks of a specific different order. In Figure 4.2, the trend characteristics are illustrated. The conditional distribution $\beta_{0t}|\beta_{0t-1},\ldots,\beta_{01}$ simplifies due to the order of the random walk and its Markov property. Choosing the stationary model corresponds to assuming a constant overall trend. A random walk of first order is used in the neighbour model,

where the distribution of y_t conditionally on the past only depends on the first lagged value. A local linear or quadratic extrapolation is assumed by the random walk of second or third order.

The variance λ^{-1} controls the width of the normal distribution and therefore the roughness of the random walk, as displayed in Figure 4.2. In the following, λ is interpreted as smoothness parameter. A small λ and therefore a large variance results in a rough path, while an increasing parameter λ results in more smooth paths.

In other contexts the type of latent model in formula (4.1) is also denoted as smoothing prior or Bayesian smoothing (Fahrmeir et al., 2009).

4.1.2 Model Representations for Different Concepts of Inference

Depending on the inference concept, the model can be seen from different points of view. In the following, the model is written in context of either generalized additive models or as a Bayesian model. Furthermore, the hierarchical time series model can be represented as a generalized linear mixed model, but it is not considered further, because this representation is not in the focus of interest for this thesis.

Generalized Additive Model

Using the setting of a generalized additive model (Fahrmeir et al., 2009), the time dependent intercepts represent the coefficients of the B-spline basis functions of order zero $B_k(t) = \mathrm{I}(k-1 \leq t < k)$, where $I(\cdot)$ is the indicator function and thus

$$\beta_0(t) = \sum_{k=1}^{K} \beta_{0k} B_k(t) = \beta_{0t}, \quad t = 1, 2, 3, \ldots.$$

Adding the parametrical term of covariates a semiparametrical model is derived as

$$\eta_t = \beta_0(t) + \boldsymbol{x}_t \boldsymbol{\beta}_x.$$

The spline coefficients are estimated by adding a penalty term as shown in section 4.2.1 to the optimization criteria, which penalizes large variation of the time series, approximated by the differences of the basis coefficients. Depending on the choice of order in the penalization term, the introduced latent models of the hierarchical time series model are obtained.

Following the theory of generalized additive models, inference can be made by penalized likelihood estimation combined with the minimization of a model selection criteria for estimation of the smoothness parameter λ. This strategy is shown in detail in section 4.2.1.

Bayesian Model

A Bayesian model is derived by defining priors for the parameters. For the covariate coefficients $\boldsymbol{\beta}_x$ a uninformative prior in form of a centred Gaussian with fixed precision matrix $\boldsymbol{B}$ is assumed.

$$\boldsymbol{\beta}_x \propto N(0, \boldsymbol{B}^{-1})$$

Special attention is attracting a specific prior for the time varying intercept as will be explained in the following. Thereby, random walks are the stochastic analogue to penalization by differences. Thus, for the parameter β_{0t} a prior model of the form in formula (4.1) is defined, which in combination with the information from the data deduces a marginal posterior probability distribution.

$$\begin{aligned} p(\beta_{0t}, \boldsymbol{\beta}_x | y_1, \ldots, y_t) &\propto p(y_1, \ldots, y_t | \beta_{0t})\ p(\boldsymbol{\beta}_0, \boldsymbol{\beta}_x) \\ &\propto p(y_1, \ldots, y_t | \beta_{0t})\ p(\boldsymbol{\beta}_0) p(\boldsymbol{\beta}_x) \end{aligned}$$

Thereby, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_x$ are stochastic independent. Note, that this marginal posterior density for $\boldsymbol{\beta}_0$ cannot be derived analytically. Thus in Bayesian models, inference is usually done by applying, e.g. Markov Chain Monte Carlo (MCMC) methods. In this thesis it will be, however, used the approach of Integrated Nested Laplace Approximation (INLA) which is characterized in section 4.3 in detail. Bayesian inference quantifies uncertainty of estimates directly and the prediction takes the uncertainty in the parameter estimates into account

4.2 Algorithm using Likelihood Inference

Regarding the different representations of the hierarchical time series model, the fit and update the model, and corresponding threshold calculation differ. In the following, the algorithm is described using likelihood inference as introduced in Heisterkamp et al. (2006).

4.2.1 Step 1: Fit Model using Generalized Additive Model Representation

For model fitting, Heisterkamp et al. (2006) proposed the iterative re-weighted least squares algorithm (IRWLS) as the model is a special case of a generalized additive model. IRWLS obtains the estimates iteratively by performing Newton-Raphson or equivalent steps in form of weighted least squares. The weight of an observation depends on the predicted value (Cox et al., 2006). In this particular case, the estimation of coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_x$ depend on the optimal smoothness parameter $\hat{\lambda}$ and are denoted with $\boldsymbol{\beta} = (\boldsymbol{\beta}_0', \boldsymbol{\beta}_x')'$.

Following the theory of generalized additive models (see Fahrmeir et al., 2009), the parameter $\boldsymbol{\beta}$ are estimated by using the penalized log-likelihood criteria

$$l_{pen}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{\lambda}{2} \boldsymbol{\beta}' \boldsymbol{A}_d \boldsymbol{\beta},$$

where the second term is a penalty term for the roughness of the estimated function $\hat{f}(t) = \hat{\beta}_{0t}$. The notation results by

$$\lambda \sum_{t=d+1}^{T} (\Delta^d(\beta_{0t}))^2 = \lambda \boldsymbol{\beta}' \boldsymbol{D}_d' \boldsymbol{D}_d \boldsymbol{\beta} = \lambda \boldsymbol{\beta}' \boldsymbol{A}_d \boldsymbol{\beta}.$$

Here, the shrinkage matrix $\boldsymbol{A}_d$ is constructed by the matrix of differences $\boldsymbol{D}_d$ corresponding to the random walk order d in the latent model with $\boldsymbol{A}_d = \boldsymbol{D}_d' \boldsymbol{D}_d$.

For illustration, if no covariates are assumed, the difference matrix of the neighbour model is

$$\boldsymbol{D}_{\text{neigh}} = \begin{pmatrix} -1 & 1 & & \dots & 0 \\ & -1 & 1 & & \vdots \\ \vdots & & \ddots & \ddots & \\ 0 & \dots & & -1 & 1 \end{pmatrix} \Rightarrow \boldsymbol{A}_{\text{neigh}} = \begin{pmatrix} 1 & -1 & & \dots & 0 \\ -1 & 2 & -1 & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & -1 & 2 & -1 \\ 0 & \dots & & -1 & 1 \end{pmatrix},$$

while for the local linear prior model results

$$\boldsymbol{D}_{\text{lin}} = \begin{pmatrix} 1 & -2 & 1 & & \dots & 0 \\ & 1 & -2 & 1 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \\ 0 & \dots & & 1 & -2 & 1 \end{pmatrix} \Rightarrow \boldsymbol{A}_{\text{lin}} = \begin{pmatrix} 1 & -2 & 1 & & \dots & 0 \\ -2 & 5 & -4 & 1 & & \vdots \\ 1 & -4 & 6 & -4 & 1 & \\ & & \ddots & & \ddots & \\ & 1 & -4 & 6 & -4 & 1 \\ \vdots & & 1 & -4 & 5 & -2 \\ 0 & \dots & & 1 & -2 & 1 \end{pmatrix}.$$

In case, covariates are included in the model, at the position of their coefficients the matrices have zeros, because they are not penalized. Thus,

$$\boldsymbol{A}_d = \begin{pmatrix} \boxed{\boldsymbol{D}_d'\boldsymbol{D}_d} & 0 \\ 0 & 0 \end{pmatrix}.$$

The estimation of $\boldsymbol{\beta}_\lambda$, with fixed λ, is given by the weighted least squares solution

$$\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \lambda\boldsymbol{A}_d)^{-1}\boldsymbol{X}'\boldsymbol{W}y,$$

where $\boldsymbol{W} = \boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}$ is a matrix of weights with $\boldsymbol{D} = \text{diag}(d_1, \dots, d_T)$ the derivative of response function, and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_T^2)$ the variance matrix of $\boldsymbol{\beta}$ which is obtained by iterative updating of a working covariance. The covariates $\boldsymbol{X}$ are centred and scaled.

For optimization of the smoothness parameter λ several choices exist. Heisterkamp et al. (2006) choose the model selection criteria ABIC. The optimal λ is chosen as the value of λ which minimizes ABIC.

$$\text{ABIC}(\lambda) = l(y|\boldsymbol{\beta}, \boldsymbol{X}, \lambda) - \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{A}_d\boldsymbol{\beta} + \frac{\text{rg}(\boldsymbol{A}_d)}{2}\log(\lambda) - \frac{1}{2}\log(\det(\boldsymbol{H})),$$

where $\boldsymbol{H}$ is the Hessian of the penalized likelihood. With the new estimate of λ the least squares equation is reweighted and the procedure is iterated.

Example. Using the simulated data of Figure 4.1 for illustration, a generalized additive model is fitted as described before while outbreaks have been excluded. The obtained model for the training period is displayed in Figure 4.3. Note, that at

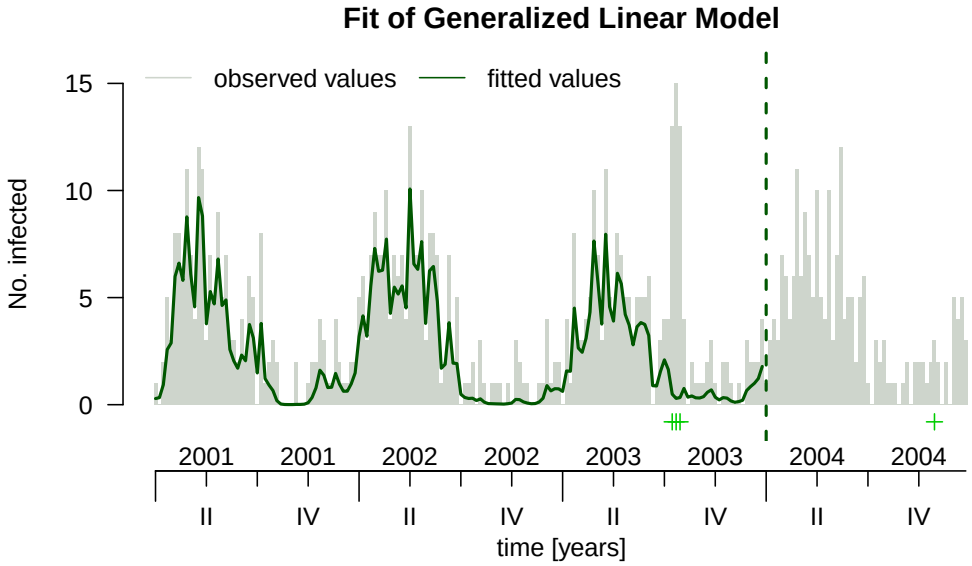

Figure 4.3: Illustrative simulated data with model fit on training data (left of the dashed line) using a random walk of order one as latent model.

the time period of outbreak the fit has a higher level than the year before, but does not follow the peak. This is caused by excluding the observed values in the weeks of outbreak while fitting. In the following, it is shown how this fit is used to update the model stepwise due to the arriving of the data.

◇

4.2.2 Step 2: Sequential Model Update

To avoid lengthy computations whenever new data arrive, the model is fitted based on the training data $y_1, \ldots, y_T$ only once. Hereafter, $\hat{\lambda}$ and the $\boldsymbol{\beta}$ coefficients including $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0T})'$ and $\boldsymbol{\beta}_x$ are kept fixed, ignoring any uncertainty, and the model is updated sequentially as soon as a new observation y_{T+1} arrive. Thus, only the intercept β_{0T+1} is updated to make application of the algorithm faster. Therefore, a penalized log-likelihood is considered to be

$$\begin{aligned} l(\beta_{0T+1}|\hat{\beta}_m, \hat{\boldsymbol{\beta}}_x, \hat{\lambda}, y_{T+1}, \boldsymbol{x}_{T+1}) &= l(y_{T+1}|\beta_{0T+1}, \hat{\boldsymbol{\beta}}_x, \hat{\lambda}, \boldsymbol{x}_{T+1}) \\ &\quad + p(\beta_{0T+1}|\hat{\beta}_m, \hat{\lambda}) \end{aligned} \tag{4.2}$$

where $\hat{\boldsymbol{\beta}}_x = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p)$ are the parameter estimates of the covariates $\boldsymbol{x}_{T+1}$ which are kept fixed, and $\hat{\beta}_m$ represents an artificial conditioning variable derived from $\boldsymbol{\beta}_0$ and depending on the chosen model for the time varying intercept.

$$\begin{aligned} \hat{\beta}_{m_1} = \hat{\beta}_{\text{stat}} &= \hat{\beta}_0 \\ \hat{\beta}_{m_2} = \hat{\beta}_{\text{neigh}} &= \hat{\beta}_{0T} \\ \hat{\beta}_{m_3} = \hat{\beta}_{\text{lin}} &= 2\hat{\beta}_{0T} - \hat{\beta}_{0T-1} \\ \hat{\beta}_{m_4} = \hat{\beta}_{\text{quad}} &= 3\hat{\beta}_{0T} - 3\hat{\beta}_{0T-1} + \hat{\beta}_{0T-2} \end{aligned}$$

The first term of the penalized log-likelihood in (4.2) is the log-likelihood of the new observation y_{T+1} given the past, and the second term is the penalty term of β_{0T+1} given the previous estimates. The model may include the covariates at time $T+1$.

In case of the Poisson distribution the log-likelihood of y_{T+1} is

$$l(y_{T+1}|\beta_{0T+1}, \hat{\boldsymbol{\beta}}_x, \hat{\lambda}, \boldsymbol{x}_{T+1}) \quad \propto \quad y_{T+1}\beta_{0T+1} - \exp(\beta_{0T+1} + \boldsymbol{x}'_{T+1}\boldsymbol{\beta}_x).$$

The penalty term of β_{0T+1} is

$$p(\beta_{0T+1}|\hat{\beta}_m, \hat{\lambda}) \quad \propto \quad \frac{1}{2}(\beta_{0T+1} - \hat{\beta}_m)'(\hat{\lambda}^{-1} + \hat{\boldsymbol{\Sigma}}_m)^{-1}(\beta_{0T+1} - \hat{\beta}_m),$$

where $\hat{\beta}_m$ is defined as above and $\hat{\boldsymbol{\Sigma}}_m$ depends on the chosen prior model of the time varying intercept as well. The covariance is $\hat{\boldsymbol{\Sigma}}_m = \boldsymbol{w}'_m\hat{\boldsymbol{\sigma}}'_m\hat{\boldsymbol{R}}_m\hat{\boldsymbol{\sigma}}_m\boldsymbol{w}_m$ with the correlation matrix $\hat{\boldsymbol{R}}_m$ estimated from the training time period, $\hat{\boldsymbol{\sigma}}_m$ the updated standard error diagonal matrix of the model parameters used in$\hat{\beta}_m$, and $\boldsymbol{w}_m$ a weighting vector corresponding to the chosen model, i.e. $\boldsymbol{w}_{\text{lin}} = (-2, 1)'$.

Example. This predictive posterior density is displayed in Figure 4.4 for the data from Figure 4.1. A prediction for the next time point $T+1 = 157$ is derived. Thus, the optimal β_{0T+1} is estimated as the mode of the predictive posterior, which is represented in Figure 4.4 by a green dot. Note, that the figure corresponds to the last time points of the fit tail in Figure 4.3.

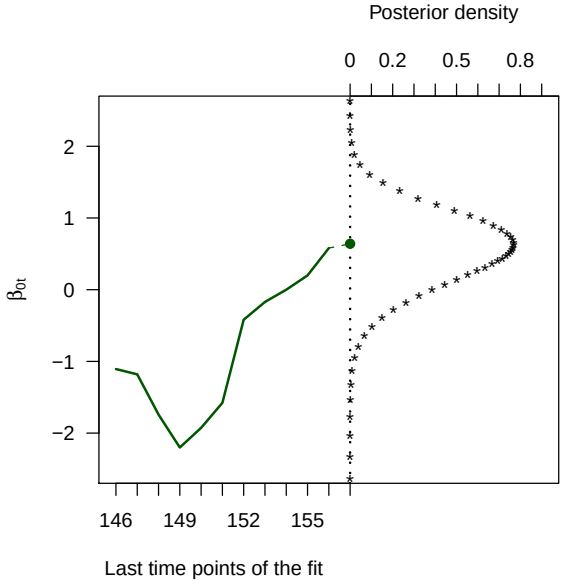

Figure 4.4: Last time point of the fit in Figure 4.3 together with the predictive posteriori density. The symbol + indicates an outbreak.

◇

Therefore, the penalized log-likelihood for β_{0T+1} from formula (4.2) derives as

$$\begin{aligned} l(\beta_{0T+1}|\hat{\beta}_m, \hat{\boldsymbol{\beta}}_x, \hat{\lambda}, y_{T+1}, \boldsymbol{x}_{T+1}) \quad \propto \quad & y_{T+1} - \exp(\beta_{0T+1} + \boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_x) \\ & -\frac{1}{2}(\beta_{0T+1} - \hat{\beta}_m)(\lambda^{-1} + \hat{\boldsymbol{\Sigma}}_m)^{-1}(\beta_{0T+1} - \hat{\beta}_m) \end{aligned} \tag{4.3}$$

To estimate the parameter of time trend β_{0T+1} the mode of this penalized likelihood is taken. Therefore, the score equation has to be solved,i.e.

$$s(\beta_{0T+1}) \stackrel{!}{=} 0 \quad \Leftrightarrow$$
$$y_{T+1} - \exp(\beta_{0T+1} + \boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_x) - (\hat{\lambda}^{-1} + \hat{\boldsymbol{\Sigma}}_m)^{-1}(\beta_{0T+1} - \hat{\beta}_m) \stackrel{!}{=} 0.$$

To determine the asymptotic variance of the estimate the observed Fisher information is needed. It can be derived from the negative inverse of the second derivative of the log-likelihood

$$\begin{aligned}\sigma^2_{\hat{\beta}_{0T+1}} &= \mathrm{Var}(\hat{\beta}_{0T+1}) = [s'(\beta_{0T+1})]^{-1} \\ &= \left[\exp(\beta_{0T+1} + \boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}}_x) + (\hat{\lambda}^{-1} + \hat{\boldsymbol{\Sigma}}_m)^{-1}\right]^{-1}.\end{aligned}$$

Substituting $\hat{\beta}_{0T+1}$, its variance estimation is derived as

$$\hat{\sigma}^2_{\hat{\beta}_{0T+1}} = \left[y_{T+1} + (\hat{\lambda}^{-1} + \hat{\boldsymbol{\Sigma}}_m)^{-1}(1 - \hat{\beta}_{T+1} + \hat{\beta}_m)\right]^{-1}.$$

With the additional feature of updating the hierarchical time series model whenever new data arrive the algorithm becomes fast and efficient. Now, the updated model is used to calculate a threshold for aberration detection.

4.2.3 Step 3: Threshold Calculation

The basic setting for the threshold computation is similar to the Farrington algorithm (see chapter 3.2): The hierarchical time series model is fitted and the estimates are used to define a threshold.

For outbreak detection again only high counts of the disease are of interest. Therefore, a one-sided $100\% \cdot (1-\alpha)$ interval for the next observation y_{T+1} is constructed at time point T. Depending on the monitored organism different types of threshold might be appropriate.

The absolute aberration detection compares the upper interval border with a fixed threshold. The relative threshold is an upper border for the variation of the time series, which is compared with the current differences in the process. The comparison with the cumulative threshold works similarly while the process variation over an aggregated time period is considered.

Absolute Threshold

A time constant and known mean of $\mathrm{E}(y_t) = \mu_0,\ t = 1, \ldots, T$, which can be computed, e.g. based on the training period, is assumed, and a fixed exceedance score $\delta > 0$ is defined. Then, an alarm is triggered if the observed value exceeds the expected one by an amount of $(1+\delta)$. Thus the absolute threshold for y_{T+1} is defined as

$$\xi_{abs} = \mu_0(1 + \delta).$$

Thereby, the exceedance score δ can be seen as percentage of exceeding, i.e. for $\delta = 0.5$ the amount of excess is 50%.

The threshold at time $T+1$ is defined by the upper border of the $100\% \cdot (1-\alpha)$ interval for the expectation at time T, $\mathrm{E}(y_T)$. Thereby, the maximum likelihood estimation of β_{0T} has the property to be asymptotically normal distributed, thus $\hat{\beta}_{0T} \overset{a}{\sim} \mathbb{N}(\beta_{0T}, \hat{\sigma}^2_{\beta_{0T}})$. As a consequence the upper limit of a corresponding $100\% \cdot (1-\alpha)$ Wald interval for $\hat{\beta}_{0T}$ is constructed by

$$\beta_{0T} + z_{1-\alpha}\sqrt{\hat{\sigma}^2_{\beta_{0T}}}$$

The upper border is retransformated by the inverse link, i.e. $\mathrm{E}(y_{T+1}) = \exp(\eta_{T+1}) = \exp(\beta_{0T+1} + \boldsymbol{x}'_{T+1}\boldsymbol{\beta}_x)$, and an alarm is triggered if

$$\exp\left(\beta_{0T} + \boldsymbol{x}'_{T+1}\hat{\boldsymbol{\beta}} + z_{1-\alpha}\sqrt{\hat{\sigma}^2_{\beta_{0T}}}\right) \geq \xi_{abs},$$

where $z_{1-\alpha}$ is the $100\% \cdot (1-\alpha)$-quantile of the standard normal distribution and the parameter estimates are substituted where necessary. Note that the uncertainty due to the estimation of the parameters $\boldsymbol{\beta}_x$ is therefore not considered.

Example. In Figure 4.5 the absolute aberration calculation on the evaluation time period in the presented data from Figure 4.1 is shown. During the whole time period of high season is triggered an alarm. The type of threshold can be used for sporadic or highly infectious diseases.

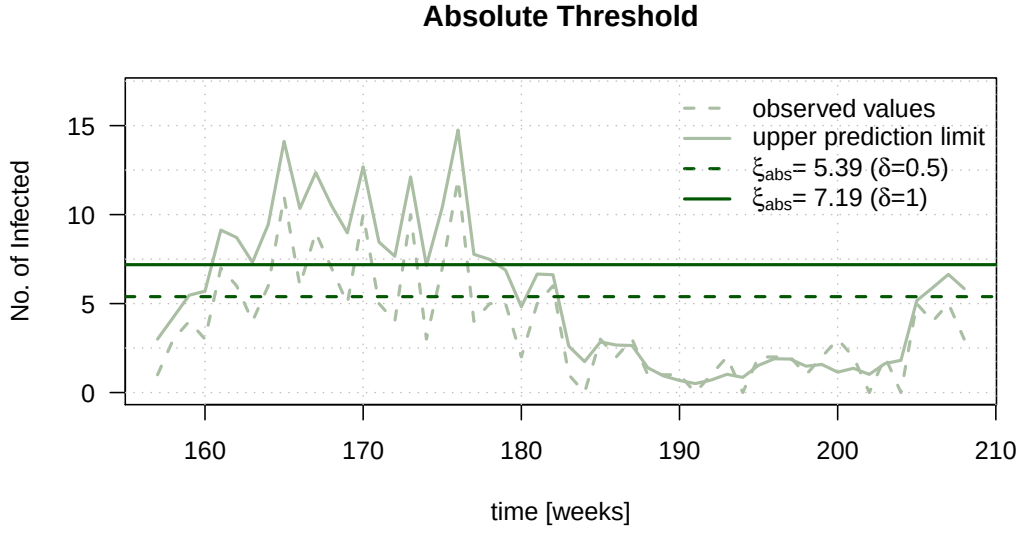

Figure 4.5: Illustrations for absolute aberration calculations

◇

Relative Threshold

A so-called relative threshold is adequate if one is interested in a sudden increase of cases compared with the normal day-to-day variation. It is assumed that in the training period the best latent model has a difference order d. The observed count y_{T+1} is not monitored directly, because it would require approximation of its marginal distribution. Instead of, the depending coefficient β_{0T+1} is used, which is included in $\Delta^d(\beta_{0T+1})$. Thereby, the estimates obtained by penalized likelihood maximization are approximately normal distributed with $\hat{\beta}_{0T} \overset{a}{\sim} \mathbb{N}(\beta_{0T}, \hat{\sigma}^2_{\beta_{0T}} + \lambda^{-1})$. A upper border of a $100\% \cdot (1-\alpha)$ Wald interval at time T is derived as

$$\hat{\beta}_{0T+1} + z_{1-\alpha}\sqrt{\hat{\sigma}^2_{\beta_{0T}} + \lambda^{-1}}.$$

The variation in form of the d^{th} difference $\Delta^d(\beta_{0T+1})$ will be monitored while the uncertainty for estimation of $\hat{\beta}_m$ and $\hat{\boldsymbol{\beta}}$ is ignored. Thus, the relative threshold is defined as

$$\xi_{rel,T+1} = \log(1+\delta) + z_{1-\alpha}\sqrt{\hat{\sigma}^2_{\hat{\beta}_{0T}} + \lambda^{-1}},$$

and an alarm is triggered if

$$\Delta_d(\hat{\beta}_{0T+1}) \geq \xi_{rel,T+1}.$$

As before, the parameter estimates are substituted where necessary.

Example. In the following Figure 4.6 the relative threshold is illustrated. It can be observed that it varies over time according to the current variation of the time series.

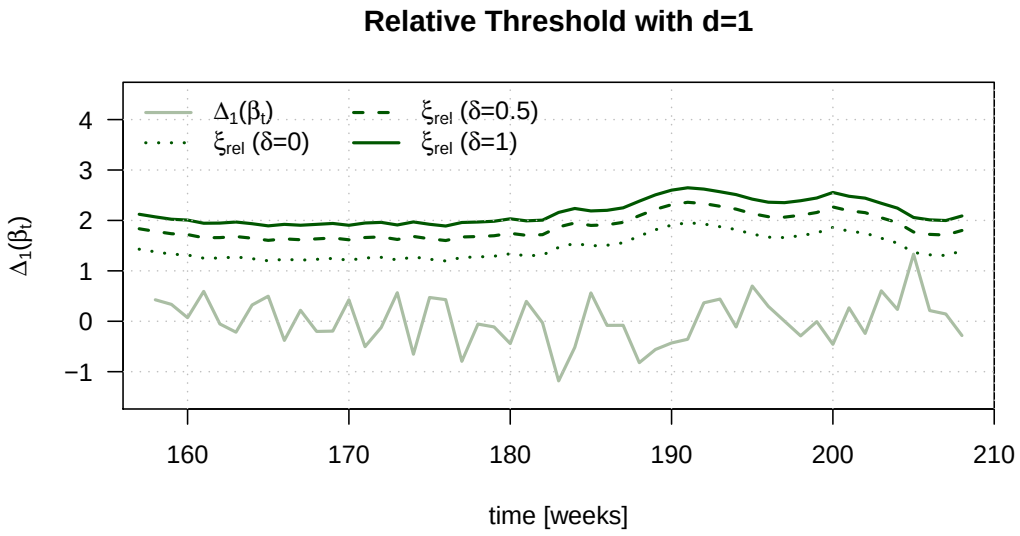

Figure 4.6: Illustrations for relative aberration calculations

◇

Cumulative Threshold

Using a so called cumulative excess over a number of time points can result in more stability of the monitoring. Assuming a given lag time k and δ like above, the threshold at time $T+1$ is based on the distribution of the difference $\beta_{0T} - \beta_{0T-k}$, which is assumed to be normally distributed with $\beta_{0T} - \hat{\beta}_{0T-k} \sim \mathbb{N}(0, \sigma^2_{\beta_{0T}-\beta_{0T-k}} + k\ \lambda^{-1})$. Thus, the cumulative threshold is defined as

$$\begin{aligned} \xi_{cum,T+1} &= \log(1-\delta) + z_{1-\alpha}\sqrt{\hat{\sigma}^2_{\beta_{0T}-\beta_{0T-k}} + k\ \lambda^{-1}} \\ &= \log(1-\delta) + z_{1-\alpha}\sqrt{\hat{\sigma}^2_{\beta_{0T}} + \hat{\sigma}^2_{\beta_{0T-k}} - 2\ \mathrm{cov}(\beta_{0T}, \beta_{0T-k}) + k\ \lambda^{-1}}. \end{aligned}$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution, δ the percentage of allowed exceeding. The parameters are replaced by their estimates. Here, the variation resulting by the estimation of the parameters is ignored, as well. An alarm is triggered, if

$$\beta_{0T} - \beta_{0T-k} \geq \xi_{cum,T+1},$$

Example. In Figure 4.7 is shown the absolute aberration calculation on the evaluation time period in the data introduced in Figure 4.1. The threshold varies much more the in the previous example.

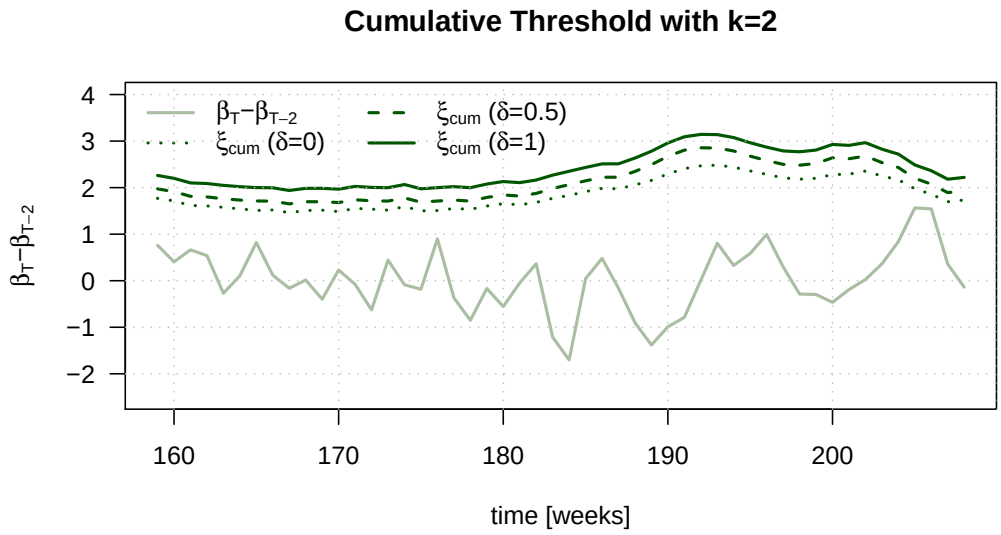

Figure 4.7: Illustrations for cumulative aberration calculations

◇

4.3 Bayesian Version of the Algorithm

Heisterkamp et al. (2006) treat the model as a generalized linear model. Therefore, the updating step and the calculation of thresholds are handled accordingly based on frequentist methods. As described, the model can be seen from a fully Bayesian point of view as well. In this case, the threshold could be derived in the context of Bayesian inference, which means directly by the predictive posterior of the observation. Thus, the threshold calculation includes directly uncertainty of prediction, and of the estimation of parameters in the model.

4.3.1 Step 1: Fit Bayesian Model

As shown before the hierarchical time series model can be represented as a Bayesian model.

$$\begin{aligned} y_t &\sim \text{Po}(\mu_t) \quad \text{or} \quad y_t \sim \text{NegBin}(\mu_t, \alpha) \quad \text{with} \\ \log(\mu_t) &= \eta_t = \beta_{0t} + \boldsymbol{x}_t \boldsymbol{\beta}_x, \end{aligned}$$

where β_{0t} is modelled by one of the specific random walk models as prior model

$$\beta_{0t} = \beta_m + u_t, \quad u_t \overset{\text{iid}}{\sim} \mathbb{N}(0, \tau^2). \tag{4.4}$$

According to the multivariate normal density, the prior density of $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0T}, \beta_1, \ldots, \beta_p)'$, when λ is given, can be written as

$$\begin{aligned} p(\boldsymbol{\beta}) &\propto \left(\frac{\lambda}{2\pi}\right)^{\frac{\text{rg}(\boldsymbol{A})}{2}} \exp\left(-\frac{\lambda}{2}\boldsymbol{\beta}'\boldsymbol{A}_d\boldsymbol{\beta}\right) \\ &\propto \lambda^{\frac{\text{rg}(\boldsymbol{A})}{2}} \exp\left(-\frac{\lambda}{2}\boldsymbol{\beta}'\boldsymbol{A}_d\boldsymbol{\beta}\right), \end{aligned}$$

and the posterior distribution is

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \lambda) = \frac{f(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{\int f(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}} \propto f(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}).$$

It can be shown that the posterior optimization problem can be written equivalently as penalized log-likelihood

$$\begin{aligned} \log(p(\boldsymbol{\beta}|y, \boldsymbol{X}, \lambda)) &= \log(f(y|\boldsymbol{\beta}, \boldsymbol{X}, \lambda)) + \log(p(\boldsymbol{\beta})) \\ &\propto l(y|\boldsymbol{\beta}, \boldsymbol{X}, \lambda) - \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{A}_d\boldsymbol{\beta} + \frac{rg(\boldsymbol{A}_d)}{2}\log(\lambda) \\ &= l_{pen}(\boldsymbol{\beta}|y, \boldsymbol{X}, \lambda), \end{aligned}$$

where the last two terms were defined in section 4.2.1, in the likelihood inference, and make up the penalty term with $\boldsymbol{A}_d$ the shrinkage matrix, and λ the given smoothness parameter.

Example. In the illustrative example the last time point of the training period $t = 156$ is used to illustrate the knowledge increase about the parameter $\hat{\beta}_{0t}$ due to the consideration of the data. Since a random walk 1 model is used, the mode of

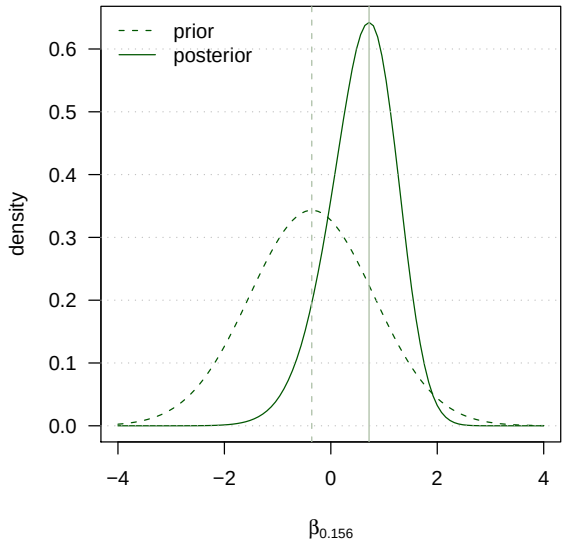

Figure 4.8: Comparison of prior and posterior for $\beta_{0t=156}$ at time point $t = 156$.

the prior at 0.37 is defined by $\hat{\beta}_{0t}, t = 154, 155$. The precision of the normal distribution, λ, is estimated as hyperparameter. The posterior distribution considers the new observation $y_{156} = 4$.

The prior has its maximum with density 0.37 at $\beta_{0t=156} = -0.36$, while the posterior reaches a density of 0.64 at $\beta_{0t=156} = 0.73$. Therefore, the posterior distribution is more steep and the variance is smaller. Thus, the uncertainty about the parameter is reduced by including the information on the latest time point.

◇

Usually, the posterior's normalization constant is unknown. Therefore, the posterior cannot be derived analytically. Thus, inference in Bayesian models can be drawn from Markov chain Monte Carlo (MCMC) methods, which could have high computing time. Therefore, the posterior distribution will be computed by a Laplace approximation instead, which is described in detail in section 4.4.2.

4.3.2 Step 2: Sequential Model Update

According to Bayesian inference, the model can be updated by using the obtained parameter posteriors based on the training period. The estimates include information about previous observations, and therefore could be included as priors in the updating step. The resulting uncertainty needs to be considered when the predictive posterior is constructed. In the context of this thesis, such work would be beyond the scope of what is possible. Therefore, the sequential model update in the Bayesian version of the hierarchical time series algorithm is not considered further. An efficient implementation for model fitting will be applied by using INLA, so that in spite of that an efficient working algorithm is obtained.

4.3.3 Step 3: Alarm Triggering using the Bayesian Approach

The strategy of triggering an alarm in a Bayesian setting follows the approach of Höhle (2008). Here, a threshold is calculated using the $100\% \cdot (1-\alpha)$-quantile of the predictive posterior distribution $f(y_{T+1}|y_1,\ldots,y_1)$. If the observed value y_{T+1} is greater than the threshold $\xi_{1-\alpha}$ an alarm is triggered for the week $T+1$.

The predictive posterior can be written as a function of the new observation's likelihood and the parameter posterior with

$$\begin{aligned} p(y_{T+1}|y_1,\ldots,y_T) &= \int f(y_{T+1},\beta_{0T+1}|y_1,\ldots,y_T)\ d\beta_{0T+1} \\ &= \int L(y_{T+1}|\beta_{0T+1})\ p(\beta_{0T+1}|y_1,\ldots,y_T)\ d\beta_{0T+1}. \end{aligned}$$

If the value is located in the upper, say, 5%-tail of the distribution, the value is unexpected high and therefore an alarm is triggered. A threshold $\xi_{1-\alpha}$ is calculated by using a quantile parameter α, so that

$$P(y_{T+1} \leq \xi_{1-\alpha}|y_1,\ldots,y_T) \geq 1-\alpha.$$

Thus, the threshold is defined as the $100\%\cdot(1-\alpha)$-quantile of the predictive posterior,

$$\xi_{1-\alpha} = \inf\{y_{T+1}|F(y_{T+1}|y_1,\ldots,y_T) \geq 1-\alpha\},$$

where $y_{T+1} \geq 0$ and $F(y_{T+1}|y_1,\ldots,y_T) = \sum_{x=0}^{y_{T+1}} f(x|y_1,\ldots,y_T)$ is the cumulative distribution function of the predictive posterior. An alarm is triggered, if

$$y_{T+1} > \xi_{1-\alpha}.$$

Example. In Figure 4.9 the Bayesian aberration calculation is illustrated using the evaluation time period of data introduced in Figure 4.1. The 95%-quantiles for the different latent models are shown. At the time point, the threshold falls below the count time series of observations an alarm would be triggered.

It can be observed that the stationary model ('iid') is very restrictive and does not allow much variability so that it does not reflect the seasonality of the time series. The quantiles using the neighbour model ('rw1') and linear model ('rw2') have more or less the same structure, while the local linear model tends to be more flexible to deal with strong variation.

◇

Corresponding to Heisterkamp et al. (2006) different kinds of thresholds can be defined, as well. Regarding the absolute threshold, an alarm will triggered, if

$$\xi_{t,1-\alpha} > \mu_0(1+\delta)$$

A kind of cumulative threshold can be derived by aggregating the k last thresholds, i.e. by triggering an alarm, if a percentage of the k last observations exceeds the corresponding calculated thresholds.

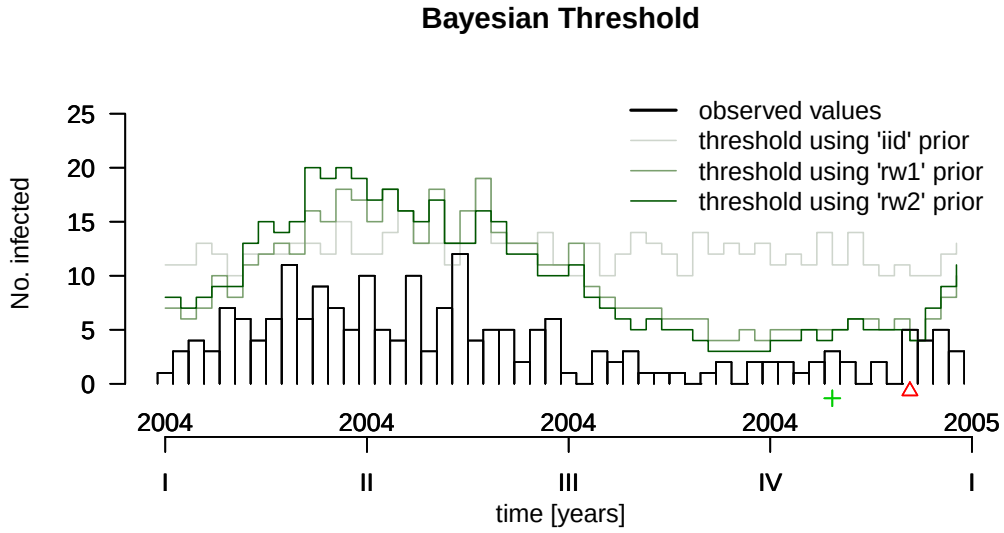

Figure 4.9: Illustrations for Bayesian aberration calculations. The symbol + indicates an outbreak and △ an alarm.

The main advantage of the Bayesian approach is that not only a binary information is given which seems to be very restrictive. Here, a probability for aberration given an observed values can be obtained, so that an easy interpretable probability for outbreak is provided.

4.4 Implementation

The implementation in software is a very important factor for the utilization of methodological proposals. The developed full Bayesian version of the hierarchical time series algorithm has several advantages such as direct threshold calculation, including prediction and estimation error, and the simple consideration of covariate processes.

For implementation, the parallelism to generalized additive models, thus likelihood inference, can be used as well. Heisterkamp et al. (2006) implemented this strategy in S-PLUS while his implementation in R is in progress.

4.4.1 `surveillance` Package for `R`

R (R Developer Core Team, 2009) is a free environment for statistical computing with increasing acceptance and popularity within the statistical community. It provides a large variety of statistical and graphical methods and is easily extensible. Since R is Open Source available under the GNU GPL licence software, it is free for anyone to use and modify.

The R add-on package `surveillance` (Höhle, 2007) offers functionality for the visualization, monitoring, and simulation of count data and categorical time series. In

this context it provides methods for online change-point detection with a focus on outbreak detection in count times series, like they are usual obtained in the public health surveillance context. The package is available under the GPL licence and downloadable from Comprehensive R Archive Network (CRAN). It provides an environment for developers of new algorithms as well.

All aberration detection algorithms in `surveillance` have the same structure of application. The first object denotes an object of class `sts` containing the observed and state time series, and the second argument `control` is a list of vector `range` specifying the time points to monitor, and algorithm specific control options (Höhle and Mazick, 2010).

In the context of this diploma thesis, the Bayesian version of the hierarchical time series algorithm was implemented in the `surveillance` framework. To fit the Bayesian models efficient approximations by INLA (Integrated Nested Laplace Approximation) will be used instead of time consuming Markov Chain Monte Carlo methods.

4.4.2 Implementation of `algo.hts()` using INLA

For implementing, the hierarchical time series algorithm the underlying model is seen as a Bayesian hierarchical model with dynamics of the intercept $\beta_0(t) = \beta_{0t}$ over time given by random walks with latent Gaussian random walks.

The general procedure is summarized in the following steps:

1. Preprocessing: Distinguish modelling and evaluation time, exclude observations of outbreak state during the modelling time, set formula, and prepare covariates.

2. Sequential steps for each time point in evaluation period:

   (i) Fit Bayesian model using `INLA` (Rue and Martino, 2009).

   (ii) Calculate predictive posterior for y_{T+1} using Monte Carlo integration.

   (iii) Compute $100\% \cdot (1-\alpha)$-quantile of predictive posterior for defining threshold $\xi_{1-\alpha}$, compare threshold of observed value, and trigger alarm if necessary.

3. Return an object `survRes` object with the modelling results including an array `alarm`, which indicates the triggered outbreaks.

In the following, the application of the algorithm using the `R` package `surveillance` is described and after that the implementation is outlined step by step. Further technical details are given in appendix A which includes the entire code of the resulting functions `algo.hts` and `algo.htsFit`.

Application of `algo.hts()`

The function is called by

```
algo.hts(disProgObj, control=list(range=NULL, co.arg=NULL, prior='iid',
         family='poisson', alpha=0.05, mc.betaT1=100, mc.yT1=10))
```

Here, `disProgObj` is an S3 object of class `disProgObj` as specified by the package `surveillance` including the observed and the state time series. Furthermore, a list of control arguments is given.

Thereby, the first argument `range` specifies the index of all time points, which should be monitored. If the range is not defined, i.e. `range=NULL`, the time points starting with the second period is monitored.

Furthermore, it is possible to include covariates in the model. With defining the control argument `co.arg` known covariates can be considered easily in the outbreak detection model. Thereby, the argument `co.arg` has to be an numerical object of the same length as the observed disease progress. Assuming the specification `co.arg=cbind(x1,x2)`, the formula follows as `observed ~f(time, model='rw1') + x1 + x2`.

Three of the four prior models, described in section 4.1.1, are implemented. The stationary model is available by choosing `prior='iid'`, the neighbour model by specifying `prior='rw1'` in the functional part of the formula argument, and the local linear model by selecting `prior='rw2'`. The quadratic model is not implemented.

The argument `family` specifies the model distribution. The options `'poisson'` for Poisson, or `'nbinomial'` for negative Binomial distribution can be used.

Furthermore, by specifying `alpha` the probability parameter for the threshold, the quantile $100\% \cdot (1 - \alpha)$ of the predictive posterior, is chosen, while `mc.betaT1` and `mc.yT1` are the number of trials for Monte-Carlo simulation deriving the predictive posterior.

Example. For the example introduced in Figure 4.1, the algorithm is applied with the neighbour prior model. First, the control argument is specified and then included in the algorithm function `algo.hts`

```
control <- list(range=157:length(disProgObj$observed), prior='rw1',
                family='poisson', alpha=0.05, mc.betaT1=100, mc.yT1=10)
modelRW1 <- algo.hts(disProgObj=disProgObj, control=control)
```

In Figure 4.10 the surveillance result is plotted. In late spring of 2004, one outbreak which is difficult to recognize, and is not detected by the algorithm. Using the `survRes` object all methods of the `R` package `surveillance` can be called.

$\diamond$

In the following, technical details of the algorithm are given step by step.

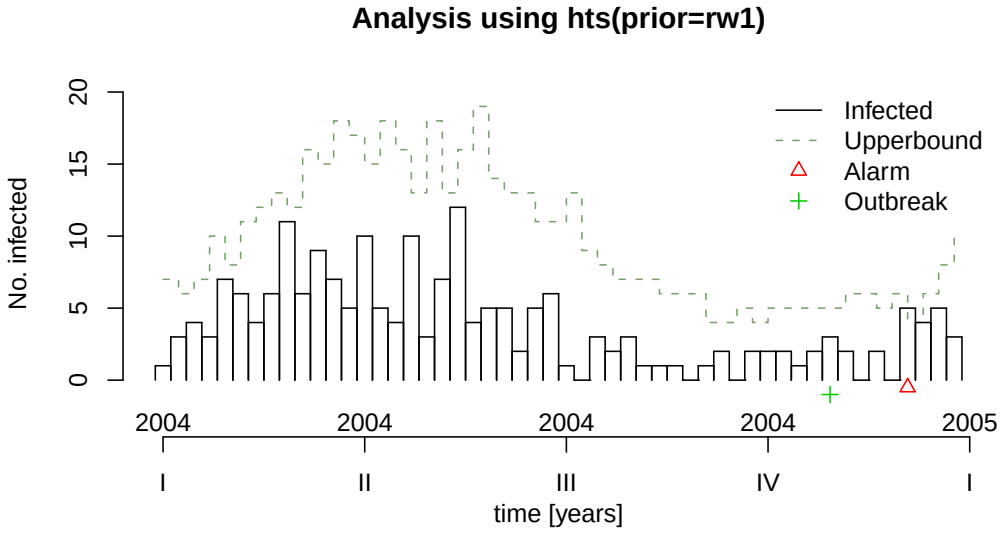

Figure 4.10: Application of surveillance by `algo.hts` using a latent neighbour model.

Step 1: Preprocessing

The model fit should represent an model of 'normality'. Thus, before fitting the model, the observations at a time point of outbreak state, which is indicated by `state=1`, are excluded by replacing their values with `NA`. This is a standard procedure which cannot turned off.

```
observed[which(state==1)] <- NA
```

As mentioned above, covariates can be included into the model by using the control argument `co.arg`. All selected variables are included as they are, which means that transformations or interactions need to be calculated in advance. Thus, the formula is specified as follows:

```
modelformula <- as.formula(paste("observed~f(time, model='",prior,")",
                                 co.arg.formula, sep="")).
```

The following steps are repeated sequentially for each time point $T + 1$ of the monitoring time which is specified by the control argument `range`.

Step 2 (i): Fit Bayesian Model using Integrated Nested Laplace Approximation

The Bayesian model is fitted using the `R` package `INLA`. Rue et al. (2009) presented this fully automatic approach for approximate inference in latent Gaussian models named Integrated Nested Laplace Approximation (INLA). Thereby, direct computations of very accurate approximations of the posterior marginals are provided, and `INLA` outperforms without comparison any Markov chain Monte Carlo (MCMC) algorithm in terms of accuracy and computational speed. Where `INLA` requires seconds and minutes to fit a model MCMC can take up to hours and days.

The marginal posterior for the latent model defined in section 4.3.1 can be written as

$$p(\beta_{0t}|\boldsymbol{y}) = \int p(\beta_{0t}|\boldsymbol{\theta},\boldsymbol{y})\, p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$$
$$\text{with } p(\theta_j|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}_{-j},$$

where $\boldsymbol{\theta}$ is the vector of hyperparameters. The equation is used to construct nested approximations. Therefore, the posterior marginals are achieved by approximations of their components according to the following three steps:

1. Approximation of the posterior marginal $p(\boldsymbol{\theta}|\boldsymbol{y})$ by using Laplace approximation.
2. Improved approximation of $p(\beta_{0t}|\boldsymbol{\theta},\boldsymbol{y})$ for selected $\boldsymbol{\theta}$ by (simplified) Laplace approximation.
3. Combining the steps 1 and 2 by using numerical integration.

Thereby, Laplace approximation is the approximation to an integral based on the second-order Taylor expansion of the log integrand and Gaussian approximation about its mode (Cox et al., 2006). In the following each step is described in more detail.

Excursus. Integrated Nested Laplace Approximation

The first step is the computation of the second part of the integrand in formula 4.5

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})}{p(\boldsymbol{\beta}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*(\boldsymbol{\theta})},$$

where $\boldsymbol{\beta}^*(\boldsymbol{\theta})$ is the mode of the full conditional for $\boldsymbol{\beta}$ given hyperparameter $\boldsymbol{\theta}$ and the normalisation constant is unknown. The denominator is computed by a Gaussian approximation, which is iteratively obtained by a quadratic Taylor approximation of the curvature around the mode. Using this approximation, the mode of the distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ is derived by using the procedure quasi-Newton. To compute the curvature of the distribution, selected points based on a normalised $\boldsymbol{\theta}(\boldsymbol{z})$ with $\boldsymbol{z} \sim \mathbb{N}(0,1)$ are computed and interpolated.

The second step is the approximation of $p(\beta_{0t}|\boldsymbol{\theta},\boldsymbol{y})$, which is derived from the already computed Gaussian marginal $p(\boldsymbol{\beta}|\boldsymbol{\theta},\boldsymbol{y})$ by

$$p(\beta_{0t}|\boldsymbol{\theta},\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})}{p(\boldsymbol{\beta}_{-0t}|\beta_{0t},\boldsymbol{\theta},\boldsymbol{y})}\right|_{\beta_{0t}=\boldsymbol{\beta}^*_{-0t}(\beta_{0t},\boldsymbol{\theta})},$$

where $p(\boldsymbol{\beta}_{-0t}|\beta_{0t},\boldsymbol{\theta},\boldsymbol{y})$ is a Gaussian approximation and $\boldsymbol{\beta}^*_{-0t}(\beta_{0t},\boldsymbol{\theta})$ its mode. The expression is modified due to computational benefits. First, the mode is approximated by $\boldsymbol{\beta}^*_{-0t}(\beta_{0t},\boldsymbol{\theta}) \approx \mathrm{E}(\boldsymbol{\beta}_{-0t}|\beta_{0t})$. Beyond, only those β_{0s} that are close to β_{0t} determine the marginal of β_{0t}. Therefore, the 'region of interest' around t is simply

defined as $R_t(\boldsymbol{\theta}) = \{s : |a_{ts}(\boldsymbol{\theta})| > 0.001\}$. Using these modifications, the approximation for $p(\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{y})$ is computed for different values and interpolated. Furthermore, a simplified Laplace approximation is introduced by doing a series expansion of $p(\beta_{0t}|\boldsymbol{\theta}, \boldsymbol{y})$ around its mode $\beta_{0t} = \mu_t(\boldsymbol{\theta})$.

Finally, the first two steps are combined in accord to the formula 4.5 using numerical integration. Further details are described in Rue et al. (2009).

◇

An implementation of the integrated nested Laplace approximation approach is provided by the `R` add-on package `inla` (Rue and Martino, 2009). Several different hierarchical Bayesian models are available. A Bayesian model is fitted by calling the function `inla()` as follows:

```
inla(formula, family=<error.distribution>, data=data.frame()).
```

For the presented approach of hierarchical time series algorithm, Poisson and negative Binomial likelihoods are used by specifying `family='poisson'` or `family='nbinomial'`. The latent model for the time effect is given by the latent random noise model `model='iid'` reflecting the stationary model, the latent random walk of order one `model='rw1'` specifying the neighbour prior model, or order two `model='rw2'` for the local linear prior model.

```
model <- inla(observed ~ f(time, model=<c('iid','rw1','rw2')>),
              family=<c('poisson','nbinomial')>), data=data}
```

Example. An exemplary model is fitted for the simulated data in Figure 4.1 at time point $t = 157$, thus the first point of the monitoring period. It is examined, that the approximation procedure is very fast and lasts took 1.5 seconds using a Duo core processor with 2.26GHz.

```
Call:
c("inla(formula = observed ~ f(time, model = 'rw1'),
        family = 'poisson', data = dati)")

Time used:
 Pre-processing    Running inla Post-processing           Total
      0.0990510       0.4091651       0.1394680       0.6476841

Fixed effects:
                mean        sd 0.025quant 0.5quant 0.975quant
(Intercept) 1.035719 0.0539935  0.9276392 1.036475   1.139566
                 kld
           0.3502116

Random effects:
Name   Model    Max KLD
time   RW1 model   0.01048
```

```
Model hyperparameters:
                   mean    sd     0.025quant 0.5quant 0.975quant
Precision for time 10.879  3.330  5.793      10.378   18.820

Expected number of effective parameters(std dev): 42.86(4.497)
Number of equivalent replicates : 3.64

Marginal Likelihood:  -351.44
Warning: Interpret the marginal likelihood with care if the
         prior model is improper.
```

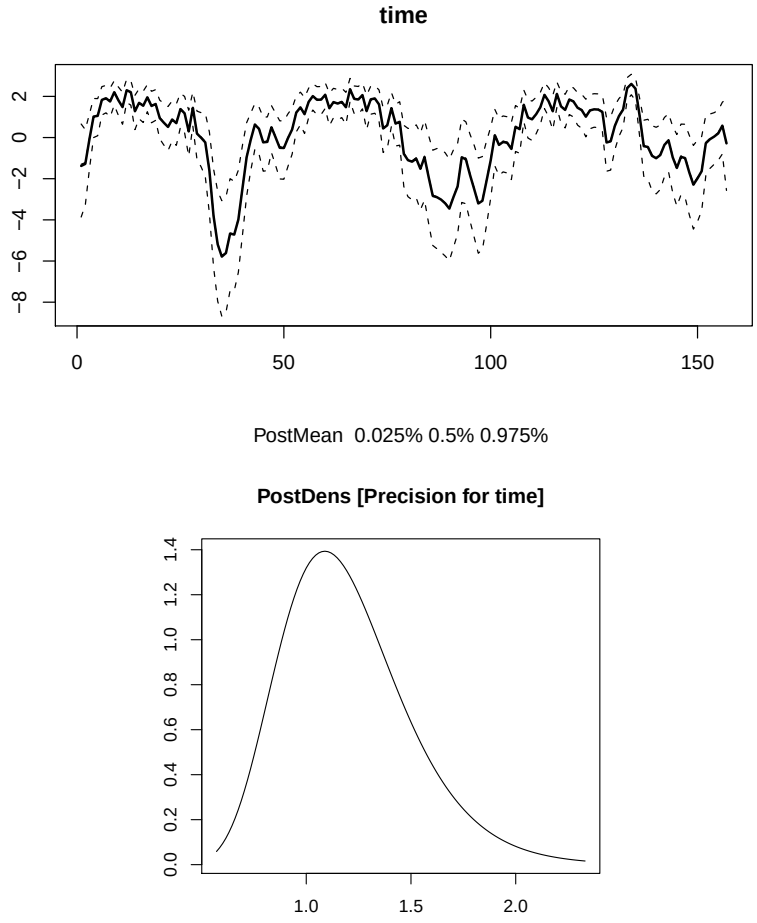

Figure 4.11: Posterior marginals by `inla()` using a latent neighbour model.

As described before, instead the model contains a fixed intercept, a latent effect for the time, which represents aggregated the time dependent intercept. It is used the random walk of order one, thus the neighbour model. The precision of this latent model λ is estimated as hyperparameter. The corresponding estimated posteriors are shown in Figure 4.11.

◇

The model is fitted for each time point during the evaluation period `range`. Since the posterior for β_{0T+1} given the observations $y_1, \ldots, y_T$ should be derived, the value y_{T+1} is unknown, hence it would be `NA` at the time point $T + 1$.

```
observed <- append(observed[1:T], NA)
```

Thus, the predictive posterior $p(\beta_{0T+1}|y_1, \ldots, y_T)$ can be obtained as

```
inla.dmarginal(x, model$marginals.random$time[[T+1]])
```

Detailed informations for the application of the `R inla` package are provided in Martino and Rue (2009).

Step 2 (ii): Calculation of Predictive Posterior

In this section, the computation of the predictive posteriori is described. Since the model is seen in a Bayesian setting, the threshold is defined in this context as well. Thus, the threshold is the $100\% \cdot (1 - \alpha)$-quantile of the predictive posterior distribution at the time point $T + 1$. It can be calculated by

$$p(y_{T+1}|y_1, \ldots, y_T) = \int L(y_{T+1}|\beta_{0T+1})\ p(\beta_{0T+1}|y_1, \ldots, y_T)\ d\beta_{0T+1}.$$

It is derived numerically by using Monte Carlo simulation, which includes the steps:

1. Sample `mc.betaT1` realizations of $\beta_{T+1} \sim p(\beta_{T+1}|y_1, \ldots, y_T)$.
2. Sample `mc.yT1` realizations of $y_{T+1} \sim f(y_{T+1}|\beta_{T+1}, y_1, \ldots, y_T)$ for each of the sampled β_{T+1}.

Thereby, random draws of $p(\beta_{T+1}|y_1, \ldots, y_T)$ are obtained using the fitted model by

```
m <- model$marginals.random$time[[T+1]]
betaT1 <- model$summary.fixed[1] + inla.rmarginal(n=mc.betaT1, m)
```

where the first term is the posterior mean of the fixed intercept β_0, and the second the time varying component. The addend of them is the time varying intercept β_{0T+1} as introduced in formula 4.4.

The sampling distribution of y_{T+1} in the second step depends on the distribution family assumed in the fitted hierarchical Bayesian model. First, the linear predictor is computed which may include covariates as well

```
etaT1 <- betaT1 + sum(dat[T1,-c(1:2)]*model$summary.fixed[-1,1]).
```

Thus, if a Poisson model was fitted, the random realizations of $f(y_{T+1}|\beta_{T+1}, y_1, \ldots, y_T)$ are computed by

```
yT1[((j-1)*mc.betaT1+1):(j*mc.yT1)] <- rpois(n=mc.yT1,
lambda=exp(etaT1[j]))
```

or, if a negative Binomial distribution was assumed,

```
yT1[((j-1)*mc.betaT1+1):(j*mc.yT1)] <- rnbinom(n=mc.yT1,
size=exp(model$theta.mode[1]), mu=exp(etaT1[j])).
```

It results in a vector of length `mc.betaT1*mc.yT1` with random draws of the estimated $\hat{f}(y_{T+1}|\beta_{0T+1}, y_1, \ldots, y_T)$.

In the strict sense of the Bayesian idea, it would be necessary to use also random draws of the so called fixed components in the model, i.e. the intercept β_0 and the covariates coefficients $\boldsymbol{\beta}_x$. Due to technical complications this variation is ignored.

Step 2 (iii): Threshold Calculation and Comparison for Alarm Triggering

The quantile of the predictive posteriori distribution is obtained by

```
qi <- quantile(yT1, probs=1-alpha, type=3, na.rm=TRUE),
```

where the function `quantile` estimates the underlying quantiles based on the order statistics of the vector `yT1` representing random draws of y_{T+1} following the estimated predictive posterior $\hat{f}(y_{T+1}|\beta_{T+1}, y_1, \ldots, y_T)$. Because of the discontinuous nature of the count data the nearest even order statistic is chosen by selecting `type=3`.

The deduce quantile is compare with the observed count, and an alarm is triggered, if necessary.

```
alarm <- disProgObj$observed > xi
```

Step 3: Return Surveillance Result

The last steps were computed for each time point of the specified range. Finally, an object `SurvRes` is returned. It includes an matrix `alarm` which indicates the the triggered alarms, and an matrix `upperbound` reflecting the threshold. Using this object several available evaluation functions of the `surveillance` package can be called, e.g. `plot()` and `algo.quality()`.

4.4.3 Error Handling

During the implementation of `algo.hts` some problems occurred, because `INLA` is still in development. Regarding the problems, an email contact with Håvard Rue, the maintainer of the R package `INLA`, was initiated. Several emails gave some answers, but could solve all problems. In the following, possible errors of the function `algo.hts` will be summarized and suggestions for their handling are given.

Operating System and Software Issues

The algorithm was implemented on the Linux system Ubuntu 9.10. The errors, described in the following, occurred on this system. Regarding an email of Havard Rue is Linux usually much faster and safer, and the operating system 'Windows could be

tricky'. There were made some tests in Windows as well, but running the procedure in other operating systems might trigger further errors.

Usually, the R version is not important, because R is used only for input-output-operations and administrative issues. The R package works as an interface to the `inla` program which uses the algorithms in the GMRFLib library which is an Open Source library in C and Fortran. Detailed informations can be found in the manual for `INLA` Martino and Rue (2009).

It could recognized irregularities due to the usage of a multiprocessor system and resulting random-bit in the computations. If the results should be reproducible the option

```
inla.setOption('num.threads', 1)
```

is suggested which allocates the `inla` to use only one processor.

Found Warnings and Errors

In this section, occurred problems will be described and possible actions will be suggested.

Firstly, the model was specified as initially introduced so that it included only a time varying intercept. The model has been very unstable and collapsed frequently. Thus, a fixed intercept as an overall mean was included.

The error message number 2 claiming that the matrix is not positive definite appeared some times during the simulation studies when negative Binomial distribution was assumed.

```
GMRFLib version 3.0-0-snapshot, has recived error no [2]
Reason    : Matrix is not positive definite
Function  : GMRFLib_factorise_sparse_matrix_TAUCS
File      : smtp-taucs.c
Line      : 698
RCSId     : file: smtp-taucs.c  hgid: 5968749cefcc  date: Fri Jul 30
```

It was figured out that the error might be resulted by the overdispersion parameter going to zero. The `INLA` upgrade on 30 July, 2010, could not solve the problem in every case of occurring. Thus, in the function `algo.htsFit` no threshold is computed and instead a missing value is returned.

Furthermore, in several cases the warning message of the Hessian having a negative eigenvalue which is set automatically to some other value was triggered.

```
*** WARNING *** Eigenvalue 0 of the Hessian is -40.6532 < 0
*** WARNING *** Set this eigenvalue to 0.532968
*** WARNING *** This might have consequence for the accuracy of
*** WARNING *** the approximations; please check!
```

In these cases the marginals were checked and examined as reasonable. Thus, the predictive posterior and the corresponding threshold is calculated. Nevertheless, it is advised to check if the result is sensible.

Another problem appeared when realizations of the time marginal were drawn. The marginal density has been zero so that the approximation did not seem to be converged. In the function `algo.hts`, this error is handled by returning `NA` as threshold value.

Very rare occurred the error number 21. Its reason could not figured out in context of this thesis.

```
GMRFLib version 3.0-0-snapshot, has recived error no [21]
Reason    : This should not happen
Message   : Condition `density->spline_Pinv != ((void *)0)' is not TRUE
Function  : GMRFLib_density_Pinv
File      : density.c
Line      : 934
RCSId     : file: density.c  hgid: 5968749cefcc  date: Fri Jul 30
```

The described errors occurred usually when negative binomial distribution was assumed. Here, a further hyperparameter is assumed. During the simulation studies of chapter 5, all Poisson distributed models converged while negative Binomial distribution models collapsed. Therefore, in case of an aborted model it is advised to refit it with same specifications, but assuming Poisson error distribution.

4.4.4 Rejected Enhancements because of Errors

In context of improving the algorithm a further seasonal component was used. Especially in the stationary model the option was tried out to allow more flexibility, if required. Furthermore, the possibility of using cyclic random walks instead of the ordinary ones were checked. It is assumed, that these models could improve the algorithm's handling of common seasonality in surveillance data. Nevertheless, these enhancements were rejected because of a high frequency of errors. Debugging the described errors might open a wide range of possible model enhancements.

4.5 Enhancements and Limitations Discussion

It could be shown, that the hierarchical time series algorithm is very flexible and different inference can be used. A special enhancement in comparison to other surveillance algorithms is the possibility to include covariate effects. All presented versions of the approach have pros and cons. In the following, they are identified and discussed.

4.5.1 Updating the Model

Because several diseases are monitored at once, the efficiency of the algorithm is very important. Therefore, it is a good solution to update the model stepwise instead of

the refitting the whole model. To update the model the likelihood (4.3) derived by Heisterkamp et al. (2006) is used. The problem with it is that uncertainty due to the previous parameter estimations is ignored. Apart from that, it will be necessary to refit the model from time to time to avoid structural change errors. This issue is not discussed, so that no time intervals of complete refitting are suggested.

It needs to be taken in account how a prior including the history of the random time effect can be hold fixed for the following updating steps. Therefore, the previously achieved parameters have to be substituted. In the presented implementation this feature is not adapted. This strategy includes a refit at each time point, which tends to be inefficient. By using the efficient fitting procedure of INLA one can cope with the computational time.

4.5.2 Threshold Computation

Heisterkamp et al. (2006) introduced three types of aberration detection based on absolute, relative, or cumulative thresholds. Note, that this approach ignores the uncertainty by the estimation, e.g. due to the covariates.

An alternative threshold calculation is given by the Bayesian version of the algorithm. Corresponding to Bayesian inference, the predictive posterior is computed, and uncertainty due to the prediction and estimation of parameters, is directly included.

Regarding the α-level of the quantiles it is necessary to specify a larger number of Monte Carlo iterations to generate an appropriate sample of the predictive posterior. This might increase the computing time, but the results will be more accurate.

Example. At the first twenty points of the evaluation period threshold of various significance levels are computed and displayed in Figure 4.12. At each time point, the 1000 realizations are drawn from the marginal of the time varying intercept ($n_{\texttt{mc.betaT1}} = 1000$), and for each resulting setting 100 values due to the Poisson distribution ($n_{\texttt{mc.yT1}} = 100$) are drawn to obtain realizations of the predictive marginal. Thus, the predictive posterior estimation is based on 100.000 Monte Carlo simulated draws which is considered to be appropriate if the α-level is small, e.g. 99% threshold. Default is $n_{\texttt{mc.betaT1}} = 100$ and $n_{\texttt{mc.yT1}} = 10$, which is usually appropriate for $\alpha = 0.05$.

$\diamond$

4.5.3 Considering of Reporting Delay

Heisterkamp et al. (2006) describes the possibility of including reporting delays and emphasizes the truly retrospective manner of outbreak detection. Nevertheless, he does not utilize it.

Therefore, $y_{it}, i = 0, \ldots, D_{max}, t = 1, \ldots, T$ is introduced where D_{max} is the maximum delay between sampling and reporting. Thus, y_{it} denotes the number of cases

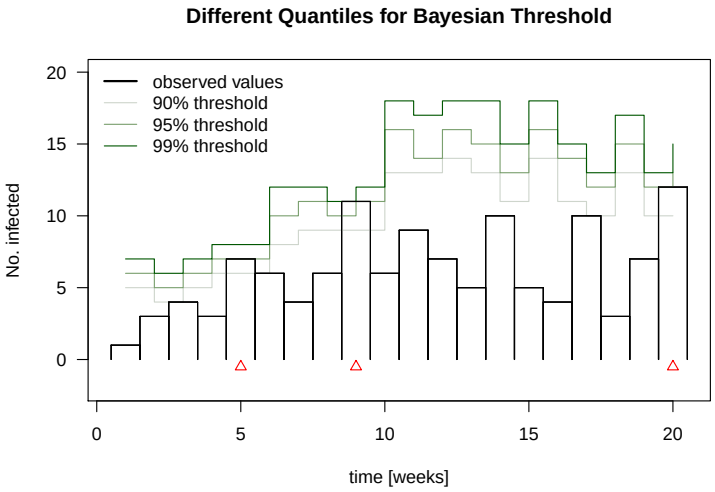

Figure 4.12: Exemplary comparison of different quantiles based on a neighbour latent model. The symbol + indicates an outbreak and △ an alarm.

originating at time $t-i$ but first reported at time t. Furthermore, $\pi_0, \pi_1 \ldots, \pi_{D_{max}}$ are the probabilities of the cases being reported after $0, 1, \ldots, D_{max}$ time units. Moreover, it can be considered that each laboratory l with a weight of f_l has specific days of reporting, e.g. every Friday. Assuming N_t to be the number of laboratories connected at each date t, the generalized model is defined by

$$\begin{aligned} y_{it}|\mu_{it} &\sim \text{Po}(\mu_{it}) \quad \text{with} \\ \mu_{it} &= \exp(\eta_t) \cdot \pi_i \sum_{l}^{N_t} f_l, \quad t = 1, \ldots, T,\ i = 0, 1, \ldots, D_{max}, \end{aligned}$$

where η_t is the linear predictor including the parameter of interest. But finally, Heisterkamp et al. (2006) ignore the reporting delay due to simplicity and the components are aggregated to $y_t = y_{\cdot t} = \sum_{i=0}^{D_{max}} y_{it}$.

Instead, this thesis introduces a slightly different ad-hoc solution for consideration of reporting delays. It does not include further information whose collection might be costly. As before, it is assumed that

$$y_t = \sum_i y_{it}, \quad i = 0, 1, \ldots, D_{\max}, t = 1, \ldots, T.$$

Therefore, π_{it} is defined as the proportion of reported data after i time points at time t, so that

$$y_{it} = \pi_{it} y_t \quad \Leftrightarrow \quad \pi_{it} = \frac{y_{it}}{y_t},$$

while $\pi_{it} \in [0, 1]$ and $\sum_i \pi_{it} = 1$. If the variation in reporting over time is small, it can be defined a generalized time-independent proportion by

$$\hat{\pi}_i = \sum_{j=0}^{i} \frac{1}{T} \sum_{t=1}^{T} \hat{\pi}_{jt} = \frac{1}{T} \sum_{t=1}^{T} \frac{y_{it}}{y_t}$$

Using this estimate for proportion of already reported data at the time point with delay i, the threshold can be adjusted correspondingly. The strategy will be applied to the Campylobacter infection reportings of the RKI in section 6.3.3.

4.5.4 Further Enhancements of the Bayesian Version

In this chapter, several further enhancements of the Bayesian hierarchical time series algorithm were introduced. In context of this thesis, it was not possible to finalized all ideas. In the following, possible enhancements are summarized to illustrate the potential of the developed version of the algorithm.

Sequential Model Update

A sequential updating procedure was introduced by Heisterkamp et al. (2006) based on likelihood inference. A Bayesian version could imaginable due to the consideration of previous observations in the parameter posteriors. These could be used as priors in a Bayesian model for updating.

Threshold Calculation

Using the predictive posterior for threshold calculation, a probability of outbreak given an observed value is obtained. Thus, other more advanced threshold are imaginable.

Seasonality

A further latent component for season to handle seasonality of surveillance time series in a proper way was added to the INLA model, but caused a high frequency of errors. Because `inla` is still development, it is assumed that these errors could be handled in future.

Multiple Time Series

With the current approach and implementation the handling of multiple time series is not possible. Therefore, possible associations between the time series and simultaneous fitting need to be considered. In principle, the fit of the Bayesian model using INLA can be extended for spatial modelling.

Chapter 5

Simulation Studies

In the previous chapters, several statistical methods for surveillance were introduced. Especially, a Bayesian version of the hierarchical time series algorithm was developed in chapter 4.3. Furthermore, in section 3.3, a variety of key parameters for evaluation were described. In the following, these will be used to evaluate the Bayesian version of the hierarchical time series algorithm in comparison to other introduced algorithms, such as the system used at the RKI (see section 3.1.4), the Farrington algorithm (see section 3.2), and the simple Bayes algorithm (see section 3.1.4).

For evaluation of surveillance methods, knowledge regarding true events of outbreaks is necessary. In real reporting data of public health organizations this information is usually not given. Instead, although it cannot represent reality in all its facets, simulated data is used.

5.1 Evaluation using `surveillance`

The `R` add-on package `surveillance`, already described in 4.4.1, provides not only several methods to detect aberration, but also procedures to evaluate their quality.

5.1.1 Simulation of Surveillance Data

One possibility of simulating data for monitoring is the usage of a hidden Markov model. Thereby, the hidden process is a Markov chain as identifier of an outbreak. Conditioning on this Markov chain, a Poisson process is simulated representing the number of disease counts.

Let be Z_t a homogeneous Markov chain defined for $t = 1, \ldots, T$ by

$$Z_t = \begin{cases} 1 & \text{outbreak at } t \\ 0 & \text{else,} \end{cases} \quad \text{and } P = \begin{pmatrix} p & 1-p \\ 1-r & r \end{pmatrix},$$

where P is the transition matrix with p the probability for no transition to an epidemic state, if the progress is in a non-epidemic state, and r the probability to stay in an outbreak state.

The number of observed counts are simulated by a Poisson process

$$Y_t \sim \text{Po}(\mu_t + \kappa Z_t),$$

where κ is the size of outbreak and μ_t models the time varying trend, e.g. by $\log(\mu_t) = \alpha + \beta t + A\sin(\omega(t+\phi))$ with level α, linear trend β, and season frequency ϕ. The hidden Markov model approach can be applied by calling the procedure `sim.pointSource()` of the `surveillance` package.

Example. In this example, the application of the `R` package `surveillance` procedure `sim.pointSource` is illustrated. It creates a disease progress object `disProgObj` by using a hidden Markov model with length of 400 time points.

```
> disProgObj1 <- sim.pointSource(p = 0.99, r = 0.5, length = 400,
+                                A = 1, alpha = 1, beta = 0, phi = 0,
+                                frequency = 1, state = NULL, K = 1.7)
```

The parameters p and r specify the hidden Markov chain: $p = 0.99$ defines the probability of the Markov chain staying in the non-epidemic state to be 99%, while $r = 0.5$ defines the probability to stay in the state of an epidemic with 50%. By using the parameter `state` the Markov chain can be predefined.

Furthermore, no trend is used since `beta=0` is used, and a season with amplitude `A=1` and oscillation frequency (`frequency=1`) is assumed. A factor to create seasonal moves along the x-axis is not specified by `phi=0`. Moreover, `K=1.7` is an additional weight for an outbreak. The period length of the season is automatically set to be 52.

◇

5.1.2 Evaluation Parameters

Function `algo.quality()` in the package `surveillance` computes the number of true positives (correct found outbreaks), false positives (number of false found outbreaks), true negatives (number of correct found non outbreaks), false negatives (number of false found non outbreaks). Furthermore, the procedure computes sensitivity, specificity, and the Euclidean distance between $(1 - \text{spec}, \text{sens})$ and $(0, 1)$. Moreover, the function computes an average value for the lag of the outbreak recognizing by the system (Höhle, 2007).

Example. The above simulated disease progress is tested with a RKI algorithm, which uses a reference set including values from the past six weeks, and its quality values are computed.

```
> survResObj1 <- algo.rki1(disProgObj1, control = list(range = 200:400))
> algo.quality(survResObj1)
           TP FP TN  FN Sens Spec      dist       mlag
rki(6,6,0) 5  5  191 0  1    0.9744898 0.02551020 0
```

Very high sensitivity and specificity can be recognized with a alarm delay of zero. Thus, the algorithm performs very well to detect the outbreaks in the data.

◇

5.1.3 Comparison between Various Algorithms

The function `algo.compare()` returns the above described quality measures for each algorithm neatly arranged, so that the algorithms can be compared. Furthermore, `algo.summary()` summarizes evaluation parameters for different data sets.

Example. The methods are explained in an example using three simulated data sets, which are separately applied to three types of RKI algorithms.

First, further two test objects are created as described above.

```
> disProgObj2 <- sim.pointSource(p = 0.99, r = 0.5, length = 400,
+                                A = 1, alpha = 1, beta = 0, phi = 0,
+                                frequency = 1, state = NULL, K = 5)
> disProgObj3 <- sim.pointSource(p = 0.99, r = 0.5, length = 400,
+                                A = 1, alpha = 1, beta = 0, phi = 0,
+                                frequency = 1, state = NULL, K = 17)
```

These objects are monitored by the RKI algorithm in the range between 200 and 400, and their quality measures are computed. Regarding the results of the first disease progress object, the results in the first line already appeared in the last example.

```
> range <- 200:400
> control <- list( list(funcName = "rki1", range = range),
+                  list(funcName = "rki2", range = range),
+                  list(funcName = "rki3", range = range))
>
> compMatrix1 <- algo.compare(algo.call(disProgObj1, control=control))
> compMatrix2 <- algo.compare(algo.call(disProgObj2, control=control))
> compMatrix3 <- algo.compare(algo.call(disProgObj3, control=control))
> compMatrix1
           TP FP TN  FN sens spec      dist       mlag
rki(6,6,0) 5  5  191 0  1    0.9744898 0.02551020 0
rki(6,6,1) 4  0  196 1  0.8  1         0.2        5
rki(4,0,2) 4  0  196 1  0.8  1         0.2        5
```

Finally, the quality measures corresponding to the different data sets are summarized using `algo.summary()`.

```
> algo.summary( list(compMatrix1, compMatrix2, compMatrix3))
           TP FP  TN FN      sens      spec       dist     mlag
rki(6,6,0) 10 18 573  2 0.8333333 0.9695431 0.16942667 0.000000
rki(6,6,1) 11  4 587  1 0.9166667 0.9932318 0.08360773 1.666667
rki(4,0,2) 11  7 584  1 0.9166667 0.9881557 0.08417085 1.666667
```

On the one hand, it can be examined that the first type of algorithm has the least alarm delays in the three settings, and on the other hand that the second algorithm has the best values regarding sensitivity and specificity.

◇

5.2 Comparison of INLA with Analytical Bayes

Before starting the simulation studies themselves, it will be examined if INLA produces appropriate results.

5.2.1 Setting

It is assumed, that the predictive posterior obtained with INLA is similar to the predictive posterior obtained analytical by using the conjugated prior following a Gamma distribution. This approach is implemented in the Bayes algorithm as introduced in section 3.1.4.

In INLA, a centred Gaussian distribution is assumed as prior, and the posterior marginals are approximated by integrated nested Laplace approximation while in the Bayes algorithm the conjugated Gamma prior is assumed, so that the predictive posterior can be obtained analytical as a negative Binomial distribution.

For the comparison, a simple setting of simulated data will be used.

```
# simulation data set
stsim <- sim.pointSource(p = 0.99, r = 0.5, length = 314, A = 1,
alpha = 1, beta = 0, phi = 0, frequency = 1, state = NULL, K = 2)
observed <- stsim$observed
freq <- stsim$freq
```

Furthermore, the Bayes algorithm uses a reference set of similar observations. For this investigation, it will be chosen the last two years ($b = 2$) with corresponding nine values symmetrical around the current observation ($w = 4$).

```
control <- list(range=157:314, b=2, w=4, actY=TRUE)
timePoint <- 314

# compute reference set
basevec <- c()
if (control$actY) {
  basevec <- observed[(timePoint - control$w):(timePoint - 1)]
}
if(control$b >= 1) {
  for (i in 1:control$b) {
    basevec <- c(basevec, observed[(timePoint - (i * freq) -
                control$w):(timePoint - (i * freq) + control$w)])
  }
}
```

To ensure comparability, the same reference set is used in INLA while the observations are only influence by a fixed intercept.

```
# compute inla model only with fixed intercept
model <- inla(basevec~1,family='poisson', data=data.frame(basevec))
```

In this setting the predictive posteriors are computed using `algo.bayes()` and using `inla()` to verify this procedure.

5.2.2 Results

The predictive posterior of the analytical setting is obtained as described in Held (2008) and summarized in section 3.1.4.

```
# compute predictive posterior based on conjugated Gamma prior
sumBasevec <- sum(basevec, na.rm = TRUE)
lengthBasevec <- sum(!is.na(basevec))
predPost <- function(x){
   dnbinom(x,sumBasevec + 1/2, (lengthBasevec)/(lengthBasevec + 1))
}
```

The predictive posterior in the function `algo.hts` is obtained by Monte Carlo methods. Thus, 1000 values for y_{T+1} are sampled corresponding to its predictive posterior.

```
# compute predictive posterior using approximation of inla
yT1 <- vector(length=100*10, mode='numeric')
m <- model$marginals.fixed[[1]]
betaT1 <- sample(m[,1], prob=m[,2], replace=TRUE, size=100)
for(j in 1:100){
  yT1[((j-1)*10+1):(j*10)] <- rpois(n=10,lambda=exp(betaT1[j]))
}
```

In Figure 5.1 the comparison is shown while the predictive posterior obtained by INLA is displayed in a barplot and the analytical obtained one is superimposed with a dark green line. It can be examined that the predictions are very similar.

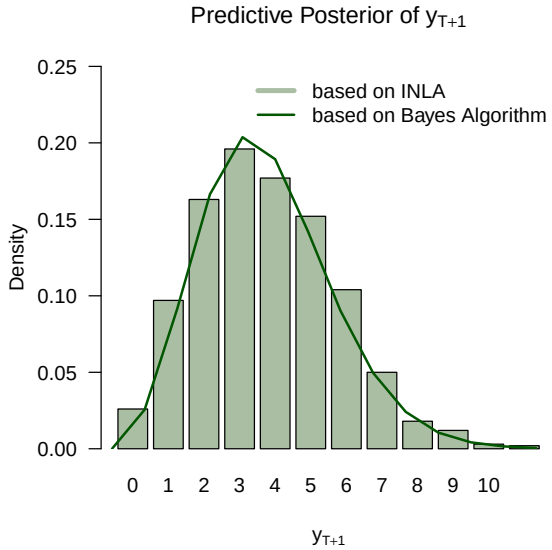

Figure 5.1: Comparison between predictive posterior obtained by INLA and analytical obtained predictive posterior using the simple conjugate prior-posterior Bayes algorithm

5.3 Observance of Significance Level

As described in section 3.1.2, infectious disease outbreak detection algorithms usually face the problem of multiple testing. The presented algorithms do not purpose a multiple adjustment procedure for the significance level. Therefore, in this section the algorithms are compared in respect to the number of triggered alarms if the time series does not include any outbreaks. It is measure corresponding to the average run length ARL^0, introduced in section 3.3.3 in detail.

5.3.1 Data Setting

For the purpose to observe the significance level, ten disease progresses for six years without outbreaks are simulated using hidden Markov models in the procedure `sim.pointSource`. Thus, the probability to get a no new epidemic at a time t is assumed to be $p = 1$ if there is no epidemic at time $t - 1$. Therefore, the hidden Markov chain transition into the state of outbreak is not possible.

```
stsim <- sim.pointSource(p = 1, length = 314, A = 1, beta = 0,
                         phi = 0, frequency = 1, K=1)
```

One obtains a suppositious disease progress for the years 2001 until 2007 with season, but without trend and any outbreaks. In Figure 5.2, one of them is shown exemplary. The years 2001–2004 are assumed to be training data while 2004–2007, i.e. after time point 157, will be monitored.

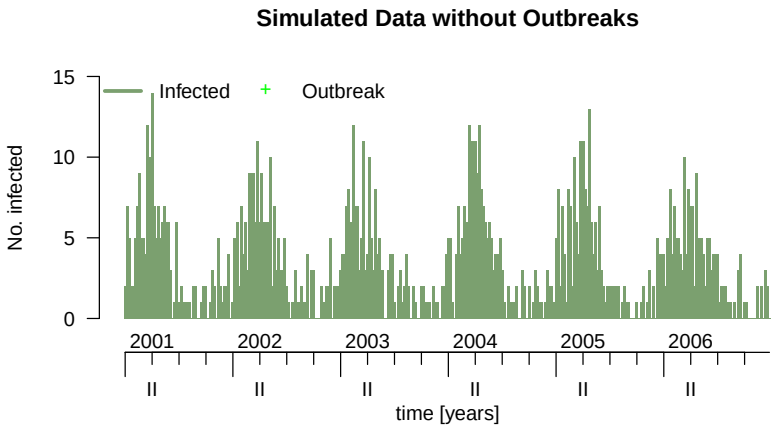

Figure 5.2: Exemplary simulated Data without outbreaks.

This evaluation procedure includes the significance levels $\alpha = 0.05$, $\alpha = 0.025$, and $\alpha = 0.01$ and is applied to the Bayesian version of the hierarchical time series algorithm with its different variations: the stationary (`iid`), neighbour (`rw2`), and linear (`rw2`) prior model with assuming Poisson or negative Binomial error distribution.

The results will be compared with the Farrington algorithm with and without option of reweighting the time series for past outbreaks, and the simple conjugate prior-

posterior Bayes algorithm. In each of them, the reference set includes symmetrically four values before and after the current week ($w = 4$) of the last two years ($b = 2$).

5.3.2 Results

In Table 5.1 the results of the simulations are shown. For each setting of algorithm and each α-level is computed the corresponding proportion of alarm as the ratio of the number of alarms n_A and the number of monitored time points n_T, i.e.

$$\alpha_{\text{observed}} = \frac{\#\text{ alarms}}{\#\text{ time points}} = \frac{n_A}{n_T}.$$

The algorithm would be optimal if the procedure has a proportion of alarms corresponding to the chosen α-setting. Thereby, it needs to distinguished between local and global α-level. Here, in case of dependent time points, the global α-level qualifies by $1 - (1 - \alpha_{local})^{n_T} \leq \alpha_{global}$. In the considered algorithms usually the local α-level is determined and no adjustments regarding multiple comparisons are made. However, in this simulation studies only the resulting global α-level is investigated.

			$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
HTS	Poisson	iid	0.038	0.017	0.004
		rw1	0.036	0.018	0.008
		rw2	0.038	0.017	0.009
	Negative Binomial	iid	0.039	0.012	0.004
		rw1	0.035	0.016	0.007
		rw2	0.039	0.021	0.013
Farrington	reweight = FALSE	(2,0,4)	0.017	0.009	0.004
	reweight = TRUE	(2,0,4)	0.030	0.016	0.007
Bayes		(2,0,4)	0.081	0.043	0.019

Table 5.1: Simulation results for setting of significance level observance: Average observed global α-level for respective algorithm in ten times series similar to Figure 5.2.

The developed Bayesian hierarchical time series algorithm yields consistent results for the different models. The α-level observations are usually lower than the defined significance level. Regarding the simulation for the significance level $\alpha = 0.01$, there is a tendency of increasing observed significance levels due to increasing allowed variation in the model. Thus, the linear prior model with negative Binomial error distribution assumption is the only model with a slightly too high observed level. Thereby, the assumption of Poisson or negative Binomial distribution did not show considerable differences. If a α-level of 5% is assumed, the algorithms keeps the level between 3.5% and 4%. The assumption of $\alpha = 0.025$ results observed levels ranging between 1.2% and 1.8%, and if α is assumed to be 1%, the observed significance level is between 0.4% and 1.3%.

The Farrington algorithm holds a much lower significance level than it is assumed. The assumed value is until three times higher than the observed one. In contrast,

the observed value of the Bayes algorithm almost doubles the assumed α-significance level. The results for the reweighted Farrington algorithm are comparable with these of the hierarchical time series algorithm. For the RKI algorithm the significance level cannot be specified. Therefore, a comparison is not possible. In the given setting a average observed significance level of 0.013 was computed.

Summarized, it can be examined that none of the methods keeps the α-level consistently while the smallest differences showed the hierarchical time series and the reweighted version of the Farrington algorithm. This observation might be reasoned in the type of simulated data which base on a Poisson Process. The method allow more variability by using quasi-Poisson or negative Binomial assumption.

5.4 Comparison Regarding Quality Key Parameters

In this section, the Bayesian hierarchical time series algorithm is evaluated regarding its quality measures sensitivity, specificity, and timeliness. These measures are compared to those obtained in other established methods of the RKI and Farrington et al. (1996).

5.4.1 Simulation Data

The simulated data is provided by the Centers for Disease Control and Prevention (CDC) in the United States(CDC Emergengy Risk Communication Branch (ERCB), Division of Emergency Operations (DEO) Office of Public Health Preparedness and Response (OPHPR), 2004). Regarding the comparison of aberration methods Hutwagner et al. (2005) introduced these simulation data.

There are given six-years time series which are obtained as follows: for 56 negative binomial parameter sets, estimated in observed diseases, are generated 1.000 iterations. Thus, 56.000 count data sets of basically 6 settings are given: mild, medium or strong seasonal progress with or without trend. Furthermore, the data has been adjusted for irregularities due to days of the week, holidays, and post-holiday periods.

For numerical reasons, the observations are aggregated per calendar week, such as the reporting data of the RKI. Furthermore, one parameter set was chosen at random of each setting, where 10 iterations were evaluated. In Table 5.4.1 the settings and their characteristics are summarized.

Ten types of outbreaks, representing various types of natural occurring events, are superimposed (see Table 5.4.1): the log normal distribution simulates a rapidly increasing outbreak, an inverted log normal distribution a slowly starting outbreak, and the spike a single day outbreak. These outbreak types are combined with different standard deviations and incubation times. During the simulation studies only the three types of distributions are distinguished, i.e. three types of outbreaks are investigated. In the following sections, aggregated and separated for each type of outbreak, sensitivity (sens), specificity (spec), and time of detection (mlag) for the selected outbreak detection algorithms are calculated.

Setting	Set file name	Trend	Seasonality	Mean	Standard Deviation
1	s03	No	Mild	7	42
2	s02	No	Medium	210	42
3	s05	No	Very	266	91
4	s01	Yes	Mild	630	231
5	s11	Yes	Medium	2107	553
6	s04	Yes	Very	42	28

Table 5.2: Chosen parameter sets of weekly aggregated simulated CDC data.

Outbreak Type	Distribution	Incubation Time (in days)	ζ	σ	Peak Size (X*SD)
1	Spike	1			3
1		1			2
2	Invert log Normal	< 7	1.3	0.4	3
2					2
2		7–14	2.4	0.3	3
2					2
3	Log normal	<7	1.3	0.4	3
3					2
3		7–14	2.4	0.3	3
3					2

Table 5.3: Outbreak types in the simulated CDC data.

The Bayesian hierarchical time series algorithm is evaluated regarding different specifications to figure out the best one for a specific outbreak setting. As in section 5.3, the Farrington, RKI, and Bayes algorithms are applied for comparison. Each of them includes 18 reference values from the previous two years, in each year nine values symmetrical around the comparable week.

5.4.2 Setting One: Mild-None Seasonality Without Trend

The first setting is characterized by no trend and mild seasonality, It has in average seven cases per week with strong variation (sd=42). Thus, it is a setting of low counts. It could be used to figure out algorithms, which have some problems with zero-observations and high variation. In Figure 5.3, the first time series of the setting is plotted.

Regarding the results in Table 5.4, the hierarchical time series algorithm has poor performance in this setting. At maximum 20% of the registered outbreaks are detected using the linear prior model. The outbreak types 1 and 3 have been detected the best by using a neighbour prior in a Poisson model while the outbreak type 2 is detected most frequently by the negative Binomial model with linear prior. The Poisson-iid-model has poor performance in detecting of all outbreak types. How-

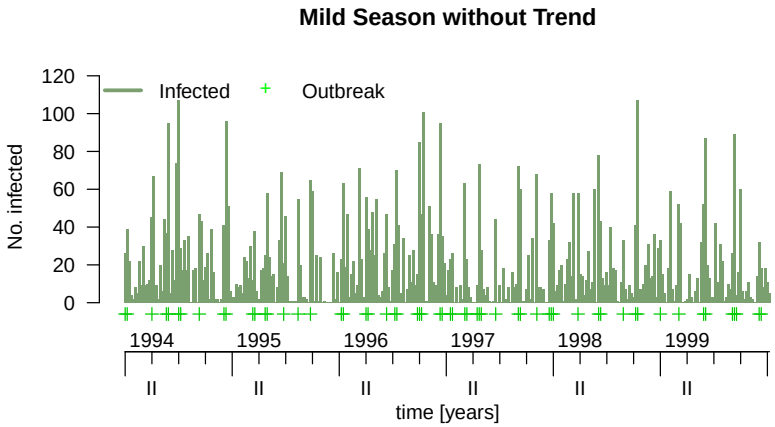

Figure 5.3: Example of CDC simulation data for setting one with mild seasonal setting without trend.

ever, the specificity, meaning the ability to detect non-outbreaks, ranges between 87% and 95%.
In comparison, the simple Bayes algorithm is examined as the best one. It has high values in sensitivity and specificity while the delay of alarm triggering is short. It has even acceptable specificity of 0.32 for the outbreak type 1, which is an one day spike and difficult to detect. The other algorithms perform in terms of their specificity a bit better than the hierarchical time series algorithm, but cannot persuade by good performance in this setting.

5.4.3 Setting Two: Medium Seasonality Without Trend

The time series of the second setting do not have trend either, but medium seasonality. In Figure 5.4, a high level of cases with a mean of 210 cases per week, low variation, and a large number of outbreaks can be recognized.

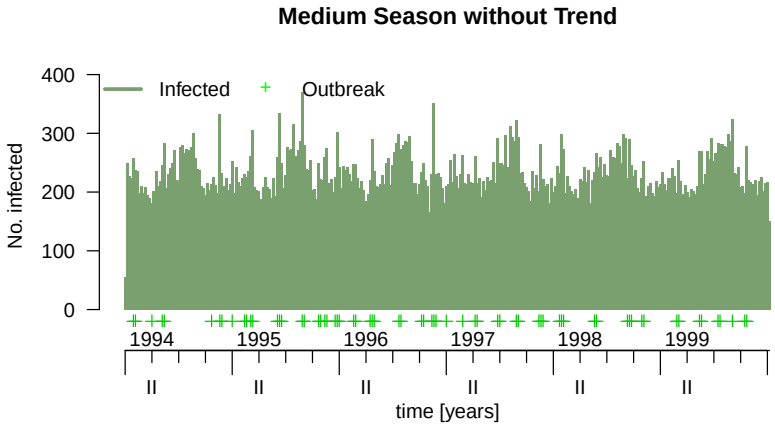

Figure 5.4: Example of CDC simulation data for setting two with medium seasonal setting without trend.

		Overall			Outbreak Type 1			Outbreak Type 2			Outbreak Type 3		
		sens	spec	mlag	sens	spec	mlag	sens	spec	mlag	sens	spec	mlag
Hierarchical Time Series													
Poisson	iid	0.06	0.96	14.32	0.07	**0.96**	13.98	0.06	**0.96**	12.48	0.04	0.96	12.80
	rw1	0.17	0.89	9.77	0.16	0.88	7.55	0.17	0.89	8.26	0.19	0.89	7.69
	rw2	0.19	0.87	7.91	0.10	0.85	6.51	0.17	0.86	7.04	0.22	0.86	6.82
neg. Binomial	iid	0.15	0.90	9.63	0.10	0.89	6.33	0.17	0.90	7.58	0.14	0.89	8.16
	rw1	0.15	0.90	9.63	0.10	0.89	6.33	0.17	0.90	7.58	0.14	0.89	8.16
	rw2	0.20	0.87	9.59	0.14	0.86	6.49	0.23	0.86	7.85	0.14	0.86	8.89
Farrington(4,0,2)	reweight=												
	FALSE	0.17	**0.99**	14.17	0.00	0.95	14.37	0.17	0.97	12.63	0.20	**0.97**	11.76
	TRUE	0.29	0.97	10.38	0.00	0.92	11.59	0.28	0.94	8.63	0.35	0.95	6.03
bayes(4,0,2)		**0.70**	0.87	**1.93**	**0.32**	0.75	**3.87**	**0.74**	0.80	**0.38**	**0.74**	0.80	**0.19**
rki(4,0,2)		0.40	0.93	7.99	0.09	0.86	8.75	0.42	0.89	6.30	0.45	0.89	4.92

Table 5.4: Simulation results for setting one with mild season and without trend

		Overall			Outbreak Type 1			Outbreak Type 2			Outbreak Type 3		
		sens	spec	mlag	sens	spec	mlag	sens	spec	mlag	sens	spec	mlag
Hierarchical Time Series													
Poisson	iid	0.11	0.89	13.48	0.10	0.89	11.70	0.11	0.89	13.25	0.11	0.89	10.88
	rw1	0.08	0.95	13.77	0.06	0.94	12.16	0.07	0.94	11.07	0.10	0.95	12.45
	rw2	0.12	0.90	10.67	0.13	0.90	**7.21**	0.13	0.90	8.51	0.12	0.90	9.62
neg. Binomial	iid	0.08	0.92	14.01	0.07	0.92	12.17	0.09	0.92	13.56	0.09	0.92	11.47
	rw1	0.06	0.95	14.75	0.06	**0.95**	13.54	0.04	0.95	13.04	0.09	0.95	13.36
	rw2	0.07	0.95	14.31	0.04	**0.95**	12.62	0.07	0.95	11.58	0.09	0.95	13.06
Farrington(4,0,2)	reweight=												
	FALSE	0.24	**1.00**	12.85	0.00	**0.95**	13.93	0.27	**0.97**	9.93	0.28	**0.97**	10.03
	TRUE	0.38	0.99	9.12	0.07	0.91	11.48	0.42	0.94	6.41	0.44	0.94	5.55
bayes(4,0,2)		**0.53**	0.95	**4.66**	**0.17**	0.86	**7.21**	**0.59**	0.90	**2.67**	**0.58**	0.90	**1.32**
rki(4,0,2)		0.26	0.99	12.29	0.00	0.94	13.76	0.28	**0.97**	9.35	0.31	**0.97**	9.01

Table 5.5: Simulation results for setting two with medium season and without trend

In the second setting (see Table 5.5), the hierarchical time series algorithm does not perform well. The outbreak detection rates are very low and do not exceed 13%. Furthermore, the average delay of alarm triggering is between 10 and 14 weeks. The model with Poisson distribution and random walk of order two performs the best. Especially, the one-day peaks are detected in comparison quite well.

Again, the simple conjugate prior-posterior Bayes algorithm shows good results. 53% of the outbreaks are detected, while 95% of the weeks without outbreaks are classified correctly. Similar quality measures can be recognized for the outbreak types two and three

5.4.4 Setting Three: Strong Seasonality Without Trend

The third setting is characterized by strong seasonality combined without any trend, The mean level is 266 cases per week. In Figure 5.5 a time series is shown exemplary.

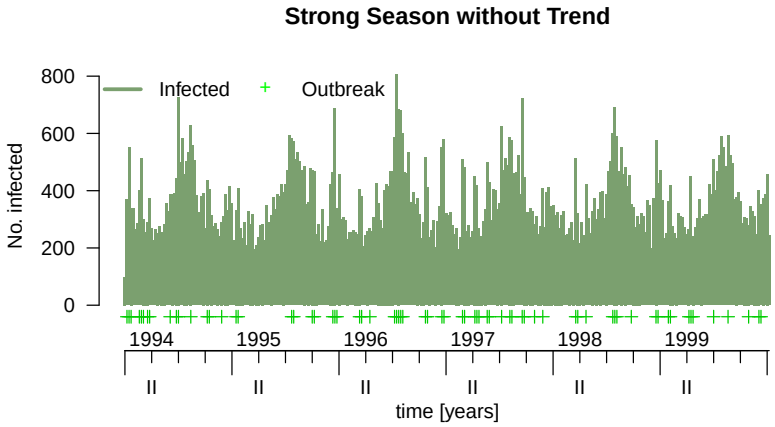

Figure 5.5: Example of CDC simulation data for setting three with strong seasonal setting without trend.

Regarding the results of the hierarchical time series model (see Table 5.6), the different models have similar results. The Poisson model with neighbour prior performs not well, but the best. It detects 17% of the outbreak weeks and 88% of the non-outbreak time points. Thereby, the delay is estimated to be eight weeks.

The performance of the Bayesian algorithm persuades by good quality values and short detecting delay of, in average, half a week regarding long-term outbreaks of type 2. The RKI and Farrington algorithms are able to detect all non-outbreak weeks while their sensitivity, especially in detection of outbreak type 1, is poor.

5.4.5 Setting Four: Mild-None Seasonality With Trend

This setting has mild seasonality with a weekly mean of 630 cases and standard deviation of 231 (see Figure 5.6). Furthermore, a strong trend is found.

		Overall			Outbreak Type 1			Outbreak Type 2			Outbreak Type 3		
		sens	spec	mlag	sens	spec	mlag	sens	spec	mlag	sens	spec	mlag
Hierarchical Time Series													
Poisson	iid	0.13	0.88	12.38	0.07	0.79	9.57	0.13	0.88	11.09	0.15	0.88	10.61
	rw1	0.17	0.88	8.05	0.09	0.78	6.05	0.23	0.87	5.54	0.17	0.87	6.80
	rw2	0.10	0.92	11.46	0.00	0.82	9.71	0.09	0.92	10.14	0.13	0.92	9.52
neg. Binomial	iid	0.15	0.88	11.65	0.13	0.79	9.31	0.15	0.87	9.65	0.16	0.88	10.35
	rw1	0.15	0.89	9.43	0.03	0.79	7.79	0.19	0.88	7.31	0.16	0.89	7.88
	rw2	0.15	0.89	9.25	0.15	0.79	6.63	0.18	0.88	7.41	0.15	0.88	8.44
Farrington(4,0,2)	reweight=												
	FALSE	0.22	**1.00**	12.90	0.00	**0.95**	14.04	0.26	**0.97**	10.52	0.23	**0.97**	11.23
	TRUE	0.36	0.99	9.51	0.00	0.91	12.47	0.42	0.95	6.58	0.37	0.95	7.06
bayes(4,0,2)		**0.71**	0.85	**1.25**	**0.37**	0.73	**3.07**	**0.76**	0.78	**0.47**	**0.72**	0.78	**0.52**
rki(4,0,2)		0.24	**1.00**	12.15	0.00	0.94	13.94	0.28	**0.97**	9.63	0.26	**0.97**	10.36

Table 5.6: Simulation results for setting three with strong season and without trend

		Overall			Outbreak Type 1			Outbreak Type 2			Outbreak Type 3		
		sens	spec	mlag	sens	spec	mlag	sens	spec	mlag	sens	spec	mlag
Hierarchical Time Series													
Poisson	iid	0.47	0.56	2.18	0.44	0.56	2.02	0.44	0.56	1.68	0.47	0.56	1.74
	rw1	0.06	**0.95**	14.34	0.06	0.94	14.61	0.06	0.94	13.54	0.07	0.94	12.78
	rw2	0.06	**0.95**	15.33	0.06	**0.95**	15.12	0.04	**0.95**	15.18	0.06	**0.95**	14.23
neg. Binomial	iid	0.50	0.53	1.29	0.58	0.52	1.14	0.47	0.52	1.12	0.52	0.52	1.32
	rw1	0.21	0.82	7.85	0.27	0.81	5.18	0.15	0.81	7.34	0.23	0.82	5.52
	rw2	0.14	0.88	10.42	0.16	0.88	8.64	0.12	0.87	9.92	0.14	0.88	8.08
Farrington(4,0,2)	reweight												
	FALSE	0.53	0.91	5.62	0.17	0.81	10.13	0.59	0.86	3.34	0.58	0.85	2.88
	TRUE	0.68	0.78	2.71	0.35	0.68	5.26	0.74	0.73	0.96	0.70	0.72	1.31
bayes(4,0,2)		**0.95**	0.29	**0.19**	**0.93**	0.25	**0.13**	**0.96**	0.26	**0.04**	**0.95**	0.26	**0.33**
rki(4,0,2)		0.58	0.89	4.84	0.23	0.79	9.21	0.62	0.83	2.61	0.62	0.83	2.46

Table 5.7: Simulation results for setting four with mild season and with trend

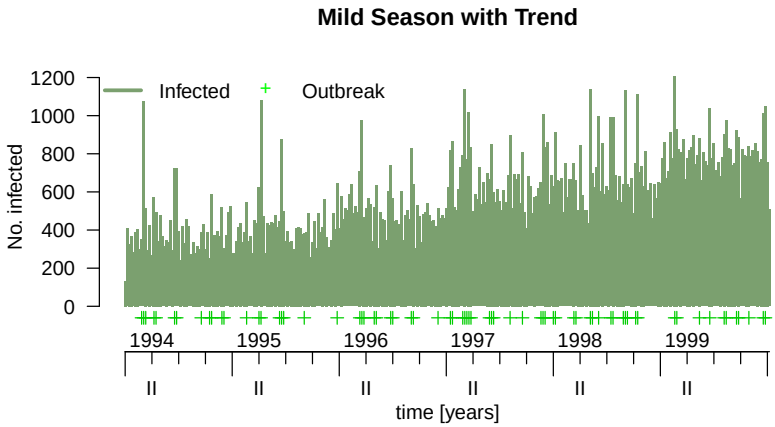

Figure 5.6: Example of CDC simulation data for setting four with mild seasonal setting with trend.

By far, the stationary model (iid) obtains the best values for sensitivity for the hierarchical time series algorithm while the negative Binomial model is a bit better (see Table 5.7). Half of the outbreaks are detected while half of weeks without outbreak are triggered as alarms as well. This result can be compared with a repeated binary decision at each time point without any foreknowledge, e.g. coin flipping. Thus, the hierarchical time series does not perform well. The expected delay is estimated to be less than 1.5 weeks.

For comparison, in terms of specificity, the other algorithms perform so-so as well. Probably the best performance obtains the reweighted Farrington algorithm by detecting 68% of outbreaks, 0.78 specificity and an expected delay of 2.7 weeks. Excluding the difficult to detect outbreak type 1, the rates are even better. The Bayes algorithm shows a high rate of sensitivity and a low expected delay, while its specificity is low. Thus, the Bayes algorithm triggers a high number of alarms.

5.4.6 Setting Five: Medium Seasonality With Trend

This setting has a high level of weekly counts, in average 2107, while the standard deviation is 553. It is characterized by trend and medium seasonality (see Figure 5.7).

Regarding the hierarchical time series algorithm, the stationary model performs the best (see Table 5.8). The decision which error distribution is assumed depends on the primary aim of detection. If higher sensitivity and less specificity is required, negative Binomial distribution is appropriate, otherwise the Poisson distribution.

The Bayes algorithm failed in this setting, because it triggered at almost all time points an alarm. The best quality has the RKI algorithm which detected 71% of outbreaks while its specificity is 0.75.

		Overall			Outbreak Type 1			Outbreak Type 2			Outbreak Type 3		
		sens	spec	mlag	sens	spec	mlag	sens	spec	mlag	sens	spec	mlag
Hierarchical Time Series													
Poisson	iid	0.27	0.76	5.86	0.24	0.68	3.78	0.22	0.75	4.50	0.27	0.76	5.76
	rw1	0.12	0.94	12.26	0.05	0.83	11.96	0.09	0.93	10.90	0.14	0.93	9.99
	rw2	0.02	**0.99**	18.46	0.00	**0.89**	17.70	0.03	**0.99**	17.75	0.03	**0.99**	17.62
neg. Binomial	iid	0.60	0.49	2.80	0.51	0.43	2.11	0.53	0.48	3.25	0.62	0.49	1.96
	rw1	0.10	0.95	12.50	0.05	0.84	12.12	0.07	0.94	10.50	0.12	0.94	9.59
	rw2	0.10	0.95	12.50	0.05	0.84	12.12	0.07	0.94	10.50	0.12	0.94	9.59
Farrington(4,0,2)	reweight												
	FALSE	0.64	0.83	4.14	0.33	0.73	4.22	0.72	0.77	2.61	0.64	0.77	2.21
	TRUE	0.89	0.56	0.88	0.73	0.47	1.80	0.95	0.50	0.44	0.88	0.51	0.43
	bayes(4,0,2)	**1.00**	0.01	**0.00**	**1.00**	0.01	**0.00**	**1.00**	0.01	**0.00**	**1.00**	0.01	**0.00**
	rki(4,0,2)	0.71	0.75	2.80	0.50	0.66	3.36	0.80	0.69	1.41	0.68	0.70	1.58

Table 5.8: Simulation results for setting five with medium season and with trend

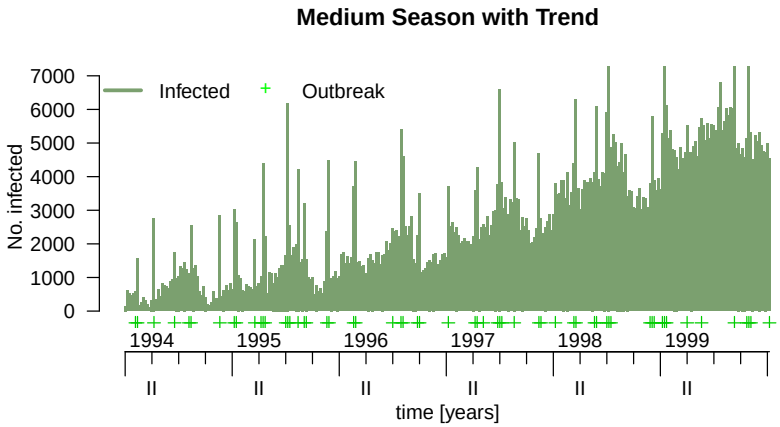

Figure 5.7: Example of CDC simulation data for setting five with medium seasonal setting with trend.

5.4.7 Setting Six: Strong Seasonality With Trend

The sixth setting is characterized by strong seasonality and the appearance of a trend. The weekly cases have mean 42 and standard deviation 28.

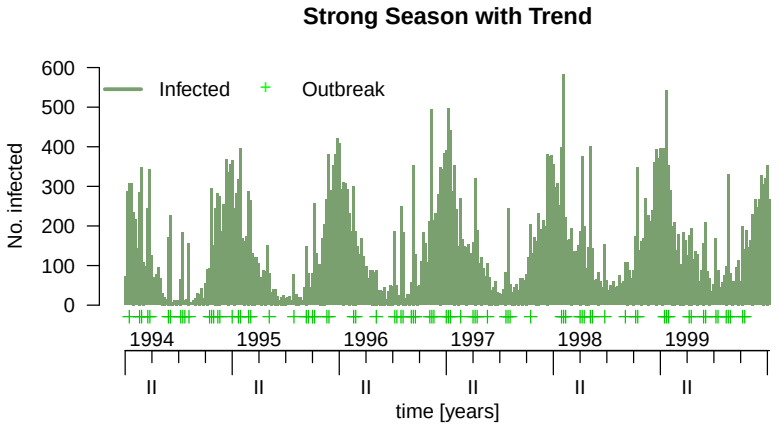

Figure 5.8: Example of CDC simulation data for setting six with strong seasonal setting with trend.

Regarding the Bayesian hierarchical time series model (see Table 5.9), it can be examined that the Poisson model with a stationary prior is not flexible enough to monitor such strong seasonality. This model almost never detected an outbreak. The random walk models perform better with sensitivity of about 16% and specificity of about 88%. All models were not able to detect properly the outbreak type 1.

The algorithms of RKI and Farrington detect all non-outbreak days correctly. Regarding the different types of outbreaks the specificity is a bit lower due to the setting of evaluation. Nevertheless, the simple Bayes algorithm is able to detect 78% of the

		Overall			Outbreak Type 1			Outbreak Type 2			Outbreak Type 3		
		sens	spec	mlag	sens	spec	mlag	sens	spec	mlag	sens	spec	mlag
Hierarchical Time Series													
Poisson	iid	0.03	**0.98**	17.70	0.00	0.88	15.12	0.02	**0.98**	17.67	0.03	**0.98**	17.00
	rw1	0.16	0.87	8.28	0.10	0.77	4.65	0.16	0.86	7.06	0.17	0.86	6.58
	rw2	0.16	0.87	8.42	0.05	0.77	7.58	0.18	0.87	6.60	0.16	0.86	6.90
neg. Binomial	iid	0.15	0.88	13.13	0.09	0.79	10.27	0.14	0.88	13.65	0.17	0.88	10.81
	rw1	0.16	0.88	8.38	0.07	0.78	7.61	0.16	0.88	6.88	0.19	0.88	6.41
	rw2	0.15	0.90	10.31	0.05	0.80	7.87	0.17	0.90	8.03	0.14	0.89	8.18
Farrington(4,0,2)	reweight												
	FALSE	0.25	1.00	12.03	0.00	**0.94**	12.84	0.27	0.97	9.26	0.27	0.96	8.70
	TRUE	0.39	1.00	8.12	0.03	0.91	11.27	0.42	0.96	5.39	0.45	0.94	4.10
bayes(4,0,2)		**0.78**	0.83	**0.92**	**0.65**	0.70	**1.56**	**0.78**	0.76	**0.44**	**0.80**	0.74	**0.18**
rki(4,0,2)		0.30	1.00	10.31	0.00	0.93	12.23	0.32	0.97	7.49	0.35	0.95	6.54

Table 5.9: Simulation results for setting six with strong season and with trend.

outbreaks while specificity is higher (0.83). The expected delay is estimated to be only one week. Its quality is good regarding all types of outbreak.

5.5 Comparison Regarding Computing Time

A usual surveillance time series, obtained by simulation using hidden Markov model (see Figure 5.2), is used to examine the algorithms regarding their computing time. All presented algorithms were applied on a 2.26GHz Duo CPU and their system time is measured.

HTS	Poisson	Negative Binomial
stationary	154.60	226.47
neighbour	140.22	258.39
linear	169.26	258.55
farrington(4,0,2)	reweight=F	reweighted=T
	3.32	4.58
RKI(4,0,2)	0.06	
Bayes(4,0,2)	0.02	

Table 5.10: Computing time of the outbreak detection algorithms in seconds.

In Table 5.10 the results are listed. The computation of the Bayesian hierarchical time series algorithm requires between 2:20 and 4:20 minutes. Regarding the usage of Bayesian models at each time point of the monitoring period, it is very fast. Alternatively, can be used Markov Chain Monte Carlo methods, whose computing time is many times higher.

Meanwhile, the Farrington algorithm requires 3 until 4.5 seconds, the RKI algorithm less than 1/10 seconds, and the Bayes algorithm takes only 0.02 seconds.

Thus, it is examined that the hierarchical time series algorithm needs much more time for computation than the others. Nevertheless, the computing time is in a manageable level. In practice, it may be acceptable as only one model is usually computed.

5.6 Conclusions of Simulation Studies

Due to the simulation studies one get an impression of the Bayesian hierarchical time series algorithm's performance in comparison to the more established RKI, Farrington and Bayes algorithm.

In this thesis, it is attached great importance to evaluate a large variety of setting. As consequence, the number of repetitions are chosen to be low (10 iterations). Comparisons to runs with 100 iterations did not show considerable differences.

At first, it is examined that the implementation of the Bayesian hierarchical time series algorithm is working for a wide variety of settings. Its performance is compared with the method used at the RKI, the established Farrington algorithm, and a simple version of Bayes algorithm.

In comparison to these methods, the hierarchical time series algorithm is able to keep the significance level the best.

The performance of the Bayesian hierarchical time series algorithm could not persuade in all settings. In a mild until medium seasonal setting a model with a stationary prior had good performance. In seasonal settings, the models with neighbour and linear prior have better results while their quality was still poor.

The simple conjugate prior-posterior Bayes algorithm shows problems in settings with trend. It performs very well in settings without trend while it specificity was very low in settings with strong trend.

The method of the RKI and the Farrington algorithm have comparable performance in the different settings. Thereby, better performance is examined in settings including trends, where detection rates up to 89% were achieved, while specificity is poor. Nevertheless, the performance, especially in data settings without trend, do not satisfy.

Apart from this, the computing time of the hierarchical time series algorithm is very slow in comparison to other surveillance methods. On the other hand, the algorithm is very fast regarding the included Bayesian models.

Chapter 6

Application to Campylobacter Data

In the second chapter medical, administrative backgrounds, as well as descriptive analysis were introduced for the Campylobacter data. Afterwards, several basic issues regarding outbreak detection and methods were introduced in chapter 3, and in chapter 4 was developed an alternative method based on Bayesian hierarchical time series models. In the following, these methods will be applied for infectious disease outbreak detection with the Campylobacter data.

The chapter starts with some surveillance-specific preparations of the data which includes the investigation of a definition for outbreak time points. Unfortunately, it is not suitable, so that the following application of the surveillance methods is done without outbreak state reference. Thus, the algorithms performance is examined qualified, and the developed Bayesian hierarchical time series algorithms results are compared to those of the other methods by the key measures sensitivity and specificity.

6.1 Preparation of Campylobacter Data for Surveillance Analysis

In the following, preprocessing steps and considerations on the preparations to the Campylobacter data for surveillance analysis are presented.

6.1.1 Restrictions on Reporting Data

The data covers the reports between 2001 and 2009. With the implementation of the reporting system in 2001, data inaccuracies were recognized. Therefore, the reports of the first year are excluded for the analysis.

```
> since02 <- which(cam.aggr$date >= as.Date("2001-12-31"))
> cam.aggr <- cam.aggr[since02, ]
```

The time period between 2002 and 2006 will be used as training period to obtain a approximation of in-control data. Consequently, outbreak cases are excluded in this time period. The defined rule is evaluated on the datasets between 2007 and 2009.

```
> inControl <- cam.aggr$cases - cam.aggr$outbreak
> range <- which(cam.aggr$date >= as.Date("2007-01-01"))
```

Furthermore, only the mainly diagnosed Campylobacter bacteria type jejuni, Campylobacter spp., and Campylobacter jejuni not differentiated are monitored to obtain reporting datasets as homogeneous as possible, which consist mostly of Campylobacter cases with subtype jejuni. Therefore, for the analysis, the category of 'others' is excluded in advance.

6.1.2 Definition of Past Outbreak Reference

For creating a disease progress object `disProgObj`, required when the framework of the R package `surveillance` is used, a time series for the state of outbreak needs to be defined. The construction of the outbreak reference set has particular importance for quality interpretation of the algorithm results. Thus, if the a priori state definition is inadequate the algorithms quality will be examined as poor as well.

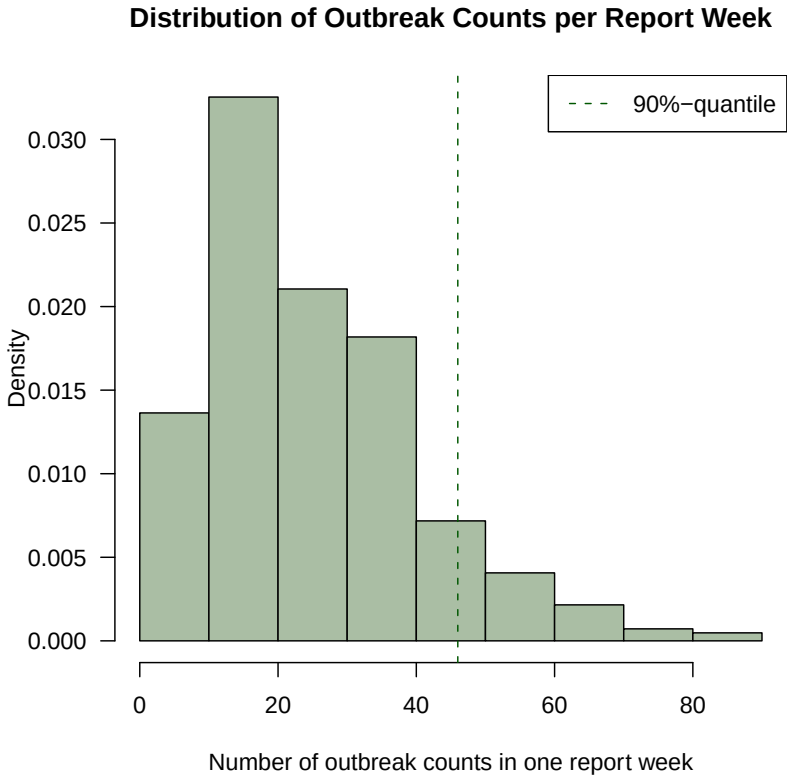

Figure 6.1: Distribution of weekly number of outbreaks

One possibility is the definition of an outbreak day as follows: If the number of outbreak cases per week is larger than a 90%-quantile threshold, the progress is defined to be out-of-control. Comparing to Figure 6.1.2, this threshold is set to 46 outbreaks per week.

Already recognized outbreak cases have been excluded, so that this information would be included to the system by the state definition. A disadvantage of this definition occurs due to the seasonal characteristics in the outbreak cases time series. Thus, the described measure includes only high season days. Especially the reporting of unusual many cases in spring 2007 mentioned in literature (Jansen et al., 2007) cannot be reproduced by this measure. Therefore, the definition is rejected and it is preferred to monitor the time series with the assumption that no past outbreaks are included in the data. The algorithm's performance will be achieved using qualified examination of the results.

6.1.3 Create Disease Progress Object

To monitor the time series a disease progress object `disProgObj` needs to be created. As discussed before, data since 2002 is included, the range represent the years 2007 until 2009. The state of outbreak is unknown and therefore no outbreak time points are defined.

```
> state <- rep(0, length(cam.aggr$cases))
> cam.disProg <- create.disProg(week = 1:length(cam.aggr$date),
+     observed = inControl, state = state, start = c(2002,
+         1), freq = 52, epochAsDate = TRUE)
```

6.2 Application of Surveillance Algorithms

This section will show the application and results of different outbreak detection methods to the Campylobacter data. The current used method for these data is the RKI algorithm. Alternatively, will be given by Farrington and the simple conjugate prior-posterior Bayes Algorithm of Höhle (2007). The developed Bayesian version of the hierarchical time series algorithm will be applied to examine its potential in the next section. Note, that there are not defined any outbreak time points regarding the missing true or suitable definition of it.

6.2.1 RKI Method

Currently, outbreak detection in Campylobacter data is done by using the RKI method which is called by `rki()` of the `surveillance` package. The reference set is constructed by the $w = 4$ weeks before and after the current week in the last two years $b = 2$.

```
> control = list(range = range, b = 2, w = 4)
> rki <- rki(disProg2sts(cam.disProg), control = control)

Running  algo.rki  on area  1  out of  1
```

Comparing Figure 6.2.1, the RKI algorithm detects two outbreak periods, one in the spring of 2007, one in late 2008, and one single week in the early beginning of 2009.

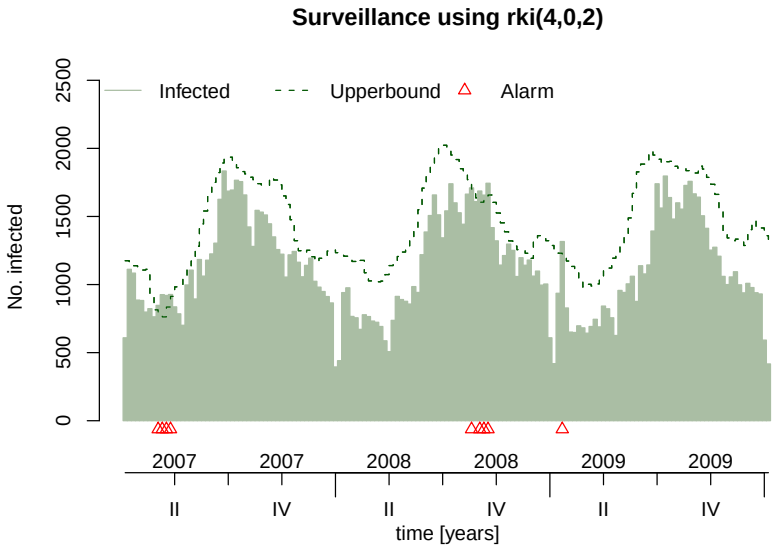

Figure 6.2: Automated Outbreak detection in the IfSG Campylobacter data by RKI method

6.2.2 Farrington Algorithm

The algorithm of Farrington et al. (1996) can be applied by calling the function `farrington()` of the `surveillance` package. The reference set is chosen as before. It is taken into consideration the last two years ($b = 2$), and corresponding four weeks before and after the current week ($w = 4$). A trend cannot be assumed, because less than 4 years of training period is available. Furthermore, the 2/3-power-transformation should be used to normalize the data, and the reweighing procedure for past outbreaks is applied, as well. Because the function is defined for a two-sided significance level, `alpha` is specified by $0.05 \cdot 2 = 0.1$ to obtain the required one-sided 5%-significance level.

```
> control <- list(range = range, b = 2, w = 4, trend = FALSE,
+     powertrans = "2/3", reweight = TRUE, alpha = 0.1)
> farr <- farrington(disProg2sts(cam.disProg), control = control)

Running  algo.farrington  on area  1  out of  1
```

In Figure 6.2.2, the result of the Farrington algorithm applied to the Campylobacter data is shown. The algorithm detects two periods in the beginning of 2007. Furthermore, in fall 2008 a sequence of alarms is triggered. At the turn of the years , there are each two single alarm weeks.

6.2.3 Simple Conjugate Prior-Posterior Bayes Algorithm

The simple conjugate prior-posterior Bayesian Algorithm bases on Poisson distributed observations in a reference set and the assumption of a Gamma prior, which leads to negative Binomial distributed predictive posterior for the current observation. It is implemented with the procedure `bayes()` and applied as follows:

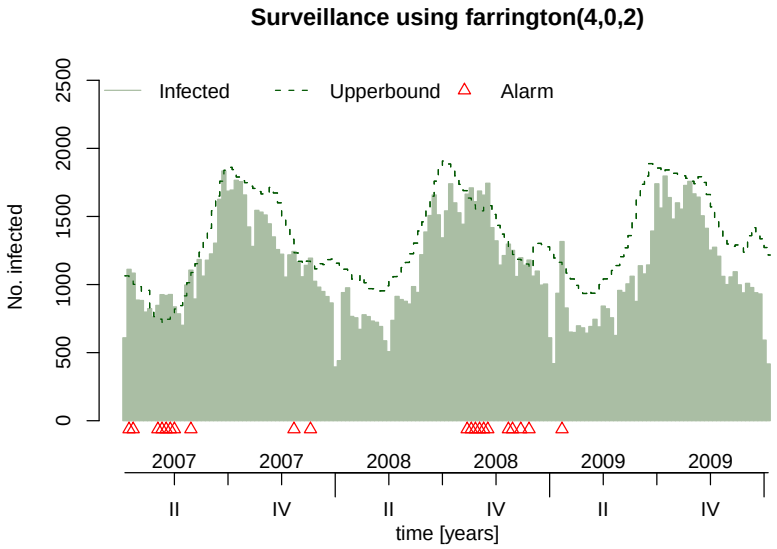

Figure 6.3: Automated outbreak detection in the IfSG Campylobacter data by Farrington algorithm

```
> control <- list(range = range, b = 2, w = 4, actY = FALSE,
+     alpha = 0.05)
> bayes <- bayes(disProg2sts(cam.disProg), control = control)

Running algo.bayes on area 1 out of 1
```

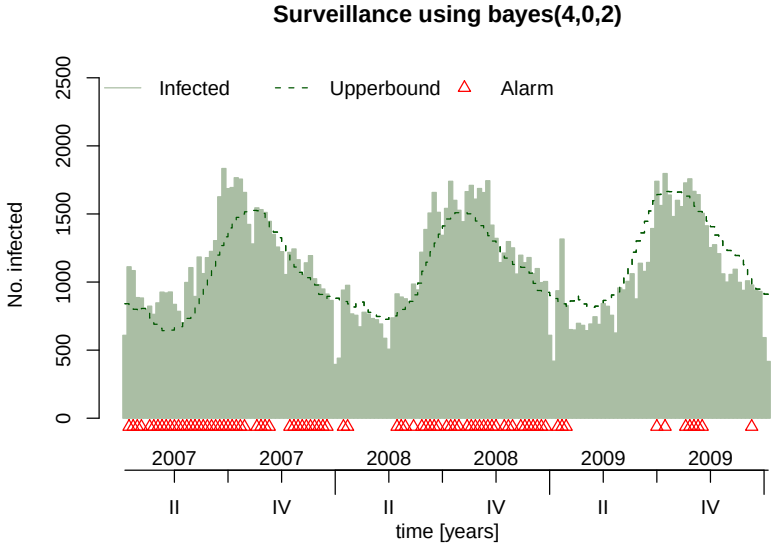

Figure 6.4: Automated outbreak detection in IfSG Campylobacter data by simple Bayes algorithm

The Bayes algorithm does not perform very well on the Campylobacter data. It detects almost the whole time period of the defined range. Only in 2009, the number of alarms decreases to a manageable level.

6.3 Application of Hierarchical Time Series Algorithm

In the following, various versions of the hierarchical time series algorithm are applied to the Campylobacter reporting data to examine the potential of the algorithm. First, the simple version corresponding to the established algorithms presented before is applied. Later on, covariates are added and the adjustments for reporting delay are tested in application.

6.3.1 Simple Application

In the model by an der Heiden et al. (2010) because of observed overdispersion a negative binomial error distribution was assumed. Thus, in the monitoring will be assumed this error distribution as well. Furthermore, it was figured out that the neighbour model is the best assumption for the varying time trend. The stationary prior cannot be applied because of strong seasonality in the Campylobacter data and the linear model seems to be too noisy.

```
> control <- list(range = range, co.arg = NULL, prior = "rw1",
+     family = "nbinomial", alpha = 0.05, mc.betaT1 = 100,
+     mc.yT1 = 10)
> hts <- algo.hts(cam.disProg, control = control)
```

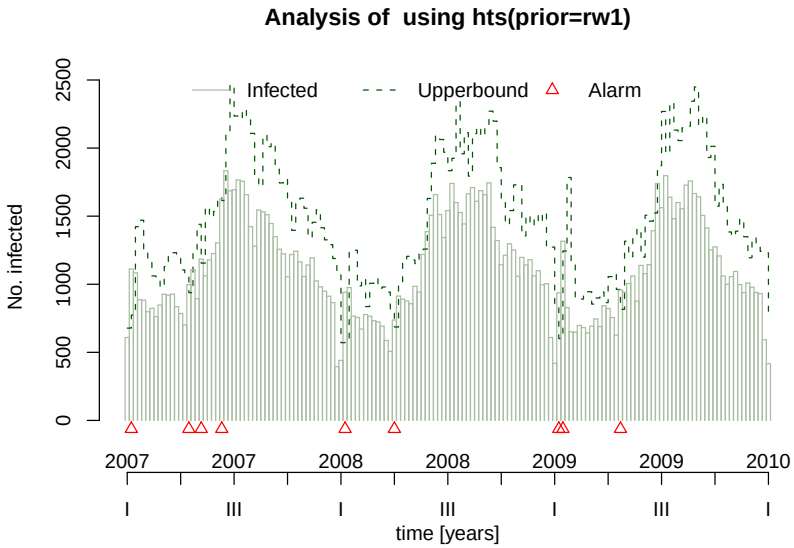

Figure 6.5: Automated outbreak detection in IfSG Campylobacter data by Bayesian hierarchical time series algorithm

The number of alarms is on a manageable level, while the higher level in early 2007 cannot detected completely. Furthermore, the algorithm is able to detect some high peaks in the following low seasons of 2008 and 2009.

6.3.2 Application including Covariates

First, a simple Bayesian model is fitted by using `inla()` examining the influence of the covariates to the observed Campylobacter cases. In section 2.4 it was assumed the influence of absolute humidity in the previous two weeks, proportion of sex and of children cases in the age under 10 years. Since the response variable in this model is the number of infected person at the current time point, and here are used the number of reported cases, it need to be considered the incubation period and the reporting delay. The incubation time is between two and five days. In chapter 2.3.3 was found that 95% of the cases were reported within three weeks. Thus, it need to be include further four weeks of absolute humidity to obtain a comparable model. It was found that sixth lag does not have any influence. Relevant variables construct the model data frame and the inla negative binomial model was applied.

```
> inla.cam <- inla(cases ~ f(time, model = "rw1") + l1.hum +
+     l2.hum + l3.hum + l4.hum + l5.hum + age + sex, family = "nbinomial",
+     data = modeldat)
> summary(inla.cam)

Call:
c("inla(formula = cases ~ f(time, model = 'rw1') + l1.hum + l2.hum +
l3.hum + l4.hum + l5.hum + age + sex, family = 'nbinomial',
data = modeldat)")

Time used:
 Pre-processing    Running inla Post-processing           Total
      0.2143540       3.3029730       0.3781068       3.8954339

Fixed effects:
                     mean          sd  0.025quant      0.5quant
(Intercept)  7.2940147274 0.262443525  6.77819626  7.2940518949
l1.hum       0.0008993357 0.005631842 -0.01016509  0.0008984256
l2.hum      -0.0033044966 0.005661695 -0.01442929 -0.0033048483
l3.hum      -0.0111122213 0.005692110 -0.02229617 -0.0111128118
l4.hum      -0.0083833875 0.005663672 -0.01951399 -0.0083830563
l5.hum      -0.0079432815 0.005601814 -0.01895273 -0.0079427401
age          1.4642870049 0.541335505  0.40023804  1.4644144940
sex         -0.6275952050 0.409668547 -1.43220401 -0.6277324843
              0.975quant          kld
(Intercept) 7.808883e+00 4.321562e-07
l1.hum      1.195355e-02 8.740793e-09
l2.hum      7.806711e-03 4.449264e-08
l3.hum      5.948503e-05 3.881157e-07
l4.hum      2.729561e-03 4.598067e-07
l5.hum      3.047383e-03 3.030373e-07
age         2.526081e+00 1.359389e-06
sex         1.766872e-01 3.563528e-07
```

```
Random effects:
Name   Model    Max KLD
time   RW1 model   9e-05

Model hyperparameters:
                     mean       sd        0.025quant 0.5quant
Overdispersion       28827.640  24251.180 4863.629   20195.642
Precision for time   36.595     2.791     31.505     36.446
                     0.975quant
Overdispersion       95792.963
Precision for time   42.467

Expected number of effective parameters(std dev): 386.30(2.775)
Number of equivalent replicates : 1.082
WARNING: The approximations could be not very accurate

Marginal Likelihood:  -2767.68
Warning: Interpret the marginal likelihood with care if the prior
model is improper.
```

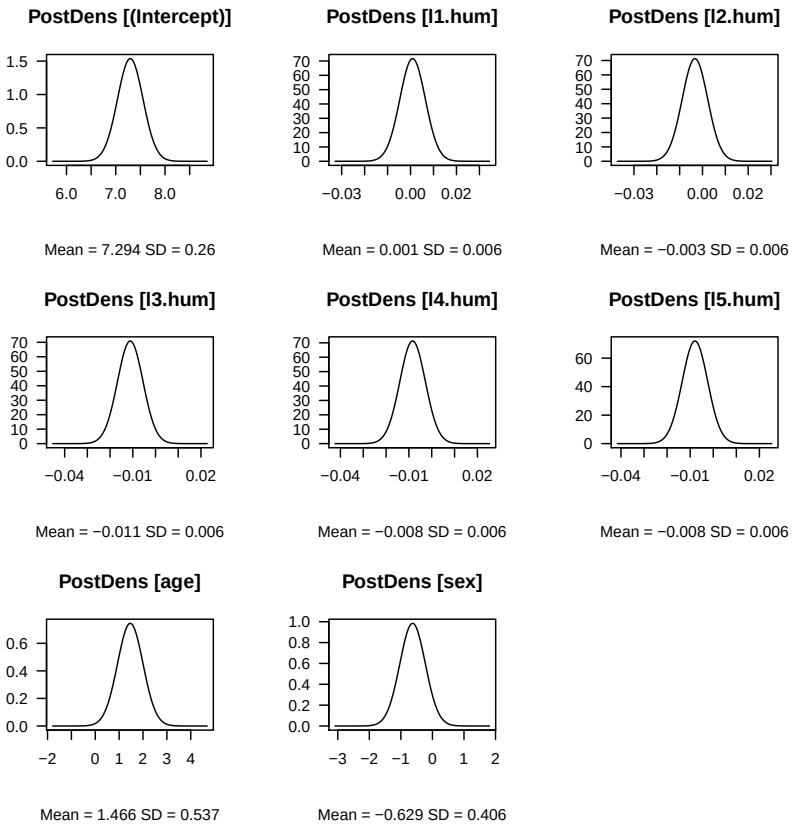

Figure 6.6: Posterior marginals for fixed effects in INLA model.

The posterior marginals approximations corresponding to the INLA model are shown in Figures 6.6 and 6.7.

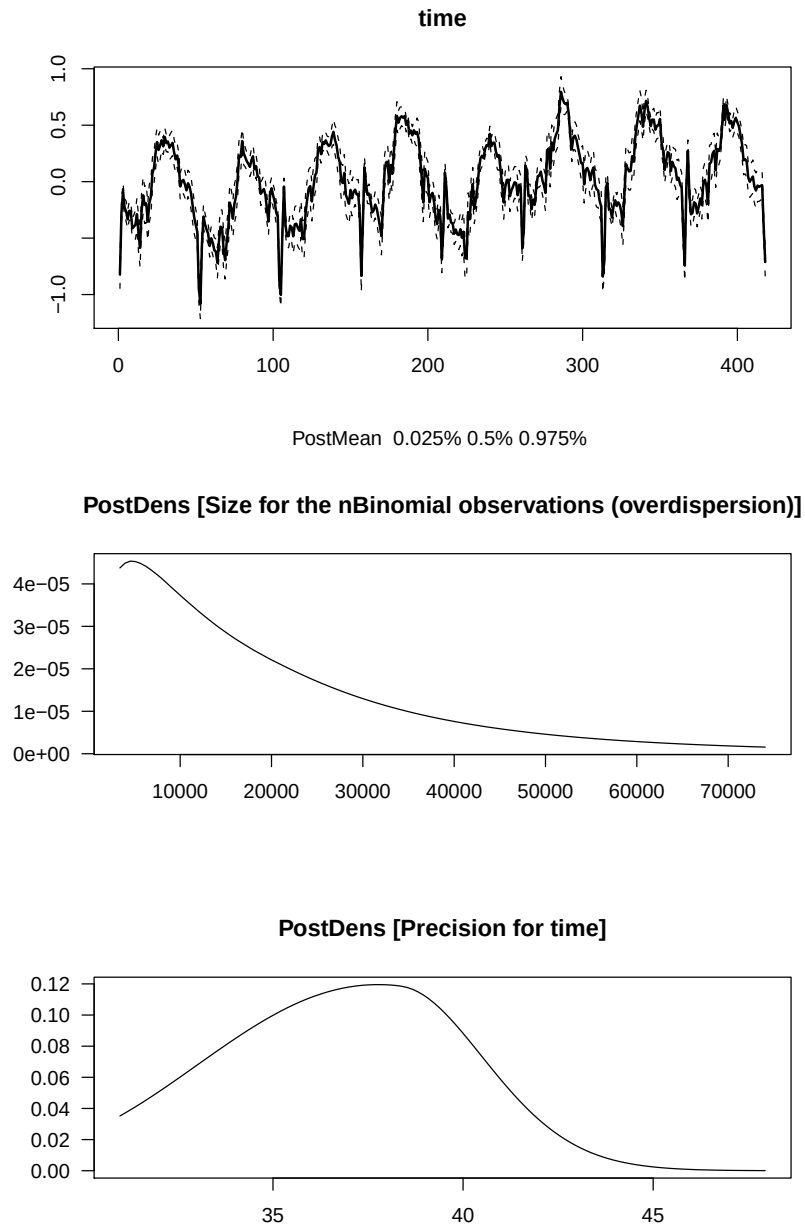

Figure 6.7: Posterior marginals for random effects and hyperparameters of INLA model.

In the following, monitoring including covariates with Bayesian hierarchical time series algorithm will be presented. Because of instabilities in the algorithm it is used Poisson instead of negative Binomial error distribution as mentioned in section 4.4.3.

```
> control <- list(range = range, co.arg = cbind(l1.hum, l2.hum,
+     l3.hum, l4.hum, l5.hum), prior = "rw1", family = "poisson",
```

```
+     alpha = 0.05, mc.betaT1 = 100, mc.yT1 = 10)
> hts.hum <- algo.hts(cam.disProg, control = control)
```

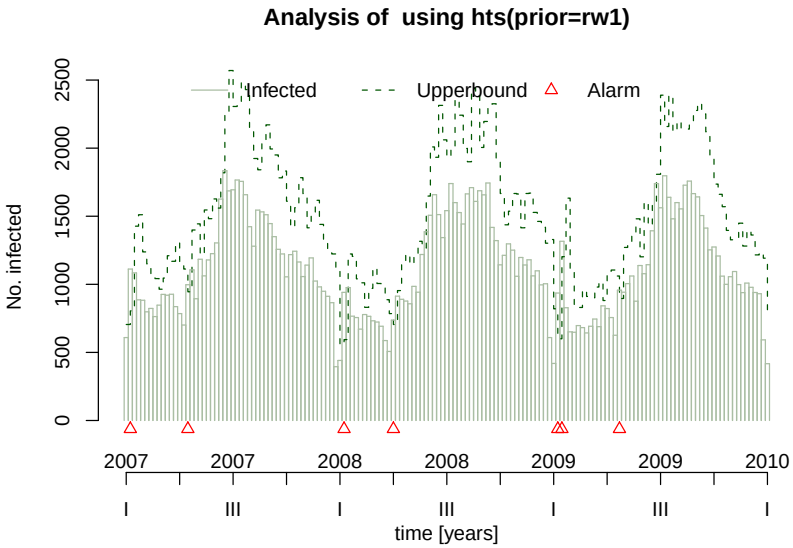

Figure 6.8: Automated outbreak detection in IfSG Campylobacter data by Bayesian hierarchical time series algorithm including covariates

The Figure 6.3.2 shows a very similar picture as the monitoring result by the simple hierarchical time series algorithm without included covariates. In the low season several outbreak alarms are triggered.

6.3.3 Adjustment for Reporting Delay

First, the proportions of reported cases are computed and examined over time to answer the question if its means can be generalized and used for prediction. As introduced in section 4.5, the means of the observed proportions during the training period are used for adjustment of outbreak detection. Thus, the proportions of delay are computed by the ratio of cases reported and total number of cases.

```
> pi1 <- cam.aggr$delay1/cam.aggr$cases
> pi2 <- cam.aggr$delay2/cam.aggr$cases
> pi3 <- cam.aggr$delay3/cam.aggr$cases
```

Figure 6.3.3 shows the proportion of reported cases within one week increases slightly over time. Furthermore, mild seasonality is detected. These characteristics disappear with increasing time of reporting and will be ignored in further analysis. The extrapolation for the time points is justifiable, because the means do not differ substantially.

Afterwards, the mean of these proportions is calculated including only the time points of the training period.

```
> mpi1 <- mean(pi1[-range])
> mpi2 <- mean(pi2[-range])
> mpi3 <- mean(pi3[-range])
```

Within one week 45.8% of the cases are reported, after two weeks the proportion increases to 77.5%, and after three weeks in average 84.7% of the cases are reported.

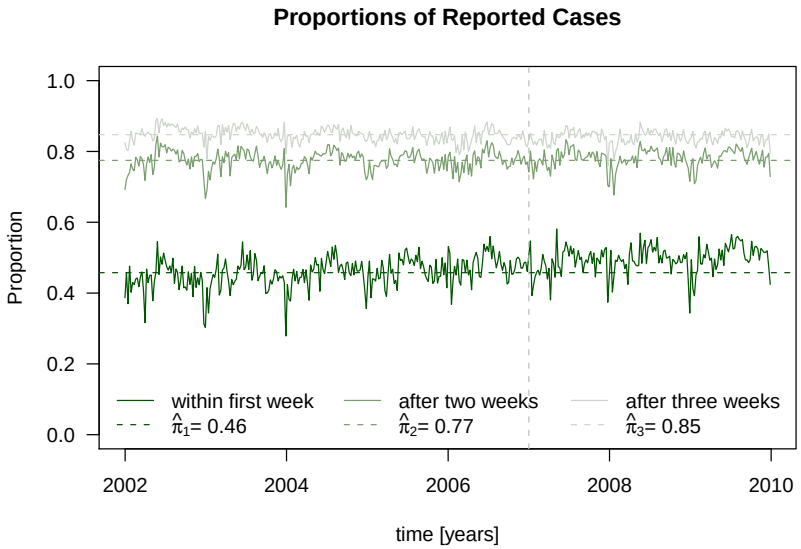

Figure 6.9: Proportions of reported cases over observed time.

At each time point, the partially observed count is adjusted by the averaged proportion of already reported cases. Thus, it is estimated how many cases would be reported when completeness is reached. The above calculated threshold ξ_{hts} is used and compared with the adjusted observed value of already reported cases. Thus, the alarms are computed corresponding to:

```
> alarm.delay1 <- (cam.aggr$delay1/pi1)[range] > hts$upperbound
> alarm.delay2 <- (cam.aggr$delay2/pi2)[range] > hts$upperbound
> alarm.delay3 <- (cam.aggr$delay3/pi3)[range] > hts$upperbound
```

To simplify the application of the adjustment for reporting delay no covariates were included in the fitting of the Bayesian hierarchical models as their integration does not change the procedure of reportng delay adjustment.

6.4 Algorithm Comparisons

In section 6.1.2 a definition for the outbreak state in the Campylobacter time series which used the number of weekly outbreak counts was investigated and proved to be unsuitable. Therefore, the surveillance methods were applied without assuming any outbreak time points. Thus key quality measures for outbreak detection as introduced in section 3.3 cannot be computed to obtain comparable performance of the algorithms. In the context of this thesis, especially the comparison with the new developed hierarchical time series algorithm is of special interest.

6.4.1 Quality of Algorithms

The algorithms results itself were examined qualified. The quality comparisons of the applied methods regarding the Bayesian hierarchical time series algorithm are done by using specificity and sensitivity. Therefore, the triggered alarms of the simple Bayesian hierarchical time series algorithm are compared with those triggered by the other algorithms (see Table 6.1).

	sensitivity	specificity
RKI	0.11	0.95
Farrington	0.10	0.95
Bayes	0.08	0.97
with Covariates	1.00	0.99
delay within 1 week	1.00	1.00
delay within 2 weeks	1.00	1.00
delay within 3 week	1.00	1.00

Table 6.1: Comparison with Bayesian hierarchical time series algorithm alarms

In comparison with the results of the RKI method a sensitivity of 0.11 and specificity of 0.96 is found. The hierarchical time series algorithm compared with the results of the Farrington algorithm results sensitivity of 0.14 and specificity of 0.97. When the alarms are compared with these triggered by the Bayes algorithm specificity of 0.97 and a very low sensitivity of 0.06 is observed. This corresponds to the high number of triggered alarms by the Bayes algorithm.

6.4.2 Quality of Algorithm with Consideration of Covariates

The INLA model was able to prove an influence of absolute humidity to the reported cases of Campylobacter infections. Thus, it would be expected that the algorithm result does change by including these covariates. This can be examined only marginal, because sensitivity and specificity equals almost one (see Table 6.1). Thus, the results are similar whether covariates are included, or not.

6.4.3 Quality of Adjustments for Delay

To assess quality of the adjustment for the reporting delay consideration the 'gold standard' is assumed to be the monitoring result which is achieved when all data is availible. Thus, the result of Bayesian hierarchical time series algorithm with complete Campylobacter data as presented in section 6.3.1.

In the Table 6.1 the above computed alarm sets are compared with the 'gold standard'. It can be examined that the adjustment is working very well regarding all types of delays. All time points are classified in the same category compared as in the case all data is given.

6.4.4 Conclusions

All applied versions of the algorithms gave reasonable results. They differ in the number of triggered alarms. Thus, the RKI method gave 9 alarms, the Farrington triggered 21 alarms, the Bayes algorithm 86 alarms, which is not on a manageable level, and the hierarchical time series algorithm 9 alarms out of the monitored 157 time points.

Final quality comparisons of the algorithms are not possible, because an objective and generally valid reference definition for true outbreak state is not available. In literature spring 2007 was examined as a time period with unusual many Campylobacter cases (Jansen et al., 2007). This period was detected the best by the Farrington algorithm.

The application of including covariates is working while the integration of absolute humidity did not change the resulting alarm set. The adjustment for reporting is reasonable and derives the same results as the procedure including the complete reporting data.

Chapter 7

Discussion

It is a capital mistake to theorize before one has the data.

Sherlock Holmes in Arthur Conan Doyle's 'Scandal in Bohemia' (Gaither and Cavazos-Gaither, 1996)

In this thesis, a particularised description of the reported Campylobacter infections in Germany between 2001 and 2009 was given. The weekly aggregated counts were examined in Germany. Regarding the spatial distribution and local incidence time series of selected districts, spatial differences could be recognized. For better understanding of the flow of reporting, especially the reporting delay in all its stages was examined. Moreover, male patients and specific age groups were found to be more exposed to Campylobacteriosis. Finally, associations for belonging to an outbreak on the factors sex, age, and bacteria type have been found.

The surveillance analysis of the Campylobacter data the algorithms of the RKI, Farrington, Bayes, and the Bayesian hierarchical time series algorithm were applied and their results compared. All methods were able to trigger off several alarm in the spring of 2007, which was considered by other sources to have unusual high counts (Jansen et al., 2007). Furthermore, the integration of absolute humidity as covariate process, reflecting the weather in Germany, into the hierarchical time series algorithm was shown. Due to the reporting delay the considered order of lags need to be enlarged.

We are making forecasts with bad numbers, but bad numbers are all we've got.

Michael Penjer, New York Times, September 1, 1989) (Gaither and Cavazos-Gaither, 1996)

There are several sources of irregularities in surveillance data. In this thesis, especially reporting delay and past outbreaks in the time series were investigated.

In general, reporting delays bias the results, and most authors emphasize the importance of their consideration, but there is not much literature about adjustment methods. In this thesis an ad-hoc adjustment was introduced to handle this problem. It investigates the proportions of reporting in dependence of the delays in the past and use them to adjust the computed threshold.

Another issue without an appropriate solution is the handling of past outbreaks in the time series. Most of the methods ignore them. Farrington et al. (1996) suggested to down weight the time series, but disadvantages of this approach were shown in chapter 3.2.

Methods cannot be better than the data they base on. Thus, a way out might be the integration of further information from other sources. In this thesis, a surveillance method based on fast available, but inaccurate Internet search query data was shortly mentioned. The combination with the delayed, but well defined public health data, might improve surveillance results.

How easy it is for unverified assumptions to creep into our reasoning unnoticed!

Beveridge, W.I.B., The Art of Scientific Investigation (Gaither and Cavazos-Gaither, 1996)

Statistical methods base on various assumptions which might need to be reconsidered.

The aim of all investigated surveillance methods is the detection of an unusual high count of infection. Thereby, only the public health definition of an outbreak is used. It is a critical concept, because the existence of an aberration is necessary, but not sufficient, for the occurrence of an epidemic (see Stroup et al., 1993). Other features of an outbreak, such as linked cases, are not considered.

The presented algorithms use an in-control process model to derive an upper prediction border. A naive approach would be the direct modelling of the threshold, e.g. by quantile regression. Thereby, a threshold could be derived by modelling directly e.g. the 95%-quantile. The problem with it could be quantile crossings and instabilities of estimations in the edges. Furthermore, it need to find a appropriate handling for the count data nature which are not standard for quantile regression (Koenker, 2005).

During the application of surveillance analysis on Campylobacter, the choice of error distribution assumption was based on a model explaining the influence of weather to the infection counts. In general, the choice of distribution is more complicated. Assuming Poisson might be too restrictive in some infectious diseases. Otherwise, negative Binomial or quasi-Poisson can be chosen.

Seasonality is a common characteristic of time series in surveillance. In the context of this thesis, their consideration through constructing reference sets or modelling via splines was investigated. Other techniques such as time series decomposition might be appropriate as well. Thereby, the time series is splitted into its components level, time, seasonality, and error (Cleveland et al., 1990). The developed Bayesian hierarchical time series algorithm could be enhanced to provide a Bayesian version of the decomposition by integrating a further seasonal component.

In the example of Campylobacter several differences regarding regions, age, sex, and argent could be recognized. According to expert knowledge of the RKI, outbreaks

usually occur in a specific strata. Thus, the simultaneous monitoring of different age groups, regions or argent might be more appropriate to detect local and national outbreaks. This leads to the multivariate surveillance. Here, questions regarding the choice of stratification, such as group size, level of location, and a meaningful association model, need to be answered (Frisén, 2003).

It is more important to predict than to estimate.

Peter Diggle (Farrington, 2010)

The thesis gave an overview of current statistical methodology in surveillance. In this context, the algorithm of Farrington, based on a generalized linear model, and hierarchical time series algorithm by Heisterkamp et al. (2006) have been described in detail. The hierarchical time series algorithm estimates the counts of observed infection cases, instead of predicting them. Thus, uncertainty is ignored when the threshold is constructed.

Moreover, choice of distribution for constructing this interval is discussable. It is based on a normal distribution due to the properties of the likelihood inference while a naive approach would be the usage of the assumed error distribution.

A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant.

Manfred Eigen, The Physicist's Conception of Nature (Gaither and Cavazos-Gaither, 1996)

The main outcome of this thesis is the developed Bayesian hierarchical time series algorithm. Based on a Bayesian model, a threshold calculation was developed which includes uncertainty of estimates and prediction simultaneously by using the predictive posterior.

This approach was implemented in the framework of the `R` package `surveillance`. Here, the fast and efficient approximation procedure of integrated nested Laplace approximation (INLA) was applicated. The resulting function `algo.hts()` is working reliable and could be verified in simulation study and in application on the time series of Campylobacter infections. It was examined that the algorithm using a stationary prior model performs well in most settings with mild season, and that the Farrington algorithm does not perform fundamentally better. The usage of the modern statistical method INLA has the disadvantage of causing problems, because INLA is still in development.

For now, the approach cannot regarded whether as right, wrong, or irrelevant. It was not possible to document superior performance, but the concept as a Bayesian model and the implementation with INLA provides a variety of enhancements, such as sequential model update, more advanced threshold calculations, a further component for explicit seasonality modelling, or spatial monitoring.

The time for this thesis project has been restricted to six month. Thus, the research needs to be focused on selected issues. If there would be more time, it would be focused on enhancements of the the Bayesian hierarchical time series algorithm. Firstly, the method would be enhanced by error handling, when negative Binomial error distributed models are assumed, and by adding a seasonal component. It is expected to improve the performance of the algorithm in application to time series with strong seasonality. Further, a proper version of the introduced adjustment for reporting delay might be added as an option in the algorithm. Additionally, if there would be much more time, the possibility of adapting the algorithm for multiple surveillance would be checked. Thereby, enhancing model complexity increases the computing time of the approach. Thus, the implementation of a sequential updating step is suggestive.

There is no such thing as the best method.

David Conesa (Farrington, 2010)

Using the example of Campylobacter infections in Germany, the thesis investigated and applied for surveillance analysis the system used at the RKI, the Farrington algorithm, the Bayes algorithm, and a Bayesian version of the hierarchical time series algorithm. In the simulation studies, an overall best performing method could not be determined. But, regarding the Bayesian hierarchical time series algorithm, and several possible enhancements, such as adding a further component for seasonality, the approach has the potential to get a very good, may be an overall best performing, surveillance method.

Thus, a full Bayesian surveillance approach, based on a time varying intercept, was developed which is able to estimate and predict threshold directly, and which has great potential due to several possible enhancements. The algorithm was implemented in the framework of the `surveillance` package, so that it is available free for everyone.

Appendix A

Code of Implementation

```
algo.hts <- function(disProgObj, control=list(range=NULL, co.arg=NULL,
prior='iid', family='poisson', alpha=0.05, mc.betaT1=100, mc.yT1=10)){

# Implementation of Hierachical Time Series Algorithm

  observed <- disProgObj$observed
  state <- disProgObj$state

  ### control arguments

  ### define range
  # missing range
  if(is.null(control$range)){
    warning('No range given. Range is defined as time from second
    period until end of time series.')
    control$range <- (disProgObj$freq+1):length(observed)
  }
  # check that range is subset of time series indices
  if(any(control$range %in% 1:length(control$observed))){
    stop("Evaluation period 'range' has to be vector of time series
    indices.")
  }
  #set order of range
  control$range <- sort(control$range)
  ### set model distribution
  control$family <- match.arg(control$family, c('poisson','nbinomial'))
  ### setting for different priors
  prior <- match.arg(control$prior, c('iid','rw1','rw2','rw3'))
  if(prior=='rw3') stop("Sorry, this prior 'rw2' is not implemented
  yet")
  ### setting for covariables
  co.arg.formula <- NULL
  if(!is.null(control$co.arg)){
    if(is.vector(control$co.arg)){
```

```
      control$co.arg <- as.matrix(control$co.arg,ncol=1)
    }
    if(nrow(control$co.arg)!=length(observed)){
      stop("Argument for covariates 'co.arg' has to have the same
      length like the time series")
    }
    for(i in 1:ncol(control$co.arg)){
      co.arg.formula <- (paste(co.arg.formula ,'+',
      colnames(control$co.arg)[i]))
    }
  }
  ### set model formula
#  if(prior=='rw1' || prior=='rw2'){
    modelformula <- as.formula(paste("observed~f(time, model='",prior,"',
    cyclic=FALSE)", co.arg.formula, sep=""))
  # setting for threshold calcuation
  if(control$alpha <= 0 | control$alpha >= 1){
    stop("significance level 'alpha' has to be a probability, thus
    between 0 and 1.")
  }
  # setting for monte carlo integration
  if(!control$mc.betaT1>0 || !control$mc.yT1>0 ||
    control$mc.betaT1!=round(control$mc.betaT1,0) ||
    control$mc.yT1!=round(control$mc.yT1,0)){
    stop('Number of Monte Carlo trials has to be an integer larger
    than zero')
  }

  ### clean model data from given outbreaks
  observed[which(state==1)] <- NA

##### sequentiell steps #####

  # progress bar
  pb <- tkProgressBar(title='progress bar', min=min(control$range),
  max=max(control$range), width=300)

  xi <- rep(NA,length(observed))
  ### calculate predictive posterior using MonteCarlo-Simulation
  for(i in control$range){

    ### prepare data frame: value for next time point yT+1 = NA
    time <- 1:i
    # data frame without covariables
    if(is.null(control$co.arg)){
      co.argi <- NULL
      dati <- data.frame(observed=c(append(observed[1:(i-1)],NA)),
```

```
      time=time)
    }else{     # data frame with covariables
      co.argi <- control$co.arg[1:i,]
      dati <- data.frame(observed=append(observed[1:(i-1)],NA),
      time=time, co.argi)
    }

    # fit model and calculate quantile
    xi[i] <- algo.htsFit(dat=dati, modelformula=modelformula,
    family=control$family, alpha=control$alpha,
    mc.betaT1=control$mc.betaT1, mc.yT1=control$mc.yT1)

    # update progress bar
    setTkProgressBar(pb, i, label=paste(round((i-min(control$range))/
    length(control$range)*100,0), '% done'))
  }
  # close progress bar
  close(pb)

  # compare observed with threshold an trigger alarm: FALSE=no alarm
  alarm <- disProgObj$observed > xi

  # return argument
  control$name <- paste('hts(prior=',prior,')',sep='')
  result <- list(alarm=as.matrix(alarm[control$range]),
  upperbound=as.matrix(xi[control$range]), disProgObj=disProgObj,
  control=control)
  class(result) <- 'survRes'
  return(result)
}

###############################################################################

algo.htsFit <- function(dat=dat, modelformula=modelformula,
family=family,alpha=alpha, mc.betaT1=mc.betaT1, mc.yT1=mc.yT1){

    # set time point
    T1 <- nrow(dat)
#print(T1)
    ### fit model
    model <- inla(modelformula, data=dat, family=family)
    if(is.null(model)){
      return(qi=NA)
    }
    ### mc simulation
    # draw sample from posteriori of betaT1
    m <- model$marginals.random$time[[T1]]
```

```
    betaT1 <- try(inla.rmarginal(n=mc.betaT1,m), silent=TRUE)
    if(inherits(betaT1,'try-error')){
      return(qi=NA)
    }
    else{
      betaT1 <- model$summary.fixed[1] + betaT1
    }
    # draw sample from yT1|yT,..,y1~f(exp(etaT1)) using sampled
    # betaT1
    nmc <- mc.betaT1 * mc.yT1
    yT1 <- vector(length=nmc, mode='numeric')
    # compute value for linear predictor
    if(ncol(dat)==2){
      etaT1 <- betaT1
    }else{
      etaT1 <- betaT1 + sum(dat[T1,-c(1:2)]*
      model$summary.fixed[-1,1])
    }
    # model poisson distribution
    if(family=='poisson'){
      for(j in 1:mc.betaT1){
        yT1[((j-1)*mc.yT1+1):(j*mc.yT1)] <- rpois(n=mc.yT1,
        lambda=exp(etaT1[j]))
      }
    }
    # model negative binomial distribution
    if(family=='nbinomial'){
      for(j in 1:mc.betaT1){
        yT1[((j-1)*mc.yT1+1):(j*mc.yT1)] <- rnbinom(n=mc.yT1,
        size=exp(model$theta.mode[1]),mu=exp(etaT1[j]))
      }
    }
    ### calculate threshold from MC samples
    qi <- quantile(yT1, probs=(1-alpha), type=3, na.rm=TRUE)
    return(qi)
}
```

Appendix B

Tables of Data Sources

initial name	new name	explanation	NA (in %)
Id	id	id of case	0
InterneRef	id.int	intern reference number	0
MeldeWoche	week	week of report	0
MeldeJahr	year	year of report	0
PersonGeschlecht	sex	sex of patient	0.1
AlterTheoretisch	age	theoretical age calculated by the birthday	0
Landkreis	districtID	id of reporting district	0
LKName	district	name of reporting district	0
Bundesland	stateID	id of reporting federal state	0
BLName	state	name of reporting federal state	0
ErkranktZeitraum			
DatumVon	start1	earliest date of illness start	11.5
DatumBis	start2	latest date of illness start	69.2
Herd	outbreak	intern id of outbreak, otherwise NA	97.1
ErregerCAM1	type1	type of bacteria type	0.4
Wert1	value1	value of bacteria type	0.5
ErregerCAM2	type2	second type of bacteria	0
Wert2	value2	value of second bacteria type	98.5
RKI_V1	arriveRKI	date inserting the first version of data set in the system of RKI	0.2
GA_V1	lastUpd	date inserting full version at local health department	0.1
IOrt			
DatumVon	locStart	start date for infection location	93.4
DatumBis	locEnd	end date for infection location	94.2
Lab_Melde	labReport	date of laboratory report	43.9
LabDiag	labDiag	date of laboratory diagnose	30.8

Table B.1: Overview for initial variables in the data frame and their meaning

station id	station name	altitude	latitude	longitude	remarks
10015	Helgoland	4	54° 01'	07° 53'	island
10020	List/Sylt	26	55° 00'	08° 24'	island
10035	Schleswig	47	54° 31'	09° 32'	
10055	Fehmarn	3	54° 31'	11° 03'	island
10147	Hamburg	11	53° 38'	09° 59'	airport
10162	Schwerin	59	53° 38'	11° 23'	
10170	Rostock	4	54° 01'	12° 04'	
10184	Greifswald	2	54° 05'	13° 24'	
10200	Emden	0	53° 23'	07° 14'	
10224	Bremen	5	53° 02'	08° 47'	airport
10270	Neuruppin	38	52° 54'	12° 48'	
10315	Muenster	48	52° 08'	07° 42'	airport
10338	Hannover	59	52° 27'	09° 40'	airport
10361	Magdeburg	76	52° 06'	11° 35'	
10379	Potsdam	81	52° 23'	13° 03'	
10384	Berlin	49	52° 28'	13° 24'	airport
10393	Lindenberg	112	52° 12'	14° 07'	
10400	Duesseldorf	37	51° 17'	06° 46'	airport
10427	Kahler Asten	839	51° 11'	08° 29'	high altitude
10439	Fritzlar	174	51° 07'	09° 17'	airport, TX missing
10453	Brocken	1142	51° 48'	10° 37'	high altitude
10469	Leipzig	131	51° 26'	12° 14'	airport
10488	Dresden	227	51° 07'	13° 45'	airport
10499	Goerlitz	238	51° 09'	14° 57'	airport
10501	Aachen	202	50° 47'	06° 05'	
10506	Nuerburg-Barweiler	485	50° 22'	06° 52'	high altitude
10548	Meiningen	450	50° 33'	10° 22'	
10554	Erfurt	316	50° 59'	10° 57'	airport
10578	Fichtelberg	1213	50° 25'	12° 57'	high altitude
10609	Trier-Petrisberg	265	49° 44'	06° 39'	
10637	Frankfurt	112	50° 02'	08° 35'	airport
10655	Wuerzburg	268	49° 46'	09° 57'	
10675	Bamberg	239	49° 52'	10° 54'	
10685	Hof	567	50° 18'	11° 52'	
10708	Saarbruecken	320	49° 12'	07° 06'	airport
10727	Karlsruhe	112	49° 02'	08° 21'	airport
10731	Rheinstetten	116	48° 58'	8° 20'	continues for Karlsruhe
10738	Stuttgart	371	48° 41'	09° 13'	airport
10763	Nuernberg	314	49° 30'	11° 03'	airport
10788	Straubing	351	48° 49'	12° 33'	
10852	Augsburg	462	48° 25'	10° 56'	airport
10870	München	444	48° 22'	11° 49'	airport
10929	Konstanz	443	47° 40'	09° 11'	airport
10946	Kempten	705	47° 43'	10° 20'	airport
10961	Zugspitze	2960	47° 25'	10° 59'	high altitude
10962	Hohenpeissenberg	977	47° 48'	11° 00'	high altitude

Table B.2: List of freely available weather stations of the German Climate Service (Deutscher Wetterdienst, DWD)

List of Figures

List of Tables

Bibliography

Adak, G. K., J. M. Cowden, S. Nicholas, and H. S. Evans (1995). The Public Health Laboratory Service national case-control study of primary indigenous sporadic cases of campylobacter infection. *Epidemiology and Infection 115*, 15–22.

Altekruse, S. F., N. J. Stern, P. I. Fields, and D. L. Swerdlow (1999). Campylobacter jejuni–An Emerging Foodborne Pathogen. *Emerging Infectious Diseases 5*(1), 28–35.

an der Heiden, M., A. Jansen, and J. Manitz (2010). Time Series Analyses for reported cases of campylobacter from 2002 to 2008 in Germany. Manuscript in preparation.

Anonymous (2006). Campylobacter-jejuni-Infektionen: Immer wieder Rohmilch als Vehikel! Bericht des zuständigen Gesundheitsamtes zu einem Ausbruch bei Schulkindern. *Epidemiologisches Bulletin 16*, 123–125.

Anonymous (2010). Thinkexist.com. Online: http://thinkexist.com/. Last access: 3 August 2010.

Buehler, J. W., R. S. Hopkins, J. M. Overhage, D. M. Sosin, and V. Tong (2004). Framework for Evaluation Public Health Surveillance Systems for Early Detection of Outbreaks: Recommendations from the CDC Working Group. *MMWR Recommendations and Reports 53*(RR05), 1–11.

Bulst, N. (1989). Krankheit und Gesellschaft in der Vormoderne. Das Beispiel der Pest. *Maladies et société (XIIe - XVIIIe siècles) : actes du colloque de Bielefeld, novembre 1986, Ed. Neithard Bulst, Paris*, 17–47.

Bundesamt für Kartographie und Geodäsie (2010, January). Geo84 verwaltungsgrenzen. Online: http://www.geodatenzentrum.de/geodaten/. Last access: 16 April 2010.

Bundesinstitut für Risikobewertung (2009). Verbrauchertipps: Schutz vor lebensmittelbedingten Infektionen mit Campylobacter. Online: http://www.bfr.bund.de. Last access: 5 Febuary 2010.

Buzby, J. C. and T. Roberts (2009). The Economics of Enteric Infections: Human foodborne Disease Costs. *Gastroenterology 136*(6), 1851–1862.

CDC Emergengy Risk Communication Branch (ERCB), Division of Emergency Operations (DEO) Office of Public Health Preparedness and Response (OPHPR)

(2004, April). Simulation Data Sets for Comparison of Aberration Detection Methods. Online: http://www.bt.cdc.gov/surveillance/ears/datasets.asp. Last access: 7 July 2010.

Cleveland, R. B., W. S. Cleveland, J. E. McRae, and I. Terpenning (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics 6*, 3–73.

Cox, S. D., D. Commenges, A. Davison, P. Solomon, and S. Wilson (2006). *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

Dengler, R. (1997). Dampf und Tropfen. Betrachtungen rund um die Luftfeuchtigkeit. *PLUS LUCIS 3*.

Deutscher Wetterdienst (2010). Climate Data of Germany. Online: http://www.dwd.de. Last access: 30 March 2010.

Fahrmeir, L., A. Hamerle, and G. Tutz (1996). *Multiariate statistische Verfahren* (2nd ed.). Walter de Gruyter.

Fahrmeir, L., T. Kneip, and S. Lang (2009). *Regression. Modelle, Methoden und Anwendungen* (2nd ed.). Springer-Verlag.

Farrington, C. P. (2010, May19^{th}). Issues raised during the conference/workshop for statistical methods for outbreak detection. Online: http://statistics.open.ac.uk/SMOD/. Conference Slides.

Farrington, C. P., N. J. Andrews, A. D. Beale, and M. A. Catchpole (1996). A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. *Journal of the Royal Statistical Society Series A*(159), 547–563.

Forßbohm, M. (2000). Die Bestimmungen des Infektionsschutzgesetzes zum Meldewesen aus Sicht des Gesundheitsamtes. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz 43*.

Frisén, M. (1992). Evaluation of Methods for Statistical Surveillance. *Statistics in Medicine 11*(11), 1489–1502.

Frisén, M. (2003). Statistical Surveillance. Optimality and Methods. *International Statistical Review 71*(2), 403–434.

Gaither, C. C. and A. E. Cavazos-Gaither (1996). *Statistically Speaking. A Dictionary of QuotationsGaither*. Institute of Physics Publishing.

Google (2010). google.org flu trends. Online: http://www.google.org/flutrends. Last access: 12 April 2010.

Grinsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinksi, and L. Brilliant (2009, February). Detecting influenza epidemics using search engine query data. *Nature 457*, 1012–1014.

Hawkins, D. M. and D. H. Olwell (1998). *Cumulative sum charts and charting for quality improvement*. Springer-Verlag.

Heisterkamp, S. H., A. L. M. Dekkers, and J. C. M. Heijne (2006). Automated detection of infectious disease outbreaks: hierachical time series models. *Statistics in Medicine* (25), 4179–4196.

Held, L. (2008). *Methoden der statistischen Inferenz - Likelihood und Bayes*. Spektrum Akademischer Verlag.

Höhle, M. (2007). `surveillance`: An R package for the monitoring of infectious diseases. *Computational Statistics 22*(4), 571–582.

Höhle, M. (2008, November). Short course on Statistical surveillance of infectious diseases. Course slides.

Höhle, M. and A. Mazick (2010). *Biosurveillance: Methods and Case Studies*, Chapter : Aberration detection in R illustrated by Danish mortality montoring. CRC Press. Eds: Kass-Hout, Taha and Zhang Xaiohui.

Hudson, J. A. et al. (2001). Seasonal variation of Campylobacter types from human cases, veterinary cases, raw chicken, milk and water. *Journal of Applied Microbiology 87*(1), 115–124.

Hutwagner, L., T. Browne, G. M. Seeman, and A. T. Fleischauer (2005, February). Comparing Aberration Detection Methods with Simulated Data. *Emerging Infectious Diseases 11*(2), 314–316.

Jansen, D. A., D. A. Käsbohrer, and D. T. Alter (2007). Campylobacter-jejuni-Infektionen treten 2007 vermehrt auf. Analyse der Situation durch RKI und BfR. *Epidemiologisches Bulletin* (36), 331–334.

Kleinman, K. P. and A. M. Abrams (2008). Assessing the utility of public health surveillance using specificity, sensitivity, and lives saved. *Statistics in Medicine 27*, 4057–4068.

Koenker, R. (2005). *Quantile regression*. New York: Cambridge University Press.

Levin, R. E. (2007). Campylobacter jejuni: A Review of its Characteristics, Pathogenicity, Ecology, Distribution, Subspecies Characteriszation and Molecular Methods of Detection. *Food Biotechnology 21*(4), 271–347.

Louis, V. R., I. A. Gillespie, S. J. O'Brien, et al. (2005). Temperature-Driven Campylobacter Seasonality in England and Wales. *Applied and Enviromental Microbiology 71*(1), 85–92.

Martino, S. and H. Rue (2009). Implementing Approximate Bayesian Inference using Integrated Nested Laplace Approximation: a manual for the `inla` program. Technical Report 2, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Morabia, A. (1996). From Disease Surveillance to the Surveillance of Risk Factors. *American Journal of Public Health 86*(5), 625–627.

Nichols, G. L. (2005). Fly Transmission of Campylobacter. *Emerging Infectious Diseases 11*(3), 361–364.

Nylen, G., F. Dunstan, S. R. Palmer, et al. (2002). The seasonal distribution of campylobacter infection in nine European countries and New Zealand. *Epidemiology and Infection 128*(3), 383–390.

Patrick, M. E., L. E. Christiansen, M. Wainø, et al. (2004, December). Effects of Climate on Incidence of Campylobacter spp. in Humans and Prevalence in Broiler Flocks in Denmark. *Applied and Enviromental Microbiology 70*(12), 7474–7480.

Peters, J., J. Lienau, G. Näther, et al. (2006). Resultate der ersten Phase des nationalen Campylobacter-Masthähnchenmonitorings 2004–2005. *Archiv für Lebensmittelhygiene 57*(5), 137–141.

R Developer Core Team (2009). *R: A Language and Enviroment for Statistical Computing*. R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robert Koch-Institut (2000). Umsetzung der Übermittlung der meldepflichtigen Infektionen nach dem Infektionsschutzgesetz. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheutsschutz 43*, 870–874.

Robert Koch-Institut (2005). RKI-Ratgeber Infektionskrankheiten: Campylobacter-Infektionen. Online: http://www.rki.de. Last access: 5 Febuary 2010.

Robert Koch-Institute (2010). Survstat@rki.de. Online: http://www3.rki.de/SurvStat.

Rue, H. and S. Martino (2009). *INLA: Functions which allow to perform a full Bayesian analysis of structured additive models using Integrated Nested Laplace Approximation*. R package version 0.0.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society 71*(2), 1–35.

Sonesson, C. and D. Bock (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society Series A*(166), 5–21.

Stark, K. (2009, March). Klimawandel und Infektionskrankheiten: Relevanz für Deutschland. Online: www.bfr.bund.de. Course slides. Last access: 19 April 2010.

Straif-Bourgeois, S. and R. Ratard (2005). *Handbook of Epidemiology*, Chapter Infectious Disease Epidemiology, pp. 1328–1362. Springer-Verlag.

Stroup, D., M. Wharton, K. Kafadar, and A. G. Dean (1993). Evaluation of a Method for Detecting Aberrations in Public Health Surveillance Data. *American Journal of Epidemiology 137*(3), 373–380.

Thacker, S. B. and R. L. Berkelman (1988). Public Health Surveillance in the United States. *Epidemiologic Reviews 10*, 164–190.

Wood, D. and C. Pooley (2007). New Campylobacter-Detecting Medium Licensed: Image Number K11505-1. Online: http://www.ars.usda.gov/is/pr/2007/071016.htm. Last access: 11 April 2010.

Woodall, W. H. (2006, April). The Use of Control Charts in Health-Care and Public-Health Surveillance. *Journal of Quality Technology 38*(2), 89–104.

World Health Organization (2007). Fact sheet no. 237: Food safety and foodborne illness. Online: http://www.who.int/mediacentre/factsheets/fs237/en/. Last access: 11 April 2010.