Mathematical Statistics
Stockholm University

A model for analysing temporal and spatial patterns of infectious diseases with an application to reported campylobacter infections

Johan Lindbäck

Research Report 2003:7
Licentiate Thesis

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2003:7**,
http://www.math.su.se/matstat

A model for analysing temporal and spatial patterns of infectious diseases with an application to reported campylobacter infections

Johan Lindbäck*

May 2003

Abstract

Many infectious diseases have incidences that vary over time with alternating high and low periods. A parametric model is formulated that makes it possible to account for such variations. The model contains parameters describing the time of the peak, and the duration and amplitude of the high incidence period. The model is fitted to weekly aggregated data on indigenous campylobacter infections reported to the Swedish Institute for Infectious Disease Control. This analysis is made with the reports geographically divided according to county and with spatial smoothing. Estimation of parameters is done using Markov chain Monte Carlo methods based on both Poisson and Negative binomial variation. The model fits data well. There are indications that the concentration of the high incidence periods varies geographically.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: johan.lindback@ucr.uu.se. Financial support from The Bank of Sweden Tercentenary Foundation is gratefully acknowledged.

Contents

1 Introduction

Infectious diseases are different from other diseases in many ways. One important characteristic is that infections might be transferred to a person from another person, an animal or the environment. All infections occur by contact in some way. Each disease has its own paths of infection. Transmission can, for example, occur via sexual contacts (e.g., HIV, chlamydia), the air (e.g., measles, influenza), contaminated food (e.g., salmonella infection, campylobacter infection) or some vector (e.g., malaria, dengue fever). The so called infectious agent can be either a virus, bacteria or parasite. Infectious disease epidemiology is in many ways similar to epidemiology in general but special attention is needed towards the characteristics, of the infectious diseases, mentioned above. The book by Giesecke (2002) gives a good introduction to the field of infectious disease epidemiology.

The Swedish Institute for Infectious Disease Control (Smittskyddsinstitutet, SMI) is a governmental expert agency with the mission to protect the Swedish population from communicable diseases (http://www.smittskyddsinstitutet.se). One of the main missions is to follow and analyse the national and international situation regarding infectious diseases and the protection against them. As a basis to carry out this mission, around fifty infectious diseases are notifiable by law in Sweden. This means that if a physician or a laboratory finds that a patient is infected with one of these diseases they are obliged to notify SMI by sending in a report. Since 1996 a computerised reporting system is successively replacing the older paper based system. The reports mainly consist of basic personal information about the patient (such as age, sex, place of residence etc.), information concerning the time of infection and other disease specific information (time, place, exposure to known risk factors etc.). In total, between 40 000 and 50 000 cases are reported each year (cf. Smittskyddsinstitutet (1999)). Traditionally, the number of reports are summed and basic statistics in form of time series or geographical distribution of the reported cases are published in monthly and yearly reports.

As is the case with most reporting systems there are severe problems with the quality of the reporting. Any analysis based on reported cases has to consider to what extent important features of the disease are reflected in the available data. One should hope that even if there is underreporting and biases in the reporting system some real effects such as temporal and spatial patterns are still seen. Revealing such patterns can be an essential contribution to the understanding of how the different infectious diseases are spread. At best the reports can be used to derive important information about the occurrence and aetiology of the notifiable infectious diseases.

When studying infectious diseases, time is often an important factor to consider. When treating patients, knowledge about the length of the incubation period and the time of infectiousness can be crucial. In the planning of public health measures (e.g., vaccination campaigns) the time of immunity is also a factor to consider. From another perspective, the individual risk of getting infected may vary over time. Seasonality is common for many infectious diseases. It is almost certain that each year there will be a peak of influenza cases at some point during the winter and a peak of salmonella and campylobacter infections in the summer. However the exact time and size of the peak may, and will, vary from year to year. Crude time series of aggregated cases often incorporates a large amount of variation. In order to discern timely patterns in the incidence it is common to smooth the time series by some function.

The geographical distribution of the disease incidence is also an important part of the aetiology of the disease. Descriptive methods usually incorporate maps, with crude incidences, divided in smaller geographical units. The purpose of the disease mapping can be, e.g., to find geographical gradients of the incidence, to generate hypotheses about disease patterns or to simply describe the geographical distribution of the disease. One popular example of disease mapping is the search for disease clusters that might be linked to some environmental source or to other sources of interest. This is used especially in the field of cancer research.

Most applications in disease mapping focuses on the underlying relative risk of contracting the disease and a common measure of interest is the standardised morbidity/mortality ratio (SMR) defined as the observed number of cases divided by the expected number of cases within a given geographical area. Often the disease under study is relatively rare and/or the geographic areas are small (as regards to the number of inhabitants). In such situations the sampling variability will be large due to a small expected number of cases and there is a considerable risk that extreme observations will occur just by chance. By smoothing the crude rates via hierarchical modelling, extreme observations can be adjusted to reflect a more realistic situation. This smoothing procedure often leads to rather complex models and parameter estimation is not straightforward. However, by using computer-intensive methods like the Markov chain Monte Carlo (MCMC), estimates can be achieved by simulation. Most of these models are set up within a Bayesian framework by putting prior distributions on the parameters, but can as well be interpreted as a likelihood random effects model in a frequentist perspective. Wakefield et al. (2001) discuss assumptions and mathematical details of Bayesian methods for disease mapping.

The aim of the present study is primarily to study how the information provided by the Swedish reporting system, with its known shortcomings, can be used to learn more about infectious diseases and their aetiology. In addition, we introduce a parametric model giving a smoothed estimate of the crude number of reported cases. The model incorporate parameters that take into account certain aspects of the seasonality of the incidence.

As a secondary aim of this study, we also have interest in a specific disease, namely campylobacter infection. We here use data on reported indigenous campylobacter infections as an example to study some of the problems and possibilities of statistical analyses regarding the kind of data discussed above.

Campylobacter infections in humans are part of a possibly complicated system of spread of the bacteria *Campylobacter*. The bacteria has been found in humans, both wild and domestic animals and in the environment. Many investigations, mainly case-control studies, have been performed to asses the risk factors for sporadic cases of campylobacter infections in humans (cf. Kapperud (1995)). However, no single risk factor appears to be able to explain more than a small fraction of the cases. It is likely that the routes of transmission differ over the year and more research is needed in this area to assess that. An analysis of the seasonal distribution of campylobacter infections in nine European countries and New Zealand was done by Nylén et al. (2002).

We fit the model to the reported data on indigenous campylobacter infections in Sweden. By studying seven years and 21 counties it is hopefully possible to discern both temporal and spatial variations regarding the parameters in the model. Examining these variations then might help in understanding the complex system of the spread of campylobacter infections and can also serve as a help in the planning of future investigations of risk factors.

From the perspective of the analysis, we do not focus on the absolute number of cases or the SMR; our interest lies in studying the functional form of the incidence curve and mainly the parameters describing it. As the resulting model is rather complex, estimation of parameters is done by using Markov chain Monte Carlo (MCMC) methods.

The structure of the thesis is as follows. In Section 2 we shortly present some basic facts about campylobacter infections. A more thorough presentation of the data is given in Section 3 together with a description of the Swedish reporting system for infectious diseases. Section 4 contains a crude analysis of temporal and spatial patterns of cases while we in Section 5 introduce the parametric model for describing the incidence of campylobacter infections and apply it to aggregated data for Sweden. We split the data by county and fit the model to individual counties both independently and dependently by geographical smoothing in Section 6. A short description of the MCMC estimation procedure is given in Section 7. Some results are presented in Section 8 and the thesis is concluded with a discussion in Section 9.

Some of the material in this thesis have been presented in a previous report (Lindbäck & Svensson (2001)). This thesis is an extension and a development of some of the ideas in that report.

2 Campylobacter – infectious agent and disease

2.1 The agent

Campylobacter infection is a bacterial disease. There are over 20 subtypes of *Campylobacter* but the main types causing gastrointestinal symptoms in humans are *Campylobacter jejuni* and *Campylobacter coli.* The bacteria can survive four weeks in water at +4 °C but less than four days at +25 °C (cf. Nothermans (1995) and Andersson & Gustavsson (1998)). To be able to multiply, the bacteria requires a temperature of +40 °C and a concentration of oxygen of at most 5 %. Therefore, the ideal place to grow is the intestines in humans and warm-blooded animals. Hence, food items are not a good place for the bacteria to grow but on the other hand, the critical infectious dose (i.e., the smallest dose of the bacteria needed to cause the disease) is very low.

2.2 The disease

Campylobacter infection is a zoonosis, which means that it is naturally transmitted between animals and man. The symptoms characterising the disease are diarrhoea, abdominal pain, malaise, fever, nausea and vomiting. The incubation period is usually one to three days but can vary between one to ten days. The illness is acute and usually over within two to five days. Campylobacter infection causes 5 %–14 % of diarrhoea worldwide and is an important cause of travellers' diarrhoea (cf. Chin (2000)). In rare cases the disease can lead to long-term consequences such as arthritis, a neurological syndrome called Guillain-Barré syndrome and sometimes even death. There is no vaccine against the disease.

2.3 The spread of the disease

Even a superficial study of statistics reveals that cases occur both in large outbreaks and as sporadic cases. An outbreak occurs, when many individuals are exposed to the same source of infection, and suffer from the disease at approximately the same time. The sporadic cases involve only one or a few individuals that are infected simultaneously. The cause of a large outbreak is often relatively easy to identify. During the period studied, 1992–1998, one can discern three major outbreaks in Sweden.

- In Kramfors (in the county Västernorrland) approximately 2 500 cases of campylobacter infections occurred during May 1994. This outbreak was caused by contaminated water (cf. Andersson et al. (1994)). Of these cases, 64 were reported to SMI (cf. Smittskyddsinstitutet (1995)).

- In Mark (in the county Västra Götaland) 3 000–4 000 campylobacter infections occurred at the end of May 1995. The cause of the outbreak was contaminated water (cf. Bresky et al. (1995)). Not more than 48 of the cases were reported to SMI (cf. Smittskyddsinstitutet (1996)).

- The third outbreak took place at a training camp for young football players in the summer of 1996. At least 123 out of 200 participants were infected after drinking unpasteurized milk. Of these cases 22 were reported to SMI (cf. Smittskyddsinstitutet (1997)). The cases came from several counties in the south of Sweden.

Even if underreporting of cases in connection with large outbreaks is severe, they are still identifiable in a crude time series of reported cases. In the following, we concentrate on cases that appear to be sporadic. This means that we have removed observations from the large outbreaks in the data we analyse.

In addition to these large outbreaks there may occur minor outbreaks involving only a few individuals. Evidently, such minor outbreaks are much more difficult to identify. The yearly report from SMI mentions seven such minor outbreaks in 1998 (cf. Smittskyddsinstitutet (1999)). These outbreaks resulted in 3–7 reported cases. The source of infection varied. Identified or suspected causes were food (unpasteurized milk, chicken, paella) or drinking contaminated water.

It has been discussed if campylobacter infections are communicable. The general understanding seems to be that there is a small risk of spread human to human, but that it is rather limited (cf. Chin (2000)). We have chosen not to include effects of such spread in the models and our analysis.

The substantial part of reported cases are sporadic, i.e., they cannot be seen to be directly associated with other cases. As for cases during outbreaks, sporadic cases are often not reported. However, one can expect the underreporting of sporadic cases to be less severe. There exists no reliable investigation of the proportion of unreported cases in Sweden. In an English study, it was established that for each reported case of *Campylobacter*, in a laboratory based surveillance system, there were 7.6 cases in the community (cf. Wheeler et al. (1999)). We cannot assume that this number is representative for Swedish conditions due to the many differences between England and Sweden concerning both the surveillance systems and the communities.

In some of the reports from the physicians, a suspected cause of the infection is mentioned. Such causes are badly prepared chicken, chicken prepared at home, eating at restaurant, secondary infections, barbecue, contact with birds, drinking unpasteurized milk or water from mountain brooks.

There are a number of investigations of risk factors associated with sporadic cases (cf. Kapperud (1995)). Many of the studies are case-control studies. Comparing to which extent cases and healthy controls have been exposed to potential risk factors one tries to identify exposures that increase the risk of getting a campylobacter infection. A few examples of risk factors studied are travel abroad, contacts with animals and food consumption. There are of course severe difficulties in managing studies of this kind. The quality of the study relies on that sufficiently good accounts of the exposures for the cases before taking ill and similar reliable accounts of exposures for the controls can be obtained. The results can be subject to recall bias since cases and controls remember or report their true exposures with different accuracy. The danger of recall bias is even larger when the participants in the study are asked to recall if they have been eating a certain food item, and to make an evaluation of the exposure (e.g., if the chicken they have consumed was undercooked or not). A well-established risk factor has been found in a series of studies from Great Britain were humans were infected due to birds (mainly magpies) pecking off the seals of milk bottles (cf. Lighton et al. (1991)). Otherwise the results presented in published studies are, as can be expected, rather diffuse. Kapperud (1995) lists identified risk factors in a number of case-control studies made in different countries. The list contains travel abroad, eating chicken, handling raw chicken, eating undercooked chicken, eating chicken at barbecues, eating poultry, eating at barbecues, drinking surface water, drinking untreated water, drinking raw milk, drinking raw goat's milk, milk bottles pecked by magpies, contact with cats, presence of a puppy in the household. This broad spectrum of risks can be taken as an indication that there are several routes of transmission of campylobacter infections to humans. Due to the large seasonal variations of (reported) campylobacter infections, it is possible that different transmission routes are open at different times of the year. Even if Tauxe (1992) calculates that 50 % of the cases are attributable to consumption of poultry products, there seems to be no single risk factor that accounts for the most of the cases. Of course, it may be the case that infections have different causes in different surroundings and at different times.

3 Description of the data and basic facts about the reporting system

3.1 Incidence of campylobacter infections

From 1992 to 1997 on average 5 000 cases per year of campylobacter infections were reported to SMI (Table 1). During 1998 and 1999, there was an increase in incidence. In 1998, infections acquired in Sweden and abroad both increased as compared with the previous year. However, the number of domestic cases was at the same level as 1994 and 1995. The increase in incidence in 1999 compared with 1998 was due to an increase in cases infected abroad. The major part of the reported cases of campylobacter infections related to persons travelling outside of Sweden. Between 31 % and 46 % of the cases each year were infected in Sweden and the rest were infected abroad. In the following analysis, we are only considering infections that have been acquired in Sweden.

Table 1: Number of reported cases of campylobacter infections in Sweden by origin of infection

Place of	Year of registration at SMI							
infection	1992	1993	1994	1995	1996	1997	1998	1999
In Sweden	1 453	1 825	2 538	2 551	1 815	1 828	2 586	2 209
Abroad	2 998	2 590	2 764	2 821	3 131	3 266	3 816	4 796
Unknown	24	70	227	208	136	212	142	132
Total	4 475	4 485	5 529	5 580	5 082	5 306	6 544	7 137

3.2 The reporting system

There are of course all kinds of quality problems associated with this compulsory notification system. Even if a disease is notifiable by law not all cases are reported properly. Any analysis of geographical and temporal patterns will rely on the precise information given in the reports and how this information is processed at SMI.

From the time an individual is infected till a report of the resulting illness ends up in the registers at SMI several steps have to be passed. First, the infected person has to go to the doctor. For diseases with mild symptoms, this can lead to a substantial underreporting because many infected persons will not seek medical help. Then the physician has to make the correct diagnosis, which may be confirmed by a laboratory test. Subsequently a report must be filled in, signed, and sent to SMI. At SMI, the reports are entered into a database. Figure 1 illustrates some critical events in the notification system.

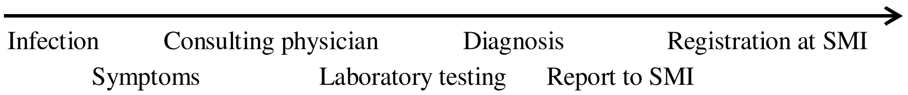


Figure 1: Critical events in the reporting chain

The most interesting event is when the patient is infected. However, it is very difficult, often impossible, to establish the exact time this event occurs, especially if the disease has a very long incubation period. The event closest to infection is the onset of disease, i.e., the first time when the patient experiences symptoms of the disease. The time of this event can also be difficult to reconstruct and if the physician reporting the case is unable to estimate the most probable time of onset, this information will be missing in the report. The only reliable time in this procedure is the date of registration, i.e., the date when the report arrives to SMI.

Of course there is a delay in reporting cases. In a register study of 20 selected notifiable infectious diseases, the time from disease onset to registration at SMI was examined (cf. Jormanainen et al. (1997)). Reporting delay was defined as the number of days between disease onset, i.e., start of symptoms, and time of registration at SMI. The median delay varied between diseases, from 15 days (meningococcal infection) to 91 days (atypical mycobacterioses) and was generally shorter for diseases of acute type and longer for diseases of more chronic type. For campylobacter infections, the median delay is 19 days and within 64 days 95 % of all reports are registered. Reports with a delay of less than one day or more than one year are considered as miscoded and therefore excluded from further analysis. For this reason 0.4 % of the reports are excluded from the analysis.

Since we are studying variations of campylobacter infections in time we need to relate each case with a date. The time of infection is what we actually are most interested in but since that date is rarely known we have to use another date instead. The closest known date to the time of infection is the time of disease onset. Because the disease onset date is missing on some of the reports we will loose information if we choose this as our time variable. The only date that is known for all cases is the date of registration at SMI. However, because of the relatively long and highly dispersed reporting delay we will in turn loose precision if we use the date of registration as time variable. There are methods, e.g., back calculation, to estimate the missing onset dates from the registration dates. However, we will, in the following analysis, use the date of onset, without trying to recreate the missing observations. Approximately 11 % of the reports will be excluded from the analysis for missing the onset date.

3.3 Age and sex

Of the 13 077 indigenous cases with information about date of onset between 1992 and 1998, about 47 % are women. The age distribution is the same for men and women with a high incidence among the youngest children and the young adults (20–35 years), cf. Figure 2.

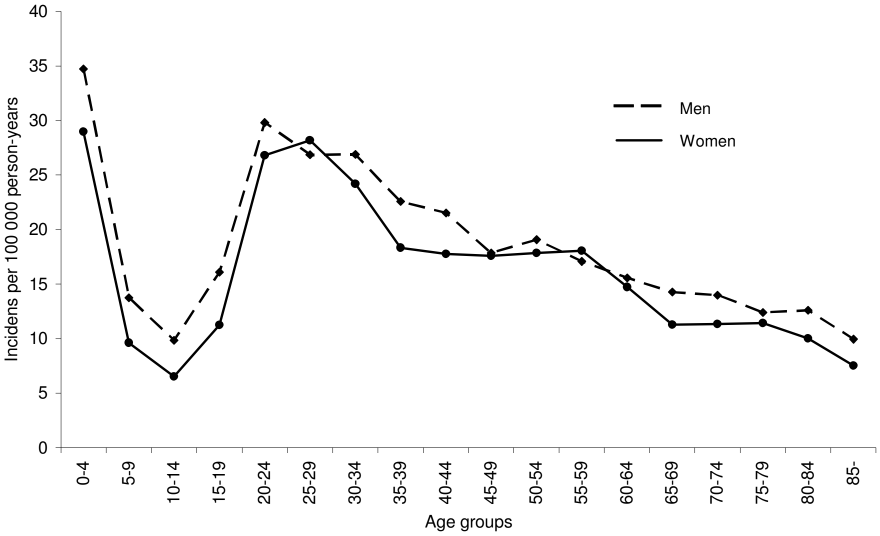


Figure 2: Age and sex distribution of incidence per 100 000 person–years of reported indigenous campylobacter infections in Sweden 1992–1998.

For adults older than 35 years the incidence is decreasing with increasing age. Notable is the dip in the incidence curve for children and youths between 5 and 20 years.

3.4 Geography

The geographical units of the analysis are counties. These are the major administrative units of Sweden. The country is divided into 21 counties. The counties organise the health care within their area. In each of these counties, a county medical officer (smittskyddsläkare) is responsible for the local supervision of infectious diseases. In Table 2 the number of indigenous cases and incidences per 100 000 inhabitants and year during the period under study are given for each county. The number of inhabitants is calculated as the mean of the number of inhabitants the last of December each of the years 1992 through 1998.

The incidence varies between 10 and 40 cases per 100 000 person-years. The extremes are Gotland with an incidence of 39.4 and Värmland with an incidence of 10.4. For most counties the population size is quite stable between and within years. However, for Gotland this is not the case. Gotland has the smallest population size, with only about 58 000 inhabitants registered. Gotland is also one of the most popular counties to visit as a tourist and many people living in other counties in Sweden have their summerhouse there. This means that the actual population in Gotland is much higher in summer than in the

Table 2: Number of indigenous cases and incidence per 100 000 person-years of campylobacter infections in Sweden, date of onset 1992–1998. (For eight of the cases the sex was unknown.)

		Population	Number of cases			Incidence per 100 000 person-years		
	County	$\times 10^3$	Women	Men	Total	Women	Men	Total
1	Stockholm	1 717	1 171	1 351	2 523	16.6	20.2	18.4
2	Uppsala	285	249	226	475	21.5	20.1	20.8
3	Södermanland	258	190	216	406	18.3	21.1	19.7
4	Östergötland	413	232	261	493	14.0	15.9	14.9
5	Jönköping	328	206	140	447	15.6	18.4	17.0
6	Kronoberg	179	127	164	291	17.7	22.8	20.3
7	Kalmar	242	196	243	439	20.2	25.3	22.7
8	Gotland	58	84	98	182	36.0	42.8	39.4
9	Blekinge	152	85	86	171	14.0	14.2	14.1
10	Skåne	1 103	1 140	1 159	2 302	25.3	26.8	26.1
11	Halland	267	211	308	519	19.7	29.0	24.3
12	Västra Götaland	1 473	915	1 120	2 037	15.4	19.2	17.3
13	Värmland	283	106	130	236	9.3	11.6	10.4
14	Örebro	275	188	182	370	16.8	16.8	16.8
15	Västmanland	260	149	157	306	14.3	15.1	14.7
16	Dalarna	289	198	209	408	17.1	18.2	17.7
17	Gävleborg	287	126	144	270	10.9	12.6	11.7
18	Västernorrland	258	190	230	420	18.3	22.4	20.3
19	Jämtland	135	79	79	158	14.7	14.6	14.6
20	Västerbotten	258	119	154	273	11.5	15.0	13.2
21	Norrbotten	265	138	213	351	13.2	19.8	16.6
	Sweden	8 785	6 099	6 970	13 077	17.2	20.1	18.6

rest of the year. The average number of visitors each year is approximately 600 000. This is a contributing cause to why Gotland has so much higher incidence than other counties. Another cause may be that the geological conditions of Gotland differ from the rest of the country in a way that influence the quality of the drinking water (cf. Andersson et al. (1998)).

In Figure 3 the counties and their incidences are indicated on a map of Sweden. At a first glance it is not easy to discern any geographical pattern. Maybe one can say that there is an over-representation of southern counties among those with higher incidence and an over-representation of northern counties among those with lower incidence.

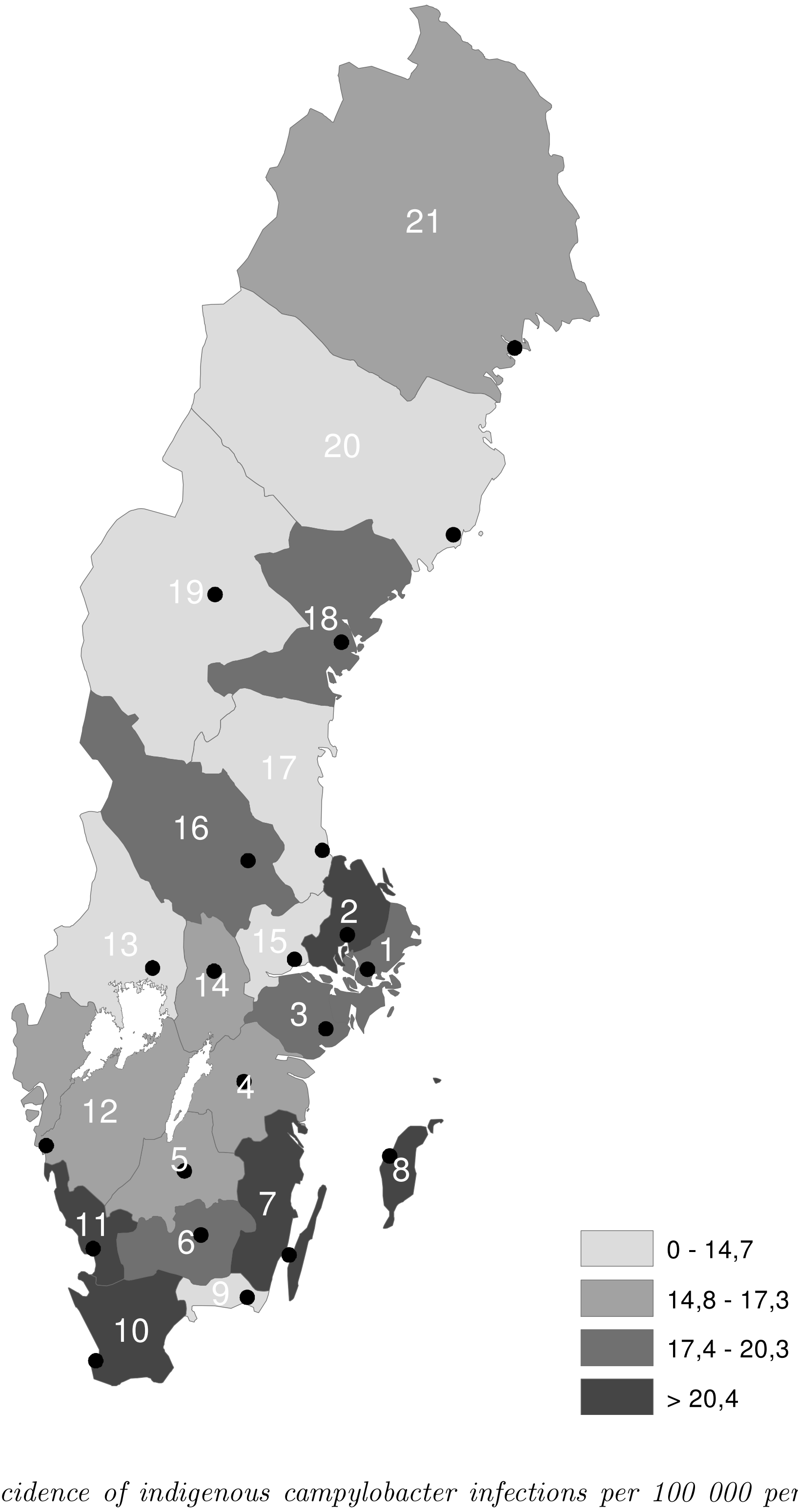


Figure 3: Incidence of indigenous campylobacter infections per 100 000 person-years in Sweden 1992 – 1998. Black dots indicate the residential cities of the counties.

3.5 Seasonality

In order to study the seasonality of the incidence of campylobacter infections we have aggregated the reported data on a weekly level. We have chosen to define the weeks as consecutive seven days periods from January 1 1992 rather than the actual week number in the calendar. Choosing weeks as aggregation level we do not have to worry about the within-week variation that might appear due to, e.g., different eating habits in weekends and weekdays. In Figure 4 the time series of the number of reported campylobacter infections in Sweden week by week during 1992–1998 is given.

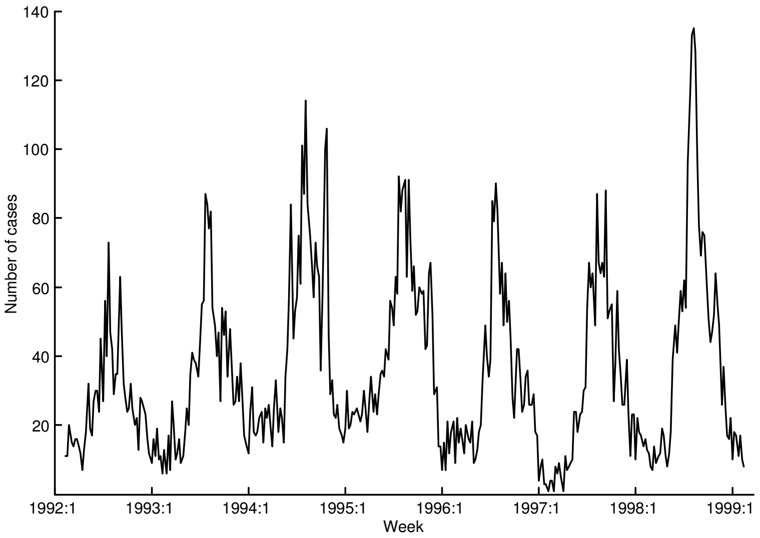


Figure 4: Weekly number of reported cases of campylobacter infection in Sweden 1992–1998.

The series reveals a large variation in the number of cases during the year and a rather stable yearly pattern. There is a high incidence period peaking approximately in late summer each year. The exact time of the peak and the duration of the high incidence period seem to vary between years.

This pattern is not only seen in the aggregated data but also in the time series for the separate counties. However, the time of the peak and the duration of the high incidence period vary between the counties. This is illustrated in Figures 11, 12 and 13 below. There the weekly number of cases is given for Stockholm, Västerbotten and Blekinge together with a smooth estimate of the mean number of reports.

4 A crude analysis of spatial and temporal distribution of cases

The data used in the following analysis are the weekly number of reported cases by date of onset in each of the 21 counties. The time series start with week 6 in 1992 and end with week 5 in 1999. We have chosen to start the analysis with week 6 mainly for technical reasons. The smoothing model used defines a yearly parameter for the lowest incidence. It is thus convenient to relate a change of parameters to the time when the incidence is lowest. This happens around week 6. The time span analysed consists of 365 weeks. Accordingly, there are 365 observations for each county.

Data from the known large outbreaks (cf. Section 2.3) have been modified to reflect a situation with only sporadic cases, i.e., the actual reported number of cases has been substituted by a mean number of reported cases in the weeks before and after the outbreak.

The time series of cases (cf. Figure 4) show large random variations. To be able to discern differences and to compare the patterns in different counties it is necessary to calculate statistics that illustrate the important features of the regular patterns. We will start by presenting some crude calculations, which are not based on any assumptions on the nature of the random variations or of the form of the temporal and spatial variations. The purpose of the analysis is to compare the time when the incidence peaks and when the high incidence period starts for the different counties.

For each county the data has been aggregated on a yearly level. That is the number of cases occurring during the i^{th} week within the year in the seven years under investigation have been summed and a mean number of cases per week have been calculated. The reports from different years are aggregated to weeks with the same position within the year. Two statistics have been calculated. The first is the number of the mid-week in the 9-week period (of consecutive weeks) with the largest number of cases. The second statistic is the first week within the year of a consecutive period of 9 weeks which has an incidence larger than the mean weekly incidence. The calculations are illustrated in Figure 5 where the 9-week moving average is given for data from Sweden. The curve has its peek in week 31 and it crosses the line indicating the mean number of reports per week for the average number of reports during week 18 to 26.

Table 3 shows the result of the calculations for the individual counties.

A general impression is that the peak week occurs earlier in the southern part of Sweden than in the northern part. The same seems to hold for the start of the high incidence period. This impression is to some extent confirmed by calculations of rank correlations between these numbers and the north-south position of the counties. The counties have been ordered according to the relative positions of their residential cities (cf. Table 3). The rank correlation between peek week and the relative north-south position of

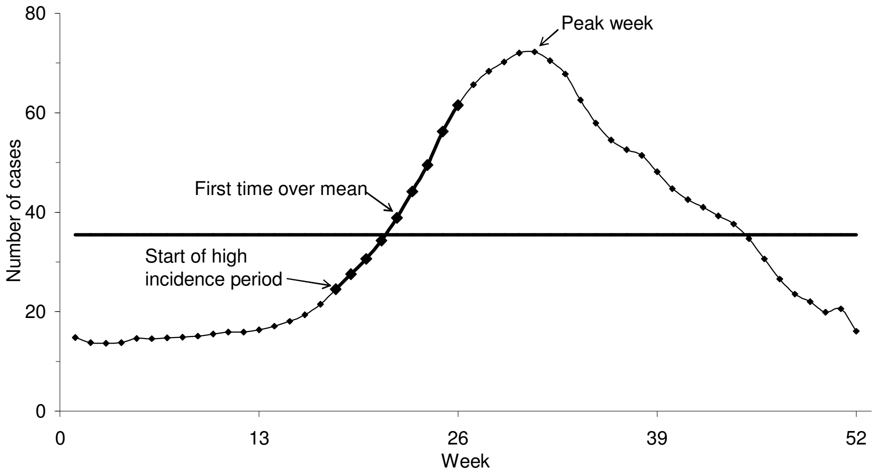


Figure 5: Nine-week centered moving average for the number of cases of campylobacter infection averaged over years for Sweden. The straight line indicates the mean number of reported cases per week. The accentuated part of the curve shows the nine-week period used for defining the start of the high incidence period.

the county is -0.42. The rank correlation for the start of the high incidence period is -0.65. The second of these rank correlations differs significantly from zero on the 5 % level, but not the first ($p = 0.055$ and $p = 0.0013$ respectively).

5 Parametric modelling for smoothing over time

There is evidently a large amount of randomness associated with the number of reported cases. In order to discern patterns as regards to variations within a year and between years we will have to smooth the time series in a convenient way. We will here do this by applying a model that describes how the mean number of cases varies in time and describes the random variations around this mean. The model that is used is in many respects very crude. It should not be regarded as a realistic stochastic model but rather as a model that produces a smoothed version of the time series. By studying the parameters in this smoothed version, we may get a better view of underlying regularities in the spread of campylobacter infections.

5.1 Model 1 – A model based on Poisson variation

The first model fitted to the data is one based on Poisson variation. This is a traditional way of modelling disease incidence data. For a given week, the number of cases is assumed

Table 3: Peak week and start of high incidence period in the counties. The last column show the relative rank of the north-south (N-S) position of the residential cities

	County	Peak week	Start of high incidence period	Rank of N-S position
1	Stockholm	31	19	10
2	Uppsala	34	18	7
3	Södermanland	27	17	12
4	Östergötland	28	18	13
5	Jönköping	30	18	14
6	Kronoberg	31	18	17
7	Kalmar	30	18	18
8	Gotland	29	18	16
9	Blekinge	29	16	20
10	Skåne	31	17	21
11	Halland	32	19	19
12	Västra Götaland	31	18	15
13	Värmland	33	20	9
14	Örebro	30	17	11
15	Västmanland	28	18	8
16	Dalarna	32	18	6
17	Gävleborg	31	20	5
18	Västernorrland	32	20	4
19	Jämtland	32	21	3
20	Västerbotten	31	20	2
21	Norrbotten	32	19	1
	Sweden	31	18	

to follow a Poisson distribution with some intensity that might depend on covariates. When the disease is rare, this is often a fair assumption. Allowing the intensity to vary over time, we try to fit a model that takes the specific patterns seen in the data into account. That is, the basic features are:

- There is a flow of cases during the entire year.
- Within each year there exists one high incidence period. The time of the peak may vary between years.
- The duration of the high incidence period may vary between years.
- The ratio between the incidence in the high and low incidence periods may vary between years.

Let Y_j be the number of cases during week j. According to the assumptions above, these numbers are stochastically independent and Poisson distributed with intensities that might depend on time. That is:

$$Y_j \sim \mathrm{Po}(\mu_j), \qquad j = 1, 2, ..., n$$

where n is the number of weeks studied.

To be able to capture the features described above we have chosen the following model for the intensities (i.e., the mean number of cases during a particular week j):

$$\ln(\mu_j) = \ln(I) + \beta'_{yr} + \tau_{yr} \left(\frac{\cos(2\pi(v_j - \theta_{yr})) + 1}{2} \right)^{\kappa_{yr}} . \tag{1}$$

I is the mean size of the population during the entire period (1992–1998). The term $\ln(I)$ serves as normation. The purpose of including I in the model is to make the parameters comparable between counties with different population sizes and is otherwise redundant.

The index yr stands for the year to which week j belongs and v_j stands for the relative position of the week within that year, i.e.

$$v_j = j \cdot \frac{7}{365} - (yr - 1992), \qquad yr = 1992, 1993, \ldots, 1998.$$

The parameters β'_{yr} describe the incidence of sporadic cases during the low incidence period of the year. In fact, $Ie^{\beta'_{yr}}$ is the average number of reported cases during the low incidence period. Instead of having a jump function that changes value for each year we have made this expression smooth by actually using a polynomial for this parameter, i.e.

$$\beta'_{yr} = v_j \beta_{yr} + (1 - v_j)\beta_{yr+1}.$$

The expression

$$\tau_{yr} \left(\frac{\cos(2\pi(v_j - \theta_{yr})) + 1}{2} \right)^{\kappa_{yr}}$$

describes the variation of the incidence over a year. It takes its highest value in year yr at time θ_{yr}. At that time the incidence is $e^{\tau_{yr}}$ higher than during a low incidence period that year. The duration of the high incidence period is related to κ_{yr}. A high value gives a short duration and a low value a long duration of the high incidence period. A graphical illustration showing what the parameters represents is shown in Figure 6.

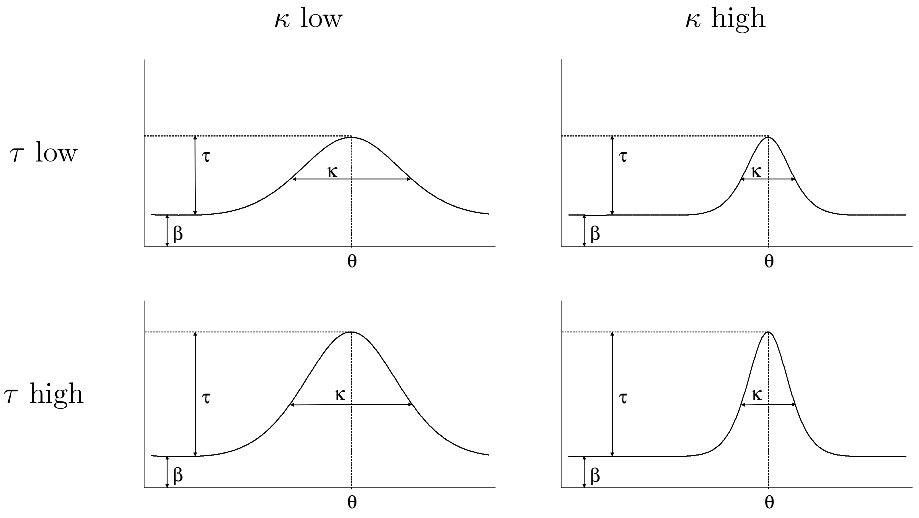


Figure 6: Graphical representation of the parameters in the model

5.2 Overdispersion

In Model 1, discussed in Section 5.1, we have assumed that the number of reported cases each week is Poisson distributed. This assumption can be questioned for several reasons. One important feature of the Poisson distribution is that its mean equals its variance, i.e., the ratio between the variance and the mean is 1. A distribution with a ratio greater than 1 is called overdispersed. There are strong reasons to believe that an overdispersed distribution should be more appropriate than the Poisson distribution. Overdispersed distributions can be motivated by the presence of

- *Minor outbreaks*

  The campylobacter cases may occur simultaneous in several persons due to exposure to the same infectious source. Even if we have discarded reported cases from major outbreaks from our analysis, there still may be reported cases from minor outbreaks. This should imply that there is dependence between cases. The most natural assumption is that the number of events when campylobacter infections are transmitted is Poisson distributed and the number of persons infected in such event is a random. The number of reported cases should in that case be modelled as a sum of a Poisson distributed number of independent random number of cases. This will yield an overdispersed distribution.

- *Secondary cases*

An infected person may spread the infection further to individuals in the neighborhood. It is often claimed in the literature that such secondary infections are uncommon. A mechanism of this kind would also cause some cases to depend on each other and imply an overdispersed distribution rather than a Poisson distribution. Another consequence could be a (stochastic) dependence of observations in subsequent weeks.

- *Dependent reporting*

  Dependencies between reported cases can also be the effect of the reporting system. If the reporting of cases is not done independently but, e.g., the reporting medical officer reports several unrelated cases simultaneously an artificial clustering, as relates reporting date, of cases occurs. Since we have chosen to use the date of the onset of the disease as the time associated with the infection we have possibly avoided this effect.

- *Heterogeneity*

  Another possible cause of overdispersion is that the relatively large areas for which the data are presented in fact consist of several sub-areas that have different patterns as regards to the temporal variation in the number of campylobacter cases. Such sub areas could, e.g., be rural and urban areas or coastal and inland parts of a county.

- *Thinning due to underreporting*

  It is well established that the campylobacter cases are severely underreported. A simple model for underreporting is that each case is reported with a certain probability, independent of other cases being reported. The observed series of reported cases is then a thinned version of the series of all campylobacter infections. A theoretical analysis shows that this kind of underreporting results in observations that are more Poisson-like than the unthinned series. Thus underreporting has an opposite effect compared with the other problems mentioned since it will make the distribution less overdispersed.

To account for overdispersion, we need to find a way to model the variance so it is allowed to be greater than the mean. This can be done in different ways, see e.g., McCullagh & Nelder (1989) or Hinde & Demétrio (1998).

Let, again, Y_j, $j = 1, \ldots, n$, represent the number of observations within week j. Y_j is assumed to follow a Poisson distribution with mean μ_j. That is, $\mathrm{E}[Y_j \mid \mu_j] = \mathrm{Var}[Y_j \mid \mu_j] = \mu_j$. One way of modelling extra variation is to simply assume a constant overdispersion and replace the variance function with

$$\mathrm{Var}[Y_j \mid \mu_j] = \phi\mu_j.$$

Another way is to assume that the parameter itself, in the Poisson distribution, follows a random distribution. This can for example be motivated by heterogeneity within the population or clustering. A commonly used approach assumes a Gamma distribution for the μ_j. This leads to a Negative binomial distribution for Y_j.

The relationship between the mean and the variance in the Negative binomial distribution can easily be derived in the following way. If we let μ_j be Gamma(α_j, δ) distributed, with a parameterisation such that

$$\mathrm{E}\left[\mu_j\right] = \frac{\alpha_j}{\delta} \qquad \text{and} \qquad \mathrm{Var}\left[\mu_j\right] = \frac{\alpha_j}{\delta^2}.$$

Then

$$\mathrm{E}\left[Y_j\right] = \mathrm{E}\left[\mathrm{E}\left[Y_j \mid \mu_j\right]\right] = \mathrm{E}\left[\mu_j\right] = \frac{\alpha_j}{\delta}$$

$$\begin{aligned} \mathrm{Var}\left[Y_j\right] &= \mathrm{E}\left[\mathrm{Var}\left[Y_j \mid \mu_j\right]\right] + \mathrm{Var}\left[\mathrm{E}\left[Y_j \mid \mu_j\right]\right] \\ &= \mathrm{E}\left[\mu_j\right] + \mathrm{Var}\left[\mu_j\right] \\ &= \frac{\alpha_j}{\delta} + \frac{\alpha_j}{\delta^2} \\ &= \frac{\alpha_j}{\delta}\left(1 + \frac{1}{\delta}\right). \qquad (2) \end{aligned}$$

The factor $(1 + 1/\delta)$ in (2) measures the overdispersion. When this factor is unity, there is no overdispersion.

The choice of the way to model the overdispersion should be subject to the underlying process generating the data. The Negative binomial distribution have the same mean value structure but different variance structure than the Poisson distribution. Using the Poisson model, we expect the same parameter estimates but underestimated variance.

To see whether it is necessary to incorporate a dispersion parameter in the model one can check the Poisson model assumption by calculating the Pearson χ^2-statistic:

$$X^2 = \sum_{j=1}^{n} \frac{\left(y_j - \hat{\mu}_j\right)^2}{\hat{\mu}_j}.$$

If the observations are in fact Poisson distributed this statistic should be approximately χ^2 distributed and approximately equal to n. The dispersion can be estimated by X^2/n.

The estimated dispersion is shown in Table 4 for individual counties and aggregated data for Sweden. Due to heterogeneity the overdispersion for Sweden is, as expected, larger than for the individual counties.

5.3 Model 2 – The Negative binomial model

The risk of being infected with Campylobacter is probably not equal across the country. One can imagine differences between rural and urban areas as well as coastal and inland areas. Clustering can occur from minor outbreaks and the possibility of person to person transmission of the infection. Although we have tried to remove the major outbreaks from the data it is impossible to identify the minor outbreaks. As previously mentioned, secondary cases are uncommon according to the literature but can still occur. All these factors can induce overdispersion.

Because of the reasons mentioned in the previous section and since there is evidence suggesting overdispersion (cf. Table 4), we have chosen the Negative binomial approach, described in Section 5.2, to extend our model to account for overdispersion. That is, the model can now be formulated as

$$Y_j \mid \mu_j \sim \mathrm{Po}(\mu_j), \qquad j = 1, 2, ..., n_j$$

where

$$\mu_j \sim \mathrm{Gamma}(\alpha_j, \delta).$$

Equation (1) is changed accordingly to

$$\ln(\alpha_j) = \ln(\delta) + \ln(I) + \beta'_{yr} + \tau_{yr} \left(\frac{\cos(2\pi(v_j - \theta_{yr})) + 1}{2} \right)^{\kappa_{yr}}. \tag{3}$$

The number of cases within a week j is still assumed to follow a Poisson distribution initially but extra variation is allowed for by assuming that the mean value parameter μ_j in the Poisson distribution is random. By assuming a Gamma distribution for μ_j we end up with a Negative binomial distribution for the number of cases per week.

The mean of the Negative binomial model is the same as in the Poisson model but now expressed in the parameters α_j and δ as

$$\mathrm{E}\left[Y_j\right] = \frac{\alpha_j}{\delta}$$

and the variance is

$$\mathrm{Var}\left[Y_j\right] = \frac{\alpha_j}{\delta}\left(1+\frac{1}{\delta}\right).$$

5.4 Prior distributions

Estimation of parameters is done within a Bayesian setting using Markov chain Monte Carlo methods as described in Section 7. This means that we will apply prior distributions to the parameters and use the means of the posterior distributions as parameter estimates. If the prior distributions are vague enough, the posterior distributions will essentially depend on the data. The following prior distributions was used for Model 2 in the final simulations:

$$\begin{aligned}
\beta &\sim \mathrm{N}\left(0, \sigma_\beta^2 = 1000\right)\\
\theta &\sim \mathrm{Uniform}\left(0, 1\right)\\
\tau &\sim \mathrm{Gamma}\left(0.001, 0.0005\right)\\
\kappa &\sim \mathrm{Gamma}\left(0.001, 0.001\right)\\
\delta &\sim \mathrm{Gamma}\left(0.001, 0.01\right)
\end{aligned}$$

6 Geographic modelling

Until now, the models discussed have only regarded the aggregated data for the whole country. However, when we look at individual counties, we can still see distinct patterns in the time series of reported cases.

County wise, the problem of overdispersion is less than when the data are aggregated for the whole country. This can partly be explained by that there should be less heterogeneity when the data are not aggregated. That is, the variation is greater between than within counties. However, calculations of the Pearson statistic for each county still suggest that there is overdispersion present when data are analysed by county. The overdispersion seems to be larger for counties with larger population. The results from the calculations, together with the estimated overdispersion, $(1 + 1/\delta)$, from the Negative binomial model, are presented in Table 4. As expected, the estimates of the dispersion parameter from the Pearson statistic and the Negative binomial model are close to each other.

We have chosen to use the Negative binomial model (Model 2) in our continued analysis of the county wise data. The cost of adding an extra parameter in the model is well motivated by the gain in precision and model fit.

Table 4: Overdispersion; as measured by the Pearson statistic and estimated from the Negative binomial (NB) model

	County	Pearson χ^2–statistic	Estimated from the NB model
1	Stockholm	1.7571	1.8302
2	Uppsala	1.0943	1.0282
3	Södermanland	1.1516	1.1492
4	Östergötland	1.2812	1.0352
5	Jönköping	1.4286	1.2915
6	Kronoberg	0.9717	1.0156
7	Kalmar	1.1645	1.0411
8	Gotland	1.1949	1.0307
9	Blekinge	1.1384	1.0306
10	Skåne	1.5994	1.6133
11	Halland	1.3696	1.2494
12	Västra Götaland	1.5748	1.4398
13	Värmland	1.0408	1.0278
14	Örebro	1.0677	1.0360
15	Västmanland	1.0792	1.0396
16	Dalarna	1.1023	1.0424
17	Gävleborg	1.2640	1.1590
18	Västernorrland	1.4834	1.2811
19	Jämtland	1.1655	1.1608
20	Västerbotten	1.4356	1.4020
21	Norrbotten	1.0512	1.0399
	Sweden	2.9354	3.0927

Fitting the model for the individual counties can simply be done by substituting the data for the whole country by each county's data respectively. This will give us, for each county, parameter estimates which then can be compared in some way. Another way is to extend the model to incorporate a geographic component by simply adding an extra index on all parameters. We have used this latter approach because, as we will see, it is then easy to modify the model to account for dependencies between the counties. Another advantage is that we only need to fit one model including all counties instead of 21 separate models. In this setting the counties are assumed to be independent of each other. Model 2 can now be expressed as:

$$Y_{ij} \sim \text{Po}(\mu_{ij}), \qquad i = 1, 2, \ldots, m, \quad j = 1, 2, ..., n$$

where i is indexing the m counties and

$$\mu_{ij} \sim \text{Gamma}(\alpha_{ij}, \delta_i).$$

The functional form describing the mean number of cases each week j is then:

$$\ln(\alpha_{ij}) = \ln(\delta_i) + \ln(I_i) + \beta'_{i,yr} + \tau_{i,yr} \left(\frac{\cos(2\pi(v_j - \theta_{i,yr})) + 1}{2} \right)^{\kappa_{i,yr}}.$$

6.1 Prior distributions

Using the same prior distributions for the individual counties as for the whole country was not possible. Making the priors too vague caused the simulation to slow down considerably and in most cases convergence was not reached. The problems occurred mainly for the parameters τ and κ. In order to be able to run the simulations, the prior distributions for those parameters had to be made more vague. The parameters had the following prior distributions in the final simulations:

$$\begin{aligned} \beta &\sim \text{N}\left(0, \sigma^2_\beta = 1000\right) \\ \theta &\sim \text{Uniform}\left(0, 1\right) \\ \tau &\sim \text{Gamma}\left(4, 2\right) \\ \kappa &\sim \text{Gamma}\left(0.3, 0.3\right) \\ \delta &\sim \text{Gamma}\left(0.001, 0.01\right) \end{aligned}$$

For counties with a large amount of data or with a strong structure in the observed incidence the data dominates the prior. However, for counties with less data, the prior distributions might strongly influence the parameter estimates. The problem of having a too informative prior distribution will be more severe for the tails than for the center of the distribution. Since we are mainly focusing on the estimates of the parameters, and not on the variances, we have chosen to set the mean of the prior distributions to be, in some sense, plausible and conservative. The prior mean for τ is set to two, suggesting that the peak incidence is $\exp\{2\} \approx 7$ times larger than the basic low incidence. For κ the prior mean is set to one, making the assumption that the high incidence period is as long as the low incidence period. The assumed means for τ and κ are actually not far from the estimated national mean of these parameters.

6.2 Model 3 – Spatial smoothing

When we split the data by counties the amount of information, per parameter to be estimated, is reduced. Also, counties with small population sizes will during most weeks have only few or no cases. It is likely that extreme observations will occur just by chance and this will influence the parameter estimates. For counties with a small population size extreme observations can easily occur just by chance but also from minor outbreaks only involving a few cases which can be hard to detect by the reporting system. We can partly overcome this problem by smoothing extreme observations towards some function of the observations for other counties. This can be done in several ways. We are here only looking at a couple of possibilities but others are easily imagined.

Assuming that all counties have some feature, regarding the flow of cases, in common, e.g., the same baseline incidence of sporadic cases, extreme local observations could be smoothed towards some global mean for all other counties. That is, if one county experiences a much higher baseline incidence one year, it is smoothed towards the mean of all other counties' baseline incidences that year. This is sometimes referred to as unstructured smoothing.

If we believe that there is a geographic component involved in the distribution of cases we can try to incorporate this in the model. That is, geographically close counties are assumed to have similar parameter values. In this case, each county's incidence is shrunk towards the mean of the neighboring counties' incidences. This is then referred to as spatially structured smoothing.

The idea, in both cases, is to punish extreme observations that might have arisen just by chance. In this way, counties with few cases can, in some sense, borrow information from other counties to possibly get better estimates of their true parameter values.

Of course, any one of the parameters in the model could be subject to smoothing. In fact, one possibility is to smooth on all parameters simultaneously. However, as a start we have chosen to look at only one parameter. The main reasons for this are the easier understanding and also to save computational time. It would, of course, be more difficult to study the effects of smoothing if we were smoothing several parameters at the same time, especially if the parameters are correlated in some sense. Also, smoothing on several parameters would increase the complexity of the model and hence probably require heavier simulations to estimate the parameters.

We have focused on smoothing on the parameter $\theta_{i,yr}$, describing the time of the peak of the high incidence period for county i in year yr. Reasons for choosing this parameter is that it is easy to interpret and also that it is interesting to study from a practical point of view.

To achieve a smoothed estimate of $\theta_{i,yr}$, we change the prior distribution from an uninformative Uniform distribution to an informative Beta distribution. The reason for choosing a Beta distribution is that we need a flexible distribution limited on the interval $(0, 1)$, preventing estimation on the wrong year. In order to smooth using information from other counties, the Beta distribution should have a mean and a variance depending on this information. We can formulate all this in the following way:

The prior distribution for $\theta_{i,yr}$ is now

$$\theta_{i,yr} \sim \text{Beta}(r, s).$$

The parameters r and s are chosen so that

$$\begin{aligned} \mathrm{E}\left[\theta_{i,yr}\right] &= \bar{\theta}_{-i,yr} = \frac{\sum_{j \neq i} w_{ij}\theta_j}{\sum_{j \neq i} w_{ij}} \qquad \text{and} \\ \mathrm{Var}\left[\theta_{i,yr}\right] &= \frac{\sigma^2 \sum_{j \neq i} w_{ij}^2}{\left(\sum_{j \neq i} w_{ij}\right)^2} \end{aligned}$$

where

$$W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{pmatrix}$$

is a weight matrix with entries w_{ij} representing the weight that county j has on county i. Hence, this implies that , $\bar{\theta}_{-i,yr}$ is the weighted average of all other counties' θ values. Note that W is not necessarily symmetric. In words this means that county j has a different influence on county i than county i has on county j.

The parameter σ^2 represents the overall variance of the smoothing function and is in this setting assumed to be equal across counties. This parameter determines the degree of smoothing between the counties. Smaller values of σ^2 increase the influence from the other counties and hence gives a more smoothed map.

The Beta distribution is parameterised in a way that

$$\begin{aligned} \mathrm{E}\left[\theta_{i,yr}\right] &= \frac{r}{r+s} \qquad \textit{and} \\ \mathrm{Var}\left[\theta_{i,yr}\right] &= \frac{rs}{(r+s)^2(r+s+1)} \end{aligned}$$

which implies that r and s should be set to

$$\begin{aligned} r &= \frac{\left(\sum_{j\neq i} w_{ij}\right)^2}{\sigma^2 \sum_{j\neq i} w_{ij}^2}(\bar{\theta}_{-i,yr})^2(1-\bar{\theta}_{-i,yr}) - \bar{\theta}_{-i,yr} \qquad \text{and} \qquad (4) \\ s &= \frac{\left(\sum_{j\neq i} w_{ij}\right)^2}{\sigma^2 \sum_{j\neq i} w_{ij}^2}\bar{\theta}_{-i,yr}(1-\bar{\theta}_{-i,yr})^2 - (1-\bar{\theta}_{-i,yr}) \qquad (5) \end{aligned}$$

respectively.

6.2.1 Choosing a weight matrix

The weight matrix defines the mutual influence the counties have on each other. If all weights are set to zero the counties are assumed to be spatially independent. On the other hand, if all weights are equal and larger than zero, all counties are expected to have the same mean and all estimates are shrunk towards that global mean. It is easy to think of other structures allowing for different kinds of spatial dependencies. We have chosen to look at two specific alternatives of weight matrices.

6.2.2 Model 3.1 – Neighbors

The first alternative for a weight matrix assumes that the parameter of interest for one specific county depends on the corresponding parameter for neighboring counties. In this setting, two counties are said to be neighbors if they share a common border. There are exceptions like the island Gotland who does not in a strict sense share a border with any county since it is surrounded by water. Gotland has as its neighbors the four counties lying on the east coast of the part of the mainland lying closest to Gotland. The weights in the weight matrix is defined, in this setting, by:

$$w_{ij} = \begin{cases} 1 & \text{if county } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases} .$$

It is easy to confirm that the expected value of the prior distribution for θ_j is the arithmetic average of the neighboring counties' θ values and that the variance is proportional to the reciprocal of the number of neighbors. The more information we have about the surrounding areas, i.e., the more neighboring counties we have, the smaller the variance of the prior distribution. In Appendix B the weight matrix showing the neighboring structure is presented.

6.2.3 Model 3.2 – Weights defined by distances

The second choice of weight matrix assumes that all counties depend on each other in a direct way. Here the weights are proportional to some function of the distance between the counties. Counties that are far apart have lower influence on each other than counties that are close. Depending on the disease and its aetiology, there are many ways of measuring the distance between two counties. If the disease is spread from man to man, one could define the distance between two counties as some measure of the flow of people travelling between the counties. This could, e.g., be realistic for a disease like influenza. Another distance could be the number of borders one has to cross to get from one county to another. In this case the dependence will diminish with the distance in some sense. This approach would not consider the size of the counties. A third distance, the approach we have chosen, is the Euclidian distance between the counties. This distance resembles the previous one but will take the sizes of the counties into account. It is however, not trivial to define the Euclidian distance between two counties. There are several possibilities such as the distance between the geographical mid-points of the counties or between the geographical means of the population densities. Both of these mid-points are difficult to establish and the gain of finding them might not be worthwhile the effort. We have instead chosen the Euclidian distance between the residential cities in the counties. In most counties (but not all) the residential city can be viewed as the mode of the population density. The weights are calculated in the following way. First the longitude/latitude position of the residential cities is established. Each degree of longitude and latitude is approximately 50 km and 110 km respectively in Sweden. After the distance in longitude and latitude has been transformed to kilometers the Euclidian distance is calculated using Pythagoras' theorem. The weights are then simply the reciprocal of the distances. In order to avoid calculating the normation constant, $\left(\sum_{j\neq i} w_{ij}\right)^2$, in the simulation, we then chose to norm the weights such that

$$\sum_{j\neq i} w_{ij} = 1.$$

As mentioned above there are other possibilities of weight matrices. Mollié (1996) study how the relative risk of death from a rare disease vary between 94 geographic areas (dèpartements) in France. In that study she uses an intermediate, so called "convolution Gaussian", distribution on log relative risks to accommodate both an unstructured prior and a purely spatially structured prior.

6.2.4 Choosing a prior distribution for σ^2

The Beta distribution has some properties that we need to consider in order to be able to choose the limits of the prior for σ^2. If $r = s$ the distribution is symmetric and the mean will be 0.5. Particularly, if $r = s = 1$ we have a uniform distribution (cf. Figure 7b). If both r and s are less than 1 the distribution is convex (Figure 7a) and if both are larger than 1 it is concave (Figure 7c). Also, the larger values of r and s the smaller the variance is.

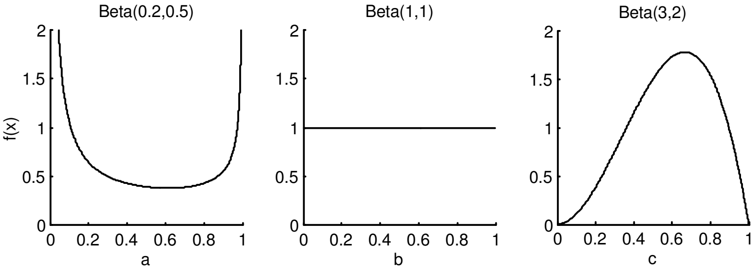


Figure 7: Examples of different beta distributions. A convex Beta(0.2,0.5) (a), a symmetric, uniform Beta(1,1) (b) and a concave Beta(3,2) (c).

Since we are interested in weighting estimates towards the mean of other counties we need a distribution with most its density close to the mean, not in the tails. This means that we want a concave distribution and hence that r and s should both be greater than 1. Clearly the variance is also a function of the mean of the distribution, taking its largest value when the mean is equal to 0.5. In fact, from Equations 4 and 5 we see that, to make sure that the distribution will be concave, σ^2 should be selected to satisfy

$$\begin{aligned}\frac{\left(\sum_{j\neq i} w_{ij}\right)^2}{\sigma^2 \sum_{j\neq i} w_{ij}^2}(\bar{\theta}_{-i,yr})^2(1-\bar{\theta}_{-i,yr}) - \bar{\theta}_{-i,yr} &> 1 \qquad \text{and} \\ \frac{\left(\sum_{j\neq i} w_{ij}\right)^2}{\sigma^2 \sum_{j\neq i} w_{ij}^2}\bar{\theta}_{-i,yr}(1-\bar{\theta}_{-i,yr})^2 - (1-\bar{\theta}_{-i,yr}) &> 1.\end{aligned}$$

Solving both inequalities for σ^2 we see that σ^2 should be less than:

$$\min\left(\frac{\bar{\theta}_{-i,yr}^2(1-\bar{\theta}_{-i,yr})\left(\sum_{j\neq i} w_{ij}\right)^2}{(1+\bar{\theta}_{-i,yr})\sum_{j\neq i} w_{ij}^2};\frac{\bar{\theta}_{-i,yr}(1-\bar{\theta}_{-i,yr})^2\left(\sum_{j\neq i} w_{ij}\right)^2}{(2-\bar{\theta}_{-i,yr})\sum_{j\neq i} w_{ij}^2}\right). \tag{6}$$

In expression 6 it can be seen, as stated above, that the upper limit of σ^2 depends on $\bar{\theta}_{-i,yr}$. In fact, the farther away $\bar{\theta}_{-i,yr}$ is from 0.5, the smaller the upper limit of the prior distribution for σ^2 should be to assure that we will have a concave distribution for $\bar{\theta}_{-i,yr}$. As previously mentioned, the parameter estimation will be done using MCMC methods. The procedure is described in detail in Section 7. One important step of the MCMC procedure is to sample candidate points of the parameter vector from its probability distribution. To allow different candidates of $\bar{\theta}_{-i,yr}$ the upper limit of the prior distribution of σ^2 must take this into account. In order to establish an appropriate upper limit, several different prior distributions for σ^2 were tried. It turned out that a Uniform$(0, 0.08)$ prior was appropriate for the "Neighbor model" (cf. Section 6.2.2) and a Uniform$(0, 0.01)$ prior was appropriate for the model where the weights were defined by the reciprocal of the distances being normed such that $\sum_{j \neq i} w_{ij} = 1$ (cf. Section 6.2.3).

7 Estimation with MCMC

In a situation with a high dimensional distribution, estimation of parameters is often not straightforward. Integration over the distribution can be difficult or even impossible. The idea behind Markov chain Monte Carlo (MCMC) is, as the name implies, Monte Carlo integration using Markov chains. Following the notation in Gilks et al. (1996), let $\pi(.)$ denote the likelihood of our vector of observed data X. Suppose we are interested in evaluating the expectation of some function $f(X)$ of the data, i.e.

$$\mathrm{E}\left[f(X)\right] = \frac{\int f(x)\pi(x)dx}{\int \pi(x)dx}.$$

One way of evaluating this expression is by Monte Carlo integration. If it is possible to draw samples $\{X_t\}$ from $\pi(.)$ one can approximate the mean by

$$\mathrm{E}\left[f(X)\right] \approx \frac{1}{n}\sum_{t=1}^{n} f\left(X_t\right).$$

If the samples are independent and n is large enough this sample mean is a good approximation of the population mean of $f(X)$. However, drawing independent samples from $\pi(.)$ is often not so easy since $\pi(.)$ can be quite non-standard. MCMC uses the fact that it turns out that the samples need not be drawn independently. If we can draw the samples from a Markov chain which has $\pi(.)$ as its stationary distribution we can still use Monte Carlo integration for estimation, hence the name *Markov chain Monte Carlo.*

7.1 The Metropolis-Hastings algorithm

How can we then construct a Markov chain that has $\pi(.)$ as its stationary distribution? It turns out that this can be done quite easily by the Metropolis-Hastings algorithm. Fifty years ago Metropolis et al. (1953) proposed a method which later on was generalised by Hastings (1970). The rather simple algorithm proceeds as follows:

We are interested in generating a sequence of random variables $\{X_0, X_1, X_2, \ldots\}$ that has the properties of a Markov chain with stationary distribution $\pi(.)$. That is, given that we are in state X_t the next state, X_{t+1}, is generated from a random distribution which only depends on the state X_t and not on the history $\{X_0, X_1, \ldots, X_{t-1}\}$.

To construct such a chain according to the Metropolis-Hastings algorithm, a candidate point Y is first drawn from a proposal distribution $q(.|X)$. This point is then accepted as the new state, X_{t+1}, with probability:

$$\alpha(X_t, Y) = \min\left(1, \frac{\pi(Y)q(X_t \mid Y)}{\pi(X_t)q(Y \mid X_t)}\right).$$

Generating a uniformly distributed random number $U \sim Uniform\,(0,1)$, we can select the next state as:

$$X_{t+1} = \begin{cases} Y & if \quad U \leq \alpha(X_t, Y) \\ X_t & if \quad U > \alpha(X_t, Y) \end{cases}$$

This "updating procedure" is then repeated a large number of times until the Markov chain has converged. The initial state, X_0, can be given any value or chosen randomly from a distribution. The choice of starting point might affect the time to convergence, but when run long enough, the chain will eventually forget its initial phase. To be sure that the Markov chain has converged to its stationary distribution, a long enough burn-in period should be run before using the iterations for estimations. A discussion on how to determine the burn-in period can be found in chapter one in Gilks et al. (1996).

All our simulations were done using the program WinBUGS (version 1.3; Spiegelhalter et al. (2000) or http://www.mrc-bsu.cam.ac.uk/bugs). For the simplest models, i.e., Model 1 in Section 5.1 and Model 2 in Section 5.3, a slice-sampling method (cf. Neal (1997)) was used. This method is appropriate when the density function is non log-concave but on a restricted range. It has an adaptive phase of 500 iterations which will not be included in the analysis. For the models in Section 6, a “current point Metropolis algorithm” was used. This method uses a normal proposal distribution. The standard deviation is tuned over the first 4 000 iterations and those will not be used in the analysis. For all models a total of 30 000 iterations was run. The last 20 000 iterations was used in the analysis and the preceding served as burn-in.

7.2 Convergence diagnostic analysis

When using MCMC methods it is important that the Markov chain converge to its stationary distribution, or at least closely enough. There are several convergence diagnostic tools developed to check whether convergence is achieved. We have not done any formal checks for convergence using any of those tools. However, we have used some *ad hoc* approaches to make sure that we have reached the stationary distribution and that the parameter estimates are valid.

As mentioned in Section 5.4, we have used somewhat informative prior distributions for the parameters in the modelling on the county level. There is evidently a strong structure in the data. When there is a large amount of data the prior distributions will not have as much influence on the parameter estimates as when there are less data. To illustrate this, prior distributions from Model 3.1 for the parameters β, τ and κ are plotted together with their corresponding posterior distributions for Västerbotten and Stockholm 1995 in Figure 8. Västerbotten is a county with a relatively small population size while Stockholm is the county with the largest population size.

Clearly, the posterior distribution is much narrower for Stockholm than for Västerbotten due to more data. The normal prior for β is too flat to be able to even be seen in the figure. That is, for β the posterior distribution is almost totally dominated by data.

During the updating procedure in WinBUGS it is relatively easy to plot the sequential iteration history to see whether the Markov chain changes states frequently or if it stays in the same state for long periods. The later would require more iterations for the chain to converge. We have also visually checked the posterior distributions for the parameters.

The time to reach the stationary distribution is also depending on the starting point for the chain. Starting points lying far away from the stationary distribution will require more iterations. We have tried different starting points, even some extreme ones, but this did not seem to affect the time to convergence.

Simulations took between 7 min (simplest model) and 6 h (most complex model) on a PC with an AMD 1.4 GHz processor and 512 Mb RAM.

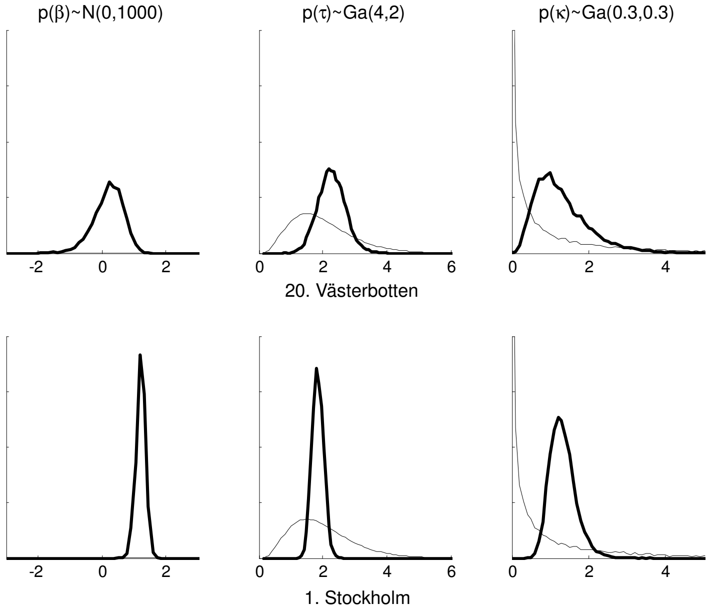


Figure 8: Prior (thin line) and posterior (bold line) distributions for three parameters for the county Västerbotten (top) and Stockholm (bottom) in 1995 from Model 3.1. Note that the prior distribution for β is too flat to even be seen in the figure.

8 Results

8.1 The fit of the models – aggregated data for Sweden

We fitted both the Poison and the Negative binomial model to the aggregated data for Sweden. In Figure 9 (top) the crude number of cases together with the estimated mean number of cases from both models are shown. Note that here and in the following analyses, as in the crude analysis, the time series start with week 6 in 1992 and end with week 5 1999 due to computational reasons. It turns out that both models fit the data well. In fact, it is almost impossible to discern any differences between the models by just looking at the figure. To show that the models do not give exactly the same fit, the absolute difference of the estimated mean number of cases between the models are also shown in the figure (bottom). The parameter estimates from the two models are shown in Table 5.

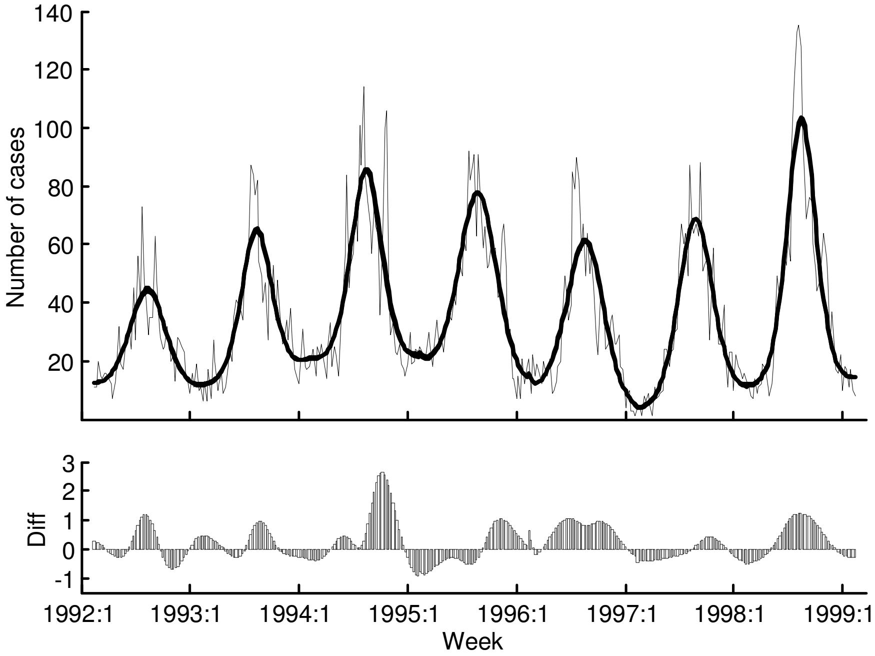


Figure 9: The two models fitted to the data (top) and the absolute difference between the estimated Poisson model and the Negative binomial model (bottom).

The absolute difference of the estimated incidences from the two models is never more than three cases and the relative difference is not more than five per cent at most. The largest difference between the models appears right after the time for the peak incidence in 1994. As can be seen in Figure 9 there appear to be a second peak during the decline of the incidence curve in this year. This is probably caused by an undetected outbreak, possibly occurring in more than just one county which would make it more difficult to discover. In fact, by taking a closer look at individual time series, there actually appear to be a second peak in late 1994 for some of the counties. Examples are Stockholm (Figure 11), Västerbotten (Figure 12), Södermanland, Västra Götaland and Örebro (all counties are shown in Appendix C). The positive differences at the time of the downslope of the curve in 1994 indicates that the Poisson model is more affected by this second peak than the Negative binomial model is.

As previously mentioned, if overdispersion is present we expect the estimates of the incidence to be similar but the variance will be underestimated if we use the Poison model instead of the Negative binomial model. This also has implications on the parameter estimates. When we fit the Negative binomial model we end up with parameter estimates with larger variance than if we fit the Poison model. We illustrate this by plotting the posterior distributions, from both models, of one of the parameters in Figure 10.

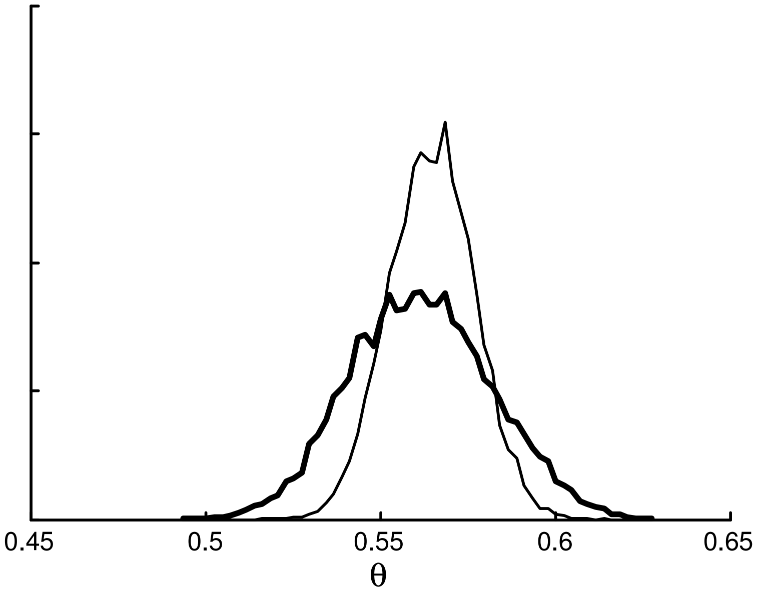


Figure 10: Posterior distributions for the parameter θ_{1996}. Thin line represents the Poisson model and bold line represents the Negative binomial model.

It is clearly seen that the Negative binomial model produce a wider posterior distribution than the Poison model. Posterior probability intervals for all parameters are shown in Table 5. Also in the table, one can see that the Negative binomial model produces estimates with larger variation, indicating the need of a dispersion parameter in the model.

Table 5: Parameter estimates with empirical 95 per cent probability intervals for the Poisson and the Negative binomial models

		Poisson model		Negative binomial model	
Parameter	Year	Mean	Prob. interval	Mean	Prob. interval
θ	1992	0.512	(0.487 – 0.539)	0.516	(0.468 – 0.568)
	1993	0.501	(0.484 – 0.517)	0.498	(0.469 – 0.530)
	1994	0.519	(0.504 – 0.533)	0.514	(0.489 – 0.540)
	1995	0.556	(0.538 – 0.573)	0.553	(0.524 – 0.584)
	1996	0.565	(0.541 – 0.588)	0.562	(0.524 – 0.602)
	1997	0.509	(0.489 – 0.529)	0.510	(0.472 – 0.543)
	1998	0.504	(0.491 – 0.517)	0.504	(0.482 – 0.529)
τ	1992	1.284	(1.240 – 1.450)	1.249	(0.987 – 1.555)
	1993	1.417	(1.289 – 1.547)	1.396	(1.172 – 1.634)
	1994	1.400	(1.295 – 1.507)	1.364	(1.187 – 1.542)
	1995	1.539	(1.417 – 1.668)	1.516	(1.313 – 1.735)
	1996	2.235	(1.999 – 2.529)	2.157	(1.812 – 2.696)
	1997	2.345	(2.124 – 2.604)	2.273	(1.948 – 2.728)
	1998	2.087	(1.938 – 2.245)	2.047	(1.806 – 2.315)
κ	1992	1.276	(0.837 – 1.856)	1.217	(0.549 – 2.328)
	1993	1.639	(1.233 – 2.142)	1.634	(0.955 – 2.658)
	1994	1.482	(1.232 – 1.761)	1.612	(1.131 – 2.220)
	1995	0.967	(0.787 – 1.173)	1.004	(0.691 – 1.394)
	1996	0.744	(0.578 – 0.941)	0.799	(0.512 – 1.184)
	1997	0.897	(0.702 – 1.119)	0.926	(0.594 – 1.331)
	1998	1.344	(1.101 – 1.622)	1.394	(0.953 – 1.948)
δ				0.478	(0.380 – 0.597)

8.2 The fit of the models – by county

By looking at the individual counties we see that the model still fits the data well even though we in some cases have substantially less data than for the whole country. As examples, the estimated mean incidence together with the observed data are shown for Stockholm (Figure 11), Västerbotten (Figure 12) and Blekinge (Figure 13).

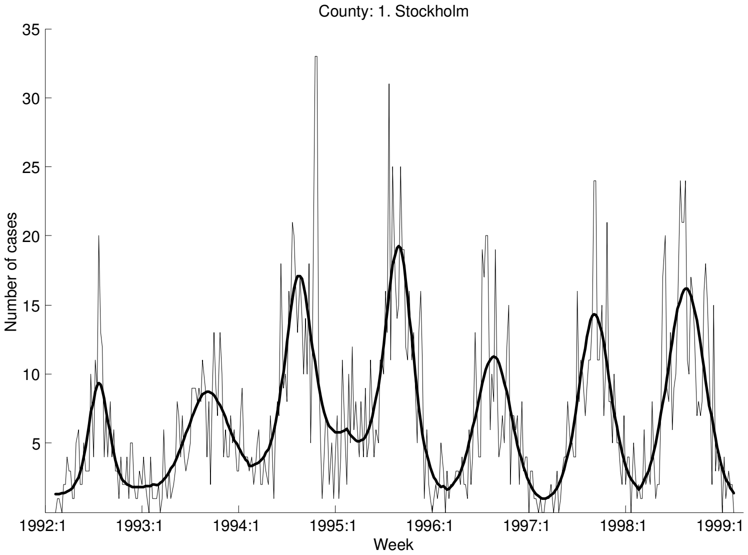


Figure 11: Observed and estimated mean number of cases week by week for Stockholm.

Stockholm is the largest county constituting approximately 20 % of the population. As a consequence of this, Stockholm is the county with the largest number of campylobacter infections. As can be seen in Figure 11, the stable structure of the time series seen for the whole country is also seen for Stockholm but he random variation around the mean is larger for Stockholm due to fewer cases.

Västerbotten and Blekinge have relatively low incidences of campylobacter infections. Still the model succeeds in finding a regular pattern in the observed time series with, in some years, very marked peaks. However, note that the peaks in 1993 and 1994 are not very prominent for Blekinge. At a first sight it may look as if the model fit the data poorly for some years, but the distinction between the high and low incidence periods appears clearer if one consider all the weeks with zero cases during the low incidence periods.

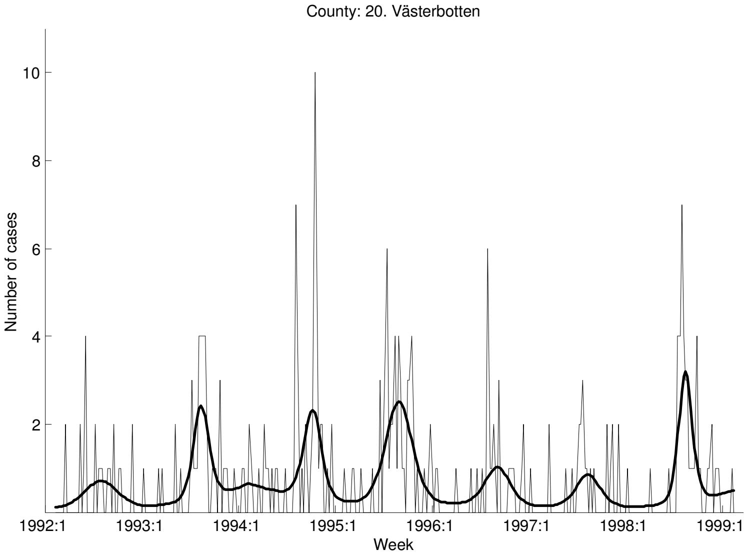


Figure 12: Observed and estimated mean number of cases week by week for Västerbotten.

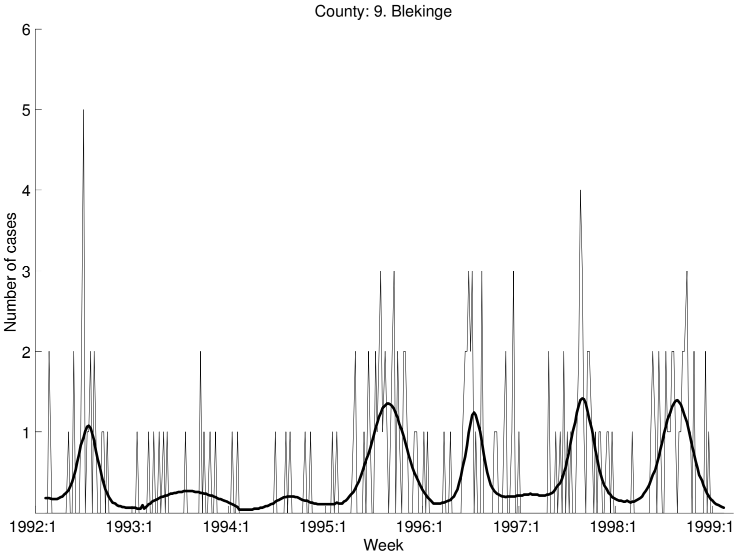


Figure 13: Observed and estimated mean number of cases week by week for Blekinge.

8.3 National results

The parameter estimates obtained from the model fitted to the aggregated data for the entire country can be used to calculate some interesting functions of the parameters. The average baseline incidence in year yr can, e.g., be calculated as $I \cdot \exp\{\beta_{yr}\}$ and the peak week is simply given by:

$$\text{Peak week} = \frac{365}{7}\theta_{yr} + 5.$$

Again, remember that we need to add five weeks since we start the time series in week 6 due to computational reasons.

The amplitudes τ_{yr} measure the logarithm of the ratio between the highest incidence in the model and the basic low incidence. That is, the actual ratio can be calculated as $\exp\{\tau_{yr}\}$.

In order to illustrate when the high incidence period starts we have calculated the number of the week in which the incidence is twice the basic low incidence. A simple calculation yields that this will be week no

$$\frac{365}{7}\left(\theta - \frac{1}{2\pi}\arccos\left(2\left(\frac{\ln(2)}{\tau}\right)^{1/\kappa} - 1\right)\right) + 5.$$

To calculate a statistic representing the concentration of the high incidence period, κ, we estimated, for each year, the proportion of cases occurring within ±2 weeks of the peak week. That is, the proportion is calculated as the estimated number of cases in the period ranging from two weeks before till two weeks after the peak week divided by the estimated total number of cases during that year. A large proportion implies a higher concentration of the peak. This measure of the size of the peak is adopted from the study by Nylén et al. (2002) in which they compared nine European countries and New Zealand regarding the seasonal distribution of campylobacter infections. As opposed to our study they used a non-parametric kernel smoother to smooth the crude observations and they also studied all campylobacter infections registered and not only the indigenous cases. As they argue in the discussion, the size of the peak can be influenced by the infections acquired abroad during holiday season which may vary between countries.

The results of all the calculations are presented in Table 6.

In two of the years, 1996 and 1997, the start of the high incidence period is earlier than in the other years. In spite of this the time of the peak does not occur sooner for these years as compared with the other years. This can seem contradictory since one might imagine that the two time points in some sense should be positively related, i.e., an early start of the high incidence period would imply an early peak week. Also, 1996 and 1997

Table 6: Baseline incidence, peak week, ratio between high and low incidence, start of high incidence period and the proportion of all cases during a year that fall ill within +/- two weeks of the peak week. Estimates from Model 2 for Sweden

	Year						
	1992	1993	1994	1995	1996	1997	1998
Peak week	32	31	32	34	34	32	31
Ratio	3.5	4.0	3.9	4.6	8.6	9.7	7.7
Starting week	21	20	21	20	17	15	18
% cases in the 5 peak weeks	17	19	19	17	19	21	22

have the highest ratios between the high and low incidence periods.

The proportion of the yearly number of cases occurring within the five peak weeks varies between years. The results presented in Table 6 suggests that the peak should be more prominent in 1998 (22 % of the cases falling within the five week period) than in 1992 and 1995 (both 17 %). This can be verified by looking at Figure 9, where it is seen that the peak in 1998 is indeed more concentrated than in 1992 and 1995.

8.4 Results by county – Geography

In the same way as for the whole country, the peak week, ratio between the high and low incidence periods, the start of the high incidence period and the density of the peak are calculated for each of the 21 separate counties. The results are presented in Tables 8–11 in Appendix A.

The time of the peak vary with a range of 13 weeks, between week 27 (Västernorrland 1993, Blekinge 1996 and Södermanland 1997) and week 40 (Västerbotten 1994). The late peak in week 40 for Västerbotten is an extreme outlier. In 1994, Västerbotten actually seem to have had two peaks. The first peak occurred approximately at the same time and magnitude as in some of the other years (cf. Figure 12). The second peak was around week 40 and had a larger magnitude than the first peak. This is actually the reason why the model fitted the second peak. As argued above, possibly there was an outbreak late in that year. The distribution of the counties' peak weeks within years is displayed in Figure 14 (top left).

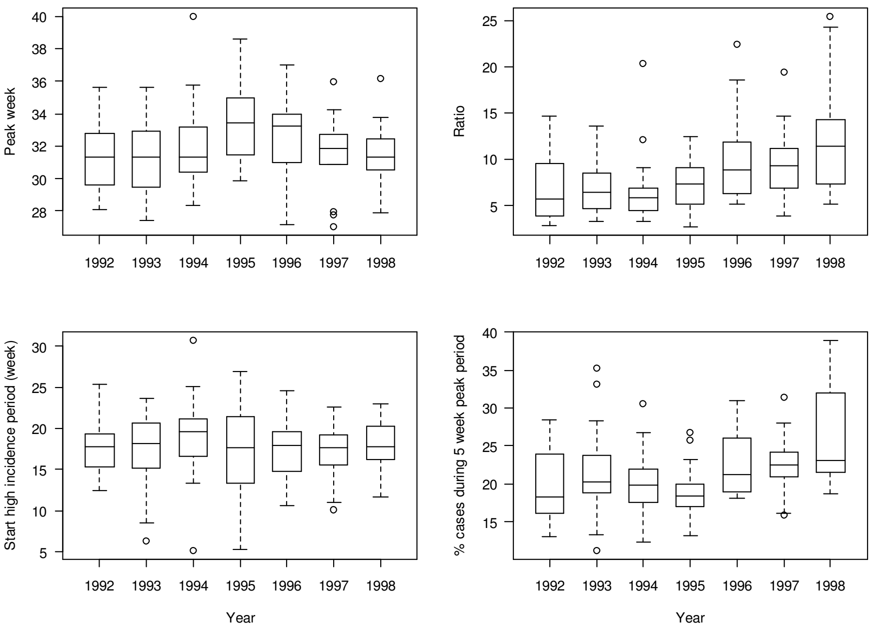


Figure 14: Distribution of peak week (top left), ratio between highest and lowest incidence (top right), start of high incidence period (bottom left) and the proportion of cases within the five peak weeks (bottom right) for all counties per year. The boxes represents the inter-quartile range and the median. Whiskers represents the minimum and the maximum values apart from outliers. Circles indicates outliers.

Generally, looking at the distribution of all counties' peak weeks per year there appear to be some differences between the years with later peaks in 1995 and 1996. However, there does not seem to be any sign of a specific trend over the years. A Friedman rank sum test gives $p = 0.072$ for testing the hypothesis of no difference in median peak week between the years.

The ratio between the highest and the lowest incidence differs significantly ($p < 0.001$) between the years (cf. Figure 14, top right) with an increasing trend over the years.

The start of the high incidence period does not seem to vary much between the years (cf. Figure 14, bottom left) and as expected, the Friedman test fails to reject the hypothesis of no difference between the years ($p = 0.9$).

Regarding the forth statistic, the proportion of cases occurring within the five peak weeks, there appear to be a trend towards higher proportions and hence more marked peaks in later years (cf. Figure 14, bottom left). As for the ratio between the high and low incidence the difference between the years was strongly significant ($p < 0.001$).

If we, instead of looking at the distribution of counties within years, look at the distribution of the yearly estimates within each county; we end up with similar boxplots but with 21 boxes, each representing the distribution for seven years (not shown). The Friedman tests in this setting suggest that there is a difference between the counties concerning the ratio ($p = 0.002$) and the density of the peak ($p < 0.001$). On the other hand, there does not appear to be any large differences between the years concerning the time of the peak ($p = 0.13$) and the start of the high incidence period ($p = 0.44$). There is no clear pattern regarding which counties are different from others with one exception; the concentration of the peak is higher for the most northern counties.

Since there might be heterogeneity in the geographical patterns between the years it is interesting to study the years separately. In Table 7 the relative north-south position of the counties are related to the estimated statistics within each year by a rank correlation coefficient.

Table 7: Spearman rank correlation between estimated functions of the parameters (peak week (peak), ratio between higest and lowest incidence (ratio), start of high incidens period (start) and proportion of cases within the five peak weeks (conc)) and the relative north-south (ns) position

	Spearman rank correlation			
Year	$\rho(peak, ns)$	$\rho(ratio, ns)$	$\rho(start, ns)$	$\rho(conc, ns)$
1992	0.42	-0.42	0.35	-0.48
1993	0.03	-0.31	-0.48	-0.65
1994	-0.21	-0.74	0.24	-0.52
1995	0.26	-0.34	0.19	-0.22
1996	-0.64	-0.28	-0.40	-0.32
1997	0.24	-0.05	-0.11	0.14
1998	0.24	-0.46	-0.33	-0.64

The only function of the parameters that seem to be somewhat consistent over the years regarding the correlation with the relative north-south position are the ratio between the highest and the baseline incidences and the density of the peak. The negative signs of the correlation coefficients means that northern counties tend to have more marked peaks than southern counties. The exception is the year 1997 in which the north-south position have a low positive correlation with the density and basically no correlation with the ratio. The differences between the years regarding the correlation of the north-south position with both the time of the peak and the start of the high incidence period explains the results of the Friedman tests for these statistics. Apparently, some years there appear to be a positive correlation while other years there seem to be a negative one.

8.5 Spatial smoothing

The spatial smoothing on the parameter θ, representing the time of the peak incidence, is done in two ways. The first (Model 3.1) is described in Section 6.2.2 and assumes that the time of the peak for a given county depends on the time of the peak for the neighbors of that county. The second (Model 3.2), described in Section 6.2.3, assumes that each county is affected by a function of all other counties depending on the relative distance between the county and the other counties. The degree of smoothing, in both models, is represented by a dispersion parameter, σ^2, in the prior distribution for θ. If the dispersion parameter is large the degree of smoothing is small and the estimate of θ is dominated by the data from the county. On the other hand, if the dispersion parameter is small, the variance of the prior distribution will be small and hence the estimate of θ will be dominated by the other counties' θ values.

Model 3.1 is first run with no constraint on σ^2, i.e., the dispersion is estimated from the data. The estimate is $\hat{\sigma}^2 = 0.0015$. It might be difficult to interpret this value; but in order to see the effect of the smoothing, the estimates of θ in this model is plotted against the estimates of θ in the model without smoothing (Model 2, cf. Section 6) in Figure 15 (left).

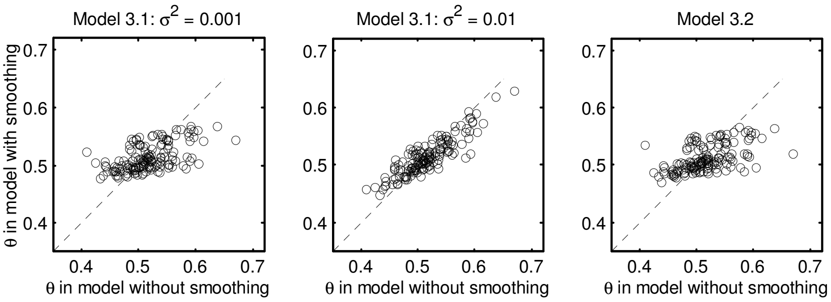


Figure 15: The effect of smoothing on the parameter θ. Estimates from Model 3.1 with $\sigma^2 = 0.001$ (left), Model 3.1 with $\sigma^2 = 0.01$ (middle) and Model 3.2 (right) against the estimates from the model without smoothing (Model 2).

The degree of smoothing appears to be quite large. The estimates from the smoothed model is much less spread than the estimates from the unsmoothed model. In fact, the variance of the estimated θs is reduced by 73 % for Model 3.1 as compared with the unsmoothed Model 2. Also, the largest part of the variance for the smoothed model can be explained by differences between years. Taking that into account the within year variance is reduced by 96 %.

The same model was also run with a fixed value of σ^2. We chose to fix the dispersion to $\sigma^2 = 0.01$, which is about seven times larger than when σ^2 was estimated from the data, in order to show the effect of reducing the degree of smoothing. The result of this can be seen in Figure 15 (middle). The estimates are still less spread than in the unsmoothed model but not as much as for the unconstrained model.

When the same plot is done for the model with weights based on distances between the counties (Model 3.2) (cf. Figure 15 (right)) we end up with almost the same effect of the smoothing as for the model with smoothing according to the neighbors (Model 3.1). In fact, by plotting the estimates from Model 3.2 against the estimates from Model 3.1 we see that they actually are very similar (cf. Figure 16).

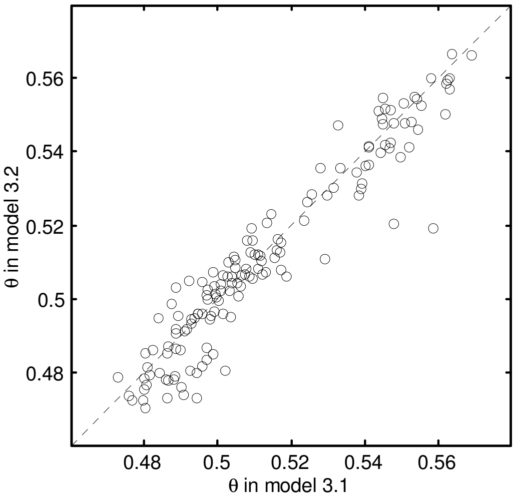


Figure 16: Estimates of θ from Model 3.2 against estimates of θ from Model 3.1. The two outliers are Norrbotten and Västerbotten, both in year 1994, for which the two models smooth the estimated time for peak incidence by a different amount.

Wakefield et al. (2001) states that using neighbors is reasonable if all regions are of similar size and arranged in a regular pattern. Our results suggest that the smoothing for the time of the peak, using neighbors works as well as using the distances between the counties, even though the regions in Sweden neither are arranged in a regular pattern nor are of the same size.

9 Discussion

Spatial and temporal modelling of disease incidence can be of great importance for a better understanding of the aetiology of a disease. Some diseases have a more or less stable seasonal pattern with alternating high and low incidence periods. This is especially true for many of the infectious diseases. Commonly, there is a yearly cycle (e.g., campylobacter infection, salmonella infection or influenza) but other cycle lengths have been observed. One famous example of the latter is the incidence of measles in Great Britain before the start of vaccination, with large outbreaks every fourth year (cf. e.g., Anderson & May (1991)).

Evidently, there is a large amount of randomness in observed data. To be able to study the seasonal patterns and make comparisons, e.g., between different geographical areas or between different periods of time, it is helpful to smooth the data in some way. One approach is to use a non-parametric method such as a moving average or a kernel smoother. For instance, Nylén et al. (2002) used a kernel smoother in their study. Non-parametric methods are very straightforward to use but lack the interpretability and flexibility of parametric models. The advantage of a parametric model is that, at best, one can represent disease specific patterns in the incidence with a small number of parameters.

We have introduced a parametric model accounting for the special aspects of diseases with the kind of seasonal pattern discussed above. The model incorporates as few as four parameters describing the functional form of the incidence curve within each cycle (e.g., year). These four parameters together with the functional form describing the incidence patterns make the model very flexible. There are of course room for improvement of the model to get an even better fit. The cosine part of the model makes the assumption that the high incidence period is symmetric around the peak week. Nevertheless, some of our data on campylobacter infections suggest that there might be a slower decline than incline around the peak incidence in some years for some counties. Therefore, another function taking this into account might result in a better fit. Figure 5 shows how this asymmetry might look. However, when data are aggregated over different areas, as in that case, the asymmetry might as well have been caused by the aggregation of several symmetric incidence curves peaking at different times.

We have also made the assumption of independence between years. In a previous report (Lindbäck & Svensson (2001)) another parameterisation of the model was tried by splitting the parameters into two parts; one that represented the average over the years and another that represented a random effect for each year. On the other hand, such parameterisation would not necessarily lead to a better model. In many situations it is more appropriate to assume complete independence between years.

This study was partially initiated due to an increasing interest in the spread of campylobacter infections. There is a need to understand the seasonal patterns seen in the observed data in order to further investigate risk factors for the disease. It is likely that the spectrum of risk factors is different during different times of the year and possibly in different geographical areas. By understanding the geographical and temporal patterns of the incidence it is maybe possible to improve the design of studies performed to investigate risk factors for sporadic cases of campylobacter infections.

The model fitted the data well. The yearly pattern of campylobacter infections is rather stable, even for areas with few cases and in spite of a possibly substantial underreporting. To make sure that the model was flexible enough to detect a peak during a period far from the expected, i.e., in late summer, we moved some of the data from late summer to early spring. The model succeeded in finding this early peak. This indicates that the model is not too constrained and that even when data are sparse there is a stable structure regarding the incidence of campylobacter infections with a peak in late summer.

Modelling disease incidence data often assumes rather simple models either based on the Poisson (aggregated data) or the binomial distribution (individual data). Often these distributional assumptions are valid or at least approximately so. However, in some situations the data tend to be overdispersed relative the assumed distribution leading to underestimation of the variance. This is perhaps even more common for infectious diseases where clustering of cases might occur easily by secondary infections. When we looked at the aggregated data for Sweden there were convincing evidence of overdispersion when we fitted the model assuming Poisson variation. To overcome the problem of overdispersion we reformulated the model to include a dispersion parameter, assuming Negative binomial variation. There appeared to be no substantial bias in parameter estimates or estimated incidence when using the Poisson model instead of the Negative binomial model but the variance was clearly underestimated. These results stress the importance to check the model assumptions regarding overdispersion in the modelling process of these kinds of data.

We have established a model to describe the incidence of campylobacter infections but no attempt have yet been made to link the different properties of the model to external factors. It is rather straightforward to incorporate other information in the model. County wise information, such as the proportion of inhabitants living in rural areas or information about the water supply, can easily be added as a county factor. Time-specific information such as the yearly (or monthly/weekly) incidence of campylobacter infections acquired abroad can also easily be added. It would also be interesting to add information about temperature and precipitation to see to which extent the climate influences the spread of *Campylobacter*. Climate data could thus be entered as a factor varying both in time and space.

The number of inhabitants differ between the counties in Sweden. As a consequence of this, variances also differs between counties. It is likely that extreme observations or extreme parameter estimates will occur in smaller areas (i.e., areas with fewer inhabitants) purely due to chance. It is possible to smooth the estimates by assuming spatial dependencies. We studied the effect of spatial smoothing on the parameter describing the time of the peak incidence. When we let the degree of smoothing be estimated from the data, the variance between the counties' estimates was substantially reduced. The results implied that all estimates were more or less smoothed towards a global mean, although the model assumed local smoothing. We tried two different kinds of spatial dependence when we performed the smoothing. The first assumed that each county's parameter value was smoothed towards the mean of the neighboring counties' parameter values. Two counties were assumed to be neighbors if they shared a common border. The second assumed smoothing towards a weighted mean of all other counties' parameter values where the weights were defined by the distance to each respective county. Both these methods gave basically the same results. Dividing Sweden by counties results in relatively few sub-areas (only 21). With too few areas it is difficult to achieve a genuine local smoothing. Smoothing one county will have effects on a relatively large proportion of the other counties as compared with a situation with more areas. This might lead to a situation where all counties are having a relatively large effect on each other and consequently all counties are smoothed towards the same value.

Estimations of the parameters in the models were done within a Bayesian setting using Markov chain Monte Carlo simulation. In order to carry out the simulations, prior distributions had to be specified for all parameters. The objective was to make the prior distributions as uninformative as possible to have the parameter estimates dominated by the data. Due to the complexity of the model, making the prior distributions too vague resulted in problems in the updating process. The somewhat informative priors, for the parameters τ and κ, is thus a weakness in this study. Nevertheless, by looking at the fit of the models and by trying different priors and starting values in the simulation process we believe that the estimates are not too much affected by the priors and that the conclusions drawn should be valid.

We have, in this report, only applied the models described on campylobacter infection data. Moreover, Model 1, the Poisson model, has also been fitted to weekly mortality data to evaluate excess mortality in relation to influenza epidemics in Sweden (cf. Figure 17).

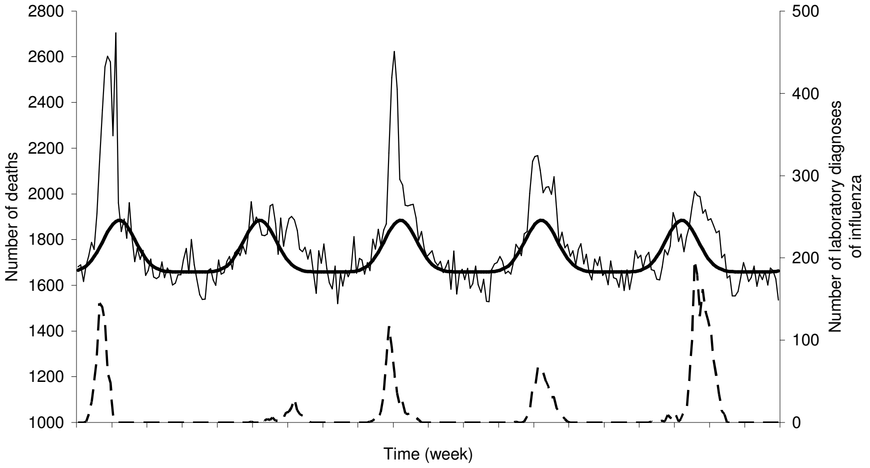


Figure 17: Excess mortality due to influenza in Sweden week 40 1993 to week 39 1998. Solid thin line represents the number of deaths per week. Bold line represents the estimated weekly mean number of deaths during weeks with no influenza diagnoses. Dashed line represents the weekly number of laboratory diagnoses of influenza.

On the side of that study, an informal comparison was done between our model and another model used to describe the variation in mortality in Scotland (cf. Gemmel et al. (2000)). By including the parameter κ, allowing for both shorter and longer peaks, our model seemed to give a better fit to the data. However, as stated above, no formal comparison was done.

A conclusion from this study is that it is indeed possible to derive interesting information from the reported data with its known shortcomings. Regarding the data on campylobacter infections it is believed that there is possibly a severe underreporting. In spite of this, there is still a very strong structure in the data. The model has also successfully been fitted to Influenza data but it should be possible to fit the model to data regarding other diseases with a similar structure of the incidence. Of course, beware of applying the model to another disease without thinking of the special characteristics of that disease.

10 Acknowledgement

My work in this project was done in part time during the last four and a half years. During this time I have had the opportunity to work with many interesting and intelligent people of whom several have influenced this work in one way or another. I would especially like to thank all my former and current colleagues at SMI, Statisticon AB, UCR and the Mathematical statistics department at the Stockholm university for interesting discussions regarding statistical theory and applications.

Some persons have had a deeper influence on this project and I would here like to take the opportunity to acknowledge them:

Yvonne Andersson and Birgitta de Jong at SMI for sharing their great knowledge about the epidemiology of campylobacter infections. Professor Johan Giesecke at SMI for initiating this project from the SMI side and for allowing me to maintain the contact with the university while working at SMI. Finally I would like to thank my supervisor Professor Åke Svensson for inviting me to work on this project and for all the interesting discussions during it. I surely have learned a lot from working on this project. I guess that was one of its purposes.

This work was partly funded by The Bank of Sweden Tercentenary Foundation.

References

Anderson, R. & May, R. (1991), *Infectious diseases of humans, dynamics and control.*, Oxford Science Publ, New York, p. 85.

Andersson, Y., Bast, S., Gustavsson, O., Jonsson, S. & Nilsson, T. (1994), 'Campylobacterutbrott [A Campylobacter outbreak]', *Epid-aktuellt* **17**, 6–9. (In Swedish).

Andersson, Y. & Gustavsson, O. (1998), Campylobacter – en vattenburen smitta att beakta vid planläggning av vattentäkt av ytvattentyp, Research report, Överstyrelsen för civil beredskap, Stockholm. (In Swedish).

Andersson, Y., Gustavsson, P., Hammarquist, C., Kulander, L., Nilsson, P. E. W. & Olsson, A.-M. (1998), Vattenkvalité och hälsa, en studie av självrapportering vid mag- och tarmsjukdom [Drinking water quality and health. A study of self-reporting of gastrointestinal illness], Research report, Godrings Tryckeri AB, Visby. (In Swedish).

Bresky, B., Studahl, A. C., Säll, C. & Luther, M. L. (1995), 'Stort utbrott av magsjuka i Marks kommun [A large outbreak of gastrointestinal illness in the municipality of Mark]', *Epid-aktuellt* **18**, 6–8. (In Swedish).

Chin, J., ed. (2000), *Control of communicable diseases manual*, 17th edn, American Public Health Association, Washington DC.

Gemmel, I., McLoone, P., Boddy, F. A., Dickinson, G. J. & Watt, G. C. (2000), 'Seasonal variation in mortality in Scotland', *International journal of epidemiology* **29**(2), 274–279.

Giesecke, J. (2002), *Modern infectious disease epidemiology*, 2nd edn, Arnold, a member of the Hodder Headline Group, London.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996), *Markov chain Monte Carlo in practice*, Chapman and Hall, London.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.

Hinde, J. & Demétrio, C. G. B. (1998), 'Overdispersion: Models and estimation', *Computational Statistics and Data Analysis* **27**, 151–170.

Jormanainen, V., Lindbäck, J. & Giesecke, J. (1997), 'Hur länge är kliniska anmälningar på väg till Smittskyddsinstitutet? [Delay of clinical notifications]', *Smittskydd* **18**, 6–8.

Kapperud, G. (1995), Descriptive and analytic epidemiology of Campylobacter enteritis in humans. WHO consulting on Emerging Foodborne Diseases, Berlin, Germany, 20-24 March 1995., Research report, Federal Institute for Health Protection of Consumers and Veterinary Medicine, World Health Organization.

Lighton, L. L., Kaczmarski, E. B. & Jones, D. M. (1991), ‘A study of risk factors for campylobacter infections in late spring’, *Public Health* **105**, 199–203.

Lindbäck, J. & Svensson, Å. (2001), Campylobacter infections in Sweden – A statistical analysis of temporal and spatial distributions of notified sporadic campylobacter infections, Research report 2001:4, Mathematical statistics, Stockholm university, Stockholm, Sweden. Available from URL (May 15, 2003): http://www.math.su.se/matstat/reports/seriea/.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), ‘Equations of state calculations by fast computing machine’, *Journal of Chemical Physics* **21**, 1087–1091.

Mollié, A. (1996), Bayesian mapping of disease, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter, eds, ‘Markov chain Monte Carlo in practice’, first edn, Chapman and Hall, London, section 20, pp. 359–379.

Neal, R. M. (1997), Markov chain Monte Carlo methods based on ’slicing’ the density function, Research report 9722, Departement of Statistics, University of Toronto, Ontario, Canada. Available from URL (May 15, 2003): http://www.cs.utoronto.ca/∼radford/papers-online.html.

Nothermans, S. (1995), Epidemiological aspects of thermophilic Campylobacter in water related environments. WHO consulting on Emerging Foodborne Diseases, Berlin, Germany, 20-24 March 1995., Research report, Federal Institute for Health Protection of Consumers and Veterinary Medicine, World Health Organization.

Nylén, G., Dunstan, F., Palmer, S. R., Andersson, Y., Bager, F., Cowden, J., Feierl, G., Galloway, Y., Kapperud, G., Megraud, F., Molbak, K., Petersen, L. R. & Ruutu, P. (2002), ‘The seasonal distribution of campylobacter infection in nine European countries and New Zealand’, *Epidemiology and Infection* **128**, 383–390.

Smittskyddsinstitutet (1995), Smittsamma sjukdomar i Sverige 1994, Epidemiologiska enhetens årsrapport. [Infectious diseases in Sweden 1994, yearly report of the dept. of Epidemiology], Research report, Smittskyddsinstitutet, Stockholm. (In Swedish).

Smittskyddsinstitutet (1996), Smittsamma sjukdomar i Sverige 1995, Epidemiologiska enhetens årsrapport. [Infectious diseases in Sweden 1995, yearly report of the dept. of Epidemiology], Research report, Smittskyddsinstitutet, Stockholm. (In Swedish).

Smittskyddsinstitutet (1997), Smittsamma sjukdomar i Sverige 1996, Epidemiologiska enhetens årsrapport. [Infectious diseases in Sweden 1996, yearly report of the dept. of Epidemiology], Research report, Smittskyddsinstitutet, Stockholm. (In Swedish).

Smittskyddsinstitutet (1999), Smittsamma sjukdomar i Sverige 1998, Epidemiologiska enhetens årsrapport. [Infectious diseases in Sweden 1998, yearly report of the dept. of Epidemiology], Research report, Smittskyddsinstitutet, Stockholm. (In Swedish). Available from URL (May 15, 2003): http://www.smittskyddsinstitutet.se/download/pdf/rapp98.pdf.

Spiegelhalter, D. J., Thomas, A. & Best, N. G. (2000), *WinBUGS Version 1.3 User Manual*, MRC Biostatistics Unit, Cambridge.

Tauxe, R. V. (1992), Epidemiology of Campylobacter jejuni infections in the United States and other industrialized nations, *in* I. Nachamkin, M. J. Blaser & L. S. Tompkins, eds, 'Campylobacter jejuni. Current status and future trends', American Society for Microbiology, Washington DC.

Wakefield, J. C., Best, N. G. & Waller, L. (2001), Bayesian approaches to disease mapping, *in* P. Elliott, J. C. Wakefield, N. Best & D. Briggs, eds, 'Spatial Epidemiology', first edn, Oxford University Press, Oxford, section 7, pp. 104–127.

Wheeler, J. G., Sethi, D., Cowden, J. M., Wall, P. G., Rodrigues, L. C., Tompkins, D. S., Hudson, M. J. & Roderick, P. J. (1999), 'Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance', *BMJ* **318**, 1046–1050.

A Interesting parameters

Table 8: Peak week by county

	County	1992	1993	1994	1995	1996	1997	1998
1	Stockholm	29	34	32	35	34	34	32
2	Uppsala	31	36	31	30	36	34	32
3	Södermanland	32	31	28	32	33	27	36
4	Östergötland	33	33	33	32	32	36	28
5	Jönköping	32	30	30	35	34	31	31
6	Kronoberg	34	33	35	36	32	33	32
7	Kalmar	35	31	31	31	31	28	30
8	Gotland	33	28	30	30	30	33	31
9	Blekinge	29	33	31	33	27	34	33
10	Skåne	36	29	30	32	32	32	33
11	Halland	29	32	36	39	31	32	32
12	Västra Götaland	35	31	31	34	36	32	31
13	Värmland	29	32	33	36	33	33	31
14	Örebro	32	28	30	35	30	31	34
15	Västmanland	28	29	30	34	33	31	28
16	Dalarna	32	33	33	31	37	28	32
17	Gävleborg	30	30	31	31	31	32	31
18	Västernorrland	30	27	32	34	34	31	30
19	Jämtland	31	29	31	30	34	31	31
20	Västerbotten	30	31	40	34	35	31	31
21	Norrbotten	30	32	34	32	36	32	32

Table 9: Ratio between high and low incidence by county

	County	1992	1993	1994	1995	1996	1997	1998
1	Stockholm	6.0	3.3	3.8	6.6	9.2	11.2	11.4
2	Uppsala	3.6	5.0	9.1	10.6	6.2	6.9	7.4
3	Södermanland	10.3	11.3	6.4	9.1	10.9	9.3	14.2
4	Östergötland	2.8	8.5	6.2	3.7	8.8	4.4	5.1
5	Jönköping	3.6	6.4	3.4	5.4	7.0	8.8	6.7
6	Kronoberg	5.0	7.1	4.9	5.1	16.7	12.1	14.1
7	Kalmar	3.8	3.5	4.7	5.7	6.3	6.7	6.5
8	Gotland	6.0	13.6	4.4	8.5	5.4	7.6	9.1
9	Blekinge	10.5	6.5	3.3	12.4	8.2	8.3	15.9
10	Skåne	3.8	4.4	4.6	2.6	5.2	9.4	5.9
11	Halland	5.0	3.3	5.0	9.8	9.7	8.3	6.7
12	Västra Götaland	3.1	4.6	3.4	4.3	10.4	12.4	9.2
13	Värmland	12.3	4.7	6.1	4.0	6.1	14.6	24.3
14	Örebro	6.3	9.3	4.3	4.8	6.8	5.6	15.1
15	Västmanland	5.2	5.5	7.6	7.0	14.2	19.4	21.2
16	Dalarna	5.5	6.1	6.9	7.3	22.4	12.7	11.2
17	Gävleborg	10.8	3.9	5.9	8.7	18.6	10.0	10.6
18	Västernorrland	14.6	6.6	12.1	8.7	13.4	9.4	12.1
19	Jämtland	8.1	10.8	20.4	10.4	11.8	3.9	14.3
20	Västerbotten	5.7	7.7	6.9	10.9	5.9	6.4	12.9
21	Norrbotten	9.5	13.3	8.6	7.8	8.5	10.2	25.4

Table 10: Start of high incidence period by county

	County	1992	1993	1994	1995	1996	1997	1998
1	Stockholm	19	20	22	21	15	16	12
2	Uppsala	19	18	5	5	20	20	18
3	Södermanland	15	15	14	12	14	16	20
4	Östergötland	18	15	17	17	20	19	19
5	Jönköping	19	18	16	23	18	16	18
6	Kronoberg	21	17	21	13	15	15	13
7	Kalmar	21	6	20	17	15	20	20
8	Gotland	16	9	21	19	23	19	16
9	Blekinge	18	8	22	13	18	22	14
10	Skåne	21	18	19	21	19	18	21
11	Halland	14	24	24	18	11	18	20
12	Västra Götaland	25	21	20	22	16	15	16
13	Värmland	15	20	25	27	19	11	16
14	Örebro	12	15	15	22	18	17	16
15	Västmanland	17	21	19	18	16	13	16
16	Dalarna	16	19	20	14	14	10	20
17	Gävleborg	13	22	19	12	14	23	21
18	Västernorrland	16	18	20	22	23	19	22
19	Jämtland	19	20	13	8	23	18	23
20	Västerbotten	15	23	31	18	23	19	22
21	Norrbotten	18	22	17	14	25	17	17

Table 11: Proportion cases within +/- 2 weeks of the peak week during each year by county

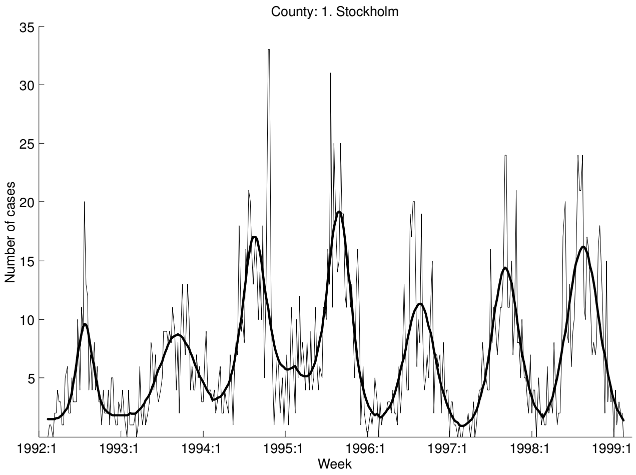
	County	1992	1993	1994	1995	1996	1997	1998
1	Stockholm	24	15	19	20	19	21	19
2	Uppsala	16	15	12	15	19	21	22
3	Södermanland	22	23	20	18	19	24	23
4	Östergötland	13	19	18	15	26	16	22
5	Jönköping	16	23	15	20	19	22	21
6	Kronoberg	18	19	18	13	24	22	21
7	Kalmar	16	11	20	20	19	28	26
8	Gotland	18	21	21	27	26	21	22
9	Blekinge	28	13	17	20	28	25	22
10	Skåne	15	19	20	14	19	24	22
11	Halland	17	18	19	17	18	23	22
12	Västra Götaland	16	20	17	18	18	23	23
13	Värmland	26	19	27	19	20	19	32
14	Örebro	16	24	16	18	23	19	23
15	Västmanland	21	25	26	19	23	25	34
16	Dalarna	18	20	22	19	18	21	28
17	Gävleborg	22	20	21	18	26	31	32
18	Västernorrland	29	24	31	26	31	26	37
19	Jämtland	25	33	26	17	30	16	39
20	Västerbotten	18	28	25	23	21	22	35
21	Norrbotten	25	35	20	19	26	24	31

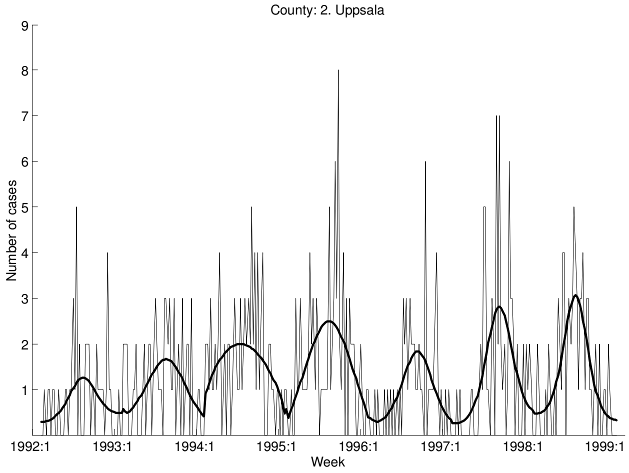
B Weight matrix for Model 3.1

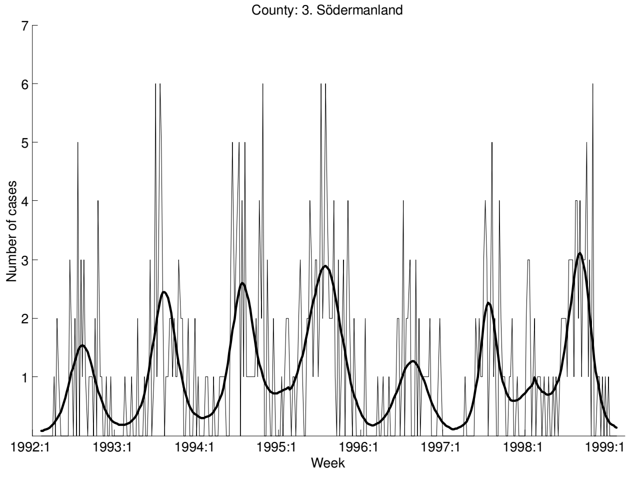
Table 12: Weight matrix for Model 3.1.

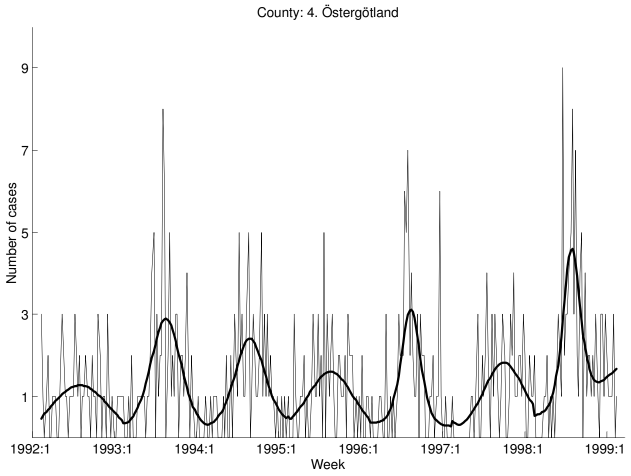
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Stockholm	-	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Uppsala	1	-	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
3	Södermanland	1	1	-	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0
4	Östergötland	0	0	1	-	1	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0
5	Jönköping	0	0	0	1	-	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0
6	Kronoberg	0	0	0	0	1	-	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0
7	Kalmar	0	0	0	1	1	1	-	1	1	0	0	0	0	0	0	0	0	0	0	0	0
8	Gotland	1	0	1	1	0	0	1	-	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Blekinge	0	0	0	0	0	1	1	0	-	1	0	0	0	0	0	0	0	0	0	0	0
10	Skåne	0	0	0	0	0	1	0	0	1	-	1	0	0	0	0	0	0	0	0	0	0
11	Halland	0	0	0	0	1	1	0	0	0	1	-	1	0	0	0	0	0	0	0	0	0
12	Västra Götaland	0	0	0	1	1	0	0	0	0	0	1	-	1	1	0	0	0	0	0	0	0
13	Värmland	0	0	0	0	0	0	0	0	0	0	0	1	-	1	0	1	0	0	0	0	0
14	Örebro	0	0	1	1	0	0	0	0	0	0	0	1	1	-	1	1	0	0	0	0	0
15	Västmanland	0	1	1	0	0	0	0	0	0	0	0	0	0	1	-	1	1	0	0	0	0
16	Dalarna	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	-	1	0	1	0	0
17	Gävleborg	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	-	1	1	0	0
18	Västernorrland	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-	1	1	0
19	Jämtland	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	-	1	0
20	Västerbotten	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	-	1
21	Norrbotten	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-

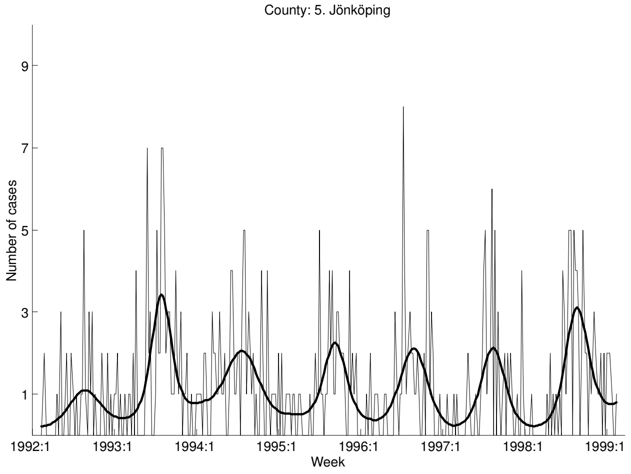
C Model 2 fitted to data – Individual counties

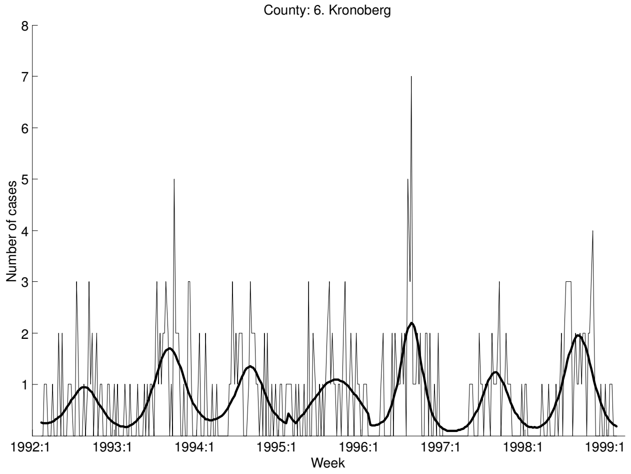


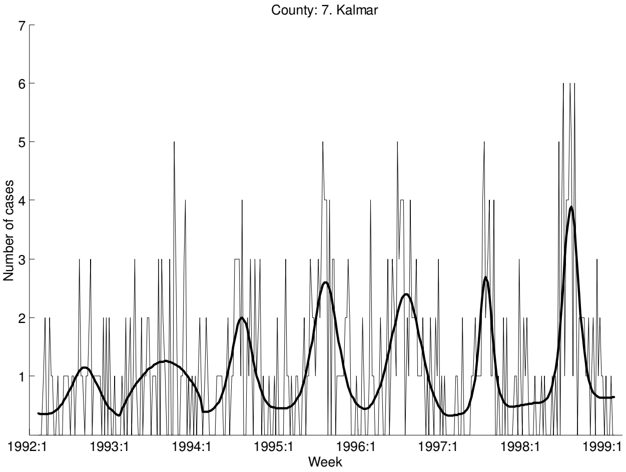


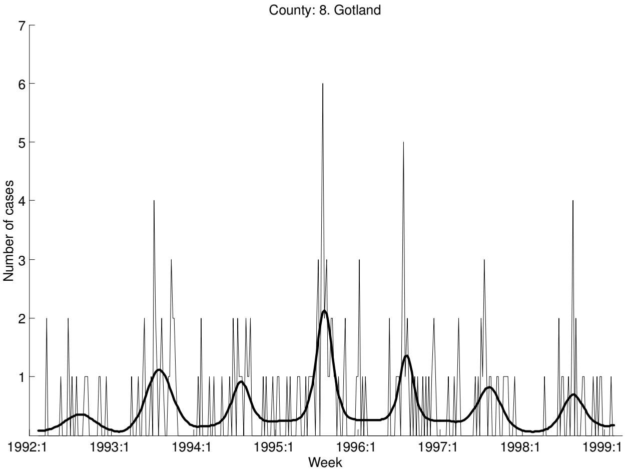


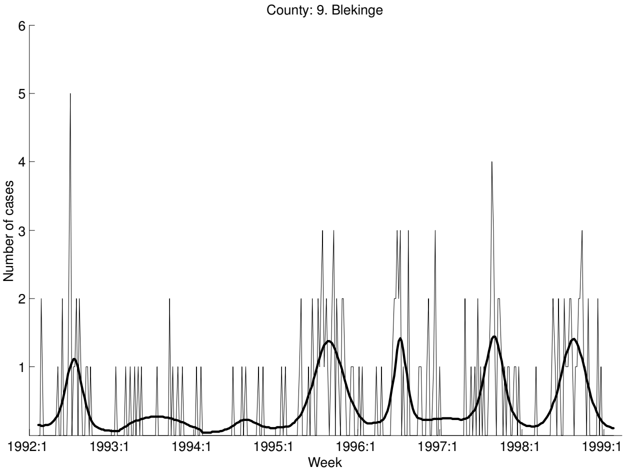

County: 4. Östergötland

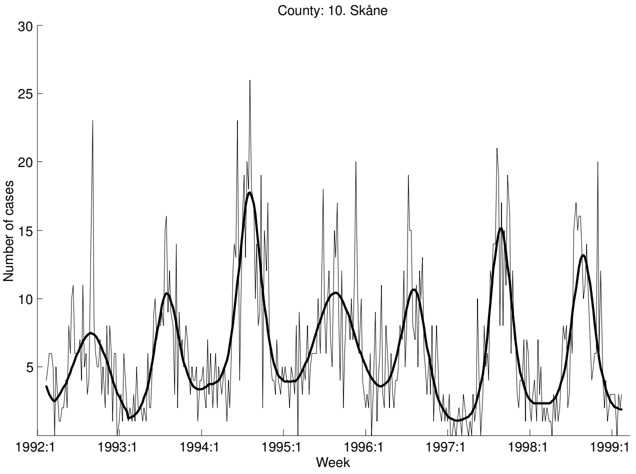

County: 5. Jönköping

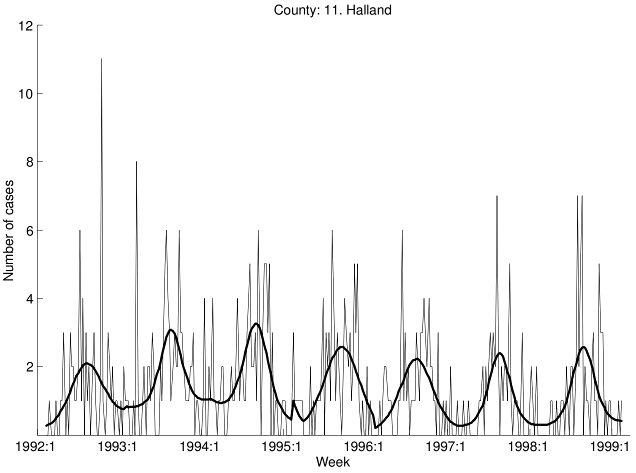

County: 6. Kronoberg

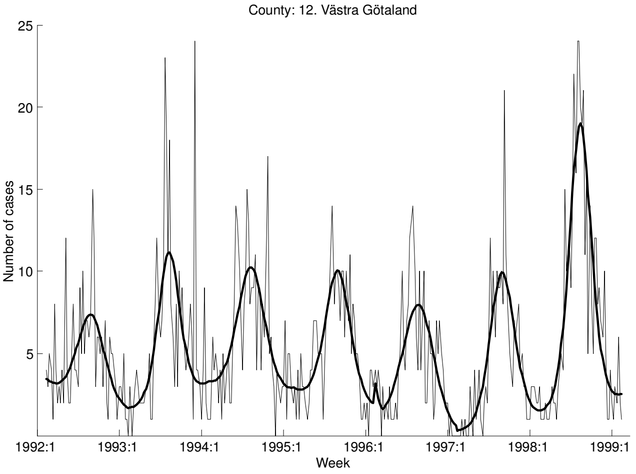

County: 7. Kalmar
Number of cases
Week
7
6
5
4
3
2
1
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1

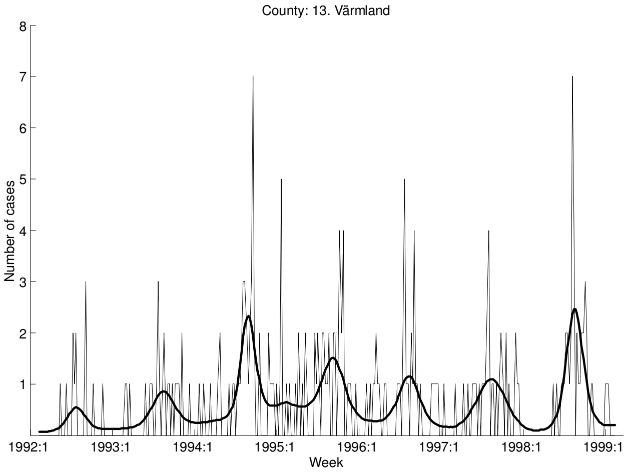

County: 8. Gotland
Number of cases
Week
7
6
5
4
3
2
1
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1

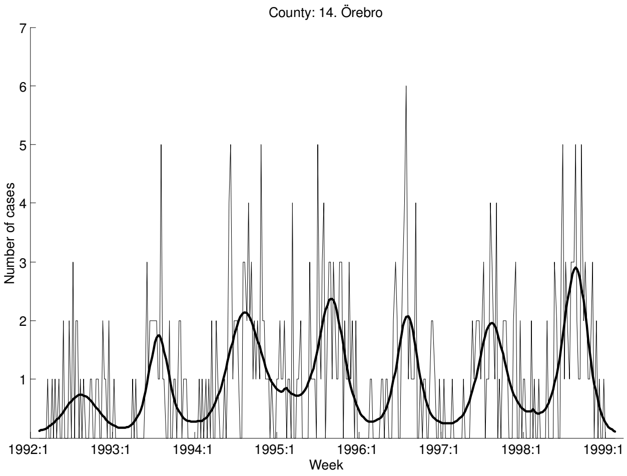

County: 9. Blekinge
Number of cases
Week
6
5
4
3
2
1
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1

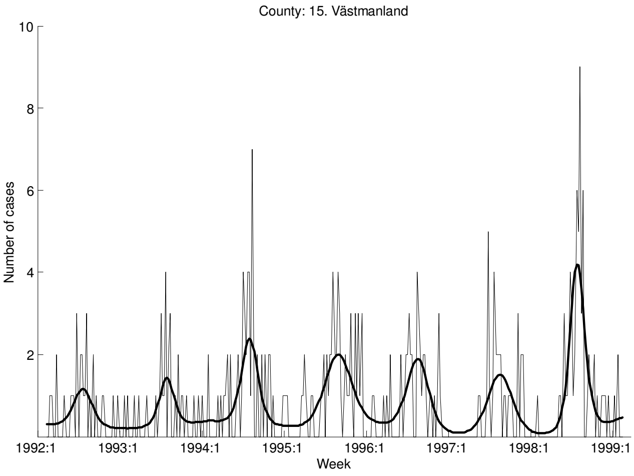

County: 10. Skåne
30
25
20
15
10
5
Number of cases
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1
Week

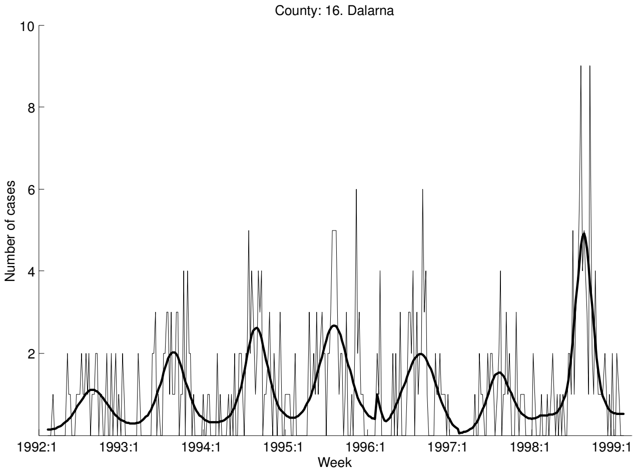

County: 11. Halland
12
10
8
6
4
2
Number of cases
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1
Week

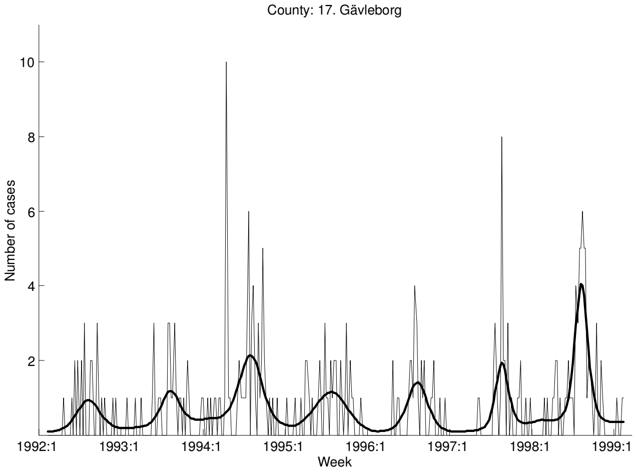

County: 12. Västra Götaland
25
20
15
10
5
Number of cases
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1
Week

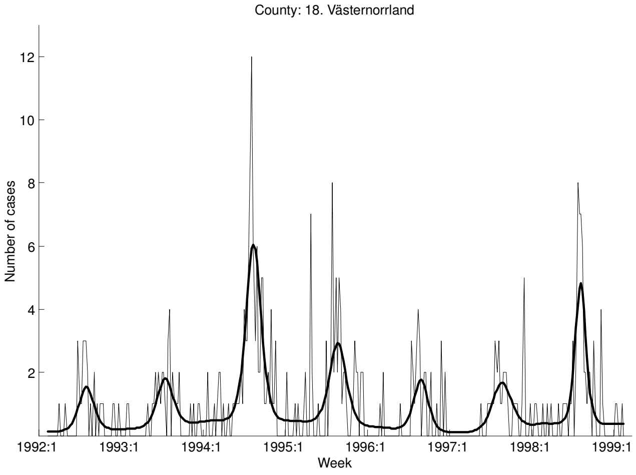

County: 13. Värmland
Number of cases
8
7
6
5
4
3
2
1
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1
Week

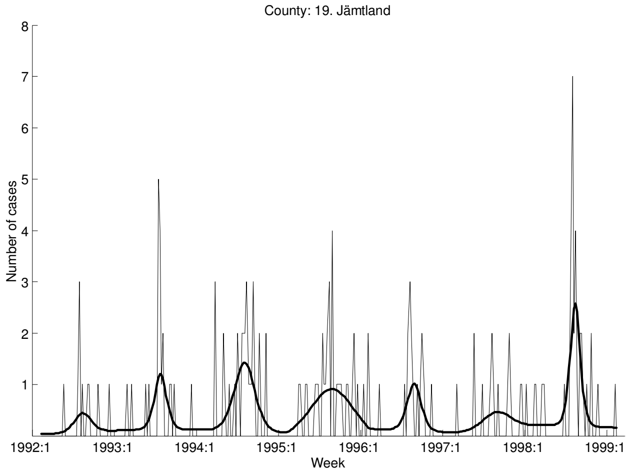

County: 14. Örebro
Number of cases
7
6
5
4
3
2
1
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1
Week

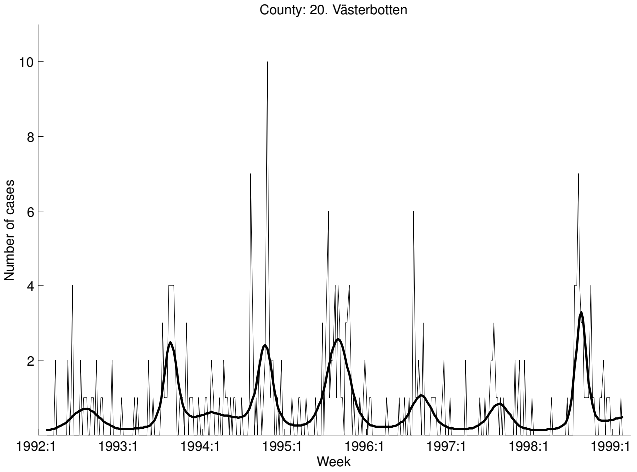

County: 15. Västmanland
Number of cases
10
8
6
4
2
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1
Week

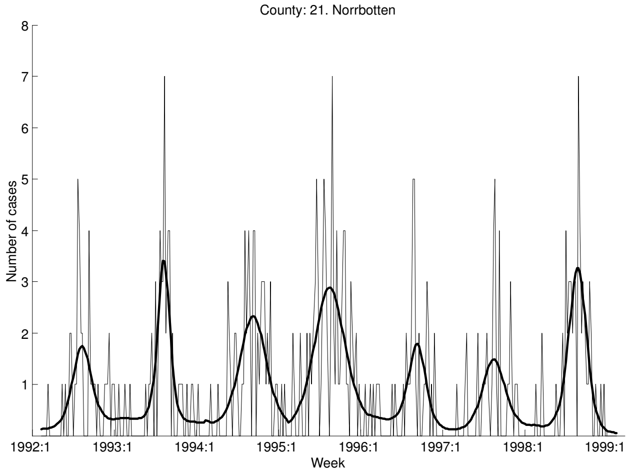

County: 16. Dalarna
Number of cases
10
8
6
4
2
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1
Week


County: 17. Gävleborg
Number of cases
10
8
6
4
2
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1
Week


County: 18. Västernorrland
Number of cases
12
10
8
6
4
2
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1
Week


County: 19. Jämtland
Number of cases
Week
8
7
6
5
4
3
2
1
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1


County: 20. Västerbotten
Number of cases
Week
10
8
6
4
2
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1


County: 21. Norrbotten
Number of cases
Week
8
7
6
5
4
3
2
1
1992:1
1993:1
1994:1
1995:1
1996:1
1997:1
1998:1
1999:1