Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Nan Amalie Videng

Joint spatio-temporal modelling of brain cancer incidence and mortality in Norway

Master's thesis in MLREAL
Supervisor: Andrea Riebler
June 2020

NTNU
Norwegian University of Science and Technology

Nan Amalie Videng

Joint spatio-temporal modelling of brain cancer incidence and mortality in Norway

Master's thesis in MLREAL
Supervisor: Andrea Riebler
June 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Preface

This thesis is the assignment of the 30-credit point course MA3950 - Mathematics Master Thesis, spring 2020. It brings my five year period at the Norwegian University of Science and Technology to an end. The thesis is the finishing work of my degree at natural science with teacher education (MLREAL). Further, it wraps up my field of specialization within statistics, giving me a complete theoretical background. This is relevant for my profession as a natural science teacher, as there is a need for a deeper understanding of subjects taught in schools.

I would like to thank my supervisor, Andrea Riebler, for her insight and guidance in the work on my thesis. She has been very patient and eager to help me with everything I have needed help with. I would also thank her for introducing me to this field, which I have found extremely educational and interesting to work with.
Thank you!
Nan Amalie Videng, June 2020

Abstract

In this thesis, we jointly model brain cancer incidence and mortality in Norway over the last 50 years, in order to discover underlying geographical and temporal patterns in the disease. The incidence and mortality counts for brain cancer can be very scarce for certain age groups or time periods. The survival rate of brain cancer in Norway is quite low. Hence, it is assumed that incidence and mortality are correlated and therefore analysed jointly in order to borrow strength between the two disease processes and increase the effective sample size. This enables the inclusion of both gender-specific and age-specific components in addition to a shared spatial random effect scaled with an outcome-specific parameter. This is useful since the data is provided for both genders in the time period 1969 to 2018, subdivided into 18 counties, 9 age groups and 10 time periods. Spatial and spatio-temporal modelling in this thesis is done using a Bayesian approach, where the inference is based on the integrated nested Laplace approximations (INLA) methodology. We have used a hierarchical structure, which takes into account the spatial dependency between neighbouring counties. The analyses were separated into a sole spatial modelling of the most recent period 2014–2018 and a spatio-temporal modelling of the entire period 1969–2018.

The age-specific component in the model was especially interesting to include, due to brain cancer being the second most common cancer in small children, after leukaemia. The results for the spatial modelling in 2014–2018 showed a change in the age effect for both incidence and mortality for the youngest children. This change was most noticeable for the mortality. These results show that the age effect is not homogeneous for all age groups. Hence, the age effect is necessary to include. For the spatio-temporal modelling in the entire period 1969–2018, the shared spatial effect showed a noticeable geographical pattern with a general increasing trend from north to south. There was clearly a lower spatial effect in the north, especially in Finnmark, and increasing to the southern counties Vest-Agder and Aust-Agder. These spatial results were not as apparent in the modelling from the recent period. For the entire period 1969–2018 the spatial effect for incidence and mortality were almost identical for both genders, with the spatial effect for women mortality being slightly larger for mortality than for incidence . However, for the single period 2014–2018, the spatial effect for women was clearly stronger for incidence than mortality. The temporal effect was higher in 2018 than it was in 1969. However, the last couple of years show an interesting decrease for the incidence.

Sammendrag

I denne oppgaven modelleres både insidens og dødelighet for hjernekreft i Norge over de siste 50 årene. Dette er gjort for å prøve å oppdage underliggende geografiske- og tidsmessige trender hos sykdommen. Insidens- og dødelighetstall for hjernekreft kan være svært små for noen aldersgrupper eller tidsperioder. I tillegg er overlevelsesraten for hjernekreft ganske lav. På grunn av dette er det antatt at det er korrelasjon mellom insidens og dødelighet, noe som gjør at man kan låne informasjon mellom de to sykdomsprosessene for å kunne øke den effektive utvalgsstørrelsen. Dette gir mulighet for å inkludere både kjønns- og aldersspesifikke komponenter i tillegg til den romlige effekten som er skalert ved hjelp av en utfallsspesifikk parameter. Dette er gunstig ettersom dataene er gitt for begge kjønn i perioden 1969 til 2018, og inndelt i 18 fylker, 9 aldersgrupper og 10 tidsperioder. Både den romlige modelleringen og rom-tid-modelleringen er gjennomført ved en Bayesiansk tilnærming, der inferensen utføres ved hjelp av "integrated nested Laplace approximations" (INLA). Videre er det brukt en hierarkisk struktur, som gir en mulighet til å ta hensyn til romlig avhengighet mellom nabofylker. Analysene er delt inn i romlig modellering av den nyligste tidsperioden, 2014–2018, og rom-tid-modellering av perioden 1969–2018.

Det er spesielt interessant å inkludere den aldersspesifikke komponenten i modellene, siden hjernekreft er den nest vanligste krefttypen for små barn, etter leukemi. Resultatene for den romlige modelleringen av perioden 2014–2018 viste en endring i alderseffekten for både insidens og dødelighet for aldersgruppene 0–9 og 10–19. Denne endringen var mest tydelig for dødelighet. Resultatene viser at alderseffekten ikke er homogen for alle aldersgrupper, som nødvendiggjør inkluderingen av denne effekten. Den romlige effekten for rom-tid-modelleringen for perioden 1969–2018 viser et tydelig geografisk mønster med en generelt økende trend fra nord til sør. Det er en tydelig lavere romlig effekt i det nordligste fylket Finnmark, som var økende til de sørligste fylkene Vest- og Aust-Agder. Disse romlige resultatene var derimot ikke like merkbare i modelleringen for den nyligste perioden. For perioden 1969–2018 så man at den romlige effekten for insidens og dødelighet var omtrent identiske for begge kjønn, mens for perioden 2014–2018 var den romlige effekten for kvinner tydelig sterkere for insidens enn for dødelighet. Videre ser en at tidseffekten er høyere i 2018 enn den var i 1969. Likevel ser en en interessant nedgang for insidens de siste årene.

Contents

1 Introduction **1**

2 Brain Cancer in Norway **3**
2.1 The most recent period: 2014–2018 5
2.2 Historic data from 1969–2018 8

3 Introduction to Spatial Statistics **11**
3.1 Bayesian inference 12
3.1.1 Integrated nested Laplace approximations (INLA) 13
3.1.2 Penalized complexity (PC) priors 14
3.2 Spatial modelling 15

4 Literature Review **17**
4.1 Etxeberria et al. (2018): Joint modelling of brain cancer incidence and mortality 17
4.2 Colonna et al. (1999): Cancer incidence prediction based on mortality . . 19
4.3 Held et al. (2005): Joint analysis of diseases with shared risk factors . . . 20
4.4 Martinez-Beneito (2013): A more general approach to multivariate disease mapping 21

5 Spatial modelling of brain cancer in Norway in 2014–2018 **23**
5.1 Applying the models from Etxeberria et al. (2018) 23
5.1.1 Modelling with a structured spatial component and age group . . 24
5.1.2 Extending by an unstructured spatial component 28
5.2 Introducing possible changes to the original models 30
5.2.1 Prior choices 30
5.2.2 Adding an unstructured component for overdispersion 31
5.2.3 Changing the age effect model 32
5.3 Using model choice criteria to choose the best models 34
5.3.1 Model choice criteria 34
5.3.2 Comparing models with structured spatial component and age group 36

5.3.3 Comparing the models with the additional unstructured spatial effect 37
5.4 Results 38

6 Spatio-temporal modelling of brain cancer in Norway from 1969–2018 43
6.1 Adding the temporal component and space-time interaction 43
6.2 Results from the spatio-temporal model 45

7 Discussion and Summary 51

Bibliography 55

Appendix 59
7.1 Challenges along the way 59
7.1.1 Challenges due to the COVID-19 pandemic situation 59
7.1.2 Discovering errors in the software 59
7.2 Learning experiences 61

Chapter 1

Introduction

National cancer registries in industrial countries possess almost complete information about the frequency of new cases, the incidence, and mortality stratified by for example gender, age and certain subnational geographical areas. The Cancer Registry of Norway (https://www.kreftregisteret.no/) is one of these nearly complete registries. In this thesis, we analyse Norwegian brain cancer data provided as gender-specific incident cases and deaths in 18 counties (fylker) for 9 age groups over a 50 year period. The objective of the thesis is divided into two main parts. The first objective is to analyse data on brain cancer incidence and mortality in Norway in the period 2014–2018, in order to understand more about its geographical distribution. The second objective is to analyse the brain cancer incidence and mortality data in Norway in the entire fifty year period 1969–2018 to learn more about the spatio-temporal distribution of the disease.

For brain cancer, there are few known causes for the disease. The main contributors to brain cancer are either genetics or high exposure to ionized radiation (Savage; 2018). The genetic risk factor concerns several gene changes that cause rare inherited syndromes. These conditions promote tumour formation, which includes for instance neurofibromatosis and Li-Fraumeni syndrome. People with these gene changes have an increased risk of brain cancer. However, these conditions only cause about 5% of tumours (Savage; 2018; American Cancer Society; 2020). Further, Savage (2018) points out that brain cancer occur most commonly in white people, and has the highest incidence in the northern Europe.

According to the latest incidence and mortality data for brain and other nervous system cancer from GLOBOCAN, there were diagnosed 162 534 new cases in males and 134 317 new cases in females worldwide in 2018 (Ferlay et al.; 2018). This represents a rate of 3.9 cases per 100 000 for men and a rate of 3.1 cases per 100 000 for women, by World standard. In Norway 440 and 445 new cases in males and females, respectively, were diagnosed in 2018 (Cancer Registry of Norway; 2019). This represents an incidence rate of 6.4 cases per 100 000 in males and 4.5 cases per 100 000 in females, by World standard (explained in chap. 2)(Ferlay et al.; 2018). In Europe there were 35 276 new cases of brain cancer for males and 29 363 new cases for females in 2018, which corresponds to rates of 6.7 cases per 100 000 in males and 4.7 cases per 100 000 in females, by World

standard (Ferlay et al.; 2018). According to Storstein et al. (2011) the Norwegian brain cancer incidence is increasing and the number of new cases per year has almost doubled since 1980, which he sees as most likely due to an increasing number of elders in the population along with better diagnostics.

The main inspiration for this thesis is an article by Etxeberria et al. (2018) which uses integrated nested Laplace approximations (INLA) to do Bayesian inference for spatial modelling of brain cancer data from two northern regions of Spain. The methodology presented in this article will be used to analyse the Norwegian brain cancer data. This includes two main models with corresponding extensions. The first model has two components, one structured spatial component and one age component. The second model includes the same components, as well as an additional unstructured spatial component. The extensions for both models include changing the modelling of the age effect and adding an overdispersion component. The Bayesian inference in this thesis is carried out using hierarchical models, which allows for the accounting of the spatial dependency between the neighbouring counties.

As in the Etxeberria et al. (2018) article, there is assumed a high correlation between incidence and mortality when it comes to brain cancer in Norway. This is done due to the relatively high mortality in brain cancer. This correlation is advantageous for the joint modelling of incidence and mortality in these small areas, as it is used to increase the effective sample size by borrowing strength from both disease processes, i.e. incidence and mortality. This means that if there is a noticeable connection between incidence and mortality, methods that make use of the correlation between these processes could be used to improve estimates and discover underlying disease patterns (Etxeberria et al.; 2018, p. 2952). This is advantageous because of the scarcity of this type of cancer. The increased sample size allowed the authors to include more variables, like age group and gender, in the models. This way of modelling is interesting because incidence and mortality are modelled jointly and then linked through a shared spatial effect, which is allowed to vary in strength by using a scaling factor, δ.

In chapter 2 the data is presented along with a preliminary explanatory analysis of the data. A theoretical introduction to spatial statistics and integrated nested Laplace approximation (INLA) follows in chapter 3. Further, several different approaches and models have been introduced in the field of disease mapping over the years. Three articles on this topic with particular relevance for this thesis, as well as the Etxeberria et al. (2018) article, will be introduced in chapter 4. The methodology introduced by Etxeberria et al. (2018) is applied to the Norwegian brain cancer data from 2014–2018, combined with relevant extensions, in chapter 5. This is followed by a spatio-temporal continuation in chapter 6, which uses the entire data set from 1969–2018. The thesis is wrapped up in chapter 7, with a discussion and summary of the main results and future work.

Chapter 2

Brain Cancer in Norway

This thesis uses data from the Cancer Registry of Norway (https://www.kreftregisteret.no/ [1]). The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by the Cancer Registry of Norway is intended nor should be inferred.

The data provided are gender specific brain cancer incidence and mortality counts from 18 Norwegian counties (fylker) over 50 years with corresponding population counts. The data was provided from the data delivery unit at the Cancer Registry of Norway on March 5th 2020. The brain cancer data includes both brain and central nervous system tumors (International Classification of Diseases-10, C70–C72), but is referred to only as brain cancer in the following sections. Both the incident cases and deaths are given by gender. Further, the data is subdivided into 9 age categories of 10 year intervals (0–9, 10–19, . . . , 70–79, 80+) and into 10 calendar periods, in 5 year intervals (1969–1973, 1974–1979, . . . , 2014–2018). These rough intervals for age and year are chosen to accommodate standards of data protection and anonymity. On January 1st 2018 Norway went from having 19 to 18 counties, with the merging of Nord-Trøndelag and Sør-Trøndelag. The last time period, 2014–2018, includes data from the entire year of 2018. Because of this, the data was provided for only 18 counties, where the data from these two counties are both found in Trøndelag, for the entire period 1969–2018.

In table 2.1 an excerpt of the brain data is presented. This table includes all columns used in the analyses in chapters 5 and 6. In the first column, *Gender*, we find the gender index, which takes the value 1 for men and 2 for women, and in the second column, *Agegroup*, the age group indices are found. Column 3, *Fylke*, includes all the names of the 18 counties. In column 4, *inci_mort*, we find the index inci_mort, which takes the value 1 if the value of the sixth column, *Cases*, is an incidence count, and the value 2 if the case is a death. Column 3 coincides with column 5, *i*, which takes the values 1 to 36, where each county takes the value 1–18 for incidence and 19–36 for mortality. The counties are numbered alphabetically. Column 7, *Population*, includes the values of the population counts for each of the age groups in each county. Each of the 10 time periods are found in column 8, *Period*, where period 1 is 1969–1973, 2 is 1974–1978, etc.

[1] The data was recieved March 5th 2020

Table 2.1: Excerpt of the BrainData matrix used in the thesis

Gender	Agegroup	Fylke	inci_mort	i	Cases	Population	Period	i_male	i_female	agegroup_male	agegroup_female	mu	i_incidence	i_mortality	z_row
1	1	Akershus	1	1	1	155683	1	1	NA	1	NA	1	1	NA	1
1	1	Akershus	1	1		154158	2	1	NA	1	NA	1	1	NA	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	Akershus	1	1	9	188720	9	1	NA	1	NA	1	1	NA	9
1	1	Akershus	1	1	9	197842	10	1	NA	1	NA	1	1	NA	10
2	1	Akershus	1	1	4	146947	1	NA	1	NA	1	2	1	NA	11
2	1	Akershus	1	1	2	147236	2	NA	1	NA	1	2	1	NA	12
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	9	Akershus	1	1	40	63437	9	NA	1	NA	9	2	1	NA	179
2	9	Akershus	1	1	23	68396	10	NA	1	NA	9	2	1	NA	180
1	1	Østfold	1	2	3	90987	1	2	NA	1	NA	1	2	NA	181
1	1	Østfold	1	2	5	88152	2	2	NA	1	NA	1	2	NA	182
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	9	Vestfold	1	18	29	35355	9	NA	18	NA	9	2	18	NA	3239
2	9	Vestfold	1	18	20	35684	10	NA	18	NA	9	2	18	NA	3240
1	1	Akershus	2	19	3	155683	1	19	NA	1	NA	3	NA	1	3241
1	1	Akershus	2	19	5	154158	2	19	NA	1	NA	3	NA	1	3242
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	9	Vestfold	2	36	6	35355	9	NA	36	NA	9	4	NA	18	6479
2	9	Vestfold	2	36	4	35684	10	NA	36	NA	9	4	NA	18	6480

This table illustrates how the data is given and how it is sorted. It also includes all columns needed for the analyses in INLA.

The data is first sorted by *inci_mort*, followed by *Fylke*, then by *Agegroup* in each county and lastly by *Gender*. As the table shows, the first half of the rows in the data set is for the incidence data and the second half contains all the information related to the mortality data. For each county, we have 180 lines of data because of the 10 time periods and the 9 age groups for each gender. The last 8 columns in the table are added to the data matrix in order to do the analyses. These will be explained further in chapter 5 and 6.

2.1 The most recent period: 2014–2018

The initial thought when analysing the data was to aggregate over all time periods. However, the way the data is subdivided into age groups and time periods made this difficult. The incident cases and deaths are not problematic, because these are only counted once. The population counts, on the other hand, will be counted twice with the use of this method. This is due to the different sizes of the age groups and time periods. Hence, initial spatial analysis focuses on the period 2014–2018. This means that for the spatial analysis, all the data from time period 10 is extracted from table 2.

One definition used in the thesis concerns the age-standardized rates. Age-standardization is a way of adjusting rates, in order to minimize the effects of differences in age composition when comparing rates for different populations (Cancer Registry of Norway; 2019). In other words, the World standard means using the World population.

In the entire period 1969–2018 a total of 34195 (47% males and 53% females) incident cases and 12938 (58% males and 42% females) deaths were reported according to the data from the Cancer Registry of Norway. For the period of 2014–2018, the reported incident cases were 5086 (47% males and 53% females) and reported deaths were 1962 (59% males and 41% females).

Figures 2.1 and 2.2 show the spatial distribution of crude incidence and mortality rates per 100 000 for men and women in the period 2014–2018. In these figures, the data is aggregated over all age groups. The purpose of these plots was to draw attention to the spatial trends in the data. In figure 2.1 the incident rates are shown, with males on the left and females on the right. Similarly, figure 2.2 shows the mortality rates on the same form.

The figures suggest that brain cancer is more common for women than men, as there seems to be more observed cases in women than in men in most of the counties. For women the rates range from 15 to 28, while the rates for men only range from 15 to 22. For Telemark and Buskerud, the rates appear higher for men than women, but otherwise the incidence rates for women are slightly above the rates of men. Further, an opposite pattern can be observed for the mortality rates, with higher mortality rates for men than women. Comparing the incidence and morality trends with the total number of observed cases and deaths in Norway in the period 1969–2018, these trends coincide. However, comparing the incidence trend to the global trend from 2018 presented by GLOBOCAN, by World standard (Ferlay et al.; 2018), these trends do not coincide as the World trend is higher incidence rates for men than women. Further, if we look at the age-standardized incidence rates per 100 000, by Norwegian standard, the general trend has been higher rates for women than men since the middle of the 90s (Cancer Registry of Norway; 2019).

The counties Telemark and Buskerud seem to have the highest incidence rates for men, followed close by Nordland. For women, Oppland and Vestfold seem to have highest rates,

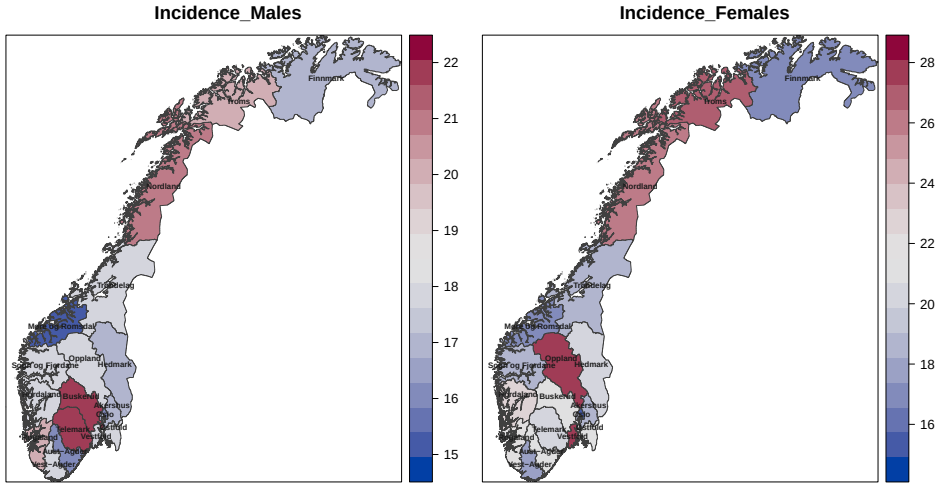


Figure 2.1: Spatial distribution of crude incidence rates per 100 000 for brain cancer in both men and women in Norway in the period 2014–2018, aggregated over the age groups

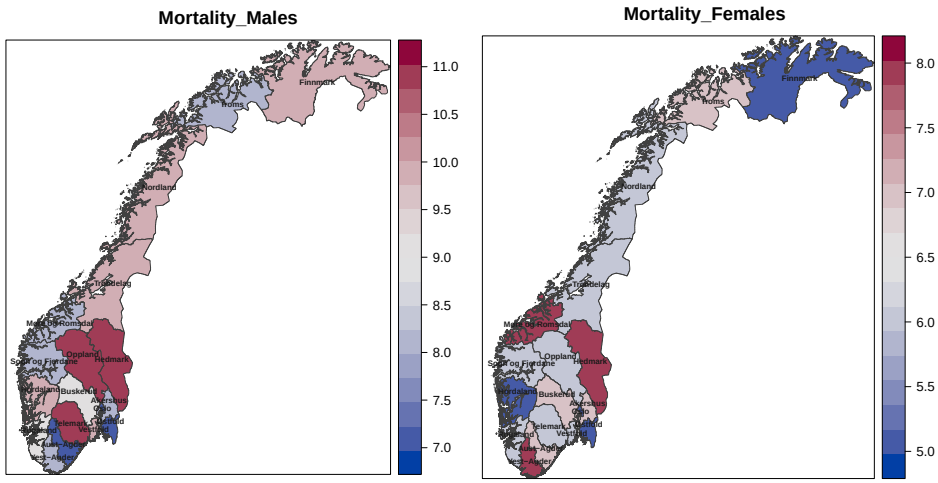


Figure 2.2: Spatial distribution of crude mortality rates per 100 000 for brain cancer in both men and women in Norway in the period 2014–2018, aggregated over the age groups

closely followed by Troms and Nordland. Further, several neighbouring counties, apart from the ones with the highest rates, seem to have quite similar rates. For the mortality

for women Vest-Agder, Møre and Romsdal and Hedmark clearly have the highest rates, while for men it is Telemark, Oppland and Hedmark. For women Finnmark tend to have low rates, which is more clear for mortality, but noticeable in both disease processes. The lowest incidence rates for men are found in Aust-Agder and Møre and Romsdal, while for mortality the lowest rates can be found in Aust-Agder, Oslo and Østfold. The incidence rate for men have a clear increasing trend from Finnmark to Nordland. The incidence rates for both genders have an increasing trend from Møre og Romsdal to Rogaland. Apart from this, it is hard to find a general trend in this explanatory plot of the data.

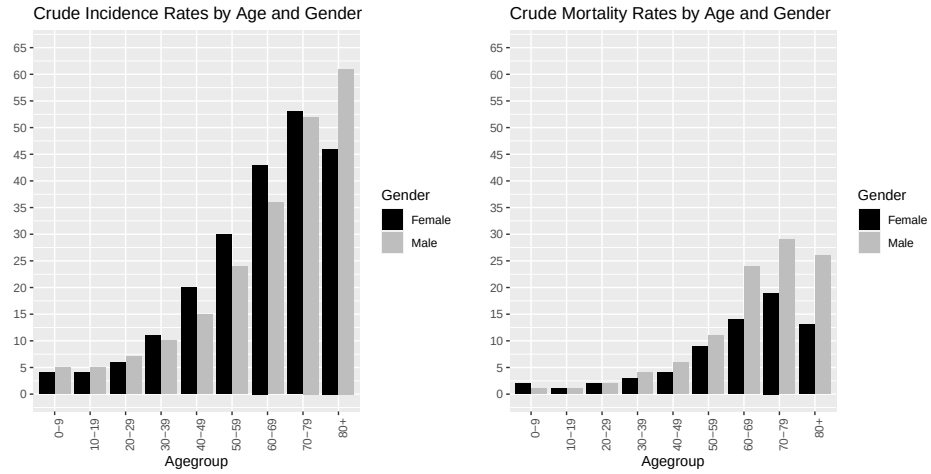


Figure 2.3: Norwegian brain cancer incidence and mortality rates per 100 000 in each age group for men and women in the period 2014-2018, aggregated over all counties.

Figure 2.3 shows incidence and mortality rates per 100 000 for each age group. In this figure, the data is aggregated over all districts in order to highlight the age trend. The rates are shown for men in the grey bars and for women in the black bars. For the age groups 10–19 to 70–79, we see a steady increase in the incidence rates for both men and women. However, between age group 70–79 and 80+ the incidence rates decrease for women. The mortality rates show a similar trend for both genders as for incidence, but here the increase is between age group 10–19 and age group 70–79 for both genders, not just women. Between the oldest two age groups, a slight decrease in the mortality trend can be seen in both genders, with a more apparent trend for women. Note that the trend is different for incidence and mortality for elderly people, which is interesting and will be investigated further. In the youngest age group we observe both higher incidence and mortality rates than in age group 10–19. This may be explained by the fact that brain cancer is the second most common cancer in young children after leukaemia, as Etxeberria et al. (2018) presents in their article. This can also be seen for Norwegian children in the period 2014–2018, as presented in *Cancer in Norway 2018* (Cancer Registry of Norway; 2019). In this incidence plot, the rates for women are higher than the rates for men in the age groups 30–39 to 70–79, which coincide with the previous plot. For mortality rates, the rates for men are higher than the rates for women, except for the youngest age group.

2.2 Historic data from 1969–2018

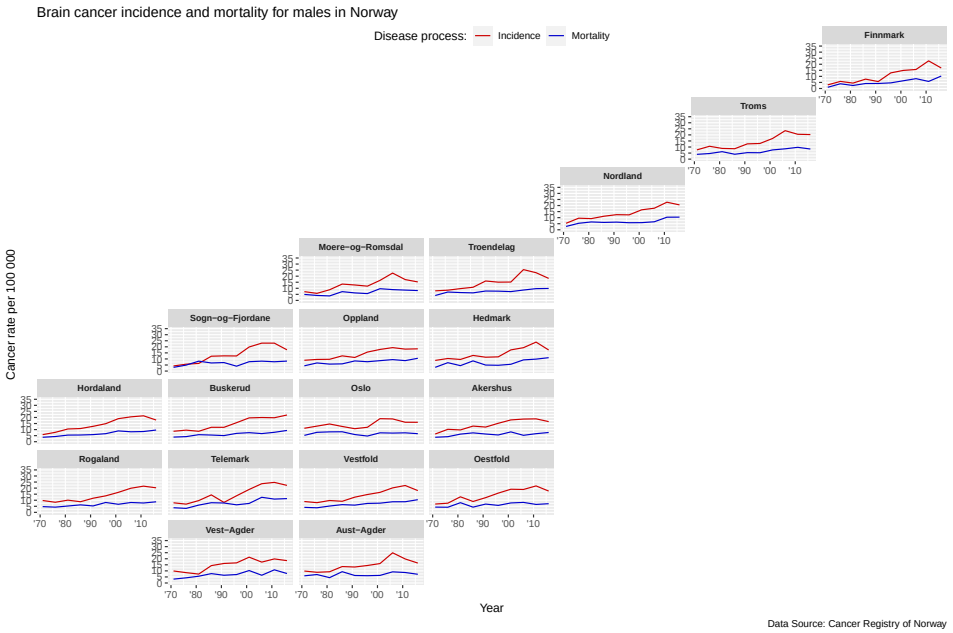
In figure 2.4a and 2.4b the brain cancer rates over time are shown for all 18 counties. These plots show the incidence and mortality rates per 100 000 for the entire period 1969–2018, aggregated over all age groups. The incidence rates are shown in red and the mortality rates are shown in blue. The plots are made using the package `geofacet` in `R`, where a grid of the Norwegian counties must be specified. The grid in these figures are visually organized to resemble the geographical location of the counties, possibly at the expense of showing the true county borders. Further, the structure of neighbouring counties is of importance, as this thesis assumes correlation between neighbouring counties. Hence, even though the three northern most counties share a border, they do not in this plot.

In both plots, the relationship between incidence and mortality is more noticeable than when we only look at one of the time periods. This can be seen in most of the counties for both genders, by similar shapes in the incidence and mortality curves, with incident rates laying above the death rates. However, there is an exception to this, where the mortality peeks above incidence for a given point in time. This can be seen in Sogn og Fjordane in the early 1980s for men. Meaning, that the total number of people who died of brain cancer is larger than people that got diagnosed with it for this exact time period.

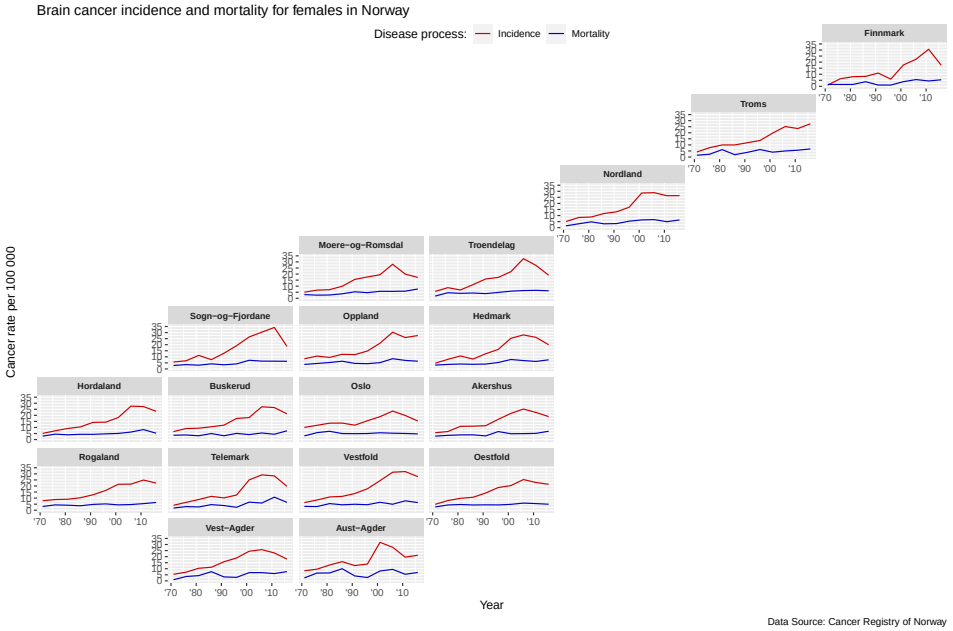
Furthermore, an increasing trend in the rates for both disease processes and both genders over time can be seen for most of the counties. Even though some of the rates seem to fluctuate from period to period, the general trend appear to be increasing from 1969 to 2018. For a couple of the counties, it is hard to see if the overall trend is increasing. This concerns Oslo for both genders, as the rates for 2014–2018 for this county do not appear to differ a lot from the rates for 1969–1973. The increase in time is harder to notice for mortality, as these rates are significantly lower than the incidence rates. It can also look as if the incidence rates are starting to decrease in the latest time periods. However, this can also just be an example of the fluctuation of the rates, rather than a sign of a decreasing rate in time.

In these plots, similarities for neighbouring regions are present. In Trøndelag for instance, the male incidence and mortality lines look similar to its neighbours, Møre og Romsdal, Oppland, Hedmark and Nordland. Another example is Hordaland for women, with the neighbours Rogaland, Telemark, Buskerud and Sogn og Fjordane. These similarities are also noticeable for several other counties.

Further in this thesis we will first investigates the spatial distribution of brain cancer in Norway in the time period 2014–2018, then look into the spatio-temporal distribution in the entire period 1969–2018. For the spatial modelling in the period 2014–2018, it would be interesting to see if there exist any underlying geographical patterns in the disease more recognizable than in figures 2.1 and 2.2. Further, the possible pattern seen in the age group plot in figure 2.3 will be investigated later in the thesis. For the entire period of 1969–2018, it would be equally interesting to discover some underlying temporal trend as looking into the spatial distribution.



(a)



(b)

Figure 2.4: Norwegian brain cancer incidence (red) and mortality (blue) rates per 100 000 in all counties for men (a) and women (b) in the period 1969–2018

Chapter 3

Introduction to Spatial Statistics

Statistical inference is the process of analysing, interpreting data and drawing conclusions from the data along with remaining uncertainty (Held and Sabanés Bové; 2014, p. vii). There are two main types of statistical inference, frequentist and Bayesian inference, where this thesis focuses on the latter. *Spatial statistics* has become popular the last few decades, mainly due to the advances in computational tools, which have increased the availability of geo-referenced data. As an example, when the interest is to evaluate the incidence of a particular disease across a country, the data can often only be available for small areas for several years. This is where spatial statistics comes in handy; by considering the possible geographical patterns of the disease, like similarity between neighbouring regions, researchers can apply this to improve the estimation of incidence in the regions (Blangiardo and Cameletti; 2015, pp. 1–3).

Before the year 2000 the Bayesian approach was not a common method to use in real case studies with spatial data, and therefore mostly found in theoretical models. This was because it did not exist any numerical, analytical or simulative computational tools to compute the posterior distribution, in the cases where it was not directly available in the form of a known distribution. This changed with the development of the Markov chain Monte Carlo (MCMC) methods around year 2000, as the first Bayesian method which could be applied to spatial data. MCMC methods enabled researchers to perform Bayesian computation on complex models on large data sets without having to simplify the structures (Blangiardo and Cameletti; 2015, pp. 2–3). The idea behind MCMC is to simulate a Markov chain, designed such that it will converge to the posterior distribution. When the convergence is achieved, one can draw random samples, which can be used to estimate posterior values (Held and Sabanés Bové; 2014, pp. 269–270).

However, because of the continuous advances in data collection, there is an increasing availability of big data sets, which has become an issue for the MCMC methods. The MCMC have a computational burden, which can lead to several days of computing time to perform Bayesian inference. To overcome this issue Rue et al. (2009) introduced integrated nested Laplace approximation (INLA), which is a deterministic algorithm for Bayesian inference for the class of latent Gaussian models which is both fast and accurate

(Blangiardo and Cameletti; 2015, p. 3). The remaining part of the section will focus on spatial modelling, Bayesian inference and the INLA methodology, which is necessary to understand the rest of the thesis.

3.1 Bayesian inference

As mentioned, the introduction of Bayesian methods for spatial data allowed the use of larger data sets and more complex models. One of the benefits of using a Bayesian approach in spatial statistics is that it accounts for the uncertainty within the estimates and predictions, as well as its flexibility and capability to deal with problems like missing data (Blangiardo and Cameletti; 2015, p. 3). Furthermore, it allows for borrowing strength over neighbouring regions by smoothing in a straightforward manner. Bayesian inference can be carried out using statistical models. The main focus in this thesis are hierarchical models. Below, Bayesian inference and hierarchical structure will shortly be explained.

Bayesian inference is one method of doing inference based on Bayes' theorem. In contrast to frequentist inference, were the focus mainly lies on the fixed, but unknown parameter $\boldsymbol{\theta}$, Bayesian inference treats the $\boldsymbol{\theta}$ as a random variable with a *prior distribution* $f(\boldsymbol{\theta})$, which contains information about prior beliefs. Hence, in Bayesian inference we can estimate parameters based both on prior knowledge as well as the data, whereas frequentist inference is only based on the data. After observing the data, Bayes' theorem is used to get the *posterior distribution*;

$$f(\boldsymbol{\theta} \mid data) = \frac{f(data \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{\int f(data \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{f(data \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(data)} \tag{3.1}$$

which is the most important quantity in Bayesian inference (Held and Sabanés Bové; 2014, pp. 167–172). In equation 3.1 we condition on $data$, which is more commonly denoted as $\boldsymbol{Y} = \boldsymbol{y}$. Here, $\boldsymbol{y} = (y_1, \ldots, y_n)$ are the observed realizations of the random variable $\boldsymbol{Y}$ with density function $f(\boldsymbol{y} \mid \boldsymbol{\theta})$. Moreover, $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the likelihood function and $f(\boldsymbol{\theta})$ is the *marginal likelihood*, which does not depend on $\boldsymbol{\theta}$. This implies that $1/f(\boldsymbol{y})$ is the normalizing constant, which ensures that the posterior distribution $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ is a valid density function and integrates to 1.

In Bayesian inference it is relatively easy to construct and estimate hierarchical models. This is one of the important qualities of the Bayesian approach. One of the purposes of hierarchical models is the methodological purpose. When data is drawn from clusters within a population, as neighbourhoods, they are no longer independent. Therefore, data observation drawn from the same cluster will be more related to each other than they will be to observations from other clusters. To atone for the biases introduced when the assumption of independence is violated, one could choose to construct hierarchical models (Lynch; 2007, pp. 231–233).

Equation 3.1 has a simple hierarchical structure of two levels. The first level is the conditional distribution for the data under the parameter, $f(\boldsymbol{y} \mid \boldsymbol{\theta})$. The second level is the marginal, prior, distribution for the parameter, $f(\boldsymbol{\theta})$.

Further, this hierarchical structure can be extended with another hierarchical level. The

updated posterior distribution would look like:

$$f(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) f(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\boldsymbol{y})}$$

where the knowledge of $\boldsymbol{x}$ is expressed through the *hyperparameters*, $\boldsymbol{\theta}$. In the Bayesian hierarchical structure, this model has three stages:

Level 1: $\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta} \sim f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$	(likelihood)
Level 2: $\boldsymbol{x} \mid \boldsymbol{\theta} \sim f(\boldsymbol{x} \mid \boldsymbol{\theta})$	(latent field)
Level 3: $\boldsymbol{\theta} \sim f(\boldsymbol{\theta})$	(hyperparameters)

where the first level is the likelihood for the data $\boldsymbol{y}$, the second level is the latent field, $\boldsymbol{x}$, and the last level is the hyperparameters, $\boldsymbol{\theta}$.

3.1.1 Integrated nested Laplace approximations (INLA)

Integrated nested Laplace approximation (INLA), is a method of doing Bayesian inference based on analytical approximations and numerical integration. This is in contrast to Markov chain Monte Carlo (MCMC), which is a sampling-based approach. One of the reasons for using INLA rather than MCMC is the computational advantage. INLA is a three stage hierarchical model with the stages being the observations, $\boldsymbol{y}$, the latent Gaussian field, $\boldsymbol{x}$, and the hyperparameters, $\boldsymbol{\theta}$. The models used in INLA-based inference are known as latent Gaussian models (LGM). LGMs are a subset of all Bayesian hierarchical models with a structured additive predictor, $\boldsymbol{\eta}$, where all latent elements are assumed to be Gaussian (Rue et al.; 2009). LGMs consist of three elements: a likelihood model, a latent Gaussian field and a vector of hyperparameters. A LGM can be written as:

$$\begin{aligned}
\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta} &\sim \prod \pi(y_i \mid \eta_i, \boldsymbol{\theta}) \\
\boldsymbol{x} \mid \boldsymbol{\theta} &\sim N(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})) && \text{(latent Gaussian field)} \\
\boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) && \text{(hyperparameters)}
\end{aligned}$$

where $\boldsymbol{Q}(\boldsymbol{\theta})$ is the precision matrix (inverse covariance matrix) of the latent Gaussian field (Martino and Riebler; 2020, pp. 1-2), and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ can be described as:

$$\eta_i = \mu + \sum_j \beta_j z_{ij} + \sum_k w_k f^k(u_{ik})$$

where μ is the intercept, $\boldsymbol{z}$ are the fixed effects with coefficients $\boldsymbol{\beta}$, $\boldsymbol{w}$ are the known weights for the unknown functions, $\{f^k\}$, of the covariate $\boldsymbol{u}$, which is used to model the random effects of $\boldsymbol{u}$ (Rue et al.; 2009). The latent components of $\boldsymbol{\eta}$ are gathered in the latent Gaussian field, $\boldsymbol{x}$, and given as $\boldsymbol{x} = \{\boldsymbol{\eta}, \mu, \boldsymbol{\beta}, \{f^1\}, \{f^2\}, \ldots\}$. This is done such that one element y_i depends on the latent field through only through η_i, which simplifies the computations needed (Martino and Riebler; 2020).

In the INLA framework the interest lies in approximating the marginal posterior components of the latent field $\pi(x_i \mid \boldsymbol{y})$ or the marginal posterior of the hyperparameters

$\pi(\theta_j \mid \boldsymbol{y})$. These can further be used to make approximate summary statistics, such as posterior means, variances or quantiles. The marginal posteriors are denoted as;

$$\pi(x_i \mid \boldsymbol{y}) = \int\int \pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{x}_{-i} d\boldsymbol{\theta} = \int \pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}$$
$$\pi(\theta_j \mid \boldsymbol{y}) = \int\int \pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{\theta}_{-j} = \int \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}_{-j}$$

Here, the integrals with respect to $\boldsymbol{x}$ is usually highly multidimensional and difficult to solve, while the integrals with respect to $\boldsymbol{\theta}$ are relatively small and solvable with numerical integration. Because of this, some of the fundamental work in INLA lies in making clever approximations to the posterior for the hyperparameters $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ and the full-conditional density $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ (Martino and Riebler; 2020, pp. 2–3).

3.1.2 Penalized complexity (PC) priors

As mentioned, for Bayesian inference we have hyperparameters, which are specified in $\boldsymbol{\theta}$. For these hyperparameters, we need to assign a prior distribution. One type of priors that has recently been proposed for Bayesian inference are penalized complexity (PC) priors (Simpson et al.; 2017). These are weakly informative and proper prior distributions, which means that they include specific prior information and have a density function that integrates to unity (Held and Sabanés Bové; 2014, pp. 180–191).

The idea of the PC prior is to penalize model complexity in order to avoid overfitting (Simpson et al.; 2017). This means that models are penalized if they include parameters that are not supported by the data. According to Simpson et al. (2017, p. 9) an overfitting prior will produce a more flexible model than might be necessary. This will make the base model, the simplest model, have almost no support in the prior and therefore in the posterior. The consequence of using an overfitting prior is that we cannot determine between flexible models supported by data and models that are flexible due to the choice of priors.

Simpson et al. (2017) introduce four principles for constructing a prior distribution for a flexibility parameter θ, e.g. precision (τ), standard deviation (σ), correlation (ρ), ... In principle 4, *User-defined scaling*, it is determined that λ, a hyperparameter selected by the user, can be selected by controlling the prior mass in the tail, which is a condition on the form: $P(f(\theta) > U) = \alpha$. Here, $f(\theta)$ is a transformation of the flexibility parameter, θ, U is a reasonable, user-defined upper bound that specifies the "tail event", and α is the weight we put on this event (Simpson et al.; 2017, p. 13). The PC prior for the standard deviation, $\theta = \sigma$, results to be an exponential prior:

$$\pi(\sigma) = \lambda \exp(-\lambda\sigma)$$

where λ determines the magnitude of the penalty for deviating from the base model. See Simpson et al. (2017) for details on the derivation. In a similar way λ for the standard deviation can be determined by the user by specifying U and α such that $P(\sigma > U) = \alpha$, which would imply that $\lambda = -\ln(\alpha)/U$. Here, $U > 0$ and $0 < \alpha < 1$.

3.2 Spatial modelling

A spatial trend is often beneficial to include in a statistical model when working with spatial data, as this extra information can improve the understanding of the data and may lead to biased estimates if ignored. Blangiardo and Cameletti specifies three types of spatial data; *area or lattice data*, *point-referenced (or geostatistical) data* and *spatial point patterns* (2015, p. 173). *Area or lattice data* is commonly used in disease mapping. Data used in disease mapping are discrete observations, as they are counts of disease incidence or deaths in pre-specified, usually non-overlapping, areas. For *area or lattice data*, the observations are found in an areal unit with well-defined boundaries. Here, we will focus on the *area data*, where the boundaries are irregular and typically based on administrative boundaries, such as districts and counties. Such spatial models can be described using a Bayesian framework, by the use of hierarchical structure, which can take into account spatial dependency based on neighbourhood structure. (Blangiardo and Cameletti; 2015, pp. 173–176). This neighbouring structure, $\boldsymbol{R}$, proposed by Besag et al. (1991), can be defined as:

$$R_{ij} = \begin{cases} n_i, & \text{if } i = j \\ -1, & \text{if } i \sim j \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

where $i \sim j$ denotes that area i and area j are neighbours, i.e. they share a border (Blangiardo and Cameletti; 2015). Further, for each row i, the column entry j is equal to 0 if areas i and j are not neighbours or -1 if i and j are neighbours. The diagonal of row i contains the number of neighbours j, denoted by n_i. By distinguishing the geographical relationship between the different districts, we can borrow strength from the neighbouring districts.

One of the methods commonly used in disease mapping is the Besag-York-Mollié (BYM) method, proposed by Besag et al. (1991). To explain this, we assume that, for each area $i = 1, \ldots, n$, the observed cases y_i conditioned on the rate, λ_i, are Poisson distributed with $E_i\lambda_i$, where E_i are the expected cases. A log linear model can be specified on the linear predictor, η_i:

$$\eta_i = \log \lambda_i = \mu + u_i + v_i \tag{3.3}$$

where μ is the intercept and $\boldsymbol{v}$, the area-specific unstructured random effect, is modelled with an exchangeable structure, meaning $\boldsymbol{v} \mid \sigma_v^2 \sim N(0, \sigma_v^2 I)$. The $\boldsymbol{u} = (u_1, \ldots, u_n)$ is an area-specific structured random effect, which can be modelled using the following distribution:

$$\pi(\boldsymbol{u} \mid \sigma_u^2) \propto \left(\left| \frac{1}{\sigma_u^2} \boldsymbol{R} \right|^* \right)^{1/2} \exp\left(-\frac{1}{2} \boldsymbol{u}^T \left(\frac{1}{\sigma_u^2} \boldsymbol{R} \right) \boldsymbol{u} \right)$$

This formulation for $\boldsymbol{u}$ is often called the intrinsic conditional autoregressive (ICAR). The ICAR together with the exchangeable random effect from equation 3.3 forms the BYM model. Even if all areas are connected, i.e. no islands, no proper joint distribution for $\boldsymbol{u}$ exists, as the covariance matrix is not positive definite and therefore does not have full rank. To resolve this issue a constraint, $\sum_{i=1}^{n} u_i = 0$, summing over all areas $i = 1, \ldots, n$, can be applied to $\boldsymbol{u}$.

Specifying a BYM model in R-INLA can be done either by `f(..., model = "bym", ...)`, or by specifying the two BYM components separately using `f(..., model = "besag", ...)` for the spatial structured component (ICAR) and `f(..., model = "iid", ...)` for the unstructured component (exchangeable) (Blangiardo and Cameletti; 2015, p. 182). The `besag` model in R-INLA is the same as the ICAR. Therefore, from now the term `besag` will be used. An extension of the `besag` model, $\boldsymbol{u}$, is called `besag2`, which is used for weighted spatial effects of two outcomes such as incidence and mortality. This model is also used by Etxeberria et al. (2018) (see section 4.1) and in this thesis. The model is described by $\boldsymbol{x} = (\delta\boldsymbol{u}, \boldsymbol{u}/\delta)$, where $\boldsymbol{u} = (u_1, \ldots, u_n)$ is the regular `besag` model and $\delta > 0$ is the weight parameter (see http://www.r-inla.org/models/latent-models for more information). An alternative formulation to model two outcomes could be $\boldsymbol{x} = (\boldsymbol{u}, a \cdot \boldsymbol{u})$, with $a > 0$ as a weight for the second outcome. However, there is no pre-specified model for this available in INLA.

Chapter 4

Literature Review

This chapter will present short summaries of four articles. These are in the field of disease mapping and are all relevant for this thesis. The first article is the main inspiration for this thesis. The methodology introduced in this article by Etxeberria et al. (2018) is further applied to the Norwegian data in chapter 5. Further, the next three articles give an overall insight in previous advancements in this field, which will be linked to the Etxeberria et al. (2018) article.

4.1 Etxeberria et al. (2018): Joint modelling of brain cancer incidence and mortality

The first article is *Joint modelling of brain cancer incidence and mortality using Bayesian age- and gender-specific shared component models* by Etxeberria et al. (2018). This article is the most important one for this thesis, as the methodology from the paper is used to analyse the Norwegian brain cancer data in the next chapter. The Etxeberria et al. (2018) paper uses a Bayesian approach to explore the possible geographical patterns for brain cancer incidence and mortality. It focuses on two northern regions of Spain, Navarre and Basque Country (BC), which are further divided into 27 districts. The article features these two regions due to the high incidence rate, which are among the highest in Europe. This is interesting because the average incidence rate in all of Spain is below the European average (Etxeberria et al.; 2018).

The data used is incident cases and deaths in brain and central nervous system cancer in both genders collected from the 27 district in Navarre and Basque Country in the period 1990-2008. The authors have organized their data from 5 year intervals into the age groups 0–9, 10–29, 30–49, 50–64, 65–74, 75–84 and 85+, because of the similar behaviour they saw in these groups. The goal of the article was to discover the geographical patterns of incidence and mortality on brain cancer for the different age-groups from the two northern districts of Spain during the period 1990–2008. In addition to geographical patterns, the high mortality in small children is highlighted. As brain cancer is the second most frequent

cancer in children, there is an interest in unveiling this age-dependence in the analysis.

The model in this article is described like this;

$$I_{ijg}|\ \lambda_{Iijg} \sim \text{Poisson}(E_{Iijg}\lambda_{Iijg}),$$

and

$$M_{ijg}|\ \lambda_{Mijg} \sim \text{Poisson}(E_{Mijg}\lambda_{Mijg})$$

where I_{ijg} and M_{ijg} are the observed number of incident cases and deaths in region i ($i = 1, \ldots, 27$), gender g ($g = 1$ for males and $g = 2$ for females), and age group j ($j = 1, \ldots, 7$).
Here E_{dijg} is the population at risk, $\log E_{dijg}$ are offsets, λ_{dijg} are the rates and $\log \lambda_{dijg}$ can be modelled using different expressions. Here, $d = I, M$ is the disease index, for incidence and mortality, respectively. In the article they introduce 10 individual models and their gender specific equivalent, noted by an asterisk. Here, two of the models are introduced in detail, as the model 3^* is the one used to analyse the Spanish data and both model 3^* and model 8^* will be used in the analysis of the Norwegian data.

$$\begin{aligned}
\text{Model 3*:}\quad \log \lambda_{Iijg} &= \delta_g u^*_{ig} + \alpha_{Ij} \\
\log \lambda_{Mijg} &= \frac{1}{\delta_g} u^*_{ig} + \alpha_{Mj} \\
\text{Model 8*:}\quad \log \lambda_{Iijg} &= \delta_g u^*_{ig} + v_{Ii} + \alpha_{Ij} \\
\log \lambda_{Mijg} &= \frac{1}{\delta_g} u^*_{ig} + v_{Mi} + \alpha_{Mj}
\end{aligned}$$

where g represents gender and δ_g is the gender-specific spatial parameter.

Here, $\boldsymbol{u}^* = (u^*_{1,m}, \ldots, u^*_{n,m}, u^*_{1,f}, \ldots, u^*_{n,f})^T$ is a spatial random effect for male and female over the $n = 27$ districts. The $\boldsymbol{u}^*$ assumed to follow a multivariate normal distribution $\boldsymbol{u}^*|\ \boldsymbol{C} \sim N(\boldsymbol{0}, \boldsymbol{C}^-)$. The covariance matrix $\boldsymbol{C}^-$ is defined as;

$$\boldsymbol{C}^- = \begin{pmatrix} \sigma^2_{u^*_m}\boldsymbol{R}^- & \boldsymbol{0} \\ \boldsymbol{0} & \sigma^2_{u^*_f}\boldsymbol{R}^- \end{pmatrix}$$

This matrix includes the two variance components $\sigma^2_{u^*_m}$ and $\sigma^2_{u^*_f}$ that allows for individual smoothing for each gender, and the spatial neighbourhood matrix, $\boldsymbol{R}$. This matrix can be explained as in equation 3.2. The $^-$ indicates the Moore-Penrose generalized inverse, which is used because $\boldsymbol{R}$ is not full rank and hence does not have a regular inverse.

Further, the $\boldsymbol{\alpha} = (\alpha_{1,1}, \ldots, \alpha_{1,7}, \alpha_{2,1}, \ldots, \alpha_{2,7})^T$ is the disease-specific age effect and $\boldsymbol{v} = (v_{1,1}, \ldots, v_{1,n}, v_{2,1}, \ldots, v_{2,n})^T$ is the spatially unstructured random effect for incidence and mortality. Both of these effects are assumed to follow a multivariate normal distribution $\boldsymbol{\alpha}\ |\ \sigma^2_\alpha \sim N(\boldsymbol{0}, \sigma^2_\alpha(\boldsymbol{I}_2 \otimes \boldsymbol{I}_7))$ and $\boldsymbol{v}\ |\ \sigma^2_v \sim N(\boldsymbol{0}, \sigma^2_v(\boldsymbol{I}_2 \otimes \boldsymbol{I}_n))$, where $\boldsymbol{I}_2$, $\boldsymbol{I}_7$ and $\boldsymbol{I}_n$ are the 2×2, 7×7 and $n \times n$ identity matrices. Here, n is the number of districts; in this article equal to 27. When analysing the data in the article, penalized complexity (PC) priors were chosen for the hyperparameters. They were chosen with the upper bound parameter $U = 1$ and weight parameter $\alpha = 0.01$.

The best model from this article was the model 3^*. By the use of this model, they estimate the gender and age-specific incidence and mortality rates and analyse of their geographical distribution in Navarre and Basque Country. They discover a geographical pattern in the brain cancer, which is more profound for incidence than mortality and for males than for females. Another key result from the article, is the ones from the youngest age group. For this age group, they observe clear West–East spatial pattern for incidence, whereas for the mortality rates the spatial pattern is more low-rate homogeneous. This increase in incidence rates for this age group is not observed in other tumours, which makes it particularly interesting. Moreover, they show that the difference between incidence and mortality rates are more noticeable for the youngest age group than for the other age groups. A final result found in the article, is the decrease in rates between the two oldest age groups.

4.2 Colonna et al. (1999): Cancer incidence prediction based on mortality

The second article is by Colonna et al. (1999), and the first ever to estimate the incidence of cancers in different regions in France. The main idea is to estimate the national cancer incidence in France by only having knowledge of the incident cases in a small part of the population. Only about 10% of the French population is covered by, so called, cancer registries. The relationship between cancer incidence and mortality, found in these cancer registries, is used in combination with national mortality data to obtain cancer incidence estimations for a given point in time. This article has a slightly different purpose than this thesis and the other articles in this chapter. However, the interesting aspect of the article lies in the assumption that there exists a relationship between incidence and mortality. This same assumption is made in both the article by Etxeberria et al. (2018) and in this thesis.

The cancer registries covers breast cancer incidence for women and colorectal cancer incidence separated by gender obtained from nine French administrative departments. The national mortality data is obtained from 21 administrative regions. The method used applies the incidence/mortality ratio from the cancer registry areas to regional mortality data to estimate regional incidence rates at three time points (1985, 1990, 1995), which is used to estimate an incidence trend. To validate their results, they use a leave-one-out method on the cancer registries for all nine departments, and compare the estimated incidence with the true incidence (Colonna et al.; 1999).

Colonna et al. (1999) conclude in this article that the breast cancer incidence has increased considerably between 1985 and 1995, where there is a noticeably higher increase in the north of France than in the south. They also notice a slight increase in colorectal incidence, but to a lesser extent. Because the approach and model from this article is not the same as in the article by Etxeberria et al. (2018) and in this thesis, I use a notation by Held et al. (2006) to explain the model. The model is described like this;

$$\begin{aligned} &M_{ijt}|\lambda_{Mijt} \sim \text{Poisson}(E_{ijt}\lambda_{Mijt}), && \text{and:} && I_{ijt}|\hat{M}_{ijt}, \lambda_{Iijt} \sim \text{Poisson}(\hat{M}_{ijt}\lambda_{Iijt}), \\ &\text{with:} && && \text{with:} \\ &\log\ \lambda_{Mijt} = u_{1i} + u_{2i}t + \alpha_{1j} + \alpha_{2j}t && && \log\ \lambda_{Iijt} = \hat{\alpha}_{1j} + \hat{\alpha}_{2j}t \end{aligned} \tag{4.1}$$

where M_{ijt} is the observed mortality counts, I_{ijt} denotes the incidence counts, E_{ijt} is the population counts and λ_{ijt} denotes the rates, in district i, for age group j and for year t. Further, $\hat{M}_{ijt}$ is the estimated deaths from the left hand side of 4.1. For the mortality, we use u_{1i} and u_{2i} to denote the district effects, and α_{1j} and α_{2j} to denote age group effects. Further, the $\hat{\alpha}_{1j}$ and $\hat{\alpha}_{2j}$ is used for the age group effect in the incidence/mortality ratio. In the log mortality, $\log \lambda_{Mijt}$, it is assumed that the time trends in each district and age group can be additively decomposed into linear trends for each district and linear trends for each age group. Equivalently, it is assumed that the log incidence-mortality ratio, $\log \lambda_{Iijt}$, is linear in each age group.

One of the advantages with the approach in this article, is the exploitation of the close relationship between incidence and mortality. Because of this, countries can get information about incidence when they lack nationwide coverage by cancer registries. This information can be applied to how the countries should distribute health resources. Even though the estimations may not be perfect, it may be beneficial for making decisions about health care. Etxeberria et al. (2018) shares this advantage of assuming that incidence and mortality are linked, but uses it to increase the effective sample size and include more variables in their analyses of geographical patterns.

Albeit the lack of incident data might be disadvantageous. In this case, only 10 % of the French population is covered by cancer registries. In an ideal world, one would have incidence data from the entire population, but then the use for estimation would be redundant. This lack of coverage could make the process of drawing proper conclusions difficult.

4.3 Held et al. (2005): Joint analysis of diseases with shared risk factors

In this next article, Held et al. (2005) uses a Bayesian approach to jointly analyse more than two diseases with shared risk factors in order to identify common geographical patterns. This is an extension of already existing frameworks, both Bayesian and non-Bayesian, for the joint modelling of two diseases. The authors use a Bayesian approach with an extended BYM model for four diseases (oral, oesophagus, larynx and lung cancers), with a log relative risk, λ_{id}, for each disease. The model can be described like:

$$y_{id} |\ \lambda_{id} \sim \text{Poisson}(E_{id}\lambda_{id}), \qquad d = 1, \ldots, 4$$

where y_{id} and E_{id} are the observed and expected number of cases in district i for disease d, with log relative risks:

$$\begin{aligned}
\lambda_{i1} &\sim N(\mu_1 + u_{1i}a_{1,1} + u_{2i}a_{2,1}, \sigma_1^{-2}) \\
\lambda_{i2} &\sim N(\mu_2 + u_{1i}a_{1,2} + u_{2i}a_{2,2}, \sigma_2^{-2}) \\
\lambda_{i3} &\sim N(\mu_3 + u_{1i}a_{1,3} + u_{2i}a_{2,3}, \sigma_3^{-2}) \\
\lambda_{i4} &\sim N(\mu_4 + u_{1i}a_{1,4}, \sigma_4^{-2})
\end{aligned}$$

The λ_{id} represents the log relative risks in area i for the four diseases d, σ_d are the corresponding standard deviations and $a_{k,d}$ are weights to allow for different risk gradient of

the shared components for the diseases. The $a_{k,d}$, where $k = 1, 2$ represents the number of shared components, is defined under the restriction $\sum_{d=1}^{n_d} \log a_{k,d} = 0$, with n_d denoting the number of relevant diseases for field k. The first three diseases, oral, oesophagus and larynx cancer, have two shared spatial components, while lung cancer only share $\boldsymbol{u}_1$ with the other diseases. In this case, $\boldsymbol{u}_1$ represents spatial differences in tobacco consumption and $\boldsymbol{u}_2$ corresponds to spatial differences in alcohol consumption.

In the article, they use the information about tobacco and alcohol consumption and the geographical variances in the consumption as a way of identifying possible common spatial patterns in the diseases themselves. They emphasize the correlation between heavy smoking and heavy drinking in the oral, larynx and oesophagus cancers, as they are known as the major risk factors and seem to act synergistically (Held et al.; 2005, p. 68).

Two of the assumptions made in the article, is that these shared risk factors have a spatial structure and that all components are independent of each other. However, if some of the components have any spatial correlation, this will manifest in the spatial patterns and the spatial analysis will not just be a result of the shared latent risk factors.

One advantage of joint modelling proposed in the article, which is similar to one of the advantages of the Colonna et al. (1999) approach, is the possibility of gaining further understanding about the diseases and in particular the spatial patterns. This could for instance be beneficial in the providing of health care services.

The approach introduced in this article, is quite similar to the one used by Etxeberria et al. (2018) in the way that similarities are exploited in order to identify common geographical patterns in diseases. Held et al. (2005) focuses on several diseases and the relationship between their risk factors, while Etxeberria et al. (2018) focuses on the relationship between the disease processes within one disease. However, for both articles, the joint analysis can be highly beneficial in order to unveil the geographical patterns.

4.4 Martinez-Beneito (2013): A more general approach to multivariate disease mapping

In the last article, by Martinez-Beneito (2013), the main idea is to introduce a framework for disease mapping that brings together already existing models. This can be seen as a more general and extended framework than the one presented by Held et al. (2005). The multivariate disease mapping problem is formulated as:

$$y_{id} \sim \text{Poisson}(E_{id}\lambda_{id})$$

where y_{id} is the observed cases, E_{id} is the expected cases and λ_{id} is the relative risk, all for the i-th geographical unit and for the d-th disease. In this model r_{id} satisfies:

$$\log \lambda_{id} = \mu_d + u_{id}$$

where the article emphasizes that different disease mapping models mostly differ in how u_{id} is defined and how spatial and multivariate dependences in diseases are defined (Martinez-Beneito; 2013). The article shows that u will follow a prior distribution describes like this:

$$\text{vec}(u) \sim N_{ID}(0, \Sigma_b \otimes \Sigma_w)$$

where $\Sigma_b \otimes \Sigma_w$ denotes the Kronecker product of the between-disease (b) and within-disease (w) covariance matrices and vec(u) for within disease dependence can be described by:

$$\text{vec}(u) \ = (\tilde{\Sigma}_b \otimes I_I)\text{vec}(R), \quad \text{where:} \quad u = \tilde{\Sigma}_w \ \epsilon \ \tilde{\Sigma}_b^T = (\tilde{\Sigma}_w \ \epsilon)\tilde{\Sigma}_b^T \ = R\tilde{\Sigma}_b^T$$

with covariance matrix:

$$\text{cov}(\text{vec}(u)) = \ \Sigma_b \otimes (D - W)^{-1}$$

where the tilde denotes the operator which returns the lower triangular matrix of the Cholesky decomposition, $D = \text{diag}(n_1, \ldots, n_I)$, where n is the number of neighbours, and W_{ij} is equal to 1 if i and j are neighbours and 0 otherwise (Martinez-Beneito; 2013). R is a matrix whose columns have the desired spatial distribution, which is assumed to follow an intrinsic conditional autoregressive (ICAR) distribution. The ICAR is often used in disease mapping models to describe a wider range of geographical patterns (Blangiardo and Cameletti; 2015).

This article proposes a general framework for disease mapping, both within and between dependence in diseases. However, in this last summary, the focus is only on the within dependence. This is due to the similarity with the models used by Etxeberria et al. (2018). For example, looking at the covariance matrix, the $(D - W)$ is equal to the neighbourhood matrix, $\boldsymbol{R}$, used in the spatial modelling by Etxeberria et al. (2018). This is because both articles use an ICAR model for the spatial distribution.

In this chapter four different articles on the subject of disease mapping have been introduced, and are therefore relevant for this thesis. The second article focused on both the geographical distribution of cancer incidence as well as the trend over time, where the most interesting aspect is the use of the link between incidence and mortality. The other two articles mainly focused on introducing more general frameworks for modelling spatial patterns in one or more diseases. These last two articles are also mentioned by Etxeberria et al. (2018), which reinforces the relevance for implementing them in this thesis.

Chapter 5

Spatial modelling of brain cancer in Norway in 2014–2018

This chapter analyses the Norwegian brain cancer data from 2014–2018 using two of the models presented in the article by Etxeberria et al. (2018), including several possible extensions. A comparison of the models will be presented, followed by the results from the best model.

Using the notation from chapter 4, I_{ijg} and M_{ijg} are the observed number of incident cases and deaths in region i ($i = 1, \ldots, 18$), gender g ($g = 1$ for males and $g = 2$ for females), and age group (j $j = 1, \ldots, 9$). The analyses in the section has been performed in the statistical computing language `R`, where the INLA methodology has been implemented using the package `R-INLA` (see www.r-inla.org).

5.1 Applying the models from Etxeberria et al. (2018)

This first section will apply the models as presented in the article by Etxeberria et al. (2018). The first part is dedicated to the main model used in their article, which is called model 3^* and contains two components, which is a structured spatial component and an age component. The second part uses model 8^*, which contains an additional component, $\boldsymbol{v}$, that includes the spatially unstructured random effects for incidence and mortality. Since these models are applied only to the last period in the data set, 2014–2018, we only use the part of the table 2.1 from section 2 where the *Period* = 10.

As mentioned in chapter 2, the column 5, *i*, is a regional index, which ranges from 1 to 36. This is because the model `besag2`, used to model the weighted spatial effect in `R-INLA`, is defined to have dimension $2n$, where n is the size of the neighbourhood structure and the number of regions. Norway has 18 counties, so $2n = 36$. In addition to columns 1–8 explained in chapter 2, the analyses in this section includes column 9, 10, 13, 14 and 15. Columns 9 and 10 include the indices *i_male* and *i_female*, which are the gender-specific regional indices. The *i_male* takes the value of *i* if the observation

corresponds to male, that is, if $g = 1$, and the value NA if the observation corresponds to female ($g = 2$). Equivalently, *i_female* takes the value of *i* if the observation corresponds to female and the value NA if the observation corresponds to male. The *mu*, found in column 13, represents intercept values needed when running the `besag2` model. This is specified to be both gender-specific and outcome-specific. In other words, $mu = 1$ if the *Cases* are both incidence and male, $mu = 2$ if the *Cases* are incidence and female, $mu = 3$ if the *Cases* are mortality and male and finally, $mu = 4$ if the *Cases* are mortality and female. Columns 14 and 15 will be further explained in subsection 5.1.2 for model 8^*.

5.1.1 Modelling with a structured spatial component and age group

Model 3^* is the simplest of the two models; containing only two components, the gender-specific spatial random effects and the age-outcome-specific effects. These components are the same as defined on page 18.

There are three different variances in this model as mentioned in section 4.1; $\sigma^2_{u^*_m}$, $\sigma^2_{u^*_f}$ and σ^2_α. INLA is working and retrieving results using the precision parameters, which are the inverse of the variances, i.e. $\tau_{u^*_m} = 1/\sigma^2_{u^*_m}$, $\tau_{u^*_f} = 1/\sigma^2_{u^*_f}$ and $\tau_\alpha = 1/\sigma^2_\alpha$ (Martino and Riebler; 2020). Internally, the log precisions are used to have a numerically stable parametrization. The priors are defined on the precisions, and the first set of priors on the hyperparameters, are PC priors. Here used with parameters $U = 1$ and $\alpha = 0.01$, i.e. $P(\sigma > 1) = 0.01$. When plotting and interpreting the results, the standard deviation will be used instead of the precisions. This is because the standard deviation is easier to interpret and the natural scale when using PC priors.

Putting this model in the LGMs framework specified in section 3, where $\boldsymbol{y}$ will correspond to the observed cases of I and M, the latent Gaussian field is $\{\delta_m, \delta_f, u^*_{1,1}, \ldots, u^*_{18,1}, u^*_{1,2}, \ldots, u^*_{18,2}, u^*_{19,1}, \ldots, u^*_{36,1}, u^*_{19,2}, \ldots, u^*_{36,2}, \alpha_{1,1}, \ldots, \alpha_{1,9}, \alpha_{2,1}, \ldots, \alpha_{2,9}\}$. The model has five hyperparameters $\{\delta_m, \delta_f, \sigma^2_{u^*_m}, \sigma^2_{u^*_f}, \sigma^2_\alpha\}$.

Implementing the model in R-INLA

In the `R-code` 5.1 most of the code to make the model 3^* in `R-INLA` can be seen. In line 4, the extraction of the data from time period 2014–2018 from the full data set, `BrainData`, is seen. This is stored in the data set `BrainData10`. Lines 7–19 covers the neighbourhood structure, which is stored in the variable `g` in line 7. This is accessed using the `inla.read.graph()`, which takes in the neighbourhood structure as an graph object. Further, line 9–19 shows the information contained this structure. In line 9 the total number of regions, $n = 18$, is found. The first number listed in lines 10–19 indicates the region, while the following numbers on each line denotes the number of neighbours followed by the neighbours. Further, line 10 shows the information about region 1. The line starts with the number 1, indicating that it is region 1. Next, we find the number 5, meaning region 1 has 5 neighbours. These neighbouring regions are then listed in increasing order. For region 1, the neighbours are regions 2, 4, 6, 10 and 11. Likewise, line 11 indicates that region 2 only has 1 neighbour, which is region 1. Moreover, line 22 defines the prior for the hyperparameters. This is a PC prior stored in `pcprec` and specified by `"pc.prec"`, with $U = 1$ and $\alpha = 0.01$ specified in `param`.

The model is specified in line 25–29. This is stored in `formulaBrain3`. `Cases` is specified as the response variable and -1 is added to remove the intercept. The `f()` is used to specify the LGM for the random effect and the hyperprior, i.e. the prior for the hyperparameters, for its corresponding hyperparameters (Martino and Riebler; 2020). The `f(i_male, ...)` and `f(i_female, ...)` corresponds to the spatial effect u^*_{ig} in the model. For these variables the `besag2` model is used, as this is a model for weighted spatial effect, where the weight is the δ_g, with $\delta_g > 0$. This is reasonable to use in order to fit the shared spatial component models. Here, the `besag2` model is described by $\boldsymbol{x} = (\delta_g \boldsymbol{u}^*, \boldsymbol{u}^*/\delta_g)$. The neighbourhood structure, `g`, for the spatial effects is included in the `graph` argument and the hyperprior, `pcprec`, is specified in `hyper`. The logical option `scale.model` determines that the model $\boldsymbol{u}^*$ should be scaled to have a generalized variance equal to 1. This is done because it makes prior specification easier. The `constr = TRUE` is a sum-to-zero constraint, needed because the covariance matrix is not full rank.

R-code 5.1: Code specifying the original model 3*

```
1  #READ R-INLA PACKAGE
2  library(INLA)
3  #EXTRACTING THE LAST TIME PERIOD FROM THE FULL DATA SET
4  BrainData10 = BrainData[BrainData$Period == 10,]
5
6  #READ THE NEIGHBOURHOOD STRUCTURE
7  g = inla.read.graph("nb-inla.txt")
8  #The neighbourhood structure contains the following information:
9  18
10 1 5 2 4 6 10 11
11 2 1 1
12 3 3 12 15 17
13 4 7 1 7 10 11 13 15 18
14 .
15 .
16 .
17 16 2 5 9
18 17 2 3 12
19 18 2 4 15
20
21 #PC PRIOR
22 pcprec = list(prec=list(prior = "pc.prec", param=c(1, 0.01)))
23
24 #FORMULA FOR THE MODEL 3*
25 formulaBrain3 = Cases ~ - 1 + mu + f(i_male, model = "besag2",
26    graph = g, hyper = pcprec, constr = TRUE, scale.model = TRUE) +
27  f(i_female, model = "besag2", graph = g, hyper = pcprec,
28    constr = TRUE, scale.model = TRUE) + f(Agegroup, model = "iid",
29    hyper = pcprec, constr = T, replicate = inci_mort)
30
31 #INLA EXECUTION
32 results3 = inla(formulaBrain3, family="Poisson", data=BrainData10,
33  E = Population, control.predictor=list(compute=TRUE))
```

Furthermore, the `f(Agegroup, ...)` is equivalent to α_{dj}, which is the same disease-specific age effect as in section 4.1. This is fitted as an independent random noise model, i.e. an iid model with $N(0, \sigma_\alpha^2)$. In this part the `constr = T` is a sum-to-zero constraint. `replicate = inci_mort` is used to generate iid replicates of this model with the same hyperparameters. Here, `inci_mort` defines how the observations are grouped into the replicated effects, that is, by incidence and mortality. The replicating is done because the age-outcome specific effect α_{dj} is assumed to be the same for both genders.

In lines 32–33, in `results3`, the INLA execution of the formula object is put in the main function `inla()`. The first argument is the formula, `formulaBrain3`. Next, the Poisson distribution is defined in the `family` argument. The data is set to `BrainData10`, as specified in line 4. The next argument is the parameter `E = Population`, which specifies the population as an offset. The last argument specifies that the marginals for the linear predictor should be computed, which is needed for the extracting of the shared spatial effect presented below.

Extracting the shared spatial effect

In `R-INLA` the default output when using the `besag2` model, is $\boldsymbol{x} = (\delta_g \boldsymbol{u}^*, \boldsymbol{u}^*/\delta_g)$. In this case, the first part of the output, $\boldsymbol{x}$, corresponds to spatial effect for incidence and the second part corresponds to the spatial effect for mortality. However, the interest in this thesis lies in the shared spatial field, not in the spatial effect for different disease processes. The only component separating the spatial field for incidence and mortality is the gender-specific weighting parameter δ_g. Therefore, we want to remove this part of the output in order to only be left with the shared spatial effect, $\boldsymbol{u}^*$. In order to do this we needed to specify the `control.predictor=list(compute=TRUE)`, which enables the usage of the `inla.posterior.sample()`, in the INLA call in the `R-code` 5.1.

R-code 5.2: Code for extracting the shared spatial field

```
1  ##function for extracting the besag2 weight
2  fun = function(){
3   a.x.male = i_male
4   a_male = theta[2]
5   a.x.female = i_female
6   a_female = theta[4]
7   nn = length(a.x.male)
8   n = nn %/% 2L
9   #undo the effect of 'a'
10  return(c(a_male, a_female, a.x.male[1:n]/a_male,
11  a_male*a.x.male[n+1:n], a.x.female[1:n]/a_female,
12  a_female*a.x.female[n+1:n]))
13 }
14
15 ##making samples of the result
16 samp = inla.posterior.sample(100, results3)
17
18 ##evalutating the samples with the function above
19 xtrct = inla.posterior.sample.eval(fun,  samp)
```

As the desired output is not already included in R-INLA, we have to extract it ourselves. The outputs for i_male and i_female are the default outputs, from which we need to extract the spatial effect without δ_m and δ_f. This is done by first generating 100 samples from the approximated posterior of the model fitted in line 31–32 of the R-code 5.1. To do this we use of the R-INLA function `inla.posterior.sample()`. These samples are then put into the function `inla.posterior.sample.eval()`, which evaluates each sample within a specific function. How this is done can be seen in R-code 5.2.

In lines 2–13 we find the actual function for extracting the weights, which is called `fun`. Lines 3–6 specifies the model results, where `a.x.male` and `a.x.female` stores the result sample values for the spatial effects and `a_male` and `a_female` stores the result samples values of the two δ_g's. Lines 7–8 stores the length of `a_male` in `nn` and half of this length in `n`. The effect of the weights are undone in lines 10–12. In line 16 we find the generating of the 100 samples of the results of the model, which is stored `samp`. These 100 samples are evaluated using the `fun`-function and stored in `xtrct`, which is found in line 19.

In other words, this function takes samples for i_male, i_female, δ_m and δ_f and returns (i_male$/\delta_m$) and (i_female$/\delta_f$) for first half of the values of i_male and i_female, and returns (i_male$\times\delta_m$) and (i_female$\times\delta_f$) for the second half of the values. The first n values and the last n values should be the same up to numerical error. When doing this, we counteract the effect of δ_g for each of the 100 samples, and are left with only the shared spatial effects for men and women. This will further be utilized to make plots of the best model in section 5.4.

Estimated posterior values of the hyperparameters

Table 5.1 shows the posterior summary information for the standard deviations and weight parameters. In particular, it shows the estimated posterior mean, standard deviation and median with corresponding 95% quantiles for the hyperparameters. The values are obtained by extracting the results on precision scale from R from `results3$summary.hyperpar`, and then transformed to standard deviation scale by the use of the R-INLA formulas `inla.tmarginal()` and `inla.zmarginal()`. Here the first formula is used to transform the estimated posterior values of the hyperparameter to standard deviation scale, while the latter is used to extract the summary information about the hyperparameters.

Table 5.1: Estimated posterior summary estimates of the hyperparameters in model 3*

Hyperparam.	Mean	SD	0.025 quant	0.5 quant	0.975 quant
$\sigma_{u_m^*}$	0.071	0.029	0.028	0.067	0.142
δ_m	0.983	0.226	0.609	0.960	1.493
$\sigma_{u_f^*}$	0.101	0.033	0.050	0.097	0.179
δ_f	1.446	0.274	0.977	1.422	2.053
σ_α	1.036	0.159	0.769	1.019	1.393

In this table, we see that the posterior means of the standard deviation of the spatial effect for male and female are 0.071 and 0.101, respectively. The estimated value of their

corresponding standard deviations are 0.029 for men and 0.033 for women. Both these standard deviations look acceptable. The means for both parameters are over twice as big as the standard deviations, which is a good sign, as the range of the values probably should not reach into negative values. Further, the posterior mean of the standard deviation of the age effect is quite close to 1 with a value of 1.036, with a corresponding standard deviation of 0.159. This is also an acceptable value. However, the value of the standard deviation for the age effect is much larger than the standard deviations for the spatial effect. This might indicate that it is more variation for age than for space and that the age effect is less smooth than the spatial effect.

Moreover, we see that the gender-specific spatial weight parameters δ_m and δ_f are equal to 0.983 and 1.446, respectively. The weight parameter for males, δ_m is much closer to 1 than the weight parameter for females. This might indicate that the shared spatial field for incidence and mortality is more similar for males than it is for females. The large value of the mean of the weight parameter for women might indicate that the spatial effects for incidence and mortality are somewhat different. As we have assumed correlation between incidence and mortality, a mean closer to 1 would be preferable. This will be further investigated in section 5.3.

5.1.2 Extending by an unstructured spatial component

Model 8^* includes an extra component v_{di}, in addition to the components we find in model 3^*. In the model comparisons in Etxeberria et al. (2018), this model only performs slightly worse than model 3^*. Because of this, it would be interesting to add $\boldsymbol{v}$ when analysing the Norwegian data to see if it could improve the estimates, or if the performance is worse, as in the article.

As mentioned in chapter 4, the v_{di} component represents the spatially unstructured random effects, which here is symmetrically added. This means that it is added for both disease processes, but it is not gender-specific as $\boldsymbol{u}$ (Etxeberria et al.; 2018). The components $\boldsymbol{u}$ and $\boldsymbol{\alpha}$ are defined in the same way as in model 3^*.

For this model we need columns 14 and 15 from table 2.1; the columns *i_incidence* and *i_mortality*. The column *i_incidence* takes the values of i when the entry in *Cases* is an incident count, meaning as long as *inci_mort* is 1. Similarly, *i_mortality* takes the values $i - 18$ when *inci_mort* is 2. In other words, both *i_incidence* and *i_mortality* takes the values from 1 to 18 for each of the 18 counties.

Putting this model in the LGMs framework, the latent Gaussian field is $\{\delta_m, \delta_f, u^*_{1,1}, \ldots, u^*_{18,1}, u^*_{1,2}, \ldots, u^*_{18,2}, u^*_{19,1}, \ldots, u^*_{36,1}, u^*_{19,2}, \ldots, u^*_{36,2}, , v_{1,1}, \ldots, v_{1,18}, v_{2,1}, \ldots, v_{2,18}, \alpha_{1,1}, \ldots, \alpha_{1,9}, \alpha_{2,1}, \ldots, \alpha_{2,9}\}$. The model has seven hyperparameters $\{\delta_m, \delta_f, \sigma^2_{u^*_m}, \sigma^2_{u^*_f}, \sigma^2_\alpha, \sigma^2_{v_1}, \sigma^2_{v_2}\}$.

Implementing the model in R-INLA

In the `R-code` 5.3, the `formulaBrain8` can be seen in line 8–14. This is just an extension of `formulaBrain3` with the additional components `f(i_incidence, model = "iid", hyper = pcprec)` and `f(i_mortality, model="iid", hyper = pcprec)`. The other components used to implement model 8^* in `R-INLA` are exactly the same as defined in section 5.1.1.

R-code 5.3: Code for impremented the original model 8*

```
1  #THE SAME NEIGHBOURHOOD STRUCTURE AS IN MODEL 3*
2  g = inla.read.graph("nb-inla.txt")
3
4  #PC PRIOR
5  pcprec = list(prec=list(prior="pc.prec", param=c(1, 0.01)))
6
7  #FORMULA FOR THE MODEL 8*
8  formulaBrain8 = Cases ~ -1 + mu + f(i_male, model = "besag2",
9     graph = g, hyper = pcprec, constr = TRUE, scale.model = TRUE) +
10  f(i_female, model = "besag2", graph = g, hyper = pcprec,
11    constr = TRUE, scale.model = TRUE) + f(Agegroup, model="iid",
12    hyper=pcprec, constr = TRUE, replicate=inci_mort) +
13  f(i_incidence, model = "iid", hyper = pcprec, constr = TRUE) +
14  f(i_mortality, model = "iid", hyper = pcprec, constr = TRUE)
15
16 #INLA EXECUTION
17 results8 = inla(formulaBrain8, family="Poisson", data=BrainData10,
18  E=Population, control.predictor=list(compute=TRUE))
```

As in model 3^*, the data used in model 8^*, is only from the period 2014–2018. Therefore, the data used here is also `BrainData10`. Since $\boldsymbol{v}$ is included for both disease processes, we need to specify one component for incidence and one for mortality in the formula. The models are specified to be `iid` since the component is an unstructured effect. The PC priors for the hyperparameters and neighbourhood structure are the same as in model 3^*. The results from this model is stored in `results8`, as seen in line 17–18.

Estimated posterior values of the hyperparameters

Table 5.2 shows all the estimated posterior summary information of the standard deviations and weight parameters for this model. This includes the estimated posterior mean, standard deviation and median with corresponding 95% quantiles. As for model 3^*, the values are found on standard deviation scale.

Table 5.2: Estimated posterior summary estimates of the hyperparameters in model 8*

Hyperparam.	Mean	SD	0.025 quant	0.5 quant	0.975 quant
$\sigma_{u_m^*}$	0.058	0.033	0.017	0.051	0.141
δ_m	0.916	0.256	0.505	0.886	1.500
$\sigma_{u_f^*}$	0.074	0.038	0.024	0.066	0.169
δ_f	1.230	0.321	0.701	1.198	1.953
σ_α	1.036	0.159	0.769	1.019	1.393
σ_{v_1}	0.083	0.032	0.037	0.078	0.160
σ_{v_2}	0.035	0.025	0.006	0.029	0.097

This table show quite different values than the model 3^*. Here, the posterior mean of the standard deviation of the spatial effect for males and females are 0.058 and 0.074,

respectively. These values are smaller and more similar than for model 3^*. However, the standard deviations are much larger compared to the means for this model than the former. Especially for men, the mean is less than twice the value of the standard deviation. Moreover, the estimated posterior value for the weight parameter for women is closer to 1 than for model 3^*, indicating that the shared spatial field might be more similar for this model.

For the standard deviations of the unstructured spatial effect, v_1 indicates incidence and v_2 indicates mortality. The mean and standard deviation for incidence have acceptable values. However, for the mortality, the standard deviation is very similar to the mean, which might indicate a problem with this model. Due to this, improvements could be made to this model as well. Moreover, the posterior values of the standard deviations for the unstructured effect are so small, almost close to zero. This might indicate that these random effects are not that important after all.

5.2 Introducing possible changes to the original models

In the previous section, it was specified that not all the estimated posterior values for the hyperparameters were ideal. This was both due to large means of the standard deviations, as well as large values for the female weight parameter. In this section different changes to the two models from section 5.1, will be presented. This includes fitting different models for the age effect, changing the priors for the hyperparameters and adding an unstructured effect to account for everything not captured by the other model components.

5.2.1 Prior choices

For model 3^* and model 8^* from section 5.1, the priors specified for the hyperparameters were the same as used in the article by Etxeberria et al. (2018). However, the parameters for these priors might be too rigid, as they impose a constraint saying that the probability of the standard deviation of the hyperparameter being larger than 1 should be equal to 0.01. In addition, the posterior values of the standard deviations in both models from the previous section show some difference in the values for the spatial effect and the age effect. Because of this, it might make sense to use the different priors in the model.

Using different priors for the different hyperparameters

One possible improvement to the models presented above might be to change the parameters for the PC priors. Below we find a proposed improvement to the original priors for the hyperparameters.

R-code 5.4: Code for implementing the new set of priors to both model 3* and 8*

```
#PRIOR FOR THE SPATIAL EFFECTS
pcprec1 = list(prec=list(prior='pc.prec', param=c(1, 0.05)))

#PRIOR FOR THE AGE EFFECT
pcprec2 = list(prec=list(prior='pc.prec', param=c(3, 0.1)))

```

```
7
8  #FORMULA AND INLA EXECUTION OF MODEL 3*
9  formulaBrain3b = Cases ~  -1 + mu + f(i_male, model = "besag2",
10    graph = g, hyper = pcprec1, constr=TRUE, scale.model = TRUE) +
11  f(i_female, model = "besag2", graph = g, hyper = pcprec1,
12    constr=TRUE, scale.model = TRUE) + f(Agegroup , model = "iid",
13    hyper = pcprec2, constr = TRUE , replicate = inci_mort)
14
15 results3b = inla(formulaBrain3b, family = "Poisson",
16  data=BrainData10, E = Population, control.predictor =
17    list(compute=TRUE))
18
19 #FORMULA AND INLA EXECUTION OF MODEL 8*
20 formulaBrain8b = Cases ~ -1 + mu + f(i_male, model = "besag2",
21    graph = g, hyper = pcprec1, constr = TRUE, scale.model = TRUE) +
22  f(i_female, model = "besag2", graph = g, hyper = pcprec1,
23    constr = TRUE, scale.model = TRUE) + f(Agegroup, model="iid",
24    hyper=pcprec2, constr = TRUE, replicate=inci_mort) +
25  f(i_incidence, model = "iid", hyper = pcprec1, constr = TRUE) +
26  f(i_mortality, model = "iid", hyper = pcprec1, constr = TRUE)
27
28 results8b = inla(formulaBrain8b, family = "Poisson",
29  data = BrainData10,E=Population, control.predictor =
30    list(compute=TRUE))
```

R-code 5.4 displays an alternative to the original priors for the hyperparameters. Here, two other PC priors are defined. The first one can be seen in line 2 and is stored in `pcprec1`. For this prior, the limitations are slightly more liberal than for the original, by defining the probability of the standard deviation of the hyperparameter being larger than 1 to be equal to 0.05, i.e. $P(\sigma > 1) = 0.05$. The `pcprec1` prior is used to the structured spatial effect in both model 3^* and model 8^*, as well as for the unstructured spatial effect in model 8^*. This can be seen in lines 8–13 and 19–26, where the new formulas for both models can be found.

Further, in line 5 an alternative prior for the age effect is shown. This is also a PC prior, which is stored in `pcprec2`. For this prior, the parameters are specified to be equal to 3 and 0.1, meaning $U = 3$ and $\alpha = 0.1$, i.e. $P(\sigma > 3) = 0.1$. In other words, the probability of the standard deviation of the hyperparameter being larger than 3 is set to be equal to 0.1. The reason for the wider definition of the prior for the age group is that posterior value of the standard deviation of the age effect is significantly larger than for the spatial effects.

5.2.2 Adding an unstructured component for overdispersion

None of the models presented so far includes a component to account for all other unstructured effects which cannot be explained by the specific components. Therefore, we now propose adding an overdispersion component. The purpose is to address unstructured heterogeneity that is not captured by the other model components. For model 3^*, this can be added like this:

$$\text{Model 3*:} \quad \log \lambda_{Iijg} = \delta_g u^*_{ig} + \alpha_{Ij} + z_{Iijg}$$
$$\log \lambda_{Mijg} = \frac{1}{\delta_g} u^*_{ig} + \alpha_{Mj} + z_{Mijg}$$

where z_{dijg} represents the overdispersion, where $\boldsymbol{z} \mid \sigma_z^2 \sim N(\mathbf{0}, \sigma_z^2 \boldsymbol{I}_N)$, with N being the number of rows in the data matrix, i.e. `BrainData10`. Equivalently, it can be added to model 8^*. How this component is added, can be seen in `R-code` 5.5 below.

R-code 5.5: Code for the original model 3* with the added overdispersion component

```
1  #FORMULA FOR THE MODEL 3* with overdispersion
2  formulaBrain3c = Cases ~ -1 + mu + f(i_male, model = "besag2",
3     graph = g, hyper = pcprec, constr = TRUE, scale.model = TRUE) +
4   f(i_female, model = "besag2", graph = g, hyper = pcprec,
5     constr = TRUE, scale.model = TRUE) + f(Agegroup, model = "iid",
6     hyper = pcprec, constr = T, replicate = Inci_mort) +
7   f(z_row, model = "iid", hyper = pcprec)
8
9  #INLA EXECUTION
10 results3c = inla(formulaBrain3c, family = "Poisson",
11  data = BrainData10, E = Population, control.predictor =
12    list(compute=TRUE))
```

The overdispersion component can be seen in line 7 in the `R-code` 5.5. It is added to `R-INLA` by the use of column 16, *z_row*, from table 2.1. This column takes the values of each row, meaning that for the full data set, it takes the values 1 to 6480. For the data set used, `BrainData10`, it takes the values 1 to 648, as this data set is ten times smaller than the full set. This is modelled using an iid model, since it is an unstructured effect.

5.2.3 Changing the age effect model

In model 3^* as specified by Etxeberria et al. (2018) an iid model has been used for the age effect when doing the analysis. The age effect has also been assumed equal for males and females. This small section will look into using other models for the age effect in both model 3^* and 8^*. The code examples shown below will only be for model 3^*, but the same changes can be made to model 8^*. This will be done when comparing all the models in section 5.3.

Introducing the random walk of order 2

A possible change to modelling the age effect is changing the iid model to an second order random walk model (rw2), assuming that effects for neighbouring time-points are similar. Given a time ordered vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)$, a random walk is defined as a model of order r such that, for a given t, α_t only depends on the previous $t-r$ elements (Blangiardo and Cameletti; 2015, pp. 132–134). Hence, the second order random walk is a model of

order 2 defined such that α_t only depends on the previous $t-2$ elements. The second-order increments in the second-order random walk are defined as:

$$\Delta^2\alpha_t = (\alpha_t - \alpha_{t+1}) - (\alpha_{t+1} - \alpha_{t+2}) \sim N(0, \sigma_\alpha^2)$$

where $t = 1, \ldots, T-2$ and σ_α is the standard deviation (Rue and Held; 2005, ch. 3). The second-order increments can be seen as an estimation of a second-order derivative of continuous time function, making the rw2 model a good model for representing smooth curves. The rw2 is flexible because of its stability to addition of a linear trend and its Markov properties, making it computationally convenient (Lindgren and Rue; 2008). The density of an second order random walk is:

$$\begin{aligned}\pi(\boldsymbol{\alpha}) &\propto \frac{1}{\sigma_\alpha^{(T-2)}} \exp\left(-\frac{1}{2\sigma_\alpha^2} \sum_{t=2}^{T-1} (\alpha_{t-1} - 2\alpha_t + \alpha_{t+1})^2 \right) \\ &= \frac{1}{\sigma_\alpha^{(T-2)}} \exp\left(\frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{Q} \boldsymbol{\alpha} \right)\end{aligned}$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)^T$ (Lindgren and Rue; 2008; Rue and Held; 2005). The precision matrix is given as (Rue and Held; 2005, ch. 3):

$$\boldsymbol{Q} = \frac{1}{\sigma_\alpha^2} \begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}$$

Modelling using a rw2 model

The first change done to the age effect, is changing the iid model to a rw2 model. This is the only change made to the model, which is done to the formula part in the `R-code`. In the `R-code` 5.6 this change can be seen in line 5–6. The same change can be made for model 8*, but this is not showed.

R-code 5.6: Code for model 3* with rw2 model for the age effect

```
1 #FORMULA FOR THE MODEL 3* with rw2 model on age effect
2 formulaBrain3d = Cases ~ -1 + mu + f(i_male, model = "besag2",
3    graph = g, hyper = pcprec, constr = TRUE, scale.model = TRUE) +
4  f(i_female, model = "besag2", graph = g, hyper = pcprec,
5    constr = TRUE, scale.model = TRUE) + f(Agegroup, model = "rw2",
6    hyper = pcprec, constr = T, replicate = Inci_mort)
7
```

```
#INLA EXECUTION
results3d = inla(formulaBrain3d, family = "Poisson",
 data = BrainData10, E = Population, control.predictor=
   list(compute=TRUE))
```

Changing to gender specific age effect

The second change done to the model for age effect, is changing from joint age effect to separate age effect for each gender. For this model, we need columns 11 and 12, *agegroup_male* and *agegroup_female*, from table 2.1 in section 2. The *agegroup_male* takes the values of the *Agegroup* if the gender is male, i.e. if *Gender* = 1, and NA otherwise, and the *agegroup_female* takes the value of the *Agegroup* if the gender is female, i.e. if *Gender* = 2, and NA otherwise.

R-code 5.7: Code for model 3* with two separate rw2 models for the age effect

```
1  #FORMULA FOR THE MODEL 3* with two rw2 models on age effect
2  formulaBrain3e = Cases ~ -1 + mu + f(i_male, model = "besag2",
3     graph = g, hyper = pcprec, constr = TRUE, scale.model = TRUE) +
4   f(i_female, model = "besag2", graph = g, hyper = pcprec,
5     constr = TRUE, scale.model = TRUE) + f(agegroup_male,
6     model = "rw2", constr = TRUE, hyper=pcprec) +
7   f(agegroup_female, model = "rw2", constr = TRUE, hyper=pcprec)
8
9  #INLA EXECUTION
10 results3e = inla(formulaBrain3e, family = "Poisson",
11  data = BrainData10, E = Population, control.predictor=
12    list(compute=TRUE))
```

The separate age effect is added to the formula for model 3^* in `R-code` 5.7. In line 5–7 these age effects can be seen, where it can be seen that the first arguments are `agegroup_male` and `agegroup_female`. For the separate age effects for men and women a random walk of order 2 was chosen for both age effects. As for the previous change in the age effect model, these separate age effect for male and female can be used for model 8^* as well.

5.3 Using model choice criteria to choose the best models

In this section, the model changes and improvements suggested in the previous section will be compared to the original models using several model choice criteria. Before starting the comparison of the different models, the model choice criteria will be briefly introduced.

5.3.1 Model choice criteria

For the comparison of the different models, three model choice criteria will be used. These are the Deviance Information Criterion (DIC), the Watanabe–Akaike Information Criterion

(WAIC) and the Logarithmic Score (LS). These are the same criteria used by Etxeberria et al. (2018).

The Deviance Information Criterion (DIC)

The first model choice criterion is the *deviance information criterion* (DIC), introduced by Spiegelhalter et al. (2002). For the use in INLA, this is defined as:

$$\text{DIC} = D(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}) + 2p_D$$

where $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{\theta}}$ are the posterior expectations of the latent effects and the hyperparameters, where posterior means are used for the latent field and posterior mode for the hyperparameters (Gómez-Rubio; 2020, ch. 2). Further, the $D()$ is the deviance, which can be set equal to -2 times the log-likelihood, and p_D is the effective number of parameters, which is defined as the posterior expected deviance (Spiegelhalter et al.; 2002). According to Spiegelhalter et al. (2002), the DIC can be considered as a Bayesian measure of fit, which is penalized by the complexity term p_D.

However, Plummer (2008) emphasizes that the approximations made using DIC only will be valid when the effective number of parameters in the models are significantly smaller than the number of independent observations. Because of this, the use of DIC in disease mapping might not be optimal, as this assumption usually does not hold. In other words, in disease mapping there are often models where p_D is the same order as n. This will further lead to under-penalization of more complex models (Plummer; 2008). Despite of this, the DIC will be used in the model comparisons, but because of its possible limitations, it will be used with care and together with other criteria. This criterion is calculated directly by `R-INLA` when setting `control.compute=list(dic=TRUE)` in the INLA function call.

The Watanabe–Akaike Information Criterion (WAIC)

The *Watanabe-Akaike information criterion* (WAIC) was introduced by Watanabe (2010) under the name widely applicable information criterion. The definition for WAIC is slightly more complicated than for the DIC, but essentially the WAIC also computes the deviance by the use of the log-likelihood of the posterior. In addition, the WAIC can also be seen as an improvement on the DIC for Bayesian models. Moreover, the posterior distribution is used in its entirety, making the WAIC fully Bayesian, and it is asymptotically equal to Bayesian cross-validation (Vehtari et al.; 2017). This criterion is calculated directly in the result part by `R-INLA` when setting `control.compute=list(waic=TRUE)`.

The Logarithmic Score (LS)

The last method *logarithmic score* (LS), is a so-called proper scoring rule (Gneiting and Raftery; 2007). A scoring rule is proper if the expected score is minimized, with respect to the true data-generating distribution, $Y_0 \sim f_0$, if the posterior distribution f is equal to the data-generating distribution f_0 (Held and Sabanés Bové; 2014, pp. 311–312). This score is defined as:

$$\text{LS}(f(y), y_0) = -\log f(y_0) \tag{5.1}$$

Here, $f(y)$ is the predictive distribution, and y_0 are the observed values.

In this thesis we do not do predictions, and therefore we do not have a predictive distribution. In order to calculate the LS, the conditional predictive ordinates (CPO) are used. In INLA the CPOs are calculated using approximate leave-one-out cross-validation. For this case the CPOs are defined as:

$$\text{CPO}_i = \pi(y_i^{obs} \mid y_{-i})$$

where y_i^{obs} are the actually observed values, the y_{-i} denotes the observations y with the i-th component omitted and $\pi()$ is the posterior predictive distribution (Held et al.; 2010, pp. 91–110). In INLA, the leave-one-out cross-validation does not actually leave out observations one by one, but rather approximate the leave-one-out cross-validation. In other words, INLA does not actually rerun the model for each observation, but approximates the resulting density. Further, we get the LS from the CPO like:

$$\text{LS} = -\sum_i \log \text{CPO}_i$$

where we sum over all CPO values and the rest of the definition is equal to the predictive definition in equation 5.1. In `R-INLA` the CPOs are calculated directly by setting `control.compute=list(cpo=TRUE)`. From the CPO values, the LS can be calculated as `LS = -sum(log(results3$cpo$cpo))`.

5.3.2 Comparing models with structured spatial component and age group

In this part, comparisons of the variations of model 3^* will be shown. The criteria shown in the previous section will be used to find the best model. The different prior values for the hyperparameters showed extremely small variations in the model choice criteria. Because the difference where so small, the original priors will first be used to do the comparison of the different models. After getting the best model, the best prior based on the estimated posterior values will be chosen.

Table 5.3: Comparing the different versions of model 3^*

Model	DIC	WAIC	LS
M3^*_{iid}	3092	3103	1552
M3^*_{rw2}	3092	3103	1552
M3^*_{sep}	3109	3120	1561
$\text{M3}^*_{iid,z}$	3078	3084	1548
$\text{M3}^*_{rw2,z}$	**3078**	**3084**	**1548**
$\text{M3}^*_{sep,z}$	3090	3096	1556

Table 5.3 shows a comparison of the various changes introduced to model 3^*, including the original model 3^*. This comparison shows every model with the PC priors with

parameters $U = 1$ and $\alpha = 0.01$. In the table, $M3^*_{iid}$ represent the original model 3^*, $M3^*_{rw2}$ represents the model with a rw2 model for the age effect, and $M3^*_{sep}$ is the model with the separate age effect. Further, $M3^*_{iid,z}$, $M3^*_{rw2,z}$ and $M3^*_{sep,z}$ represent the same models with the overdispersion component, z_{dijg}, added.

From this it can be seen that the model $M3^*_{iid,z}$ and $M3^*_{rw2,z}$ performs equally well. The performance of the iid models for the age effect are almost identical as the rw2 models. However, we will move forward with $M3^*_{rw2,z}$ as this is one of the best performing models and the rw2 model is a smoother model than the iid. This model has an overall good performance in all model choice criteria, implicating that it is probably the best alternative to the model 3^*. In the table, the criteria for this model is highlighted in bold.

The different prior values for the hyperparameters showed extremely small variations in the model choice criteria. For this reason, the estimated posterior mean and standard deviation for standard deviation of the hyperparameters for the $M3^*_{rw2,z}$ are displayed with this prior change side by side with the original prior. This is done in order to check if the prior changes have an impact on the values rather than the model choice criteria. This can be seen in table 5.4.

Table 5.4: Comparing the estimated posterior means and standard deviations of the hyperparameters of the best model with different priors

	PC(1, 0.01)		PC(1, 0.05) &	PC(3, 0.1)
Hyperparameter	mean	SD	mean	SD
$\sigma_{u^*_m}$	0.059	0.030	0.061	0.030
δ_m	0.994	0.250	0.992	0.247
$\sigma_{u^*_f}$	0.083	0.033	0.085	0.034
δ_f	1.319	0.289	1.312	0.287
σ_α	0.288	0.072	0.312	0.084
σ_z	0.119	0.022	0.119	0.022

This table displays the original PC prior with parameters $U = 1$ and $\alpha = 0.01$ on the left and the updated PC priors on the right, with parameters $U = 1$ and $\alpha = 0.05$ for the spatial effect and the overdispersion, and $U = 3$ and $\alpha = 0.1$ for the age effect. For all hyperparameters, it is clear that the estimated posterior means and standard deviations are quite similar for both sets of priors. It is only small difference in the last decimal place for most of the values. Hence, the results are not particularly prior sensitive, which is a positive result. As the results are so similar, the second set of priors, $P(1, 0.05)$ and $P(3, 0.1)$, are chosen for the remainder of the thesis as they are more tailored towards the application.

5.3.3 Comparing the models with the additional unstructured spatial effect

In this part, comparisons of the variations of model 8^* will be shown. The criteria presented in section 5.3.1 will be used to find the best model. The results from this comparison can be seen in table 5.5, with all 6 different combinations of the model. As the best

model from the previous comparison showed improvement with the two different PC priors, we use the same priors when comparing the models in this section. The unstructured and structured spatial effects have PC priors with parameters $U = 1$ and $\alpha = 0.05$. Like in the previous comparison, the PC prior with $U = 3$ and $\alpha = 0.1$ is fitted for the age effect.

In these tables, the indices are changed in comparison to the previous section. This is done in order to make the tables easier to read. Here, the index A takes the values A_i, A_r and A_s for the iid model, the rw2 model and the separate model for the age effect, respectively. Further, the index v_i is added for the unstructured spatial effect, which can be modelled using an iid model. Finally, the z indicated the overdispersion component as previously.

Table 5.5: Comparing the different versions of model 8^*

Model	DIC	WAIC	LS
$M8^*_{A_i,v_i}$	3093	3104	1553
$M8^*_{A_r,v_i}$	3093	3104	1553
$M8^*_{A_s,v_i}$	3109	3121	1561
$M8^*_{A_i,v_i,z}$	3087	3099	1551
$M8^*_{A_r,v_i,z}$	**3079**	**3087**	**1548**
$M8^*_{A_s,v_i,z}$	3090	3098	1556

Out of all 12 different models, the best model can be seen in bold in table 5.5. This is the model 8^* with rw2 model for the age effect, iid models for the unstructured spatial effect and the overdispersion component. In the table it is clear that many of the models have almost equal performance. However, this model has the lowest WAIC and lowest LS, making it the best performing model. This could be debatable since many of the models have almost the same performance, but this model is chosen on the basis of the slightly lower values.

5.4 Results

In this section the results of the two best models from section 5.3 can be seen. Even though model 8^* has one component more than model 3^*, these versions of the models performed so similarly that only the spatial effect of the version of model 3^* will be plotted, as no additional information on the spatial effect was seen in plot of the more complicated model. Further, when talking about model 3^* and model 8^*, it implies the best models from the previous section.

Figure 5.1 shows the average of the samples for the shared spatial effect for the brain cancer incidence and mortality per 100 000 in both men and women. In the figure, we find the shared spatial effect for men on the left and for women on the right. These values are extracted as explained in section 5.1.1. In other words, the plot shows $\exp(\boldsymbol{u}^*)$. Further, the figures show the plots on the same colour scale, with values ranging from 0.85 to 1.15. Both genders are plotted on the same scale as a way to emphasize the differences in the shared spatial effect between the genders.

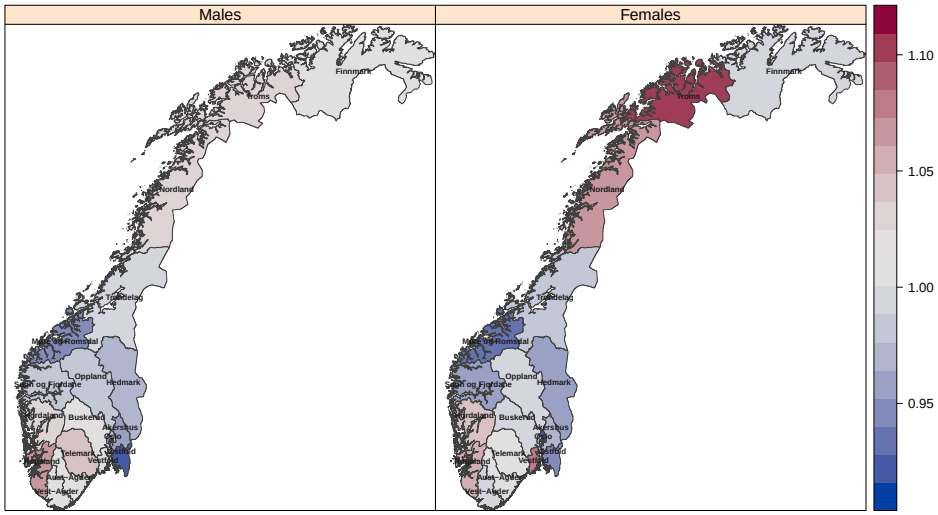


Figure 5.1: Estimated posterior mean of the gender-specific shared spatial random effects for men and women for all counties in 2014–2018

The figure clearly shows a larger variation in the spatial effect for females than for males. This can be seen from the plots, as the spatial effect has values in the entire range from 0.85 to 1.15 for women, whereas for men most of the values are all closer to 1. This suggests that the spatial effect for men is not as strong as for women. It should be pointed out that the range for women is not huge, but significantly larger than for men. Signs of regional trends, with increasing rates from north to south, can be observed for both genders. For both genders, this trend can be observed along the western coast, from Møre og Romsdal to Rogaland. For males, this trend is also slightly apparent in the northern most counties, from Finnmark to Nordland. However, this is not equally noticeable. No general regional trend can be seen for either gender.

However, one trend found in the plots for both genders, is the similarities among bordering counties. In other words, the plots show almost no sharp increases or decreases in rates between neighbouring counties, but rather smooth transitions. This coincide with the assumption of similarities between neighbouring regions. There is, however, some exception from this smoothness as well. For women, there are noticeable exceptions between Trøndelag and Nordland and between Troms and Finnmark, where the transitions between the counties are more rough. Overall, the similarities between the counties are more visible for men than for women.

Table 5.6 presents the estimated posterior summary information of the hyperparameters of the adapted model 3^*, which includes a rw2 model for the age effect and overdispersion. Here we find the posterior mean, standard deviation and median with the corresponding 95% quantiles. For the standard deviation of the spatial effect we have a posterior mean of 0.061 with a standard deviation of 0.030 for men and a mean of 0.085 with standard

Table 5.6: Estimated posterior summary estimates of the hyperparameters in the adapted model 3* with rw2 model for age effect and overdispersion

Hyperparam.	Mean	SD	0.025 quant	0.5 quant	0.975 quant
$\sigma_{u_m^*}$	0.061	0.030	0.018	0.056	0.134
δ_m	0.992	0.247	0.588	0.965	1.554
$\sigma_{u_f^*}$	0.085	0.034	0.034	0.080	0.166
δ_f	1.312	0.287	0.830	1.286	1.952
σ_α	0.312	0.084	0.185	0.299	0.511
σ_z	0.119	0.022	0.079	0.118	0.165

deviation of 0.034 for women. Hence, both means are over twice as large as their standard deviations. The small standard deviations of the spatial effect corresponds to high precisions. This would indicate that we have a smooth spatial effect. Moreover, the standard deviations for the age effect and overdispersion look rather good, with reasonable means and standard deviations significantly smaller than their means. Comparing the standard deviation of the age effect of this updated model with the values of the original model 3^*, this standard deviation is smaller than in the original. Hence, the age effect of the updated model is smoother than the age effect in the original model.

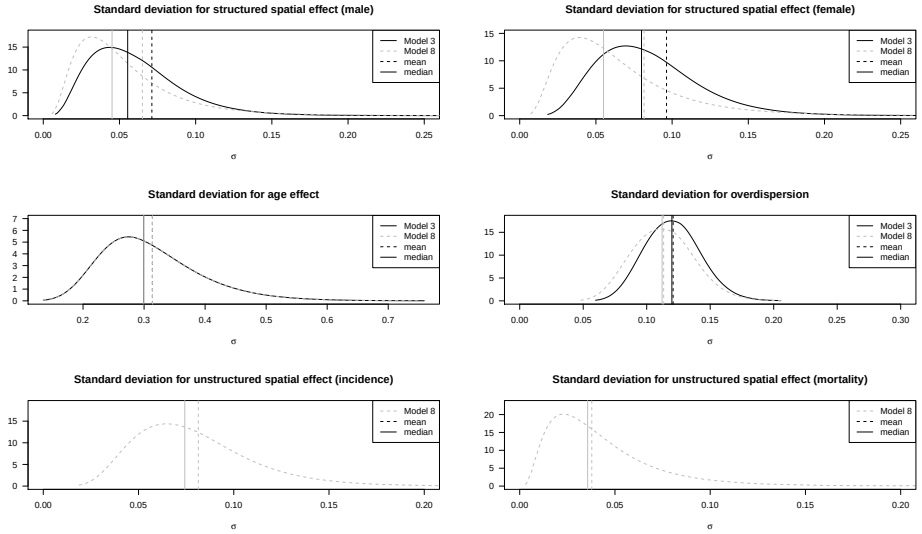


Figure 5.2: Posterior marginal distribution for the hyperparameters of the adapted model 3^* (solid black curves) and the adapted model 8^* (dotted grey curves) on the standard deviation scale, with vertical lines for the posterior means (dotted) and medians (solid).

Further, the estimated means for δ_m is almost equal to 1, indicating that the shared spatial field for men seem to be near identical for incidence and mortality. However, for females, δ_f is reasonably larger than 1, as in the original model from section 5.1.1. Never-

theless, for this improved model the estimated value is closer to 1 than the original model. This indicates that the shared spatial field for women in the improved model is more similar than in the original. As we have assumed that a relationship between incidence and mortality exists, this is a positive result and strengthens the improved model.

In figure 5.2 we find the posterior marginal distributions of the hyperparameters on standard deviation scale. In these plots, the solid black curves represent the model 3^* and the dotted grey curves represent model 8^*. As the model 3^* performed better than model 8^*, this figure is included in order to show some of the differences between the two models. In the bottom two plots, we find the standard deviations of the unstructured spatial effect, v_1 and v_2 , which is only found in model 8^*. In the top row, the posterior marginal distributions for the standard deviation of the spatial effects are displayed and in the middle row we find the standard deviation of the age effect and the overdispersion. Further, in the vertical lines, the posterior means and medians are found, where the means are the dotted lines and the medians are the solid lines, again separated by colour for the two models.

For the hyperparameters for the spatial effect for females, we see that the median and mean lie close together for model 3^*, but has a larger gap for model 8^*. For the spatial effect for males, this distance between the median and mean is similar, with the gap being slightly narrower for model 3^*. For the age effect the means and medians appear to be equal for both models. For the overdispersion the distances between the means and medians seem almost the same, where the distribution of model 3^* is shifted further to the right than model 8^*. However, both curves show signs of being symmetrical. Both the hyperparameters for the structured and the unstructured spatial effect show posterior marginals that are right skewed. For the structured spatial effect, model 3^* has a much more symmetrical shape than model 8^*. For the age effect and the overdispersion, the curves are practically equal in shape for both models.

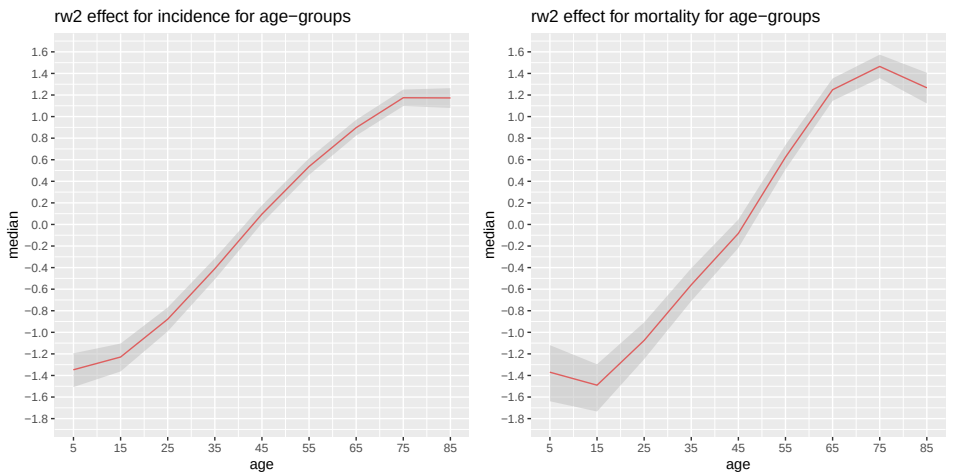


Figure 5.3: Estimated posterior median with 95% quantiles of the rw2 effects for age groups in the adapted model 3^*, for incidence (left) and mortality (right)

In figure 5.3 we see the median of the rw2 effect for the age groups, for both incidence and mortality. As with figure 5.1, this figure is shown only for model 3^* because of the similarities in the results for model 3^* and model 8^*. This is made using `results3$summary.random$Agegroup`, where `result3` is the result of the best model 3^*. Because the same age effect is assumed for men and women, these figures regards both genders. In these plots the red line indicates the median and the grey area surrounding this, is the 95% quantile region. In the plot, the ages 5–85 on the x-axis corresponds to the mid-values of the age groups 0–9, . . ., 80+ in the data, where 5 = 0–9, 15 = 10–19, etc.

In the incidence plot on the left, there is clearly an increasing trend between age group 10-19 and age group 70–79. Between age group 70–79 and 80+ the trend looks like it has slightly flattened out. Looking back at the explanatory plot of the age groups, the trend kept increasing for men, but decrease for women from age group 70–79. This might explain the flattening of the curve. A similar trend can be seen in the mortality plot, on the right, between age groups 10–19 and 70–79. This is a similar pattern as seen in the exploratory plot of the crude incidence and mortality rates shown in figure 2.3 in section 2. This might suggest that the oldest men and women are more likely do die of other causes than of brain cancer.

For both disease processes there is a change in the effect between the youngest children and the children in age group 10–19. For incidence the age effect is increasing in a slower rate than for the the following age groups. For mortality the age effect for the youngest children is higher than for the children in the next age group. This can be seen by the small dip in the curve in the left part of the lines. This trend is also shown in the article by Etxeberria et al. (2018), who points out that most cancer types usually have increasing rates from age group 0–9 to 10–19. The shape of the curve for the youngest age groups is interesting, as there are usually not many cancer types where this trend is present. However, since brain cancer is the second most common cancer type in small children, this trend might be reasonable. The sharp curve for mortality will also implicate that the youngest children are more likely to die of brain cancer than the adolescents in age group 10–19.

Chapter 6

Spatio-temporal modelling of brain cancer in Norway from 1969–2018

The previous chapter investigated the spatial patterns of brain cancer in Norway for the period 2014–2018, which does not allow us to say anything about the temporal variations. However, investigating the temporal variations from 1969–2018 could be equally interesting and important.

Therefore, this chapter will analyse the Norwegian brain cancer data over all time periods, from 1969–2018, using the best of the models presented in the spatial modelling in section chapter 5, with the added time component. For this section, the index t is needed for the modelling. This index represents time and have the range $t = 1, \ldots, 10$, for the 10 time periods in 1969–2018.

6.1 Adding the temporal component and space-time interaction

For this analysis, the entire data set, `BrainData`, is required. From table 2.1 column 8, *Period*, is used in order to account for the time. In this column, 1 represents the period 1969–1973, 2 is 1974–1978, etc. The model used in this section can be described like this:

$$\begin{aligned} \text{Model T:} \quad \log \lambda_{Iijgt} &= \delta_g u^*_{ig} + \alpha_{Ij} + \beta_{It} + z_{Iijgt} \\ \log \lambda_{Mijgt} &= \frac{1}{\delta_g} u^*_{ig} + \alpha_{Mj} + \beta_{Mt} + z_{Mijgt} \end{aligned}$$

where β_{dt} is the time component, which is assumed to follow a multivariate normal distribution $\boldsymbol{\beta} \mid \sigma^2_\beta \sim N(\mathbf{0}, \sigma^2_\beta(\boldsymbol{I}_2 \otimes \boldsymbol{I}_{10}))$. The β_{dt} is, as mentioned, simply added to the best model 3^* from the previous chapter. The only interaction between the spatial and temporal part accounted for, is the overdispersion component, making a quite simple

spatio-temporal model. One could of course use a more complex spatio-temporal model, for instance by adding a more complex space-time interaction. The temporal component is only specific for disease and time. In other words, the same time effect is assumed for both genders, for all counties and all age groups. The model has seven hyperparameters $\{\delta_m, \delta_f, \sigma^2_{u^*_m}, \sigma^2_{u^*_f}, \sigma^2_\alpha, \sigma^2_\beta, \sigma^2_z\}$

The `R-code` 6.1 contains the code for this model. As in all previous models, the neighbourhood structure is the same as first explained in section 5.1.1, and can be seen in line 2 in this code. Further, the priors for the hyperparameters are shown in lines 4–5. Because the best model 3^* from section 5.3 used the two priors `pcprec1` and `pcprec2`, this will be done for this time model as well. The `pcprec1`, with parameters $U = 1$ and $\alpha = 0.05$, will be used for the spatial component and the overdispersion, while `pcprec2`, with parameters $U = 3$ and $\alpha = 0.1$, will be used both for the age and time components.

R-code 6.1: Code for implementing the model with the added time component over all time periods

```
1  #NEIGHBOURHOOD STRUCTURE
2  g = inla.read.graph("nb-inla.txt")
3  #PC PRIOR
4  pcprec1 = list(prec=list(prior='pc.prec', param=c(1, 0.05)))
5  pcprec2 = list(prec=list(prior='pc.prec', param=c(3,0.1)))
6
7  #FORMULA WITH TIME
8  formulaTime = Cases ~ -1 + mu + f(i_male, model="besag2",
9     graph=g, hyper=pcprec1, constr=TRUE, scale.model=TRUE) +
10  f(i_female, model="besag2", graph=g, hyper=pcprec1,
11    constr=TRUE, scale.model=TRUE) + f(Agegroup, model="rw2",
12    hyper=pcprec2, constr=TRUE, replicate=inci_mort) +
13  f(Period, model="rw2", hyper=pcprec2, constr=TRUE,
14    scale.model=TRUE, replicate=inci_mort) +
15  f(z_row, model = "iid", hyper = pcprec1)
16
17 #INLA EXECUTION
18 resultsTime = inla(formulaTime, family = "Poisson",
19  data = data.frame(BrainData), E=Population,
20    control.predictor = list(compute=TRUE))
```

In lines 8–15 the formula for the model is found, which is stored in `formulaTime`. As previously, both -1 and `mu` is added for the intercept, where -1 is added in order to remove the default intercept and `mu` is the pre-specified intercept needed in the model. Further explanation on `mu` can be found in section 5.1.1. The models for the spatial effect are modelled with the `besag2` as before, and can be seen in lines 8–11. The age effect can be seen in lines 11–12 and are modelled using a rw2 model, as this was used in the best performing model in the spatial section. In lines 13–14 the time component can be seen, `f(Period, model = "rw2", ...)`. This is also modelled using a rw2, and uses the same prior as the age effect. Time is also modelled using a sum-to-zero constraint, `constr = TRUE`, for the same reasons as the other components. In addition, `replicate = inci_mort` is used to generate replicates of this model with the same hyperparameters. Here, `inci_mort` defines how the observations are grouped into the

replicated effects, that is, by incidence and mortality. The replicating is done because we assume that the time effect is disease-specific, but the same for both genders. Lastly, in line 15 the overdispersion component is added. This is modelled using an iid model, with `pcprec1` used as the prior.

In lines 18–20, the INLA execution of `formulaTime` is stored in `resultsTime`. Here, the data is specified to be the entire data set `BrainData`. The rest of the arguments are defined in the same way as in chapter 5.

6.2 Results from the spatio-temporal model

Table 6.1 comprises the estimated posterior values of the hyperparameters from this model. These posterior values are the posterior mean, standard deviation and median with corresponding 95% quantiles.

Table 6.1: Estimated posterior summary estimates of the hyperparameters in the spatio-temporal model on standard deviation scale

Hyperparam.	Mean	SD	0.025 quant	0.5 quant	0.975 quant
$\sigma_{u_m^*}$	0.062	0.019	0.032	0.059	0.104
δ_m	0.997	0.163	0.712	0.985	1.352
$\sigma_{u_f^*}$	0.061	0.021	0.029	0.058	0.111
δ_f	0.933	0.184	0.643	0.908	1.361
σ_α	0.407	0.085	0.269	0.396	0.602
σ_β	0.141	0.029	0.093	0.137	0.208
σ_z	0.188	0.008	0.172	0.188	0.205

For this model, the estimated posterior mean of the standard deviation of the spatial effect for males is 0.062 with a standard deviation of 0.019. This is very similar to the values for females, which are 0.061 for the mean with a standard deviation of 0.021. This might indicate that the spatial effects for men and women are quite similar. Further, the standard deviations are significantly smaller than their means, indicating that the model is a good fit. As in section 5.4, these small values of the standard deviations of the spatial effect will correspond to high precisions, indicating a smooth spatial effect. The values for the age effect, the overdispersion and the time effect are also satisfactory. The standard deviations of the time effect and the overdispersion have quite small values, which might indicate that these effects are quite smooth. The standard deviation of the age effect is the largest of the standard deviations and has almost the same value as for the model for the single time period. This might indicate that it is more variation in age than the other components and that the age effect is still not as smooth as the other effects.

One thing worth noticing is that the gender-specific spatial weight parameters for male and female are more similar than for the spatial model of a single time period. These values are very high and close to 1 for both genders, which reinforces our assumption of the spatial field being similar for incidence and mortality. In the spatial analyses the weight parameter for females was larger than 1, which was not ideal. However, now the results for this parameter are more satisfactory.

Figure 6.1 displays the posterior marginal distributions of the hyperparameters of the time model on standard deviation scale. In this figure we find the standard deviations of the spatial effects on the top row, the standard deviaitons of the age effect and the time effect in the middle row and the overdispersion in the bottom row. As for the posterior marginal figure in section 5, the dotted line represents the mean and the solid line represents the median. For all these plots, the lines for the mean and the median lie very close together and close to laying on top of each other. The curves of these marginal distributions are more symmetrical than the ones from the spatial results, even though some of the curves still are slightly right skewed. Overall the shape of these curves look rather good, suggesting that the fitted model might be a good fit.

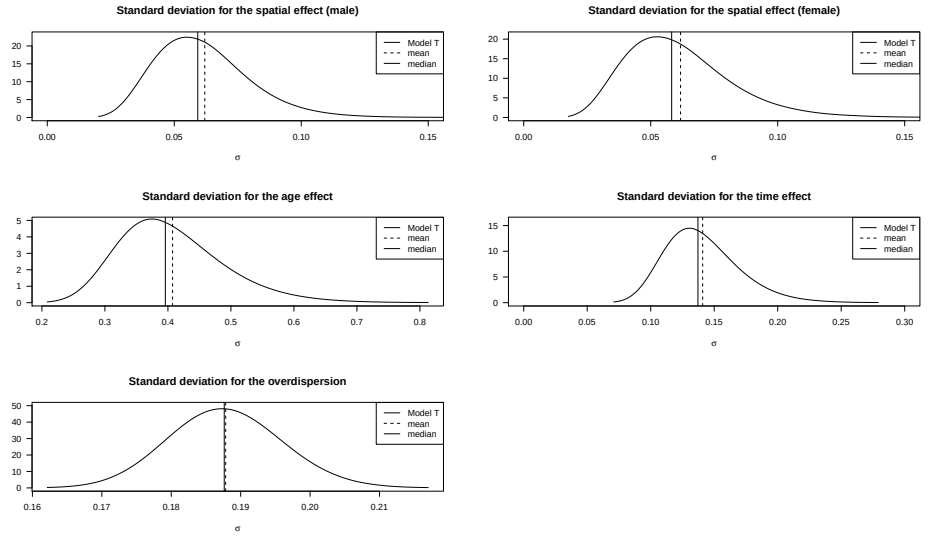


Figure 6.1: Posterior marginals distribution for the hyperparameters of the spatio-temporal model on the standard deviation scale, with vertical lines for the posterior means (dotted) and medians (solid).

Figure 6.2 displays the posterior mean of the gender-specific shared spatial effect for both men and women. This values are extracted from the result in the same way as explained in section 5.1.1, in order to only show the spatial effect without the gender-specific spatial parameter δ_g. As mentioned before, this parameter is the only parameter separating the spatial field for incidence and mortality. These plots show clear signs of a regional trend, which is increasing from north to south. This can be seen in both plots from Finnmark and all the way down to Aust-Agder and Vest-Agder. For men this trend is more noticeable, with only one clear exception in Trøndelag. Apart from Trøndelag, we see that the effect is gradually increasing from very low in Finnmark to quite high in Vest-Agder. For women, the trend is not equally visible, but it looks as if the effect is increasing from Finnmark to Aust-Agder, with slightly lower effect in Buskerud and Telemark. Interestingly, one of the lowest effects are found in Østfold for both genders, which is one of the few thing that corresponds to the explanatory plots in section 2.

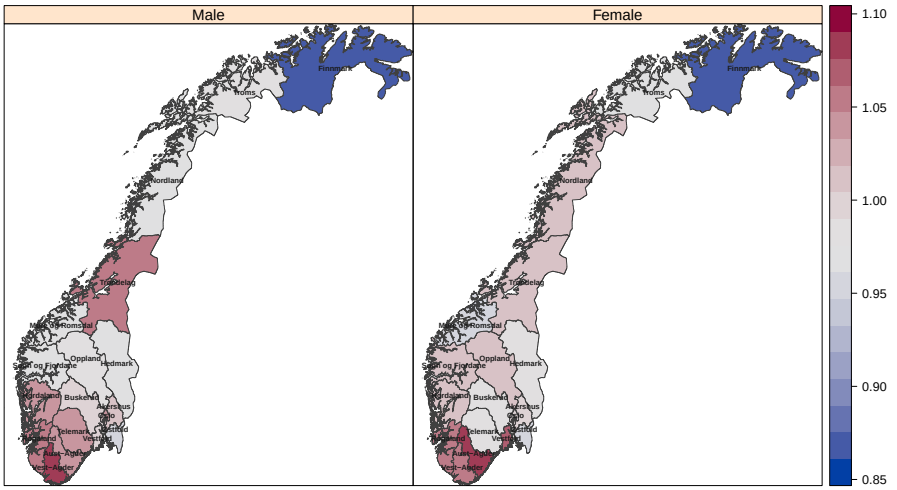


Figure 6.2: Estimated posterior mean of the gender-specific shared spatial random effects for the spatio-temporal model, for men and women for all counties.

Moreover, the similar effects between neighbouring counties is very visible in these plots. This can be seen be the smooth colour changes between the neighbouring counties. Of course, Trøndelag stands out here as well for men, as it did for the regional trend. However, most of the other counties share this pattern. There is also a big increase in the effect for males between Finnmark and Troms, seen as a sharp transition from one county to the other.

Figure 6.3 displays the rw2 effect for the age groups, for incidence on the left and mortality on the right. As in the age effect plot in the previous section, the red line represents the estimated posterior median of the age effect for each age group and the grey area is the corresponding 95% quantile region. On the x-axis the 9 age groups are shown, denoted by the mid-values of each age group, i.e. 5 is age group 0–9, 15 is age group 10–19, etc. This plot is made from `resultsTime` in the same way as previously.

In the plot for incidence, the effect is clearly increasing between age group 10–19 and age group 60–69. Between age group 60–69 and 70–79 the curve is still increasing, but at a much slower rate than in the previous age groups. This is followed by a decrease in the effect between the two oldest age groups. This is a different result than the one from the spatial analysis from 2014–2018, where the effect only flattened out between these two age groups. This is quite interesting and might indicate over time, the effect is lower for the oldest age group, but in 2014–2018 there were perhaps a slight increase in the disease for the oldest men and women. Further, between age group 10–19 and 80+, the trend for mortality is almost identical to the trend from 2014–2018, with a steady increasing effect that decreases between the oldest age groups. This might indicate the the mortality trend has been overall stable, or perhaps the period 2014–2018 coinciding so well could just be coincidental, and choosing another period might not fit so well with the overall trend. The

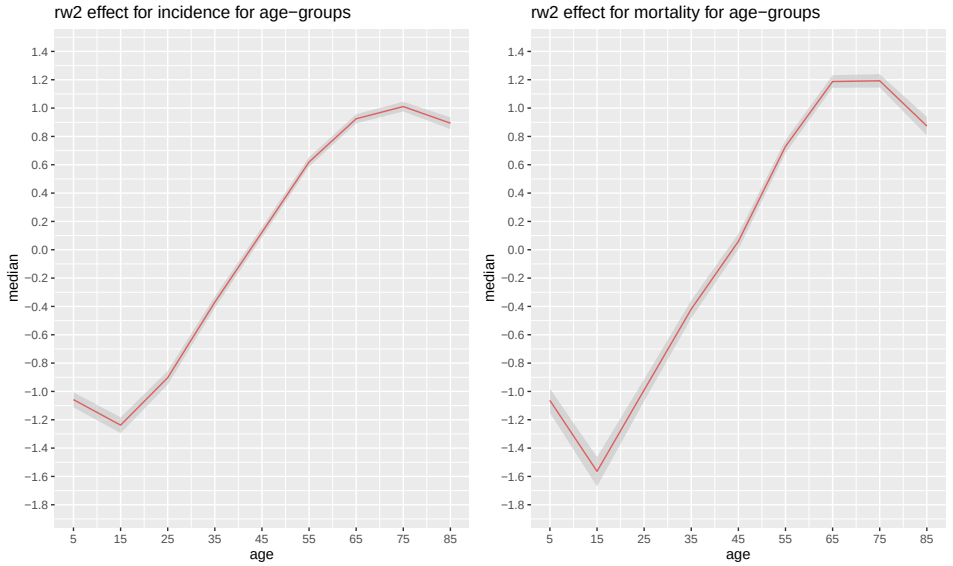


Figure 6.3: Estimated posterior median with 95% quantiles of the rw2 effects for age groups, for incidence (left) and mortality (right)

decreasing trend in both incidence and mortality between the to oldest age groups are quite similar to the ones displayed by Etxeberria et al. (2018).

Further, the sharp decrease in the mortality effect between the two youngest age group is more noticeable in this figure than in the one from 2014–2018. However, both curves have the same shape, with the higher effect for the youngest children. Moreover, the trend between these age groups in the incident effect is slightly different in this plot compared to the plot in the previous section. For 2014–2018 the effect was increasing slowly between these age groups, whereas for this plot, the effect is decreasing between age group 0–9 and 10–19. This v-shaped pattern is similar to the rates shown by Etxeberria et al. (2018), which, according to them, is not a common pattern in other cancer types.

Overall, the 95% quantile band of this model is much narrower than for the single period model in the previous chapter, indicating less uncertainty in the effect. Moreover, the range of the age effect for both incidence and mortality for the single period model is slightly larger than for this model. This might indicate that is is more variability between the age groups in the single period model than in this model.

Figure 6.4 shows the estimated posterior median with corresponding 95% quantiles for the time effect. In the left plot the time effect for incidence is shown, and in the right we see the mortality. These time effect are naturally plotted against the time, which are shown in the x-axis. In the plot, the mid-values of the time periods are shown on the x-axis, i.e. 1971 corresponds to the period 1969–1973, 1976 corresponds to period 1974–1978, etc. Because the same time effect is assumed for men and women, these figures regards both genders. In these plots the red line indicates the median and the grey area surrounding this, is the 95% quantile region.

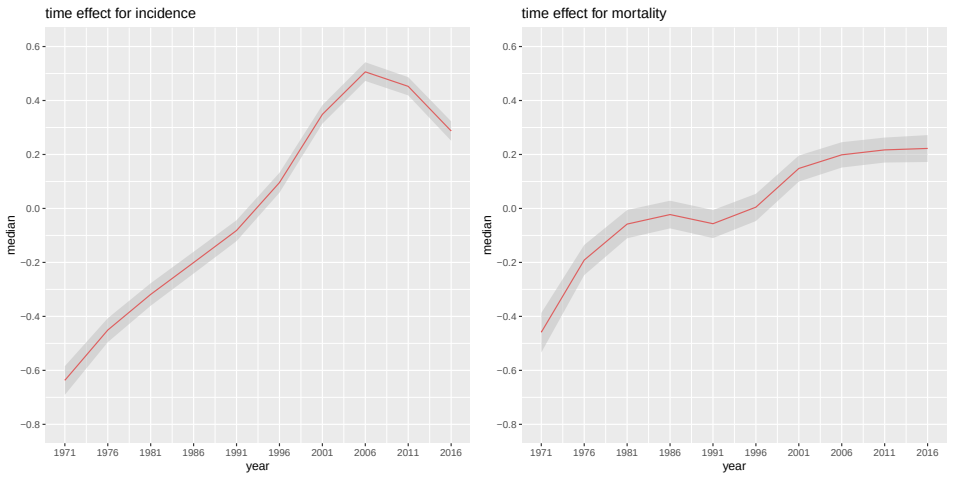


Figure 6.4: Posterior median with corresponding 95% quantiles of the time effect for both incidence and mortality.

In the time plot for incidence the effect is steadily increasing between the first and the eighth period, meaning from 1969–2008. However, from 2008 and until 2018 the effect is gradually decreasing. This might correspond to the decrease seen in the last and second to last period in the time plots in section 2. In 2011 Storstein et al. remarked that the incidence rates in Norway were increasing, which correspond to the increasing effect up to 2008 in this incidence plot. From this plot, it could look the the effect is decreasing, as it has decreased the last couple of time periods. However, it is hard to say if it will continue to decrease, since we do not know how the disease will evolve in the future.

For the mortality on the right side of the figure, the trend is slightly different. Overall the effect is increasing from 1969 to 2018, but it has a very different curve than the incidence. Here, we see a steep increase between 1971 (1969–1973) and 1986 (1984–1988), followed by a slight decrease to 1991 (1989–1993). Then the effect is increasing to 2006 (2004-2008), followed by a flattening of the curve. This suggests that the death rates might have started to stabilise in the last ten years, but it might also just be a saddle point as in period 1989–1993.

Chapter 7

Discussion and Summary

The main novelty of this thesis is both to understand more about the geographical distribution of brain cancer in Norway in the period 2014–2018, as well as understanding more about both its geographical and temporal distribution in the entire 1969–2018 period. The high correlation between incidence and mortality in brain cancer has been taken advantage of, allowing us to borrow strength from both disease processes.

The main results in this thesis are separated into the spatial results from the recent period 2014–2018 and the spatio-temporal results from the entire period 1969–2018. This is done because one of the main interests was to apply the methodology presented by Etxeberria et al. (2018) for Norwegian data. However, since we could only use the recent period when doing this, we also wanted to extend the models by adding a temporal component and space-time interaction. As mentioned earlier, the entire data set could not be used to apply the Etxeberria et al. (2018) methodology, as this is based on aggregating their data over all time periods. Aggregating was not possible in our case as the time periods and age groups were given on different scales, causing a problem with the population counts with the aggregation.

For the main spatial models for the period 2014–2018, we compared several models with and without an unstructured spatial effect using three different model choice criteria, i.e. DIC, WAIC, LS. The best versions of these comparisons were named model 3^* and model 8^*. Model 3^* included a structured shared spatial component scaled by a outcome-specific weight parameter, an age effect and an overdispersion component, and model 8^* had the same components as model 3^*, but with an added unstructured spatial effect. In both these models the outcomes are modelled jointly and then linked through a shared spatial effect, which is allowed to vary in strength by using the outcome-specific weight parameter, δ_g. The estimated posterior values of δ_g were given for both males and females, and a value close to 1 would indicate that the spatial effects for incidence and mortality are similar. Most of the results for both models were similar enough that presenting them for both models seemed redundant. Model 3^* had also slightly better model choice criteria performance, therefore the results were only presented for this model.

For the structured spatial effect in all models, we have assumed that the neighbouring

counties are correlated in such a way that we can borrow strength from the neighbouring regions in order to distinguish underlying geographical patterns. The assumed neighbouring patterns were hard to notice in the explanatory plots in the data section, but in the results from the spatial analysis the effects between neighbouring counties have clearly smooth transitions. Overall, for the period 2014–2018, the spatial effect was stronger for females than for males. However, the range of effects for both genders were not particularly big, making quite a flat effect nevertheless. The small range of the effects may be due to shared spatial effect, which is assumed equal for incidence and morality. The reported incident cases in the period 2014–2018 were larger for females than males, but the reported deaths in the same period were larger for males. This opposite trend in the reported cases might cause the flatness in the shared spatial effect for both genders.

Another interesting result from model 3^* are the age effect results. Here, a change of rates between the youngest age groups was observed. For incidence this was seen by a more gentle increase between the two youngest age groups than the following age groups, and for mortality the effect was decreasing between the same two age groups. However, in the modelling of the entire period 1969–2018, this decrease was seen in both incidence and mortality. This trend is similar to the trend found by Etxeberria et al. (2018), which is interesting and might indicate that the trend from the north of Spain and Norway are quite similar. This might be reasonable since brain cancer is the second most common cancer type in small children in both countries. Further, the curve observed for mortality might implicate that the youngest children are more likely to die of brain cancer than the adolescents in age group 10–19. Moreover, the mortality effect showed a decrease between the two oldest age groups. Brain cancer has a 5 year relative survival rate of around 60 percent for men and 75 percent for women in Norway today (Cancer Registry of Norway; 2019). These are not terrible survival odds, but this survival rate includes both non-malignant and malignant brain tumours, where the latter has a much lower survival rate than the former. As a result of this, perhaps a big part of the patients in the 80+ group who die of brain cancer have a malignant type while those diagnosed with the other type rather die of other causes. Both the results for the youngest and oldest age groups are interesting and the reason for the changes in the curves could possible be further explained by an epidemiologist.

The second analysis performed in this thesis, was the spatio-temporal modelling over all time periods. This model showed much better estimations of the hyperparameters than the sole spatial models. This could be seen by the standard deviations being smaller than in the spatial analysis and the weight parameter for females being much closer to 1. The temporal component and the spatio-temporal interaction used in this model made quite a simple model. Due to the short period of time this thesis had to be completed (one semester), there was not enough time to compare other models with for example more complex space-time interaction and choose the best one. However, in future work this might be a natural extension of this spatio-temporal approach.

In the time model, the results were quite interesting. Firstly, the shared spatial effect for men and women were very similar. The standard deviations of the spatial effect for both genders were practically identical and both outcome-specific weight parameters were close to 1. For men, the mean of the weight parameter was 0.997, indicating that the spatial effect for incidence and mortality were almost equal. For women, this value was 0.933,

indicating a slightly stronger effect for mortality than for incidence. Secondly, the shared spatial effect for this model showed a clear regional trend, from north to south. This could be seen in the figures by the clear increasing effects from Finnmark in the north to Vest-Agder in the south. In the introduction, it was mentioned that the highest incidence was found in northern Europe. By World standard rates, it could be seen that Norway in 2018 had slightly lower rates than the rest of Europe, but significantly higher than the World average. It is interesting that Norway, which lies in northern Europe, had a lower rate than Europe as a whole, and maybe the low rates in northern Norway can explain this. As the only known causes of brain cancer are genetics and high exposure to ionized radiation, it is hard to say anything about the reasons for the trend seen in Norway. This could however possibly be explained by an epidemiologist, which may be a natural next step further on.

The second interesting result from the spatio-temporal section, is the temporal effect. For incidence it was shown that the time effect is strictly increasing up to 2008, before it started to decrease to 2018. This increase in incidence trend can be linked to the claim by Storstein et al. (2011) presented in the introduction. In 2011 they said that the incidence of Norwegian brain cancer is increasing and that the number of cases per year has almost doubled since 1980. In the time trend, we see an steeply increasing curve from 1980 to 2008. Storstein et al. (2011) explained the increase by an increasingly older population in Norway and better diagnostic tools. After 2008 in our time effect plot, the trend is decreasing. As mentioned earlier, it is hard to say if this is the start of a decreasing trend or just a bump in the general increasing trend. The Norwegian population is probably still getting older, as the age wave of the so-called boomers is only in the beginning. Because of this, it might be strange that the time trend is decreasing, and might be explained by the second reason mentioned by Storstein et al. (2011), i.e. diagnostics. In other words, improvements done to diagnostics today may not be as substantial as they were in the past. This might explain the decrease in the trend. However, I have not found any literature to support this claim.

Moreover, the mortality time trend also showed an increasing trend from 1969–2018, but not as dramatic as the incidence. As mentioned previously, the trend for mortality showed a small saddle point in the early 1990s, followed by an small increase, which seemed to be flattening out for the last two time periods. The flattening of the curve might be explained by the continuous medical advancements in cancer treatments. However, since this type of cancer is quite low survival rate and the incidence show signs of a decrease, this might also explain the flattening of the curve for mortality. In other words, since incidence and mortality are assumed linked in this thesis, the decrease in incidence would probably cause a decrease or flattening in mortality.

One issue found in the thesis, was the large value of the weight parameter for females in the sole spatial model. As the assumption was that the shared spatial field between incidence and mortality was very similar, the value of 1.312 indicated quite a large difference in the spatial effect for incidence and the spatial effect for mortality. However, the results were not too outrageous, which meant that the results could still be presented as the shared spatial field. In the plot of the spatial effect the smaller range of the spatial effect for men than women was mentioned. This might be explained by this difference in the female spatial effects for incidence and mortality. Moreover, this small issue with the weight parameter was gone in the spatio-temporal modelling. As the spatio-temporal

model takes in more data than the spatial model, the problem with the weight parameter might just be caused by the small amount of data used.

Further future work would be to investigate alternative models for the spatial structure. The way the `besag2` is defined, i.e. $(\delta_g \boldsymbol{u}, \boldsymbol{u}/\delta_g)$, makes interpreting the results a bit confusing. The interpretation of the weight parameter is quite complex, because of the weights being multiplied to both parts of the results. The interpretation of this is not completely intuitive, as it might had been with a model with weight parameter appearing for only one outcome. As mentioned in section 3, one alternative could be to use a model, where only the second outcome in the result is scaled by a weight parameter, i.e. $(\tilde{\boldsymbol{x}}, a\tilde{\boldsymbol{x}})$. Such a model would make the interpretation of the results somewhat easier, as the results for mortality, in this case, would only be the results for incidence scales by a factor $a > 0$. However, such a model does not exist pre-specified in `R-INLA`. Another interesting model could be to assume a separate `besag` or ICAR model for incidence and mortality, but group them using a correlation parameter (see Riebler et al. (2012) for an application of correlated random walk models).

Summing up, the models in this thesis include gender-specific, age-specific and time-specific components in the analyses of the geographical and temporal distribution of the disease. The assumed relationship between the two disease processes are beneficial for investigating the amount of the spatial patterns common for both disease processes and the amount specific for each one. However, the models presented are not perfect and show some problems, as specified above. This is especially true for the sole spatial model, while the spatio-temporal model appear to quite good for the purpose of this thesis. Future work would be beneficial in order to improve the results.

Bibliography

American Cancer Society (2020). What causes brain and spinal cord tumors in adults?, *American Cancer Society* .

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics* **43**(1): 1–20.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*, John Wiley & Sons.

Cancer Registry of Norway (2019). Cancer in Norway 2018 - Cancer incidence, mortality, survival and prevalence in Norway, *Cancer Registry of Norway* .
URL: *https://www.kreftregisteret.no/globalassets/cancer-in-norway/2018/cin2018.pdf*

Colonna, M., Grosclaude, P., Faivre, J., Revzani, A., Arveux, P., Chaplain, G., Tretarre, B., Launoy, G., Lesec'h, J. M., Raverdy, N., Schaffer, P., Buémi, A., Ménégoz, F. and Black, R. J. (1999). Cancer registry data based estimation of regional cancer incidence: application to breast and colorectal cancer in French administrative regions., *Journal of Epidemiology & Community Health* **53**(9): 558–564.

Etxeberria, J., Goicoa, T. and Ugarte, M. (2018). Joint modelling of brain cancer incidence and mortality using Bayesian age- and gender-specific shared component models, *Stochastic Environmental Research and Risk Assessment* **32**(10): 2951–2969.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I. and Bray, F. (2018). Global Cancer Observatory: Cancer Today., *International Agency for Research on Cancer.* . Visited on: 2020-04-19.
URL: *https://gco.iarc.fr/today*

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American statistical Association* **102**(477): 359–378.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*, Chapman and Hall/CRC Press, London, United Kingdom.

Held, L., Natário, I., Fenton, S. E., Rue, H. and Becker, N. (2005). Towards joint disease mapping, *Statistical Methods in Medical Research* **14**(1): 61–82.

Held, L., Rue, H., Morgenstern, V. and Becker, N. (2006). Statistical extrapolation of background incidence for process quality assurance in mammography screening, *Technical report*, University of Munich, Department of Statistics.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference: Likelihood and Bayes*, 2014 edn, Springer Berlin Heidelberg, Berlin, Heidelberg.

Held, L., Schrödle, B. and Rue, H. (2010). 'Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA' in Kneib, T. and Tutz, G (ed.), *Statistical modelling and regression structures*, Springer, pp. 91–110.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations, *Scandinavian Journal of Statistics* **35**(4): 691–700.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*, Springer Science & Business Media.

Martinez-Beneito, M. A. (2013). A general modelling framework for multivariate disease mapping, *Biometrika* **100**(3): 539–553.

Martino, S. and Riebler, A. (2020). 'Integrated Nested Laplace Approximations (INLA)' in N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels (ed.), *Wiley StatsRef: Statistics Reference Online*, American Cancer Society, pp. 1–19. **URL:** *https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08212*

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison, *Biostatistics* **9**(3): 523–539.

Riebler, A., Held, L., Rue, H. et al. (2012). Estimation and extrapolation of time trends in registry data—borrowing strength from related populations, *The Annals of Applied Statistics* **6**(1): 304–333.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*, CRC press.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2): 319–392.

Savage, N. (2018). Searching for the roots of brain cancer, *Nature* **561**(7724): 50–51.

Simpson, D., Rue, H., Riebler, A., Martins, T. G. and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors, *Statistical science* **32**(1): 1–28.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4): 583–639.

Storstein, A., Helseth, E., Børge Johannesen, T., Schellhorn, T., Mørk, S. and van Helvoirt, R. (2011). Høygradige gliomer hos voksne, *Tidsskrift for den Norske Legeforening* **131**(3): 238–241.

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Statistics and Computing* **27**(5): 1413–1432.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* **11**(116): 3571–3594.

Appendix

7.1 Challenges along the way

Working with this thesis, I've encountered a few challenges out of my control. These are presented below.

7.1.1 Challenges due to the COVID-19 pandemic situation

This year the world has become the victim of a pandemic, namely the COVID-19 virus. On February 26th the virus reached Norway, with the first detected case in Tromsø. Approximately two weeks later, Norway went into sort of a shutdown. The government introduced several interventions in order to be able to control the virus, one of these was the closing of the universities. This meant that I, along with all my fellow master degree students, lost access to my reading room place and had to work with the thesis from home.

In the beginning this turned out a bit difficult. Being a person who likes to separate work from leisure, I've become used to going to school at the same time every day and doing my work there, before going home and having free time. However, now I had to work with my thesis in the same room I use to relax. This required some adjustment, since there was no environmental change between work and other things. At school I have a large desk and a comfortable chair in a quiet room, while in my small apartment, I only have a tiny dining table and dining chairs. This is not an optimal working environment for my taste, but after some time I adjusted to the situation just fine.

In addition to this, all the communication with my supervisor now had to be done by Skype. This was also an adjustment in the beginning, as we used to meet almost weekly in person. Not all supervision is equally easy over Skype, but after some time we found a way that worked.

All in all, apart from a few bumps in the road in the beginning, the new everyday life has grown on me and it has been a learning experience to deal with these unexpected circumstances.

7.1.2 Discovering errors in the software

The second challenge I encountered in this thesis was an error in the programming software `R`. As explained in the thesis, all my results are produced using `R-INLA`. Within this package, I used a model called `besag2` in order to model the structured spatial effect in the models. However, when starting to analyse the first results made using this model, we discovered that the precisions for spatial effect were extremely small. They were so

small that we suspected something was wrong. At first I only looked for errors in my own code, trying to figure out if the low precisions were made by me somehow. Yet, when I found no errors in my part of the code, we started suspecting the `besag2`. This was reinforced when I tested the model using the regular `besag` model instead of `besag2`, and got much larger, and more normal, precision values. When comparing the regular `besag` to the weighted version, `besag2`, the differences in precisions ranged from $1.75 \cdot 10^{-3}$ for the `besag2` to around 1.5 for the `besag`. The difference should not be that large, so we reached out to Håvard Rue, the creator of `R-INLA`, who found an error for the `besag2` model. Therefore, I had to wait for him to fix the errors in the model before I could continue my analyses.

Table 7.1: Estimated posterior values of the hyperparameters in model 3* before the fix

Hyperparam.	Mean	SD	0.025 quant	0.5 quant	0.975 quant
$\tau_{u_m^*}$	0.00175	0.00057	0.00083	0.00169	0.00302
δ_m	0.97015	0.00251	0.96507	0.97020	0.97495
$\tau_{u_f^*}$	0.00161	0.00053	0.00073	0.00156	0.00277
δ_f	0.97070	0.00305	0.96416	0.97095	0.97606
τ_α	1.11285	0.34314	0.56623	1.07263	1.90328

In table 7.1 we see the precision values before the fix of the `besag2` model in `R-INLA`. And in figure 7.1 the standard deviation for the hyperparameters is shown, when using the errored `besag2` model. Here we see the spatial components have standard deviations with means around 24, which corresponds to variances of 576. This is very big and not ideal, as it probably suggest that something is wrong; which it was.

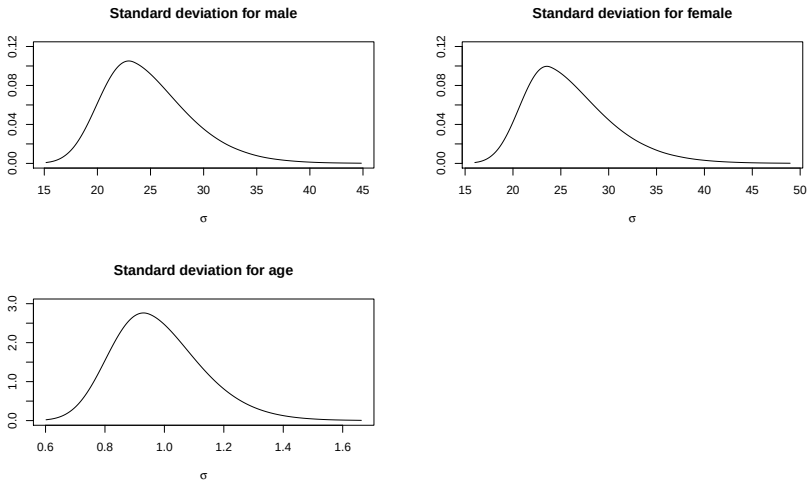


Figure 7.1: The standard deviation of the flawed besag2 model

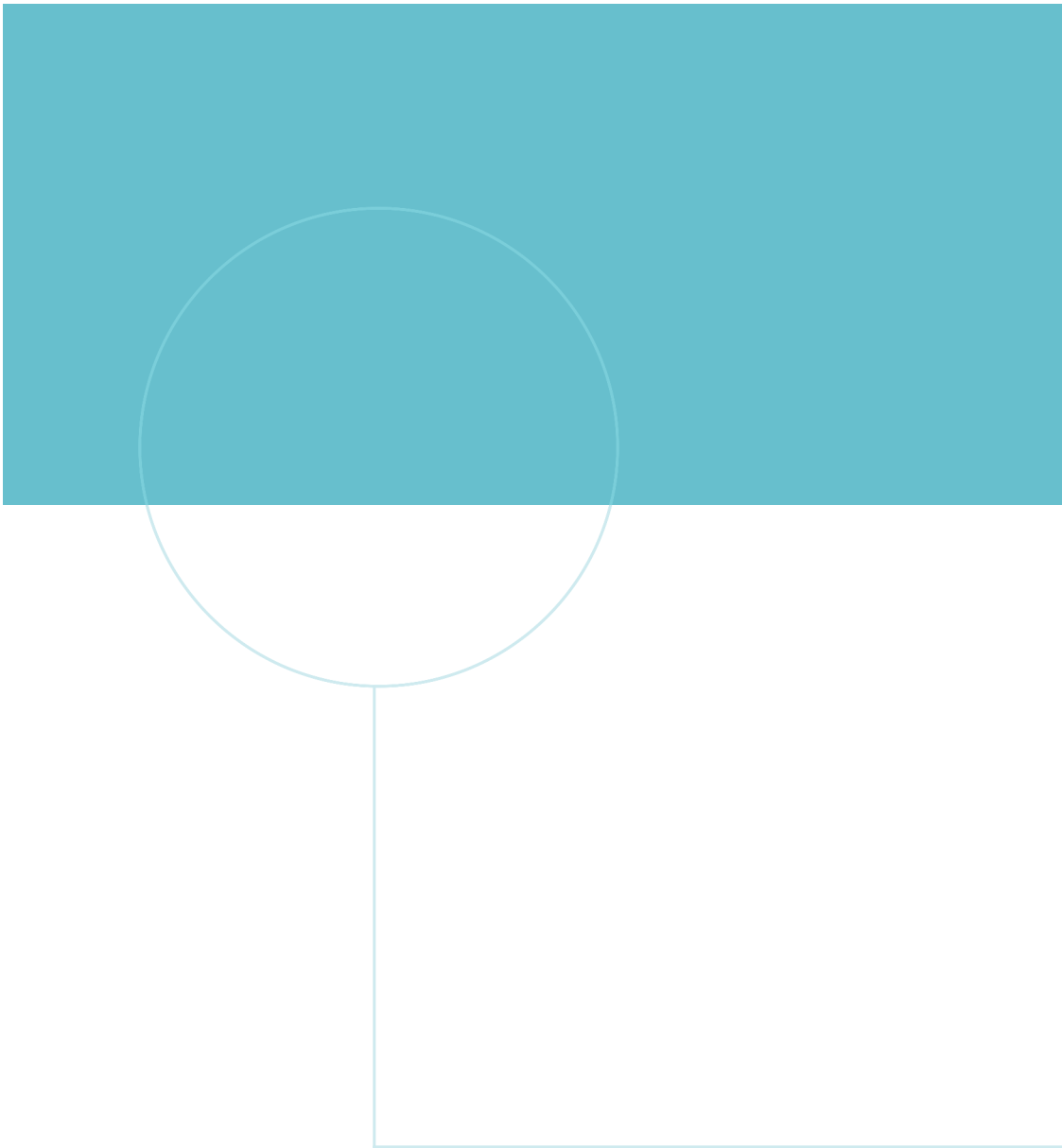
7.2 Learning experiences

As mentioned in the preface, this thesis is only a 30 point master thesis. Before this thesis, I had one reading course, which introduced me to Bayesian inference, spatial statistics and INLA. Before this, all other courses taken in my five year period at NTNU, have not been as relevant for my thesis and more relevant for my general understanding of statistics.

As most of this thesis was all new to me in the fall 2019, I have learned a lot by working on it. However, it took some time getting comfortable with everything being used in the thesis.

Before this fall, I had not heard of INLA, even though I thought I was quite familiar with `R`. It was especially the INLA part that took some time to learn, both theoretical background, but also to understand every component needed to be used in `R-INLA` for the thesis.

Working with this thesis has taught my a great deal about spatial statistics, Bayesian inference, priors and INLA. I have probably almost learned something new every day working with the thesis.



NTNU

Norwegian University of
Science and Technology