Modeling Complex Data with Spatial Correlation, Zero-inflation and Overdispersion: the Combined Modeling Approach

Thomas Neyens

Promotor: Prof. dr. Christel Faes
Co-Promotor: Prof. dr. Geert Molenberghs

Acknowledgements

This work would not have been accomplished without the direct or indirect help of many people. While I cannot mention everyone, I would like to thank some of them in particular.

Prof. dr. Christel Faes, my promotor, has always given me a lot of guidance, support and understanding. Christel, thank you for the numerous occasions when we were able to discuss statistical issues, but also for those times we just chatted about real life. I am very grateful to have had you as a mentor, as you had perfect insight in my capabilities, knowing that my background did not lay in pure mathematics. I would like to thank my co-promotor Prof. dr. Geert Molenberghs too, to find the time to respond to my large amount of questions, in person or by e-mail, at any time of the day or night. Also the other members of the jury, Prof. dr. Marc Aerts, Prof. dr. Clarice Demétrio, Prof. dr. Helena Geys, Prof. dr. Andrew Lawson, Prof. em. dr. Noël Veraverbeke and Prof. dr. Geert Verbeke are greatly acknowledged for their help in any direct or indirect way throughout the years.

Of many of the statisticians I've worked with on a scientific level, dr. Wondwosen Kassahun has been the one I owe the most. In the course of writing five papers together, we became close friends. I remember how hard we laughed when calculations told us that our too ambitious simulation study would take 80 years. As a matter of fact, I'm having a little laughter attack when writing this.

It has also been a pleasure to work with so many other wonderful people throughout the years. I would especially like to thank my fellow (former) assistants Candida, Kim, Lisa, Leen, Ruth and An, since I believe we worked very well together as a team. Kim and Lisa, thank you for being the great room mates you were! A very special and warm

"Thank You!" goes to Candida, who during those 6 years has become one of my closest friends. From the times we both had an office in Block C to just a few weeks ago, when we discussed "Het semi-conditioneel tweezijdigheidsbewijs" during our little walks, and all the statistical interventions in between, those were moments I valued a lot and will never forget.

Paulien, thank you for the love you give me every day again and again, for the beautiful times we had on our trips to all corners of the world, for the Wallander evenings and for so many other things. Thank you for being you!

Lastly, I would like to thank my family. My sister is the sweetest sister in the world and the only person I know that genuinely laughs every single time I (try to) make a joke. My mother has been the most caring and loving woman in my life, giving me tender but good advice whenever I needed it. And finally, I would like to thank my father: I still remember how we used to watch nature documentaries when I was just a little child, or when you explained me things about the stars and planets. I strongly believe that those moments were pivotal in my early life and that they are most likely the reason why I am standing here today. I hope that when I ever become a father, I will become one like you were for me!

Thank you all very much!

Thomas Neyens
Diepenbeek, June 26 2015

Publications

The material presented throughout this thesis is based on the following publications:

Neyens, T., Faes, C., and Molenberghs, G. (2012) A generalized Poisson-gamma model for spatially overdispersed data. *Spatial and Spatio-temporal Epidemiology*, **3**, 185–194.

Neyens, T., Lawson, A. B., Kirby, R. S., and Faes, C. (2015a) The bivariate combined model for spatial data analysis. *Statistics in Medicine*, *under revision*.

Neyens, T., Faes, C., and Molenberghs, G. (2015b) Hierarchical Bayesian inference using integrated nested Laplace approximation for the combined model: a simulation study. *Computational Statistics and Data Analysis*, *submitted*.

Neyens, T., Faes, C., and Molenberghs, G. (2015c) The zero-inflated combined model for spatial data analysis. *Work in progress*.

Kassahun, W., **Neyens, T.**, Molenberghs, G., Faes, C., and Verbeke, G. (2012) Modeling overdispersed longitudinal binary data using a combined beta and normal random-effects model. *Archives of Public Health*, **70**: 7.

Kassahun, W., **Neyens, T.**, Molenberghs, G., Faes, C., and Verbeke, G. (2014). A zero-inflated overdispersed and hierarchical Poisson model. *Statistical Modelling*, **14**, 439–456, DOI:10.1177/1471082X14524676.

Kassahun, W., **Neyens, T.**, Molenberghs, G., Faes, C., and Verbeke, G. (2014) Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. *Statistics in Medicine*, **33**, 4402-–4419, DOI: 10.1002/sim.6237.

Kassahun, W., **Neyens, T.**, Molenberghs, G., Faes, C., and Verbeke, G. (2015). A joint model for hierarchical continuous and zero-inflated overdispersed count data. *Journal of Statistical Computation and Simulation*,**84**, 552–571, DOI: 10.1080/00949655.2013.829058.

Kassahun, W., **Neyens, T.**, Molenberghs, G., Faes, C., and Verbeke, G. (2015) The zero-inflated and hurdle combined model for count data in a Bayesian context. *Work in progress*.

Additional publications:

Stevens, AS., Pirotte, N., Plusquin, M., Willems, M., **Neyens, T.**, Artois, T., and Smeets, K. (2015) Toxicity profiles and solvent-toxicant interference in the planarian *Schmidtea mediterranea* after dimethylsulfoxide (DMSO) exposure. *Journal of Applied Toxicology*, **35**, 319–326.

Contents

List of Tables

List of Figures

List of abbreviations

Here, a list is given of the most often used abbreviations in this thesis.

AED	:	anti-epileptic drug
AIC	:	Akaike information criterion
BB	:	beta-binomial
CAR	:	conditional autoregressive
CARCON	:	conditional autoregressive convolution
CH	:	correlated heterogeneity
COM	:	combined
COPD	:	chronic obstructive pulmonary disease
CPO	:	conditional predictive ordinate
DIC	:	deviance information criterion
GLM	:	generalized linear model
GLMM	:	generalized linear mixed model
GMRF	:	Gaussian Markov random field
GOF	:	goodness-of-fit
HCOM	:	hurdle combined
HNB	:	hurdle negative binomial
HP	:	hurdle Poisson
HPN	:	hurdle Poisson-lognormal
INLA	:	integrated nested Laplace approximation
LIKAR	:	Limburgs Cancer Registry
M	:	marginal predictive likelihood

MCAR	:	multivariate conditional autoregressive
MCMC	:	Markov chain Monte Carlo
ML	:	maximum likelihood
MSE	:	mean squared error
MSPE	:	mean squared predictive error
NB	:	negative binomial
PCAR	:	proper conditional autoregressive
P	:	Poisson
pD	:	effective number of parameters
PG	:	Poisson-gamma
PN	:	Poisson-lognormal
RE	:	random effect
RR	:	relative risk
SIR	:	standardized incidence rate
UCAR	:	univariate conditional autoregressive
UH	:	uncorrelated heterogeneity
ZICOM	:	zero-inflated combined
ZINB	:	zero-inflated negative binomial
ZIP	:	zero-inflated Poisson
ZIPN	:	zero-inflated Poisson-lognormal

Chapter 1

Introduction

1.1 Introduction

During the last 25 years, the vast increase in computational possibilities has taken modern statistical mathematics from early standard analyses, based on simple data settings, to more computationally demanding applications of multi-dimensional and large sample data structures. This evolution has been accompanied by successful integrations of statistics in different scientific fields, such as large sample survey data (Bethlehem, 2009) or complex type-III clinical studies in which cohorts are followed up through time (Bliddal et al., 2011, Kazemi et al., 2013). Other examples can be found in environmental sciences where multiple samples are taken on different, possibly correlated locations (Ruthsworth et al., 2014, Wang et al., 2014), among many other examples. Indeed, while traditional statistical practices involved data coming from study designs that were based on strong assumptions, such as independence between observations within populations, recently developed methods can cope with the introduction of different forms of correlation structures and subsequent hierarchies within the data, such as in e.g. spatio-temporal data modeling (Cressie and Wikle, 2011). This thesis builds upon those statistical modeling techniques that take into account one or more of these data complexities, while this introductory chapter specifically gives an overview of the data types and settings on which emphasis is placed further on.

1.2 Correlated data

Throughout this thesis, all topics will based on different ways to take variability in data into account. A lack or an excess of variability can often be attributed to the presence

of correlation within the data, which can be best illustrated with an example: Imagine a study in which one wants to assess if the egg weights of two closely related bird species differ. Due to possible genetic parent factors affecting birth weight, common sense tells that it is good to measure only one egg from a number of nests for each species in order to obtain random and representative samples. If however multiple eggs are weighted within each nest, not all observations are independent any more, as eggs will be genetically more related to eggs from the same nests than to eggs from other nests. In other words, some observations are correlated in a way that the weight of the first egg can vaguely predict the weight of the other eggs in the same nest. This induces a hierarchy in the data, which causes problems when the goal is to obtain a random sample. Off course, in many cases, these problems can be avoided by thinking through and setting up a thorough study design, but sometimes these correlations are necessary due to regulatory or practical reasons (e.g. when the number of nests to be sampled from is small) or due to the scientific question itself. E.g., when one wants to do a clinical follow-up study, patients are measured on multiple occasions through time, making the observations within an individual correlated. Or when the occurrence of a certain disease is investigated on a spatial scale, observations on locations that lie close to each other are possibly more correlated than when the distance between both is large. In this thesis, the main focus is pointed towards spatial data structures, but also data with a time structure will be investigated.

1.2.1 Longitudinal data

In the bird eggs example, observations within nests were correlated, but there was no reason to assume that there was a direction in the correlation structure in the sense that egg A was not correlated in a different way to egg B than to egg C. This is in contrast with a follow-up study in which individuals are measured a number of times throughout the study. Indeed, these so-called longitudinal data also induce a data hierarchy where observations are nested in individuals, but here, observations are structured in a way that the first measurement can be differently correlated with the second measurement than with the third. A lot of research has been done towards the implementation of longitudinal data structures in statistical methods, resulting in an extensive amount of literature devoted to the topic (Verbeke and Molenberghs, 2000, Molenberghs and Verbeke, 2005, Fitzmaurice et al., 2009). This thesis will focus on the longitudinal setting in a number of cases, especially in the chapters 3, 4, 5 and 7.

1.2.2 Spatial data

As the name already implies, spatial data are data which have a certain location in space. Let $\boldsymbol{s} \in \mathbb{R}^d$ be a data location in $\boldsymbol{d}$-dimensional Euclidean space and suppose the potential datum $\boldsymbol{Z}(\boldsymbol{s})$ at spatial location $\boldsymbol{s}$ is a random quantity. When $\boldsymbol{s}$ varies over an index set $D \subset \mathbb{R}^d$, a multivariate random field (or random process)

$$\{\boldsymbol{Z}(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^d\}$$

is formed (Cressie, 1991). Therefore, the correlation structure here is multi-dimensional, in contrast to a longitudinal data structure, which only has one dimension. Different types of spatial data exist and they can be subdivided in three classes, namely geostatistical, lattice and point pattern data.

(1) Geostatistical data are defined as data with a spatial index $\boldsymbol{s}$ that can vary continuously over $\mathbb{R}^d$. These data have an origin in mining engineering sciences, in which geostatistics emerged in the early 1980's as a discipline that combined engineering with mathematics and statistics. Matheron (1963) is known as one of the founders of this scientific niche and largely used these techniques to predict the ore grade in a mining block from observed samples, a still very popular prediction process which he named kriging. (2) Lattice data comprise of data that are measured on a collection of subsets of $\mathbb{R}^d$, called lattices. When these locations are regularly spaced points in $\mathbb{R}^d$, these can be referred to as regular lattices. This type of data is used in many scientific fields, such as in remote sensing from satellites where the earth's surface is divided in a grid of small rectangles, called pixels. Irregular lattices on the other hand do not have displacements that follow a predictable pattern and do not have obvious geometrically linkages. An example can be found in disease mapping, where e.g. the counts of newly diagnosed cases of a certain disease within a specified time period are mapped per municipality. (3) When the variable to be analyzed is the location of a set of events $\boldsymbol{Z}(\boldsymbol{s}_i)$, one is interested in point patterns, e.g. when the interest lies in the clustering through space of a certain disease.

While methods from one class of spatial data can be borrowed from methods associated with another class (for an overview, consult Cressie, 1991), the spatial analyses of this thesis are built around irregular lattice data analysis on a flat plane, in other words, with $d = 2$. The setting will be that of disease mapping, a scientific field which has become increasingly popular in recent years (Elliott at al., 2000, Lawson, 2013). An extension here exists in that spatial observations can be done on multiple occasions through time. These so-called spatio-temporal or space-time data have gained considerable popularity in recent

years, partly due to the post-millennium computational developments that have made it possible to analyze these mostly large and complex data sets. Cressie and Wikle (2011) give a complete and in-depth overview of these modeling techniques. Spatial analyses are considered in chapters 6-8.

1.3 Discrete data

In spatial and longitudinal data analysis, different types of data can be analyzed, such as continuous data, categorical data, time-to-event data, etc. This thesis however specifically investigates the use of binary and count data: (1) Binary data are a categorical data type that can take only two possible values, such as "success" versus "failure" or "yes" versus "no". Note however that by dichotomization, continuous data or categorical data with more than 2 categories can be converted into a binary variable. (2) Count data on the other hand are numerical data that can only take non-negative integer values. It is important to add that these integers arise from counting, not by ranking. Ordinal data may also consist of integers, but in the latter, individual values are subject to their location on an arbitrary scale while only the relative ranking is important. This is in contrast to count data, which have a quantitative value, rather than a qualitative. Agresti (2002) delivers a thorough overview of the statistical methodology concerning these and other types of discrete data. As the next chapters will explain more clearly, there are some issues with binary and count data the practitioner has to deal with, which mainly have to do with known and/or unknown structures within the data. Therefore, this manuscript will focus on the implementation of longitudinal or spatial correlation structures in binary and count data, with special emphasis on disease mapping with count data.

Chapter 2

Data sets

2.1 Introduction

In the following chapters, a number of data sets will be used to support the methodological concepts. All of these data sets exhibit one or more hierarchical characteristics, which will be highlighted and dealt with in the remainder of this thesis. Most of the data are confidential and therefore not publicly available. A division is made between count and binomial data, as later chapters will show that these types of data need different modeling approaches.

2.2 Data sets with count responses

2.2.1 Likar data

The Limburgs Cancer Registry (LIKAR) withholds a number of data sets that will be used in this thesis. LIKAR is designed to register all cancers in the province of Limburg (Belgium), making it possible to consult information about a specific type of cancer per region, age and gender (http://likas.edm.uhasselt.be/). The LIKAR database contains the numbers of new, histologically or cytologically proven invasive cancers within male or female inhabitants of Limburg between the years 1996 and 2005. The data collected and stored in the LIKAR cancer registry are obtained from participating laboratories and practitioners.

The area of the Limburg Cancer Registry consists of the province Limburg, situated in the north east of Belgium. It consists of 44 towns, with the largest populations centred

Table 2.1: Summary Statistics for different cancer types for the 44 municipalities of Limburg.

	Kidney	Prostate	Bladder	
	Male	Male	Male	Female
Mean	12.21	128.14	23.82	5.46
Standard deviation	11.09	107.23	29.85	8.21
Minimum	0	1	0	0
Maximum	57	591	192	48

in the middle of the province. Although Limburg is known as a less urbanized province in the upper part of Belgium (Flanders), the northern towns have suffered from elevated soil concentrations of heavy metals. Also fruit cultivation which is centred around St.-Truiden in the southwestern part of Limburg, is notorious for causing elevated levels of chemicals in soil and water. Herbicides and pesticides have been indicated as risk factors for kidney cancer (Mellemgaard et al., 1994) and prostate cancer (Ferris-i-Tortajada et al., 2011), so it is therefore not unlikely to notice higher cancer aggregations in particular agricultural areas.

In chapters 6 and 8, a number of disease mapping methods will be illustrated by using male kidney cancer, prostate cancer and male and female bladder cancer LIKAR counts during the ten-year period of 1996-2005. Summary statistics are given in Table 2.1 and a map of the observed number of cancer cases in Figure 2.1. Prostate cancer cases are much more encountered than the other cancer types. In fact, it is the most prevalent cancer type among men in Limburg (an average of 128 cases per town). Male bladder and kidney cancer on the other hand have mediocre numbers, with respective averages being 24 and 12 per town, while female bladder cancer is relatively rare (an average of 5 cases per town).

2.2.2 Flemish mesothelioma data

The Flemish mesothelioma data consist of counts of newly diagnosed mesothelioma cases for males in 1999 in Flanders (without Brussels). This very rare but highly aggressive cancer that affects the membrane lining of the lungs and abdomen is typically linked to asbestos exposure. In Flanders, mainly Eternit NV, a corporation involved with the production of house coating materials, which was based in Kapelle-op-den-Bos, a small town close to Antwerp, has been notorious for using asbestos until 1994. These data

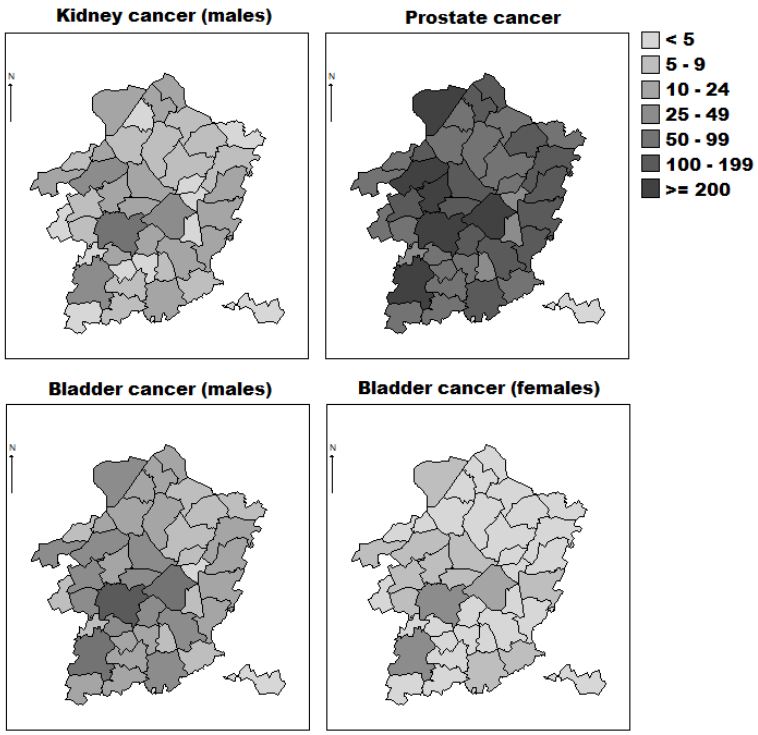


Figure 2.1: Maps of observed counts for male kidney cancer, prostate cancer and male and female bladder cancer in the 44 municipalities of Limburg between 1996 and 2005.

are part of a larger study aimed at determining long-term effects of asbestos exposure and contamination, since after 1994 mesothelioma incidences in and around Kapelle-op-den-Bos have remained frequent, possibly caused by Eternit-sourced asbestos exposure. Due to the disease being very rare, many zero counts occur, with most cases seen in and around Kapelle-op-den-Bos (Table 2.2, Figure 2.2). A specific statistical approach to deal with excessive zeros will be proposed in Chapter 7, where these data will be further investigated.

2.2.3 Georgia asthma and COPD data

The Georgian asthma and chronic obstructive pulmonary disease (COPD) data set represents counts of new cases of asthma and COPD in all 159 counties of Georgia (USA) in 2005 (Figure 2.3). Note that these respiratory diseases are likely to show similar spatial or non-spatial tendencies since it is expected that they may have common etiological factors

Table 2.2: Male mesothelioma cases in the 308 municipalities of Flanders in 1999: summary statistics.

Statistics	Mesothelioma
n	308
mean	0.32
sd	1.24
median	0
min	0
max	19

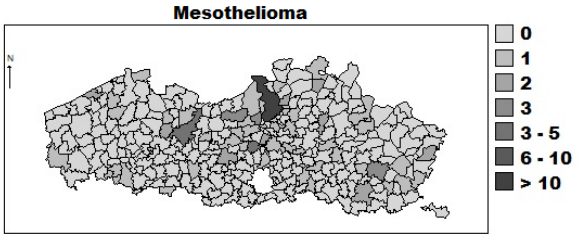


Figure 2.2: Observed male mesothelioma counts in the 308 municipalities of Flanders in 1999.

or determinants. The two diseases have similar mean numbers and standard deviations, while maximum asthma values are almost twice as high as the COPD counts (Table 2.3). It is clear that the observed numbers of both diseases are drastically increased in and around Georgia's capital, Atlanta, although it is not clear if this is truly caused by increased risks of getting the diseases in those areas or if this is just an artefact from the larger population sizes compared to e.g. the south western municipalities. In Chapter 8, methods will be discussed to simultaneously investigate the risks of getting asthma or COPD. This data set was also analyzed by Lawson (2013).

2.2.4 Epilepsy data

The epilepsy data set (Faught et al., 1996) has been studied already frequently in statistical literature (Booth et al., 2003, Molenberghs et al., 2010). The main goal of this study

Table 2.3: Summary statistics for asthma and COPD counts in the 159 counties of Georgia in 2005.

	Asthma	COPD
Mean	72.15	92.99
Standard deviation	138.48	112.93
Minimum	0	0
Maximum	1105	697

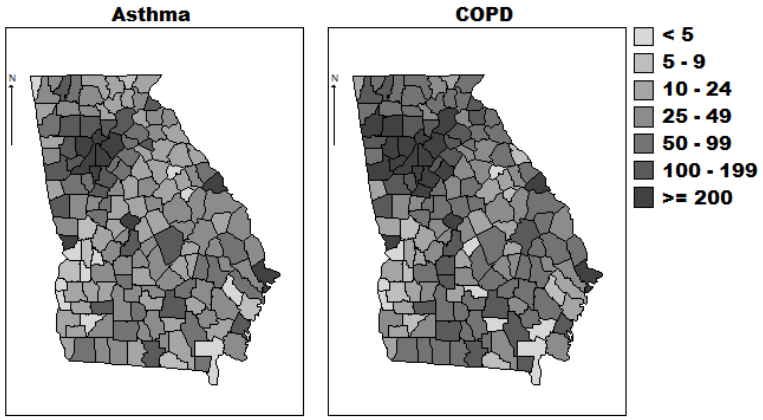
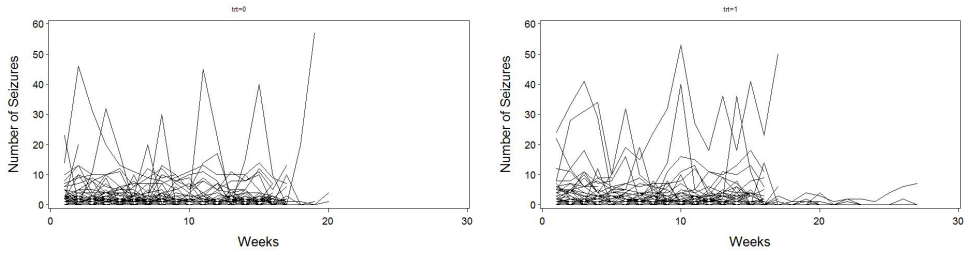


Figure 2.3: A map of observed counts for asthma and COPD in the 159 counties in Georgia (USA) in 2005.

was to investigate if a new anti-epileptic drug (AED), in combination with one or two other AEDs reduced the number of epileptic seizures. After a 12-week baseline period, 45 patients were randomized into a placebo group, with 44 patients assigned to the active (new) treatment group. Patients were followed (double-blind) during 16 weeks and measured weekly, after which they became part of a long-term open-extension study, with some individuals being followed for 27 weeks (Figure 2.4). The primary outcome variable is the number of epileptic seizures per week. In this manuscript, emphasis is placed on the relation between the number of these seizures in relation to treatment type and time as explanatory variables and will be used in chapters 5 and 7.

2.2.5 Flemish contact data

The Flemish contact data set (Goeyvaerts et al., 2014) is part of a large-scale study on social contact behaviour in households with young children in the Flemish geographic region including Brussels. From April to November 2011, participants were recruited to complete a paper diary of their contacts during one randomly assigned day without changing their usual behaviour. The data set used here, which comes directly from that survey, gives the number of contacts (physical contact involving skin-to-skin touching, with or without conversation, or a two-way conversation at less than 3 meters distance) a person reported during 1 day. In this survey, individuals are clustered within households and the households are clustered within municipalities. In total, data were collected from 1312 participants, from 336 households within 211 municipalities. One of the questions



(a) Control **(b)** Treatment

Figure 2.4: Longitudinal profiles displaying the number of epileptic seizures per patient throughout the study in the epilepsy study.

of interest is whether the number of contacts differed between males and females. Table 2.4 and Figure 2.5 give summary statistics and boxplots which already indicate that differences in the number of contacts between males and females are small. In contrast to the previous data sets, more than two data hierarchies, caused by collecting data within households and sampling households within municipalities, have to be taken into account in this setting. This data set will be analyzed in Chapter 5.

2.3 Data sets with binomial responses

2.3.1 Jimma infant growth study

The Jimma infant survival differential longitudinal growth study is an Ethiopian study, set up to establish risk factors affecting infant survival and to investigate socio-economic, maternal, and infant-rearing factors that contribute most to the child's early survival. Children born in Jimma, Keffa and Illubabor, located in south western Ethiopia were examined for their first year growth characteristics. At baseline, there were a total of 7969 infants enrolled in the study, whereby 4317, 1494, and 2158 were from rural, urban, and semi-urban areas, respectively. The children were followed-up every two months, until the age of one year. Of special interest in this thesis is the risk factor for overweight in children. Overweight among infants is associated with various risk factors.

It is of particular interest to identify these risk factors in early life through weight and height measurements, which helps in prevention and treatment of overweight and obesity to reduce the incidence of several adulthood diseases (Freedman et al., 1999). This outcome is defined by dichotomization of the body mass index (BMI), with a BMI

Table 2.4: Summary statistics for the number of contacts per person for females and males in the Flemish contact study.

Statistics	Female	Male
Sample size	660	652
Mean	15.63	15.22
Standard deviation	11.24	10.76
Median	13	12
Minimum	0	0
Maximum	90	90

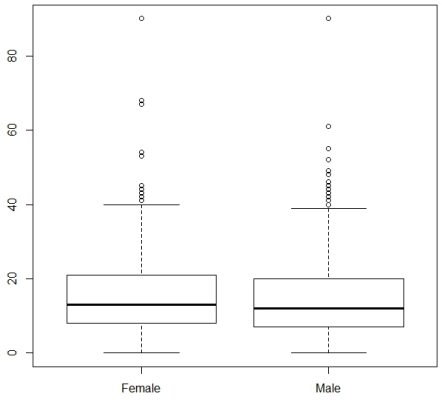


Figure 2.5: Boxplots for the number of contacts per person for females and males in the Flemish contact study.

over the 85th percentile for his or her age referring to overweight. The 85th percentile for age- and sex-specific BMI classification of overweight is used based on Center for Disease Control recommendation (Mei et al., 2002). The first question of interest is whether the percentage of overweight infants changes over time, and whether the evolution differs between genders and the place of residence (rural, urban and semi-urban), as well as the breastfeeding behaviour. Table 2.5 gives a summary of the percentage of overweight infants as a function of gender, location and follow-up time (age in months). Figure 2.6 shows that differences in evolution can probably be found between children with or without breastfeeding. It is important to note though that numbers in the group without breastfeeding were very low when compared to the group with breastfeeding, especially in the early months, e.g. in month 0, there were only 6 children that received no breastfeeding, while 7866 did receive it. The rather steep increase through time for the group without breastfeeding can therefore be an artefact of the small sample size in the early months. Also the place of residence is likely to have an effect, while both gender trends seem to behave alike. These data will be investigated in Chapter 4.

2.3.2 Jimma longitudinal family survey of youth

The Jimma longitudinal family survey of youth is another Ethiopian study where data were collected from households. The study began in 2005, and was repeated in 2007. More than 90% of the study subjects present at baseline were visited and willing to respond

Table 2.5: Percentage of overweight male and female infants by place of residence for each of the seven follow-up times in the Jimma infant growth study.

	rural		urban		semi-urban	
Time	female	male	female	male	female	male
0	11.5	12.2	16.5	14.5	20.3	21.5
2	12.1	12.7	13.4	13.5	20.6	22.4
4	12.1	12.4	12.7	16.4	22.5	20.2
6	13.4	12.3	13.8	14.9	18.3	21.0
8	12.7	11.8	14.9	19.5	20.2	23.1
10	13.4	11.4	14.9	14.9	19.5	22.6
12	13.8	14.1	16.9	16.0	17.6	18.2

in the second round. The study population is representative of the relatively large town of Jimma, the small towns of Yebu, Serbo, and Sheki, and nearby rural areas. The sample includes 3700 households as well as 700 adolescents. The outcome of interest is the adolescents' current school attendance coded as 0 (not currently attending) or 1 (currently attending). Current school attendance was 90.2% and 91.1% in the first round survey and 93.5% and 92.8% in the second round for male and female adolescents, respectively. The research question is to examine whether or not the percentage of school attendance depends on the adolescents' involvement in work to support themselves or their families to earn money, whether they are living in urban towns or rural areas, as well as on gender and age (Belachew et al., 2011). An overview of the summary statistics (Table 2.6) does not give a clear view of possible important factors, although it is likely that the work status is important. Furthermore, the evolution for the categorical covariates does not seem to differ among their categories. It is clear that statistical modeling will be needed to obtain a better insight into these data, as will be done in Chapter 4.

Table 2.6: Numbers for adolescents that were or were not at school at both survey rounds in the Jimma longitudinal family survey of youth. Also the mean and variance for the age of the adolescents at both rounds is given.

Covariate	Category/	Round 1		Round 2	
	Statistic	In school	Not in school	In school	Not in school
Work	Yes	393	79	462	73
	No	1331	99	1320	58
Gender	Male	845	92	874	68
	Female	879	86	908	63
Place	Rural	582	110	615	70
	Urban	667	32	665	39
	Semi-urban	475	36	502	22
Age	Mean	13.658	13.517	15.093	15.015
	Variance	1.500	1.347	1.620	2.031

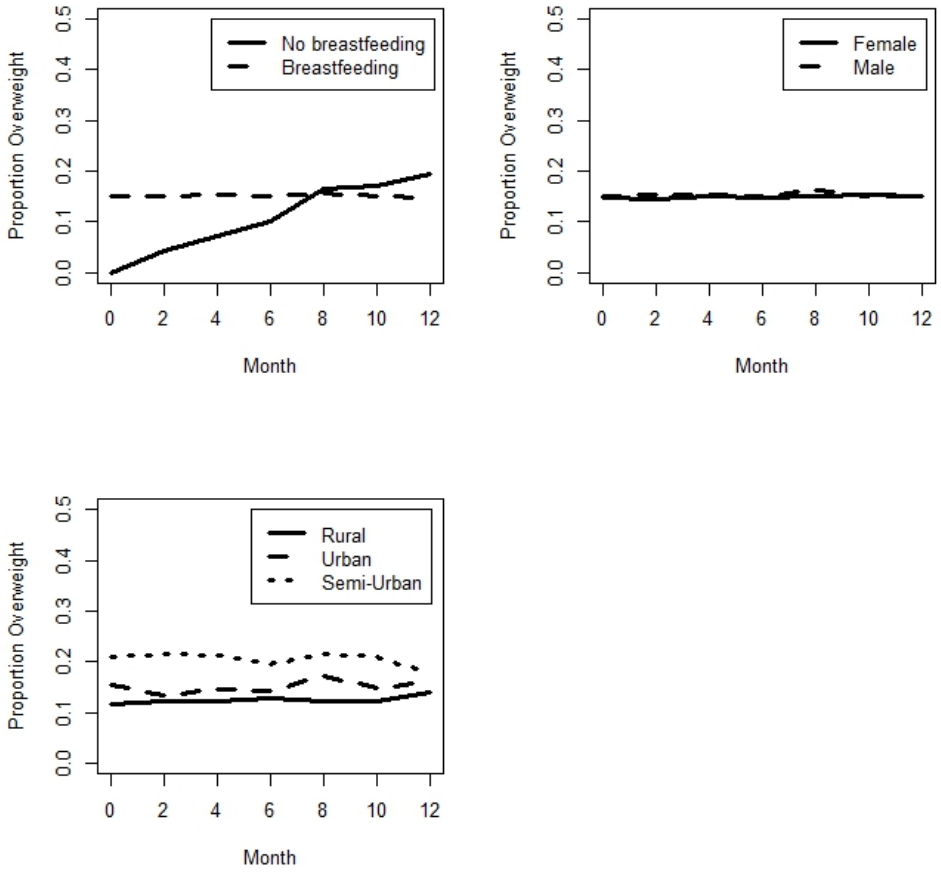


Figure 2.6: Longitudinal profiles for the proportion overweight in children for the Jimma infant growth study.

Chapter 3

Extra-variance

3.1 Introduction

In Chapter 1, a selection of count and proportion data sets was presented. Typically, count data are modeled through the use of the Poisson distribution, while grouped binary (or binomial) data are assumed to follow a binomial distribution. An issue frequently encountered in analyses focusing on these forms of data, is that the data show more variation than what would be expected from the assumed underlying distributions, which can be caused by a number of reasons. It is important to take this additional variability into account, because it may be originated from structural aspects in the data or unknown factors that influence the outcomes. This chapter first discusses basic concepts about distributions for binomial and count data. The second part of this chapter is dedicated to the specification of different models that provide ways of dealing with so-called extra-variance.

3.2 Basic concepts

As already was indicated in Chapter 1, the modeling of binomial and count data may not always be straight-forward. Binomial data comprise grouped binary data, which contain a minimum of information as they only give information on whether an event occurred or not. Count data contain substantially more information but are also limited as they are bound to be non-negative integers. Due to the omnipresence of these data types, many methods have been developed to use them in statistical inference. Before discussing these techniques, the following paragraphs give an overview of a number of the concepts on which these models were built.

3.2.1 Binomial distribution

Suppose $y_1, y_2, ..., y_n$ denote binary responses (1 = "success", 0 = "failure") for n trials. Those trials are assumed to be identical, meaning that the probability of success $P(Y_i = 1) = \pi$, and therefore also the probability of failure $P(Y_i = 0) = 1-\pi$, is the same for each trial, and secondly independent, meaning that $\{Y_i\}$ are independent random variables. The total number of successes, $Y = \sum_{i=1}^{n} Y_i$ then follows a binomial distribution with index n and parameter π, denoted by bin(n,π). The probability mass function for the possible outcomes y for Y is

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y},$$

with $y = 0, 1, 2, ..., n$. Since $E(Y_i) = E(Y_i^2) = \pi$, and $Var(Y_i) = \pi(1-\pi)$, the binomial distribution has mean and variance

$$\begin{aligned} \mu &= E(Y) = n\pi \\ \sigma^2 &= Var(Y) = n\pi(1-\pi), \end{aligned}$$

while skewness is described by $E(Y-\mu)^3/\sigma^3 = (1-2\pi)/\sqrt{n\pi(1-\pi)}$. The distribution converges to normality, as n increases.

3.2.2 Poisson distribution

When dealing with count data not resulting from a fixed number of trials, observation y_i will be a non-negative integer. A well-known distribution that places mass on that range is the Poisson distribution. It expresses what the probability is that a count of events occurs in a fixed interval of time and/or space, while it assumes that one event is independent of the others. Say, Y is a discrete random variable and has a Poisson distribution with parameter $\lambda > 0$ and $y = 0, 1, 2, \ldots$, then the probability mass function of Y, denoted by $Poi(\lambda)$ (Poisson, 1837, p. 206), is given by

$$p(y) = \frac{e^{-\lambda}\lambda^y}{y!},$$

with $\lambda = E(Y) = Var(Y)$. It is unimodal with the mode equal to the integer part of λ. Its skewness is described by $E(Y-\lambda)^3/\sigma^3 = 1/\sqrt{\lambda}$. The distribution approaches normality as λ increases.

3.2.3 Extra-variance

In practice, the assumptions for the previous distributions that the mean and variance depend on a single parameter, are frequently violated. For example, count observations

often have more variability than predicted by a Poisson distribution. When looking at the kidney or prostate cancer data example (Section 2.2.1), this is indeed the case. One can think of a geographical factor, caused by for example industrial pollution or other, unknown explanatory variables that introduce heterogeneity. In order to put this issue in a more general context, suppose that Y is a random variable with parameter λ, but λ can vary because of known or unknown factors. If $\mu = E(\lambda)$, than unconditionally,

$$\begin{aligned} E(Y) &= E\left[E(Y|\lambda)\right], \\ Var(Y) &= E\left[Var(Y|\lambda)\right] + Var\left[E(Y|\lambda)\right]. \end{aligned}$$

If Y is conditionally Poisson (given λ), then $E(Y) = E(\lambda) = \mu$ and $Var(Y) = E(\lambda) + Var(\lambda) = \mu + Var(\lambda) > \mu$.

A similar phenomenon can be seen in analyses assuming binomial data. Recall the egg data example from Section 1.2. Suppose one wants to investigate the predation probability on the nests. Let n_i denote the number of eggs in nest i ad π the probability for an egg to be predated. In practice, π might vary between nests due to e.g. location or individual parent behaviour. In such cases, extra-variation is typically seen due to this additional nest-to-nest variation, with the distribution of the number of egg predation per nest ranging from 0 to n_i, which is more than when there was only a single value of π.

These forms of extra-variance are typically called overdispersion. Note that underdispersion, when the variance is smaller than expected, also exists, but it is only rarely seen in practice. For example, in the egg data example it is possible that parental behaviour generally improves during the nesting season, which might decrease predation. The decrease in predation across nests can make the nests more similar than they would be by chance. In this thesis, a slightly specific nomenclature is applied when the variance assumption is violated: (1) overdispersion is defined as all extra-variance caused by unknown factors, also known as uncorrelated heterogeneity (UH), while (2) all extra-variance caused by a known structural aspect of the data (e.g. a time structure in a longitudinal analysis, a spatial structure in a disease mapping) is denoted as correlated heterogeneity (CH). In the following sections, an overview is given to take these issues into account, with a final description of the so-called combined model that literally combines methods to deal with CH and UH.

3.3 Models dealing with extra-variance

While ordinary linear regression is typically used to investigate the effect of one or more explanatory variables on Gaussian data, generalized linear models (GLMs) are often

employed for modeling univariate non-Gaussian data. GLMs include a wide range of statistical models that relate outcome variables such as counts, binary rates and ratios, etc, to a linear combination of predictor variables (McCullagh and Nelder, 1989, Agresti, 2002, Molenberghs and Verbeke, 2005). A GLM typically has three components: (1) a random component identifies the probability distribution of a vector of observations of Y, (2) a systematic component specifies the vector μ in terms of a vector of p fixed unknown parameters $\boldsymbol{\xi}$ and (3) the link function specifies the function of E(Y) that the model equates to the systematic component.

Let Y be a random variable from an exponential family distribution. The density can then be written as

$$f(y) \equiv f(y|\eta,\phi) = \exp\left\{\phi^{-1}[y\eta - \psi(\eta)] + c(y,\phi)\right\},$$

with unknown parameters η (natural parameter) and ϕ (dispersion or scale parameter), and $\psi(\cdot)$ and $c(\cdot,\cdot)$ known functions. Recall from Section 3.2.1 that for grouped binary responses, $Y \sim \text{Bin}(\pi, n)$. When the goal is to relate the variability in the outcome to a set of covariates, a density function

$$f(y|\eta,\phi) = \pi^y(1-\pi)^{1-y} = \exp\left[y\ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right], \tag{3.1}$$

is used and thus $\eta = \ln[\pi/(1-\pi)]$, $\psi(\eta) = \ln[1+\exp(\eta)]$, $\phi = 1$ and $c(y,\phi) = 0$. As indicated before, the mean is given by $\mu = \psi'(\eta) = \pi$ and the variance, var$(Y) = \phi\psi''(\eta) = \pi(1-\pi)$ (Nelder and Wedderburn, 1972). Now let $Y_1,\ldots,Y_N$ be a set of independent binary outcomes, and let $\boldsymbol{x_1},\ldots,\boldsymbol{x_N}$ be the corresponding p-dimensional vectors of covariate values. With a logit link function, π_i is linked to the linear predictor by taking the natural logarithm of the odds of π, $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x_i}'\boldsymbol{\xi}$, with $\boldsymbol{\xi}$ a vector of unknown regression coefficients, resulting in the well-known logistic regression model. Other link functions are available such as the probit or inverse normal, $\Phi^{-1}(\pi) = \boldsymbol{x_i}'\boldsymbol{\xi}$ or the complementary log-log function, $\ln\{-\ln(1-\pi)\} = \boldsymbol{x_i}'\boldsymbol{\xi}$. When turning to statistical inference, let $L_i = f(y_i|\eta_i,\phi)$ denote the contribution of y_i to the likelihood for N independent observations, $Y_1,\ldots,Y_N$. The likelihood function L then is

$$L(\boldsymbol{\xi}) = \Pi_{i=1}^N L_i = \Pi_{i=1}^N f(y_i|\eta_i,\phi),$$

where $f(y_i|\eta_i,\phi)$ is as defined by (3.1) for observation i. Without going into much detail, it is noteworthy to add that general-purpose iterative methods, such as Newton-Raphson or Fisher Scoring, can be applied to obtain the maximum likelihood estimates of the unknown model parameters (McCullagh and Nelder, 1989, Agresti, 2002).

For count outcomes, the model of interest is: $Y \sim$ Poisson(λ). A GLM density function is

$$f(y) = \frac{e^{-\lambda}\lambda^{y}}{y!},$$

with $\eta = \ln\lambda$, $\psi(\eta) = \exp(\eta) = \lambda$, $\phi = 1$ and $c(y, \phi) = -\ln y!$. The key assumption of the Poisson distribution is the mean-variance equality. Note that when one defines $Var(Y) = \phi v(\lambda)$, the assumption translates to $\phi = 1$. If $Y_1, \ldots, Y_N$ is a set of independent count outcomes, and if $\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}$ are the p-dimensional vectors of covariate values, then the Poisson regression model with $\boldsymbol{\xi}$ a vector of p fixed, unknown regression coefficients is given by $\log(\mu_i) = \boldsymbol{x_i}'\boldsymbol{\xi}$. For observation y_i the contribution to the log-likelihood is proportional to $L_i = y_i \log\omega_i - \omega_i$. The likelihood function L then becomes (McCullagh and Nelder, 1989)

$$L(\boldsymbol{\xi}) = \Pi_{i=1}^{N} L_i = \Pi_{i=1}^{N} \exp(y_i \log\lambda_i - \lambda_i).$$

Here also, the Newton-Raphson or Fisher Scoring can be used to obtain the maximum likelihood estimates of unknown model parameters (McCullagh and Nelder, 1989, Agresti, 2002).

3.3.1 Overdispersion models

Recall that the previous basic models for binomial and count data assume the mean and variance to depend on a single parameter. A way to avoid those stringent assumptions, is to use so-called overdispersion models. Here, the dispersion parameter ϕ is allowed to differ from 1, which induces the possibility to have larger ranges in the variability, e.g. for count data $Var(Y) = \phi v(\lambda)$, while still specifying a relation between the mean and the variance. The general idea is use a two-stage approach in which one assumes the parameter of interest (π for binomial data, λ for count data) to be a random variable that introduces an extra form of variation. This is easily illustrated for the Poisson case: Assume that $Y_i|\lambda_i \sim Poi(\lambda_i)$ and that λ_i is a random variable with $E(\lambda_i) = \mu_i$ and $Var(\lambda_i) = \sigma_i^2$. As in Section 3.2.3, it follows that

$$\begin{aligned} E(Y_i) &= E\left[E(Y_i|\lambda_i)\right] = E(\lambda_i) = \mu_i, \\ Var(Y_i) &= E\left[Var(Y_i|\lambda_i)\right] + Var\left[E(Y_i|\lambda_i)\right] = E(\lambda_i) + Var(\lambda_i) \\ &= \mu_i + \sigma_i^2. \end{aligned}$$

Similarly, this can also be shown for binomial data. Note however that a set of i.i.d. Bernoulli data cannot contradict the mean-variance relationship, but a violation is possible for hierarchical Bernouilli cases and the related binomial data setting.

More generally, the two-stage approach can be defined as follows: There is a distribution for the outcome, given a random effects $f(y_i|\theta_i)$ which, combined with a model for the random effect, $f(\theta_i)$, produces the marginal model

$$f(y_i) = \int f(y_i|\theta_i)f(\theta_i)d\theta_i.$$

Typically, full distributional assumptions about the random effects are made, which are bound together by the property of conjugacy, in the sense of Cox and Hinkley (1974, p. 370) and Lee et al. (2006, p. 178). Conjugacy refers to the fact that hierarchical and random-effects densities have similar algebraic forms, producing a general and closed-form solution for the corresponding marginal distribution. Common choices are the beta distribution for π_i and the gamma distribution for λ_i. For the former, this leads to the so-called beta-binomial model, in which the binomial model is combined with a beta distribution (Molenberghs and Verbeke, 2005, Skellam, 1948, Hinde, and Demétrio, 1998a, Hinde, and Demétrio, 1998b, Kleinman, 1973). For Poisson data, the unconditional distribution of the outcome turns out to be a negative binomial distribution (Breslow, 1984, Hinde and Demétrio, 1998a, Hinde and Demétrio, 1998b). Indeed, if $\sigma_i^2 > 0$ in (3.2), the variance is larger than the mean, implying that the negative binomial allows for overdispersion. When $\sigma_i^2 = 0$, the Poisson model results as a special case.

3.3.2 Generalized linear mixed models

When non-Gaussian data are hierarchically organized (repeated measures, spatial clustering, etc.), the GLM is usually extended to a generalized linear mixed model (GLMM). This model type has one or more subject-specific random effects, usually a Gaussian type, added in the linear predictor to capture the correlation (Engel and Keen, 1992, Molenberghs and Verbeke, 2005, Pinheiro and Bates, 2000) and/or multiple hierarchies (Goldstein, 2002).

Let Y_{ij} be an outcome for the i^{th} subject measured at the j^{th} time point. Let the elements of $\boldsymbol{b}_i$, being the q-dimensional vector of the random effects, be normally distributed with mean $\mathbf{0}$ and variance-covariance matrix D, that is $\boldsymbol{b}_i \sim N(\mathbf{0}, D)$, with $\mathsf{E}(\boldsymbol{b}_i) = \mathbf{0}$ and $\mathsf{Var}(\boldsymbol{b}_i) = D$. Then, it is assumed that the conditional distribution of the response, $Y_{ij}|\boldsymbol{b}_i$ is independent and belongs to the following exponential family density

$$f_i(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \phi) \;=\; \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y, \phi)\right\}.$$

The expectation is, $\mathsf{E}(Y_{ij}|\boldsymbol{b}_i) = \mu_{ij} = h^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i)$, where $\boldsymbol{h}(.)$ is a known link function, $\boldsymbol{x}_{ij}$ is a p-dimensional design matrix of the fixed effect parameters $\boldsymbol{\xi}$, and $\boldsymbol{z}_{ij}$

is a q-dimensional design matrix of the random effects $\boldsymbol{b}_i$. The marginalized likelihood contribution of subject i is

$$f_i(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \phi) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i.$$

From this, the marginal likelihood for $\boldsymbol{\xi}$, D and ϕ is given as

$$L(\boldsymbol{\xi}, D, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \phi) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i, \tag{3.2}$$

Numerical approximations are needed, since in general, expression (3.2) does not have an analytical solution. An extensive overview of different approximation techniques can be found in Molenberghs and Verbeke (2005) and Skrondal and Rabe-Hesketh (2004).

For the case of binomial data Y_{ij}, a multi-stage presentation is

$$\begin{aligned} Y_{ij} &\sim \text{Bin}(n, \pi_{ij} = \kappa_{ij}), \\ \kappa_{ij} &= \frac{\exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right)}{1 + \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right)}, \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, D) \end{aligned}$$

Turning to count data, a similar approach yields

$$\begin{aligned} Y_{ij} &\sim \text{Poi}(\lambda_{ij}), \\ \lambda_{ij} &= \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right), \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, D). \end{aligned}$$

Note that the difference between a GLMM and an overdispersion model is vague. Both models have emerged from two disparate strains of methodological research, but indeed, an overdispersion model can introduce correlation structures and it can be specified such that it becomes a GLMM.

3.4 The combined model

In many cases however, known structural aspects are present in the data while at the same time extra-variance can be caused by unknown factors. In other words, a model could benefit from taking into account the previously introduced correlated and uncorrelated heterogeneity. While in the previous section models were presented that roughly dealt with one or the other form of additional variability, in this section, we introduce the combined model as proposed by Molenberghs et al. (2007). The model literally combines

an overdispersion model with a GLMM, and by bringing both the overdispersion effects as well as the normal random effects towards the generalized linear model framework, a general family is produced (Molenberghs et al., 2010).

Let Y_{ij} be the jth binomial/Poisson outcome measured for subject i ($i = 1, \ldots, N$, $j = 1, \ldots, n_i$) and the n_i measurements are grouped into a vector $\boldsymbol{Y}_i$. Then, conditionally on the q-dimensional random effects $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$, one assumes the outcomes Y_{ij} to be independent and to have densities of the form

$$f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}, \phi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij}, \phi)\right\},$$

with the conditional mean being

$$E(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}) = \mu_{ij}^c = \theta_{ij}\kappa_{ij}.$$

The conditional mean is factorized into an overdispersion component θ_{ij} where the random variable $\theta_{ij} \sim \mathcal{G}_{ij}(\vartheta_{ij}, \sigma_{ij}^2)$, and again the GLMM component $\kappa_{ij} = g(\eta_{ij}) = g(\boldsymbol{x}_{ij}'\boldsymbol{\xi} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i)$ with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ p-dimensional and q-dimensional vectors of known covariate values, with $\boldsymbol{\xi}$ a p-dimensional vector of unknown fixed regression coefficients, while $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$. Note that the linear predictor and/or the natural parameter can be referred to with two notations, η_{ij} and λ_{ij} since the former refers to the 'GLMM part', while the latter encompasses the random variables θ_{ij}.

The likelihood contribution of subject i is

$$f_i(\boldsymbol{y_i}|\boldsymbol{\xi}, D, \boldsymbol{\vartheta}_i, \Sigma_i) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \boldsymbol{\theta}_i)\ f(\boldsymbol{b}_i|D)\ f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \Sigma_i)\ d\boldsymbol{b}_i\ d\boldsymbol{\theta}_i, \quad (3.3)$$

and from this, the marginal likelihood is given as

$$\begin{aligned} L(\boldsymbol{\xi}, D, \boldsymbol{\vartheta}, \Sigma) &= \prod_{i=1}^{N} f_i(\boldsymbol{y_i}|\boldsymbol{\xi}, D, \boldsymbol{\vartheta}_i, \Sigma_i) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \boldsymbol{\theta}_i)\ f(\boldsymbol{b}_i|D)\ f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \Sigma_i)\ d\boldsymbol{b}_i\ d\boldsymbol{\theta}_i, \end{aligned}$$

where $\boldsymbol{\vartheta}_i = E(\boldsymbol{\theta}_i) = E\left[(\theta_{i1}, ..., \theta_{in_i})'\right]$, $\text{Var}(\boldsymbol{\theta}_i) = \Sigma_i$ and $\boldsymbol{\xi}$ again groups all covariate parameters. Note however that in the model types used in this thesis, such as the Poisson-gamma model for count data, the components θ_{ij} of $\boldsymbol{\theta}_i$ are assumed to be independent, such that Σ_i reduces to a diagonal matrix. Although this is natural in many cases, since the correlation is taken up by the normal random effects, it is possible to allow for covariance structures within $\boldsymbol{\theta}_i$. Furthermore, it is possible to allow for

dependence between $\boldsymbol{\theta}_i$ and $\boldsymbol{b}_i$, but in most cases, such as throughout this thesis, independence is assumed. Also note that the conjugacy here is not interpreted any more as described in Section 3.3.1. For the combined model, strong conjugacy, which is defined as conjugacy conditional on the normal random effect $\boldsymbol{b}_i$, is needed. Recall that only a few distributions allow for strong conjugacy. The Poisson distribution is one of them, but the Bernoulli and related binomial distributions only satisfy conjugacy in the sense of Section 3.3.1, or in other words, conjugacy does not "survive" the inclusion of the normal random effects in the Bernoulli and binomial distributions. In Chapters 4 and 5 however, a number of methods will be presented which make it possible to fit both Poisson and binomial combined models.

For binomial data, we obtain

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(\pi_{ij} = \theta_{ij}\kappa_{ij}), \\ \kappa_{ij} &= \frac{\exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right)}{1 + \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right)}. \end{aligned}$$

When considering $\theta_{ij} \sim \text{Beta}(\alpha, \beta)$, then $\phi = \alpha/(\alpha+\beta)$, and

$$\sigma^2_{ij} = \sigma_{i,jj} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \qquad \sigma_{i,jk} = \rho_{ijk}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Closed forms for neither mean nor variance follow when normal random effects are present, which may result in problems in practice.

For the Poisson case, a multi-stage model presentation will mostly be used throughout the following chapters. It can be written as

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\theta_{ij}\kappa_{ij}), && (3.4) \\ \kappa_{ij} &= \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right), && (3.5) \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, D), && (3.6) \\ \theta_{ij} &\sim \text{gamma}(\alpha, \beta). && (3.7) \end{aligned}$$

Remember that it is implicitly assumed that the components θ_{ij} of $\boldsymbol{\theta}_i$ are independent. See also Molenberghs et al. (2007) and Aregay et al. (2013), amongst others for extensive investigation of this model. Note that the Poisson-gamma model can be partially marginalized as,

$$\begin{aligned} Y_{ij} &\sim \text{NegBin}(\alpha, \kappa_{ij}\beta), && (3.8) \\ \kappa_{ij} &= \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right), && (3.9) \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, D), && (3.10) \end{aligned}$$

as will be done in parts of this thesis. Indeed, the above negative binomial model is obtained from the Poisson-gamma-normal model, by integrating out the gamma random effect (see also e.g. Molenberghs et al., 2007, Neyens et al., 2012). Note that to avoid overparametrization problems, a restriction has to be applied to shape parameter α and scale parameter β. Therefore, in the following chapters, $\alpha = 1/\beta$ is used when the combined model is applied. Also note that in most literature and in a selection of software packages (such as SAS) the gamma distribution is defined in terms of the shape and scale parameters. Others, such as WinBUGS, use the parametrization with the shape and rate (1/scale) parameters. While this thesis is written using the former notation, this has to be taken into account when consulting program codes in the appendices.

Chapter 4

Bayesian Estimation of the Combined Model for Binomial Data

4.1 Introduction

In Chapter 3, an overview of different model families was given. When an appropriate model is chosen, the decision of what estimation method to use, forms the following step. Initially, inference for the combined model was conducted in the likelihood framework by the use of partial integration, in which the conjugate random effect is integrated out analytically while the normal random effect is integrated out numerically (Molenberghs et al., 2007). Others also worked within the likelihood framework, such as Njagi et al. (2013), who used the combined model in the joint modeling context, and Efendi et al. (2013) and Kassahun et al. (2014) in the context of marginalized combined models. Kalema and Molenberghs (2014) proposed the use of pseudo-likelihood as an alternative estimation method. Others however, such as Aregay et al. (2013) and Ghebretinsae et al. (2012b) embedded the combined model for count and time-to-event data, respectively, in a Bayesian framework, with estimation done with Markov Chain Monte Carlo methods (MCMC). Neyens et al. (2012) also used MCMC to introduce the combined model into a spatial data analysis. MCMC is a simulation-based method that, although it also works with a likelihood function, is based on a vastly different set of assumptions than the likelihood techniques. In practice, Bayesian estimation is popular when the data hierarchies become complex or when the analyst wants to include prior knowledge in the

analysis. It is important however to assess if likelihood and Bayesian analyses for the combined model yield the same results.

This chapter presents a comparison between the combined and traditional models via both partial likelihood and MCMC. In Section 4.2, partial likelihood and MCMC are introduced, while Section 4.3 focusses on model comparison techniques. In Section 4.4, two binary data case studies will be investigated, namely the Jimma infant growth study and Jimma longitudinal survey of youth, which were introduced in Sections 2.3.1 and 2.3.2 respectively, while a discussion is given in Section 4.5.

4.2 Estimation methods

As seen in Chapter 3, the combined model for binomial data contains a set of normal and beta random effects. In the next sections, partial integration, a likelihood-based technique, and MCMC, a Bayesian method, will be introduced. It was mentioned earlier that the lack of strong conjugacy might be problematic for estimation in the likelihood framework, but it still remains feasible in the binomial case, as will be shown next. MCMC, which is a sampling-based method, does not build on the conjugacy characteristic as it does not need a closed-form posterior distribution.

4.2.1 Partial integration

Partial integration, or partial marginalization, is a straight-forward method in the likelihood framework and is recommended in many data settings, although computational issues arise when working with very large data sets and/or when the data contain multiple hierarchies. In this method, the conjugate random effect is first integrated out analytically from the likelihood, while leaving the normally distributed random effect embedded in the predictor. The fully marginalized likelihood is then obtained by numerically integrating out the normal random effect using software such as the SAS procedure NLMIXED or the R function nlme. Details concerning its implementation when working with the combined model can be found in Molenberghs et al. (2010), but an overview is given here.

The likelihood contribution for subject i is given by (3.3) with $f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \boldsymbol{\theta}_i)$ corresponding with the hierarchical distribution, e.g. a Poisson or binomial distribution, $f(\boldsymbol{b}_i|D)$ with the normal random effects distribution and $f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \Sigma_i)$ with the conjugate random effects distribution, such as a gamma or beta random effects distribution which is conjugate to respectively the Poisson or binomial distribution, and $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{in_i})$.

Partial integration leads to the likelihood

$$f_i(\boldsymbol{y_i}|\boldsymbol{\xi}, D, \boldsymbol{\vartheta}_i, \Sigma_i) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \boldsymbol{\vartheta}_i, \Sigma_i) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i.$$

For the binomial data, Molenberghs et al. (2010) investigated the logit and probit links, while in this thesis, only the former will be looked at, mainly due to its omnipresence in univariate data analyses. Because there is lack of strong conjugacy, it is impossible to exploit the conjugate form, such as the beta-binomial here. It is however possible to integrate out the beta random effects that are assumed to be independent within a subject. The corresponding probability becomes

$$f(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\xi}, \alpha, \beta) = \frac{1}{\alpha_j + \beta_j} . (\kappa_{ij}\alpha_j)^{y_{ij}} . \left[(1-\kappa_{ij})\alpha_j + \beta_j\right]^{1-u_{ij}},$$

which can be used afterwards to obtain the fully marginalized likelihood by numerical integration. Indeed, while strong conjugacy does not apply for the binomial case, integration over θ is fairly simple, pointing out that partial integration brings a simple way to overcome the issue of defeating strong conjugacy.

4.2.2 Markov chain Monte Carlo

An attractive alternative to likelihood-based estimation is the Bayesian method, which still builds upon the likelihood principle, stating that the information content of the data is solely and entirely expressed by the likelihood function, but which assumes all parameters - so both types of parameters that were defined in the likelihood framework as fixed or random - to be stochastic. Inference is based on the posterior distribution, which is proportional to the product of likelihood and priors

$$p(\boldsymbol{\xi}, \boldsymbol{b}_i, D, \boldsymbol{\vartheta}_i, \Sigma|\boldsymbol{y_i}) \propto \left[\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \boldsymbol{\theta}_i) f(\boldsymbol{b}_i|D) f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \Sigma)\right] \left[p(\boldsymbol{\xi}) p(D) p(\boldsymbol{\vartheta}_i) p(\Sigma_i)\right],$$

where $f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \boldsymbol{\theta}_i)$, $f(\boldsymbol{b}_i|D)$ and $f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \Sigma)$ are as before, while $p(\boldsymbol{\xi})$, $p(D)$, $p(\boldsymbol{\vartheta}_i)$ and $p(\Sigma_i)$ are the prior densities for $\boldsymbol{\xi}$, D, $\boldsymbol{\vartheta}_i$ and Σ_i respectively. Note that in the binomial case,

$$p(\boldsymbol{\xi}, \boldsymbol{b}_i, D, \alpha, \beta|\boldsymbol{y_i}) \propto \left[\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \boldsymbol{\theta}_i) f(\boldsymbol{b}_i|D) f(\boldsymbol{\theta}_i|\alpha, \beta)\right] \left[p(\boldsymbol{\xi}) p(D) p(\alpha) p(\beta)\right],$$

follows. Sampling algorithms are used, with Markov Chain Monte Carlo (MCMC) being most widespread (Ripley, 1987, Gelman and Rubin, 1992) to obtain the normalizing

constant of the posterior distribution. MCMC methods use iterative simulation of parameter values within a Markov chain. When this chain runs over a long period, it will converge towards a stationary distribution (posterior distribution) and one is then able to generate a sample from that posterior distribution. However, although the simulation-based method behind MCMC is said to be able to tackle very complex models, it is known to hamper a lot of the more complex analyses, mainly by a computational burden. Indeed, MCMC is an exact method, in the sense that in theory, when you sample throughout a large amount of iterations, convergence will always be met. It should be noted though that in practice however convergence of complex models can be highly time-consuming, due to associations among the parameters, which forces the user to end the MCMC sampler at one point, making it not exact any more.

An important feature of Bayesian inference is the inclusion of prior knowledge. Although this prior knowledge can be very informative, in this thesis, weakly informative priors will mostly be used. In this chapter particularly, the following prior distributions are used, in accordance to Kassahun et al. (2012): $\xi \sim N(0, 10^{-6})$, $b_i \sim N(0, \tau_i)$, with $\tau_i = 1/\sigma_i^2$, as also suggested in the literature (Gilks et al., 1996; Gelman et al., 2004) and $\theta_{ij} \sim$ beta(α, β), is unimodal and concave, when $\alpha > 1,\ \beta > 1$ (Agresti, 2002). Note that if one or both of these parameters would be less than 1, then the probability mass function would go to infinity near its boundaries, 0 and 1, and hence would not be concave. As a result, the mode would not exist, leading to computational problems in MCMC. For this reason, the restriction $\alpha > 1,\ \beta > 1$ is used, such that the density is always concave and unimodal whereby it is always finite over the support $[0, 1]$. For the hyper parameters τ_i, the inverse-gamma prior $IG(0.001, 0.001)$, and for α and β, an improper uniform prior is used, as also suggested by Gelman et al. (2004).

4.3 Model comparison

In order to assess fit, the well-known log-likelihood and Akaike Information Criterion (AIC) were used when working in the likelihood framework. In Bayesian statistics, this issue is more controversial. Spiegelhalter et al. (2002) suggest the use of the so-called deviance information criterion (DIC). Assume a probability model $P(y|\theta)$. The effective number of parameters with respect to a model with parameter Θ is given by $pD\{y, \Theta, \widetilde{\theta}(y)\} = E_{\theta|y}[-2\log p(y|\theta)] + 2\log[p\{y|\widetilde{\theta}(y)\}]$. Note that the arguments $\{y, \Theta, \widetilde{\theta}(y)\}$ are usually dropped from the notation. In general, $\widetilde{\theta}(y) = E(\theta|y)$, the posterior mean of the parameters is used for estimation purposes. For $f(y)$ being a fully specified standardizing term that is a function of the data alone, pD, defined as a 'mean

deviance minus the deviance of the means', is given by $pD = E[D(\theta|y)] - D(E[\theta|y])$, where $D(\theta) = -2\log P(y|\theta) + 2\log f(y)$ is the Bayesian deviance, used as a measure for goodness-of-fit. The deviance information criterion, defined as the classical estimate of fit plus twice the effective number of parameters $DIC = D(E[\theta|y]) + 2pD = E[D(\theta|y)] + pD$ is used for model comparison. According to this criterion, the model with the smallest DIC is to be preferred. pD and DIC are easily computed using the available MCMC output by taking the posterior mean of the deviance to obtain $E[D(\theta|y)]$ and the plug-in estimate of the deviance $D(E[\theta|y])$ using the posterior means $E[\theta|y]$ of the parameter θ. In non-hierarchical models, pD approximates the effective number of parameters to be estimated. However, for hierarchical models, pD is a measure of model complexity instead of being merely the number of effective parameters to be estimated. In general, it is difficult to say what would constitute an important difference in DIC for model comparison. Spiegelhalter et al. (2002) suggested models receiving DIC within 1-2 of the 'best', deserve consideration, and 3-7 have considerably less support. These rules of thumb appear to work reasonably well (Spiegelhalter et al., 2002). For the best model preferred based on DIC, the important risk factors could be identified looking at the credible intervals. In the case of a single parameter and data that can be summarised in a single sufficient statistic, the credible interval and the confidence interval can be treated equivalently. Hence, to identify the risk factor, the consideration whether zero is in or outside of the credible interval was made. In terms of parameter interpretation, it is important to refer back to the beneficial properties that come with the conjugacy property. Indeed, because the θ_{ij} follow a conjugate distribution, the interpretation of the parameters is the same as in a classical generalized linear mixed model. Precisely, this means that the effect on the regression parameters only comes from the normal random effects in the linear predictor, a fact well documented. For a review, see for example Molenberghs and Verbeke (2005).

4.4 Case studies: Jimma infant growth study and Jimma longitudinal survey of youth

For the Jimma infant growth study, assuming independence, the sample average probability of success and the sample variance are 0.150 and 0.128, respectively, indicating that the prescribed mean-variance link is maintained. Note that this is always true for the Bernoulli case. In the binomial setting however, which takes the hierarchical structure into account, the sample average and the sample variances are 0.141 and 2.107, respectively, thus implying the mean-variance relationship for these data is violated. When looking at the binomial case for the Jimma longitudinal survey of youth data, the sample average probability of success was 0.919 and the sample variance was 0.168, which indicates that

the results are in line with the prescribed mean-variance relationship. This may suggest, at first sight, that these data are not prone to exhibit strong extra-variance, even in the hierarchical binomial setting. In addition to the exploratory analysis, tests for overdispersion were conducted. The commonly used approach is to compute the ratio of the residual deviance to the residual degrees of freedom, which approximates the overdispersion parameter ($\hat{\phi}$). When the ratio is appreciably larger than 1, overdispersion is said to occur. It is pointed out that this approach could be misleading when $n_i p_i$ is not sufficiently large, where p_i is the probability of the success event. This is because it is based on asymptotic theory. As a result, a better approach is based on a quasi-binomial model, which allows extra dispersion (Skellam, 1948). The approximated overdispersion ($\hat{\phi} = 2.37$), computed as the ratio of the residual deviance to the residual degrees of freedom in the binomial, and the one estimated in the quasi-binomial model, based on a χ^2 statistic instead of residual deviance ($\hat{\phi} = 2.47$), for the Jimma infant growth data are very similar, both suggesting the presence of strong overdispersion. However, a similar analysis for the Jimma family survey data, does not suggest considerable overdispersion, with values 0.765 and 1.129, approximated by the ratio of the residual deviance to the residual degrees of freedom in the binomial, and estimated by the quasi-binomial, respectively.

4.4.1 Jimma infant growth study: estimation via partial integration

In order to model the binary BMI data in function of a set of covariates, the following combined model specification can be used for subject i and measurement j:

$$\begin{aligned}
Y_{ij} &\sim \text{Bernoulli}(\theta_{ij}\kappa_{ij}),\\
\text{logit}(\kappa_{ij}) &= \xi_0 + b_{0,i} + (\xi_1 + b_{1,i})Time_{ij} + \xi_2 Sex_i +\\
&\quad \xi_3 Rural_i + \xi_4 Urban_i + \xi_5 Breastfed_{ij} + \xi_6 Sex_i * Time_{ij}\\
&\quad + \xi_7 Rural_i * Time_{ij} + \xi_8 Urban_i * Time_{ij} + \xi_9 Breastfed_{ij} * Time_{ij},\\
b_{0,i} &\sim N(0, d_0),\\
b_{1,i} &\sim N(0, d_1),\\
\theta_{ij} &\sim \text{beta}(\alpha, \beta).
\end{aligned}$$

$Time_{ij}$ is the time point at which the j^{th} measurement is taken for the i^{th} subject, which is centred at month six. Spatial differences between rural, urban and semi-urban places were investigated. The infant growth data set is analyzed with (i) a simple logistic model, so the model above without $b_{0,i}$, $b_{1,i}$ and θ_{ij}, (ii) a beta-binomial model introducing only the overdispersion parameter θ_{ij}, (iii) a random-effects logistic model with only $b_{0,i}$ and $b_{1,i}$ and (iv) the combined model, as specified above.

Table 4.1: Jimma infant growth study. Parameter estimates, standard errors, and p-values for the regression coefficients in (1) the logistic model, (2) the beta-binomial model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.

		Logistic	Beta-binomial
Effect	Parameter	Estimate (s.e., p)	Estimate (s.e., p)
Intercept	ξ_0	$-1.896(0.128, 0.001)$	$-0.448(1.099, 0.683)$
Time	ξ_1	$0.127(0.031, 0.001)$	$0.188(0.090, 0.037)$
Gender: male	ξ_2	$0.027(0.025, 0.294)$	$0.029(0.039, 0.456)$
Place: rural	ξ_3	$-0.602(0.029, 0.001)$	$-0.949(0.501, 0.058)$
Place: urban	ξ_4	$-0.376(0.037, 0.001)$	$-0.628(0.381, 0.099)$
Breastfeeding	ξ_5	$0.545(0.128, 0.001)$	$0.788(0.347, 0.023)$
Slope gender: male	ξ_6	$-0.003(0.006, 0.602)$	$-0.007(0.011, 0.534)$
Slope place: rural	ξ_7	$0.018(0.007, 0.014)$	$0.029(0.020, 0.161)$
Slope place: urban	ξ_8	$0.016(0.009, 0.097)$	$0.026(0.022, 0.251)$
Slope breastfeeding	ξ_9	$-0.133(0.031, 0.001)$	$-0.199(0.098, 0.041)$
Std. dev. random intercept	$\sqrt{d_0}$	—	—
Std. dev. random slope	$\sqrt{d_1}$	—	—
Ratio	α/β	—	$1.827(1.622, 0.259)$
-2log-likelihood		41286	41286
AIC		41306	41308

Parameter estimates of the logistic model and the beta-binomial model are presented in Table 4.1 and the corresponding estimates of the logistic-normal model and the combined model are given in Table 4.2. Clearly, the logistic-normal model is an important improvement, in terms of fit (AIC), relative to both the ordinary logistic model and the beta-binomial. Moreover, considering the combined model, there is a very strong improvement in fit when the beta and normal random effects are simultaneously allowed for. The overdispersion term in the combined model is significant ($p < 0.001$), implying the presence of considerable extra-variability due to the grouped nature of the data, which is beyond what can be accommodated by the commonly used logistic-normal model.

The logistic-normal model ignores the overdispersion that results from the grouped nature of the data. On the other hand, the beta-binomial model accommodates overdis-

Table 4.2: Jimma infant growth study. Parameter estimates, standard errors, and p-values for the regression coefficients in (1) the logistic-normal model, and (2) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.

		Logistic-normal	Combined
Effect	Parameter	Estimate (s.e., p)	Estimate (s.e., p)
Intercept	ξ_0	-2.741(0.186, 0.001)	-2.661(0.215, 0.001)
Time	ξ_1	0.132(0.042, 0.002)	0.147(0.049, 0.003)
Gender: male	ξ_2	0.010(0.054, 0.852)	0.020(0.064, 0.751)
Place: rural	ξ_3	-0.908(0.064, 0.001)	-1.058(0.082, 0.001)
Place: urban	ξ_4	-0.581(0.082, 0.001)	-0.689(0.099, 0.001)
Breastfeeding	ξ_5	0.635(0.179, 0.001)	0.764(0.209, 0.001)
Slope gender: Male	ξ_6	-0.003(0.010, 0.728)	-0.005(0.012, 0.660)
Slope place: rural	ξ_7	-0.015(0.011, 0.167)	0.024(0.014, 0.085)
Slope place: urban	ξ_8	-0.011(0.014, 0.432)	0.015(0.017, 0.377)
Slope breastfeeding	ξ_9	-0.149(0.044, 0.001)	-0.167(0.049, 0.001)
Std. dev. random intercept	$\sqrt{d_0}$	1.774(0.034, 0.001)	2.107(0.088, 0.001)
Std. dev. random slope	$\sqrt{d_1}$	0.193(0.007, 0.001)	0.237(0.014, 0.001)
Ratio	α/β	—	0.234(0.045, 0.001)
$-$2log-likelihood		37000	36971
AIC		37024	36997

persion which is assumed independent, implying independence between repeated measurements. Again, this is not realistic and therefore the combined model is the more viable candidate, supported further by the aforementioned AIC and likelihood comparison. The combined model suggests a significant time interaction with breastfeeding. Although the interaction between time and place of residence was not significant, the main residence effects were however. Furthermore, the main effect and slope of gender were not significant, implying that the proportion of overweight seems to be invariant among male and female infants over time. This means that early initiation of breastfeeding has a protective effect against the risk of being overweight in late infancy ($\widehat{\xi_9} = -0.167$, $p = 0.001$). Next to that, infants living in rural and urban areas are at lower risk of being overweight as compared to those in semi-urban areas with ($\widehat{\xi_3} = -1.058$, $p = 0.001$), and ($\widehat{\xi_4} = -0.689$,

$p = 0.001$), respectively, as shown in Table 4.2.

4.4.2 Jimma longitudinal survey of youth: estimation via partial integration

In order to model the current school attendance in function of a set of covariates, the following model, with $Y_{ij} \sim$ Bernoulli$(\theta_{ij}\kappa_{ij})$ and

$$\begin{aligned} \text{logit}(\kappa_{ij}) &= \xi_0 + b_{0,i}\xi_1 Age_{ij} + \xi_2 Urban_{ij} + \xi_3 Semi_u rban_{ij} + \xi_4 Work_{ij} \\ &\quad + \xi_5 Sex_{ij} + \xi_6 Round_{ij}, \end{aligned}$$

is considered, where Y_{ij} denotes the school attendance status of individual i at time point j and $b_{0,i} \sim N\left(0, \sigma_0^2\right)$ and $\theta_{ij} \sim$ beta(α, β). Also here, differences between rural, urban and semi-urban places were investigated. Again, a model without random effects, an overdispersion model, a model with only the normal random effects and lastly the combined model were fitted.

Results from fitting all four models can be found in Tables 4.3 and 4.4. AIC and likelihood comparison of the beta-binomial with the standard logistic model shows no improvement in fit, implying absence of strong evidence for overdispersion. This can be noted from likelihood comparisons of the simple logistic and the beta-binomial on the one hand, as well as the logistic-normal and the combined, on the other. One can easily see, however, that the commonly used logistic-normal and the combined models are significant improvements over the standard logistic model. We further observe, while the logistic-normal model suggests a significant intercept ($p = 0.045$), that the same does not emerge when the combined model is considered ($p = 0.099$), implying the beta random effect has some impact on the p-values. For these data, with two repeated measures per subject, the logistic-normal model seems adequate and the overdispersion term in the combined model is not significant ($p = 0.29$), strengthening what has been mentioned in the earlier sections. Further extension by adding a random slope did not improve the fit of neither the logistic-normal nor the combined models (details not shown).

Based on the logistic-normal model in Table 4.4, adolescents living in urban and semi-urban areas have a higher school attendance than those living in rural areas, with $\widehat{\xi_2} = 1.098$ ($p = 0.001$) and $\widehat{\xi_3} = 1.092$ ($p = 0.001$), respectively. Gender is also significantly associated with school attendance, while this is lower for female adolescents ($\widehat{\xi_4} = -1.241$, $p = 0.001$). There is evidence that school attendance increases in the second round visit compared to the first ($\widehat{\xi_6} = 0.398$, $p = 0.010$).

Table 4.3: Jimma longitudinal family survey of youth. Parameter estimates, standard errors, and p-values for the regression coefficients in (1) the logistic model, (2) the beta-binomial model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.

		Logistic	Beta-binomial
Effect	Parameter	Estimate (s.e., p)	Estimate (s.e., p)
Intercept	ξ_0	1.171(0.626, 0.061)	1.155(0.702, 0.099)
Age	ξ_1	0.039(0.049, 0.414)	0.044(0.055, 0.421)
Place: urban	ξ_2	0.971(0.148, 0.001)	1.089(0.266, 0.001)
Place: semi-urban	ξ_3	0.979(0.159, 0.001)	1.104(0.284, 0.001)
Gender: female	ξ_4	-1.111(0.123, 0.001)	-1.226(0.237, 0.001)
Work	ξ_5	0.134(0.122, 0.274)	0.146(0.138, 0.288)
Round	ξ_6	0.341(0.141, 0.016)	0.390(0.178, 0.029)
Std. dev. random effect	$\sqrt{d}$	—	—
Ratio	α/β	—	0.009(0.014, 0.528)
-2log-likelihood		1987.7	1987.4
AIC		2001.7	2003.4

4.4.3 Bayesian estimation

For comparison with the previously applied estimation method in the likelihood framework, the same models were applied to the two surveys, but now in a Bayesian framework. Convergence was checked using the Gelman-Rubin diagnostic as well as by visual inspection of the trace and QQ plots (Brooks and Gelman, 1998).

The posterior summaries of logistic and beta-binomial for the Jimma infant growth data set are given in Table 4.5, while the corresponding estimates of the logistic-normal and combined models are presented in Table 4.6. Similarly, for the Jimma longitudinal family survey of youth, estimates of these four models are shown in Tables 4.7 and 4.8. The parameter estimates are fairly similar to what was obtained previously in the likelihood approach in both cases, except for differences in the case of the beta-binomial for the Jimma Infants data in Table 4.5 when compared with Table 4.1. In terms of significance of the parameters, the same conclusion is reached for the two case studies in both approaches, except that the beta-binomial for the intercept and time

Table 4.4: Jimma longitudinal family survey of youth. Parameter estimates, standard errors, and p-values for the regression coefficients in (1) the logistic-normal model, and (2) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.

		Logistic-normal	Combined
Effect	Parameter	Estimate (s.e., p)	Estimate (s.e., p)
Intercept	ξ_0	1.443(0.719, 0.045)	1.463(0.888, 0.099)
Age	ξ_1	0.046(0.056, 0.408)	0.058(0.070, 0.408)
Place: urban	ξ_2	1.098(0.178, 0.001)	1.379(0.393, 0.001)
Place: semi-urban	ξ_3	1.092(0.189, 0.001)	1.339(0.368, 0.001)
Gender: female	ξ_4	-1.241(0.147, 0.001)	-1.499(0.339, 0.001)
Work	ξ_5	0.153(0.144, 0.287)	0.189(0.182, 0.296)
Round	ξ_6	0.398(0.155, 0.010)	0.519(0.237, 0.028)
Std. dev. random effect	$\sqrt{d}$	1.138(0.188, 0.001)	1.342(0.318, 0.001)
Ratio	α/β	—	0.013(0.013, 0.293)
-2log-likelihood		1972.9	1972.1
AIC		1988.8	1990.1

effects in the Jimma infants study shows significance in the likelihood framework as given in Section 4.4.1, while the same does not emerge from the Bayesian analysis, as observed from the 95% credible interval which include zero for these effects. The various models using DIC were compared. For both studies, there is a significant reduction in the DIC of the logistic-normal and the beta-binomial, as compared to the simple logistic. We observe a rather high degree of model improvement by combining beta and normal random effects simultaneously, to allow for both the overdispersion and the data hierarchy. Moreover, the logistic and the beta-binomial ignore the correlation stemming from the data hierarchy on the one hand, and the logistic-normal does not allow for the overdispersion, on the other, which altogether make the combined model the preferred one.

According to Spiegelhalter et al. (2002), in comparing complex hierarchical models where the number of parameters is not clearly defined, pD, defined as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest, not only measures the effective number of parameters but also the model complexity. These authors further noted that the contribution pD_i of

each observation i turned out to be its leverage, defined as the relative influence that each observation has on its own fitted value. For y_i conditionally independent given θ_i, pD_i shows its interpretation as the difficulty in estimating θ_i with y_i. This shows the connection between the sample size, the parameters to be estimated, and the pD. The Jimma infants ($n = 7969$) and the Jimma longitudinal family survey ($n = 2100$) data have a large number of subjects followed longitudinally, where each subject was measured seven and two times, respectively. For these reasons, the pD values, as presented in Table 4.6 and Table 4.8, appeared to be larger for the infant growth study than for the longitudinal family survey of youth as the by-product of the MCMC estimation to obtain leverage of each observation. The two competing models, i.e., the logistic-normal model and the combined model resulted relatively in larger values of pDs in both of our case studies.

Unlike the Jimma infants study in Table 4.6, the pD of the combined model for the Jimma longitudinal family survey of youth in Table 4.8 ($pD = 211.9$) is lower than that of the logistic-normal ($pD = 241.5$). This implies that for the Jimma longitudinal family survey of youth there is reduced dimensionality suggesting that the combined model is less complex to fit than the logistic-normal, although this is not what we usually expect, as the combined model seems more complex, since it includes both beta and normal random effects, while the logistic-normal includes only the normal random effects. However, for these specific data, this resulted likely because there is less conflict between the specific data set, and the prior distributions which could be associated to the conjugacy of the beta random effects, as well as the peculiar data features including the number of subjects and repeated measurements per subject.

4.5 Concluding remarks

In this chapter, a comparison between partial integration and MCMC was given for the combined model for binary data, while its modeling flexibility was illustrated too. The analysis of the case studies shows that, in the presence of overdispersion and clustering, the combined model results in an improvement in model fit, which is similar to the finding in Molenberghs et al. (2010). Maximum likelihood estimation through partial integration was considered by using the SAS procedure NLMIXED, while Bayesian inference was applied via WinBUGS. Note for the latter that information about the parameters induces correlation, which then leads to reduced effective dimensionality although the reduction depends on the available data (Spiegelhalter et al., 2002). Complexity reflects the

Table 4.5: Jimma infant growth study. Estimated posterior mean and standard deviation in (1) the logistic model, (2) the beta-binomial model.

		Logistic	Beta-binomial
Effect		Mean(s.d.)	Mean(s.d.)
Intercept	ξ_0	$-1.894(0.123)$	$-1.486(1.488)$
Time	ξ_1	0.126(0.031)	0.155(0.207)
Gender: male	ξ_2	0.027(0.026)	0.003(0.066)
Place: rural	ξ_3	$-0.602(0.029)$	$-2.486(1.290)$
Place: urban	ξ_4	$-0.377(0.037)$	$-1.973(1.210)$
Breastfeeding	ξ_5	0.543(0.123)	1.126(0.294)
Slope gender: male	ξ_6	$-0.003(0.006)$	$-0.015(0.016)$
Slope place: rural	ξ_7	0.018(0.007)	0.160(0.178)
Slope place: urban	ξ_8	0.015(0.009)	0.1610.182)
Slope breastfeeding	ξ_9	$-0.132(0.030)$	$-0.289(0.097)$
Std. dev. random intercept	$\sqrt{d_0}$	—	—
Std. dev. random slope	$\sqrt{d_1}$	—	—
Ratio	α/β	—	3.222(0.524)
DIC		41310.0	40390.0
pD		9.9	2511.0

difficulty in fit and hence it seems reasonable that the measure of complexity may depend on both the prior information concerning the parameters under scrutiny and the specific data that are observed. This can be elucidated from the Jimma longitudinal family survey of youth result, where the combined model is less complex in fit, which likely results from the conjugacy of the beta random effect and the number of subjects as well as the repeated measurements per subject (Kassahun et al., 2012). In this chapter, we have shown that Bayesian estimation via MCMC provides similar results as partial integration-based inference for the combined model for binomial data. The Bayesian methodology provides a useful alternative for parameter estimation of the combined model.

Although this study mainly focussed on the technical aspects of the scientific question at hand, the interpretation of the results remained largely untouched. The Jimma infant growth study revealed however that early breastfeeding lowers the risk of overweight

Table 4.6: Jimma infant growth study. Estimated posterior mean and standard deviation in (1) the logistic-normal model, and (2) the combined model.

		Logistic-normal	Combined
Effect		Mean(s.d.)	Mean(s.d.)
Intercept	ξ_0	$-2.773(0.191)$	$-2.755(0.258)$
Time	ξ_1	0.137(0.042)	0.169(0.062)
Gender: male	ξ_2	0.020(0.054)	0.026(0.069)
Place: rural	ξ_3	$-0.915(0.065)$	$-1.115(0.085)$
Place: urban	ξ_4	$-0.606(0.083)$	$-0.749(0.103)$
Breastfeeding	ξ_5	0.666(0.185)	0.903(0.253)
Slope gender: male	ξ_6	$-0.003(0.010)$	$-0.006(0.012)$
Slope place: rural	ξ_7	0.015(0.011)	0.026(0.015)
Slope place: urban	ξ_8	0.011(0.014)	0.017(0.018)
Slope breastfeeding	ξ_9	$-0.144(0.041)$	$-0.192(0.061)$
Std. dev. random intercept	$\sqrt{d_0}$	1.783(0.035)	2.212(0.074)
Std. dev. random slope	$\sqrt{d_1}$	0.193(0.007)	0.250(0.013)
Ratio	α/β	—	0.288(0.031)
DIC		33605.1	33377.6
pD		5400.7	6218.3

at late infancy. This finding is in line with Bergmann et al. (2003), who showed that breastfed infants had lower BMI's after 3 months from birth than bottle-fed infants, though the BMIs at birth were nearly identical in both groups. Owen et al. (2005), who reviewed 61 studies, states that initial breastfeeding protects against obesity in later life, although the precise magnitude of the association remains unclear. Unlike Owen et al. (2005), the present study showed that infants in the breastfed group were fatter, at birth, as compared to those who were not breastfed. This is likely because of the unmeasured maternal history, such as maternal BMI, and socio-cultural aspects, which are considered to be the risk factors of overweight in children (Gillman et al., 2006). In addition, it is a common practice in the study area that mothers provide additional liquid or solid food starting from early infancy, in addition to breastfeeding. This is probably because they believe that a child with more weight is considered as healthy, which is likely to have its own impact on the BMI in the early infancy. In this

Table 4.7: Jimma longitudinal family survey of youth. Estimated posterior mean and standard deviation in (1) the logistic model, (2) the beta-binomial model.

		Logistic	Beta-binomial
Effect		Mean(s.d.)	Mean(s.d.)
Intercept	ξ_0	1.185(0.624)	1.151(0.731)
Age	ξ_1	0.039(0.049)	0.047(0.057)
Place: urban	ξ_2	0.977(0.148)	1.134(0.183)
Place: semi-urban	ξ_3	0.987(0.161)	1.161(0.202)
Gender: female	ξ_4	$-1.113(0.123)$	$-1.266(0.148)$
Work	ξ_5	0.133(0.122)	0.154(0.140)
Round	ξ_6	0.343(0.142)	0.404(0.165)
Std. dev. random effect	$\sqrt{d}$	—	—
Ratio	α/β	—	0.0111(0.0029)
DIC		2002.0	2001.0
pD		6.97	13.77

study, it is also shown that place of residence does not have a long term effect in the risk of being overweight. Instead it is the mode of feeding, which is more important. Spatial differences, in the sense of differences observed in the risk of overweight among infants living in urban versus semi-urban areas, might be attributable to other family related factors like social class, family income, educational level of the parents, and other socio-cultural variables, which are indicated to affect the nutrition of young children and women in Ethiopia (Macro, 2008). Future studies on early growth of children could benefit from careful measurement of a wider range of potential confounders of overweight.

In investigating school attendance among adolescents, it was shown that girls have a lower rate of current school attendance than boys, which is a common situation in most Sub-Saharan African Countries. According to the World Health Organization (WHO, 2009), there was a clear gender gap observed in primary or secondary school enrolment when the Gender Parity Index (GPI), the ratio of female to male enrolment, is considered. Between the years 1999 and 2003, the GPI was found to be 0.7, indicating that there were only 7 girls enrolled at primary schools for every 10 boys. This gender gap increases as

Table 4.8: Jimma longitudinal family survey of youth. Estimated posterior mean and standard deviation in (2) the logistic-normal model, and (2) the combined model.

		Logistic-normal	Combined
Effect		Mean(s.d.)	Mean(s.d.)
Intercept	ξ_0	1.452(0.732)	1.272(0.953)
Age	ξ_1	0.047(0.057)	0.077(0.078)
Place: urban	ξ_2	1.107(0.180)	1.427(0.270)
Place: semi-urban	ξ_3	1.104(0.192)	1.382(0.269)
Gender: female	ξ_4	$-1.247(0.149)$	$-1.528(0.214)$
Work	ξ_5	0.155(0.145)	0.199(0.184)
Round	ξ_6	0.401(0.157)	0.521(0.203)
Std. dev. random effect	$\sqrt{d}$	1.148(0.203)	1.417(0.266)
Ratio	α/β	—	0.013(0.003)
DIC		1943.0	1915.0
pD		241.5	211.9

the level of education increases. This study showed spatial differences, with adolescents in urban and semi-urban area having a higher rate of school attendance than those in the rural areas, which is in line with report of the World Bank (2005), where it was stated that among children in rural areas with a school in the neighbourhood, less than 44 $\%$ registered for school; in urban areas, the percentage is much higher (up to 86 $\%$). According to the report, the distance to the nearest school, household characteristics, and the learning environment were among the possible reasons of the gap in school attendance. Further efforts should be made to fill the gap in school attendance among boys and girls, also in urban and rural areas by focusing on the potential causes, such as lagging experience in primary schooling, which is then exacerbated by such factors as the practice of early marriage among Ethiopian women and families' reluctance to invest in girls' education. Situating schools closer to children's homes in rural areas, and an improvement of the quality of the services is necessary. Longitudinal studies with a larger number of repeated measurements per subject should indeed be conducted to get better insights in these long-time school enrolments.

Chapter 5

Integrated Nested Laplace Approximation for the Combined Model for Count Data

5.1 Introduction

In Chapter 4, partial likelihood and MCMC were compared and gave similar results. While especially the Bayesian framework is attractive in several applications in this thesis, such as multi-hierarchical and spatial designs, MCMC-based computation can be time-consuming. Rue et al. (2009) proposed integrated nested Laplace approximation (INLA) as an alternative estimation method for Bayesian computing to overcome the computational burden of MCMC. The method has already been proven successful in many situations (Paul et al. 2010, Schrödle and Held 2010, Riebler et al. 2011), but others, e.g. Taylor and Diggle (2014) have criticized the claim that INLA is more robust than MCMC (Paul et al. 2010). Indeed, as INLA is an approximation method, it is of interest to investigate whether or not the method is useful for estimation of the combined model in terms of (1) shortened computation time and (2) quality of the parameter estimation, in comparison with MCMC.

In this chapter, INLA will be introduced within the context of the combined model.

Section 5.2 will introduce partial integration, MCMC and INLA within the context of count data, while Section 5.3 will focus on case studies and a simulation study. Finally, concluding remarks are given in Section 5.4.

5.2 Estimation methods

Before introducing integrated nested Laplace approximation (INLA), I will recall partial integration and MCMC. In Chapter 4, both estimation techniques were already investigated for binomial data, but the Poisson case remained untouched. The focus in this chapter will be put on count data, since the binomial combined model cannot be estimated yet via integrated nested Laplace approximations. The reason for this is fairly simple: INLA only provides a number of likelihood functions, such as the negative binomial or beta-binomial, but the combined model is not yet a part of this selection. Due to strong conjugacy, the combined model for Poisson data can be formulated though as a negative binomial model with a normal random effect. For binomial data however, the lack of strong conjugacy inhibits the combined model formulation as a beta-binomial model with a normal random effects. This issue will be made more clear in Section 5.2.3.

5.2.1 Partial integration

Recall that in partial integration, the conjugate random effect is first integrated out analytically from the likelihood, while the normal random effect is integrated out numerically. Thus, while partial integration leads to the likelihood given in (3.3), the gamma random effects, that are assumed to be independent within a subject are integrated out, leading to the probability

$$f(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\xi}, \alpha) = \binom{\alpha + y_{ij} - 1}{\alpha - 1} . \left(\frac{1/\alpha}{1 + \kappa_{ij}/\alpha}\right)^{y_{ij}} . \left(\frac{1}{1 + \kappa_{ij}/\alpha}\right)^{\alpha} \kappa_{ij}^{y_{ij}},$$

corresponding with the specification of the model in (3.8)-(3.10).

5.2.2 Markov chain Monte Carlo

A general Bayesian formulation of the combined model was also already presented in Section 4.2.2. In the Poisson case, the posterior distribution becomes

$$p(\boldsymbol{\xi}, \boldsymbol{b}_i, D, \alpha|\boldsymbol{y_i}) \propto \left[\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b}_i, \boldsymbol{\theta}_i) f(\boldsymbol{b}_i|D) f(\boldsymbol{\theta}_i|\alpha)\right] [p(\boldsymbol{\xi}) p(D) p(\alpha)],$$

where $\alpha = 1/\beta$ denotes the gamma shape parameter. While MCMC is the most popular Bayesian estimation technique, the computational burden of its sampling-based nature can become problematic as was mentioned earlier, in practice leading to (1) long computing times and (2) sampling from non-converged posteriors. These issues have been the main reason for researchers to search for approximate methods which are as accurate as MCMC but with much faster computation times.

5.2.3 Integrated nested Laplace approximation

Rue and Held (2005) proposed an alternative method for the traditional MCMC-based Bayesian analyses, namely INLA (integrated nested Laplace approximation). INLA is an alternative Bayesian estimation method for models with a Gaussian Markov random field (GMRF). In order to use INLA to estimate the combined model (3.4)-(3.7), partial integration of the conjugate random effect is again required. The posterior distribution after partial integration is given by

$$p(\boldsymbol{\xi}, \boldsymbol{b}_i, \boldsymbol{\vartheta}_i, D, \Sigma_i | \boldsymbol{y}) \propto \left[\prod_{j=1}^{n_i} f_{ij}(y_{ij} | \boldsymbol{\xi}, \boldsymbol{\vartheta}_i, \Sigma_i, \boldsymbol{b}_i) f(\boldsymbol{b}_i | D) \right] [p(\boldsymbol{\xi}) p(D) p(\boldsymbol{\vartheta}_i) p(\Sigma_i)] ,$$

where now $f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{\vartheta}_i, \Sigma_i, \boldsymbol{b}_i)$ corresponds with the distribution resulting from integrating out the conjugate random effects (e.g. the negative-binomial distribution in the Poisson case), as before. Interest is in the posterior distribution of the random effects $p(\boldsymbol{b}_i|\boldsymbol{y})$ and of the so-called hyperparameters $p(\boldsymbol{\xi}|\boldsymbol{y})$, $p(D|\boldsymbol{y})$, $p(\boldsymbol{\vartheta}_i|\boldsymbol{y})$ and $p(\Sigma_i|\boldsymbol{y})$. Again, the Poisson case follows from replacing $\boldsymbol{\vartheta}_i$ and Σ_i by α. Instead of deriving these marginal posteriors via sampling, INLA approximates them using Laplace approximations, using three steps.

First, the marginals of the hyperparameter vector are approximated. Say we use $\boldsymbol{z} = (\boldsymbol{\xi}, \boldsymbol{\vartheta}_i, \Sigma_i, D)$, then this marginal can be expressed as

$$p(\boldsymbol{z}|\boldsymbol{y}) = \frac{p(\boldsymbol{b}, \boldsymbol{z}|\boldsymbol{y})}{p(\boldsymbol{b}|\boldsymbol{z}, \boldsymbol{y})}$$

for any vector $\boldsymbol{b}$. A Laplace approximation is now performed by replacing the denominator by a Gaussian approximation

$$\tilde{p}_G(\boldsymbol{b}|\boldsymbol{z}, \boldsymbol{y}) \propto \exp\{-\frac{1}{2}(\boldsymbol{b} - \boldsymbol{\mu}(\boldsymbol{z}))' \boldsymbol{Q}^*(\boldsymbol{z})(\boldsymbol{b} - \boldsymbol{\mu}(\boldsymbol{z}))\},$$

by matching the mode and curvature of the approximation with a series expansion of the original density, followed by a evaluation for each $\boldsymbol{z}$ at the mode of the Gaussian approximation. This results in

$$\tilde{p}(\boldsymbol{z}|\boldsymbol{y}) = \frac{p(\boldsymbol{b}, \boldsymbol{z}|\boldsymbol{y})}{\tilde{p}_G(\boldsymbol{b}|\boldsymbol{z}, \boldsymbol{y})} |_{\boldsymbol{b}=\boldsymbol{b}^*(\boldsymbol{z})},$$

in which $\boldsymbol{b}^*(\boldsymbol{z})$ denotes the mode of the iteratively obtained Gaussian approximation.

In a second step, approximations to $p(\boldsymbol{b}_i|\boldsymbol{y},\boldsymbol{z})$ are calculated, with three ways to go forward: (1) a simple but notoriously inaccurate method (termed the 'gaussian' option when using R-INLA) is to derive the marginals as the univariate marginals of the Gaussian approximation $\tilde{p}_G(\boldsymbol{b}|\boldsymbol{z},\boldsymbol{y})$, by $\tilde{p}(\boldsymbol{b}_i|\boldsymbol{z},\boldsymbol{y}) = N(\boldsymbol{b}_i;\mu_i(\boldsymbol{z}),\sigma_i^2(\boldsymbol{z}))$. (2) A second method uses the Laplace approximation, as described above, but the marginals of the latent field are now estimated by

$$\tilde{p}(\boldsymbol{b}_i|\boldsymbol{z},\boldsymbol{y}) \propto \frac{p(\boldsymbol{b},\boldsymbol{z}|\boldsymbol{y})}{\tilde{p}_{GG}(\boldsymbol{b}_{-i}|\zeta_i,\boldsymbol{z},\boldsymbol{y})}|_{\boldsymbol{b}_{-i}=\boldsymbol{b}^*_{-i}(\boldsymbol{b}_i,\boldsymbol{z})},$$

where $\tilde{p}_{GG}$ is the Gaussian approximation to the distribution $\boldsymbol{b}_{-i}|\boldsymbol{b}_i,\boldsymbol{z},\boldsymbol{y}$. Due to the fact that this method, which is denoted as 'laplace' in R-INLA, requires an iterative search of the mode $\boldsymbol{b}^*_{-i}(\boldsymbol{b}_i,\boldsymbol{z})$ for all $\boldsymbol{b}_i$'s, it is known to become time-consuming. (3) The default in R-INLA, termed 'simplified.laplace', performs a series expansion of the latter Laplace approximation and then fits a skew-normal distribution to this series expansion. The use of the skew-normal brings more flexibility than when working with a Gaussian distribution. On top of that, it can be understood as an approximation with improved location and skewness errors, which is claimed to be very fast and still very accurate (Rue et al., 2009).

The resulting estimates of the marginals can thus be given as

$$\begin{aligned}\tilde{p}(\boldsymbol{b}_i|\boldsymbol{y}) &= \int \tilde{p}(\boldsymbol{b}_i|\boldsymbol{y},\boldsymbol{z})\tilde{p}(\boldsymbol{z}|\boldsymbol{y})d\boldsymbol{z}\\ \tilde{p}(z_j|\boldsymbol{y}) &= \int \tilde{p}(\boldsymbol{z}|\boldsymbol{y})d\boldsymbol{z}_{-j},\end{aligned}$$

with $\tilde{p}(\boldsymbol{b}_i|\boldsymbol{y},\boldsymbol{z})$ found by one of the three approximation strategies mentioned above. Because the main interest lies in the estimation of the marginals of the latent field, the third and final step is to perform numerical integration with respect to $\boldsymbol{z}$, by finding the sum

$$\tilde{p}(\boldsymbol{b}_i|\boldsymbol{y}) \approx \sum_k \tilde{p}(\boldsymbol{b}_i|\boldsymbol{y},z_k)\tilde{p}(z_k|\boldsymbol{y})\Delta_k,$$

where Δ_k denotes the area weight corresponding to the integration point z_k. There are two ways of choosing these points. In this manuscript, we only use the so-called Central Composite Design (CCD) strategy, which also is the default in R-INLA. Here, a small amount of support points in the m-dimensional space of the hyperparameter vector are used and each center point is augmented with a group of points used to estimate the curvature of $\tilde{p}(\boldsymbol{z}|\boldsymbol{y})$, yielding a fast and precise method (Rue et al., 2009).

5.3 Estimation method comparison

In this section, data analyses are described for both the epilepsy data (Section 2.2.4) and the Flemish contact data sets (Section 2.2.5). For both analyses, results for the MCMC and INLA estimation are provided. Three versions of the INLA estimation are provided, as introduced in Section 5.2.3: the simplified Laplace approximation, the full Laplace and the Gaussian approximation. Likelihood-based results are only provided for the epilepsy data. Here, estimation could be easily done with partial integration. The Flemish contact data set however has multiple hierarchies, as observations were nested within house households and households were nested within towns. Hence, a multi-hierarchical model was needed, which is easy to fit using Bayesian techniques, but with no straight-forward procedure in the likelihood setting. The likelihood, MCMC and INLA analyses were done in SAS 9.4 (proc NLMIXED), WinBUGS 14 (via BRUGS in R 3.0.1) and R 3.0.1 (R-INLA package), respectively. INLA's default prior specifications were used, namely $N(0, 1000)$ for the 'fixed' effects, $\gamma(1, 10000)$ for the precision of structured normal random effects and $\gamma(1, 1)$ for the overdispersion parameter in the negative binomial model. The same prior specification was used in the MCMC analysis.

In order to make a comparison between the INLA and MCMC estimation methods, the agreement statistic, in line with the accuracy statistic as proposed by Faes et al. (2011), is proposed. If $f(\zeta|\boldsymbol{y})$ is defined as the posterior density of a parameter ζ estimated by MCMC and $\tilde{f}(\zeta|\boldsymbol{y})$ as the posterior density estimated by INLA, then the integrated absolute difference is defined as $IAD(\tilde{f}, f) = \int_{-\infty}^{+\infty} |\tilde{f}(\zeta|\boldsymbol{y}) - f(\zeta|\boldsymbol{y})| d\zeta$, which is scale-independent between 0 and 2 and invariant to monotone transformations on the parameter ζ. The agreement statistic used in this thesis is defined as

$$\text{Agreement}(\tilde{f}, f) = 1 - \{IAD(\tilde{f}, f) / \sup_{\tilde{f}, f \text{ is density}} IAD(\tilde{f}, f)\} = 1 - IAE(\tilde{f}, f)/2,$$

which lays in the interval $[0, 1]$. The agreement statistic can be interpreted as the percentage of overlap between the posterior densities $f(\zeta|\boldsymbol{y}$ and $\tilde{f}(\zeta|\boldsymbol{y})$.

Table 5.1: Parameter estimates (and s.d.) for the epilepsy data and the Flemish contact data. Three INLA-based, likelihood and MCMC results are provided.

Epilepsy data				
Model	Computation time	ξ_0 Est. (s.d.)	ξ_1 Est. (s.d.)	ξ_2 Est. (s.d.)
INLA (Simp. Laplace)	5.787 sec.	0.858 (0.170)	-0.147 (0.234)	-0.018 (0.005)
INLA (Laplace)	1.287 min.	0.858 (0.170)	-0.147 (0.234)	-0.018 (0.005)
INLA (Gaussian)	6.283 sec.	0.864 (0.170)	-0.143 (0.234)	-0.018 (0.005)
Likelihood	18.550 sec.	0.858 (0.167)	-0.173 (0.267)	-0.018 (0.005)
MCMC	25.430 min.	0.915 (0.172)	-0.157 (0.260)	-0.018 (0.005)
Model		α Est. (s.d.)	σ Est. (s.d.)	
INLA (Simp. Laplace)		2.446 (0.208)	1.067 (0.087)	
INLA (Laplace)		2.446 (0.208)	1.065 (0.087)	
INLA (Gaussian)		2.446 (0.208)	1.065 (0.089)	
Likelihood		2.462 (0.212)	1.060 (0.087)	
MCMC		2.456 (0.207)	1.069 (0.088)	
Flemish contact data				
Model	Computation time	ξ_0 Est. (s.d.)	ξ_1 Est. (s.d.)	ξ_2 Est. (s.d.)
INLA (Simp. Laplace)	10.527 sec.	2.784 (0.079)	-0.004 (0.029)	-0.043 (0.015)
INLA (Laplace)	2.134 min.	2.784 (0.079)	-0.004 (0.029)	-0.043 (0.015)
INLA (Gaussian)	9.366 sec.	2.773 (0.079)	-0.004 (0.029)	-0.043 (0.015)
MCMC	58.027 min.	2.787 (0.080)	-0.005 (0.029)	-0.044 (0.015)
Model		α Est. (s.d.)	σ_0 Est. (s.d.)	σ_1 Est. (s.d.)
INLA (Simp. Laplace)		5.817 (0.368)	0.010 (0.005)	0.487 (0.025)
INLA (Laplace)		5.817 (0.368)	0.010 (0.005)	0.487 (0.025)
INLA (Gaussian)		5.817 (0.368)	0.010 (0.005)	0.487 (0.025)
MCMC		5.813 (0.367)	0.014 (0.014)	0.489 (0.026)

5.3.1 Epilepsy data

Let Y_{ij} be the number of epileptic seizures for the i^{th} person ($i = 1, ..., 89$) in week j ($j = 1, ..., n_i$). The following combined model was fitted to the data:

$$\begin{aligned} Y_{ij} &\sim \text{Poi}(\theta_{ij}\kappa_{ij}), \\ \kappa_{ij} &= \exp\left(\xi_0 + b_{0,i} + \xi_1 Trt_i + \xi_2 Time_{ij}\right), \\ b_{0,i} &\sim N\left(0, \sigma^2\right), \\ \theta_{ij} &\sim \text{gamma}(\alpha, 1/\alpha) \end{aligned}$$

with ξ_0, ξ_1 and ξ_2 being the intercept, the treatment and time effect respectively, $b_{0,i}$ the random intercept effects with standard deviation σ and θ_{ij} the conjugate random effect

with parameter α. Indeed, due to strong conjugacy in the Poisson case, the combined could be written as a negative binomial model that includes a normal random effect, as is done in INLA.

Results are given in the top panel of Table 5.1. MCMC results are based on 10000 runs after a burn-in of 10000 runs. All estimation methods agree on a non-significant treatment effect, while there is a slightly negative significant time effect. Only small differences between the simplified Laplace, full Laplace and Gaussian options in INLA were observed, with the first almost exactly the same. Figure 5.1 gives a visual representation of the posterior and likelihood estimates for the epilepsy data example. INLA and MCMC posteriors are very similar for the treatment and time effects ξ_1, ξ_2, the random effects standard deviation σ and the overdispersion parameter α ($> 94\%$) while the agreement was lower for the intercept, ξ_0 (85%). Also note that the ML estimates coincide with the MCMC and INLA posterior's mode for all parameters, except for a slight difference for the intercept ξ_0. In conclusion, in this example, all methods give similar results, while a major time gain is obtained with INLA as compared to MCMC.

5.3.2 Contact data

The number of contacts Y_{ijk} in municipality i ($i = 1, ..., 211$) in household j ($j = 1, ..., n_i$) for individual k ($k = 1, ..., n_{ij}$) was modeled using a multi-hierarchical combined model. The model can be summarized as follows:

$$\begin{aligned}
Y_{ijk} &\sim \text{Poi}(\theta_{ijk}\kappa_{ijk}),\\
\kappa_{ijk} &= \exp\left(\xi_0 + b_{0,i} + b_{1,ij} + \xi_1 Sex_{ijk} + \xi_2 Time_{ijk}\right),\\
b_{0,i} &\sim N\left(0, \sigma_0^2\right),\\
b_{1,ij} &\sim N\left(0, \sigma_1^2\right),\\
\theta_{ijk} &\sim \text{gamma}(\alpha, 1/\alpha).
\end{aligned}$$

In analogy with the epilepsy data model, ξ_0, ξ_1 and ξ_2 are the intercept, the gender (female is reference) and time effect, respectively. Now, $b_{0,i}$ and $b_{1,ij}$ are random effect terms modeling the correlation within municipalities and within households, respectively. And again, α is the overdispersion parameter.

Results are given in the lower panel of Table 5.1 and in Figure 5.2. For the contact data, there was no significant gender effect, but a negative time effect was present. Again, these results were in good agreement between all estimation methods. The agreement statistics show that the INLA and MCMC posteriors are very alike, except for σ_0. The

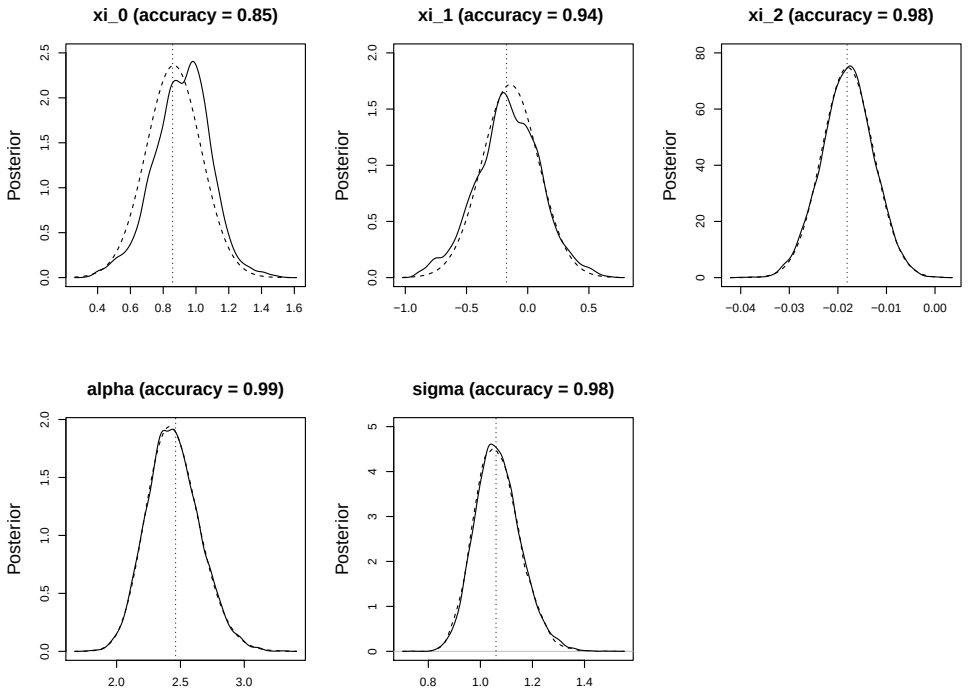


Figure 5.1: Visual representation of the posterior/likelihood parameter estimates for the epilepsy data example. Full line = MCMC, dashed line = INLA (strategy = simplified Laplace), dotted line (vertical) = MLE.

INLA method provides smaller estimates for the heterogeneity amongst the municipalities as compared to the MCMC method.

5.3.3 Simulation study

Next to the results given by the case studies, it is useful to conduct a simulation study in order to investigate the properties of the different methods further on. In accordance to the epilepsy case study model, data were simulated as coming from a combined negative binomial model for count data,

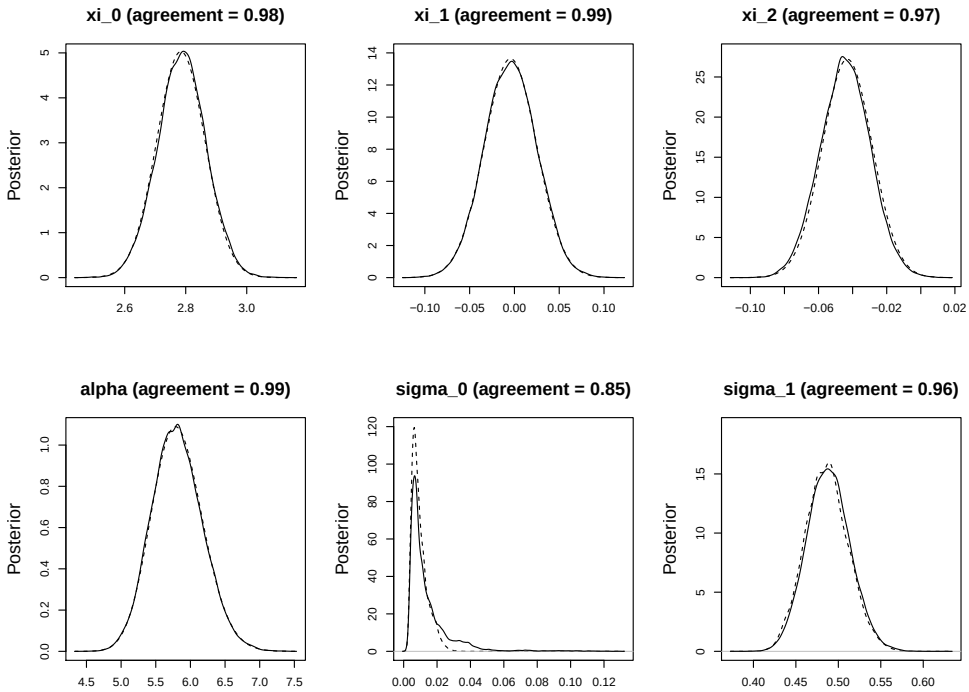


Figure 5.2: Visual representation of the posterior parameter estimates for the contact data example. Full line = MCMC, dashed line = INLA (strategy = simplified Laplace).

$$
\begin{aligned}
Y_i &\sim \text{Poi}(\theta_i \kappa_i), && (5.1)\\
\kappa_i &= \exp\left(\xi_0 + b_{0,i} + \xi_1 Trt_i + \xi_2 Time_i\right), && (5.2)\\
b_{0,i} &\sim N\left(0, \tau^{-1}\right), && (5.3)\\
\theta_i &\sim \text{gamma}(\alpha, 1/\alpha) && (5.4)
\end{aligned}
$$

For each individual, observations at 5 time points were generated according to (5.1)-(5.4). In total, 4 different simulation settings were considered, with either 50 or 200 individuals generated on one hand, and with σ either 0.5 or 2 on the other hand. For each setting, 100 data sets were created and analyzed with INLA (here only the simplified Laplace approximation strategy was applied), partial integration and MCMC. In INLA, default priors were used for ξ_0, ξ_1, ξ_2 and τ. In MCMC, normal priors with mean$= 0$ and precision$= 0.0001$ were given to ξ_0, ξ_1 and ξ_2 and a gamma distribution with parameters 0.1 and 1000 to τ. A normal prior with a

mean $= 0$ and precision $= 0.01$ was specified for log(α) in both INLA and MCMC analyses. Mean bias $= \sum_{n=1}^{i}(\widehat{\theta}_i - \theta^{true})/n$, estimated variance $= \sum_{n=1}^{i} s.e.(\widehat{\theta}_i)/n$ and mean squared error (MSE) $= \sum_{n=1}^{i}(\widehat{\theta}_i - \theta^{true})^2/n$ with $i = 1, .., n$ were used as summary statistics to compare the results. Next to that, the agreement statistic, as explained earlier, was applied to investigate in what way the INLA and MCMC results were alike. Table 5.2 summarizes the results.

The average calculation times in INLA were up to several hundreds of times faster than calculations done with MCMC and also slightly faster but still within the same time frame of the partial integration's computation times. Exact numbers are not given, since in an attempt to obtain research results within a manageable time frame, the simulation study was divided among several computers with differing processor strength ($1.9 - 2.3$ GHz) and memory ($4096 - 16382$ MB RAM), which would make an in-depth comparison in computation times unfair. It suffices to say however that when an INLA analysis (INLA analyses never took longer than a few seconds) took around 5 seconds, the analyses via MCMC would take approximately half an hour. Indeed, the enormous shortening of computing time has been argued before to be the major advantage of INLA over MCMC.

Before comparing the different estimation techniques, a few observations appear immediately when looking at the INLA results (Table 5.2) only (with similar results for MCMC and the likelihood approach). First of all and as expected, results become more accurate when the sample size increases. Indeed, the MSE values for all parameter estimates clearly show a downward trend as sample sizes increase. On the other hand, it is apparent that when the standard deviation of the normal random effect increases, the parameter estimates change in terms of their precision, but not all in the same way. Fixed effect and normal random effect's variance parameters tend to become estimated less accurately when the standard deviation increase. In contrast, the overdispersion parameter is estimated more accurately when the variability in the normally distributed extra-variance increases. This is also to be expected, since estimation of fixed effects is known to be more precise when the variability in the data is small, while the overdispersion parameter estimation benefits from larger extra-variance in the data.

When a comparison between INLA, the likelihood approach and Bayesian MCMC analysis (Table 5.2) is made, some other properties are observed: Overall, INLA behaves approximately the same as the likelihood and Bayesian MCMC methods. While MSE values are almost alike, they often slightly favour the likelihood approach as compared to the Bayesian approach. However, focusing on the overdispersion parameter α, con-

siderable differences exist between the estimation techniques, especially when the sample size and standard deviation are both small (n$= 50$ and sd$= 0.50$). In this case, partial integration performs best, while both Bayesian estimation methods have large bias, with MCMC performing worse than INLA. Further, focusing on the variance parameter σ^2, it is observed that INLA produced larger (negative) bias, especially for large sample sizes. This corresponds with the observation seen in the data application, that INLA often leads to too small estimated standard deviation. When comparing INLA and MCMC, one can conclude that there is no indication to favour one of the estimation methods, which is compelling since INLA is an approximation technique that was shown earlier to work hundreds of times faster than MCMC. Furthermore, boxplots of the agreement statistics (Figure 5.3) give an idea of how much overlap there is between the INLA and MCMC posterior densities. Interestingly, very high overlap (between 0.9 and 1) was seen for all covariates and the overdispersion parameter, indicating that INLA and MCMC give almost the same results. The standard deviation of the normal random effects term had a consistently lower agreement in settings with low heterogeneity ($\sigma = 0.5$), which is in line with the results from the case studies performed earlier.

5.4 Concluding remarks

In this chapter, it has been made clear that when working with the combined model for count data, INLA provides a very good alternative to MCMC, with computation times that are shortened up to 600 times and parameter estimates that are as precise as those given by the more conventional methods, such as PROC NLMIXED in SAS or MCMC in WinBUGS. Small differences between the estimation options do sometimes occur, most notably for standard deviations of the normal random effects. The agreement statistics also show that MCMC and INLA provide similar results, but that differences in the estimated standard deviations can be considerable, with an underestimation of the variability by INLA. This is mainly observed when the standard deviation is small.

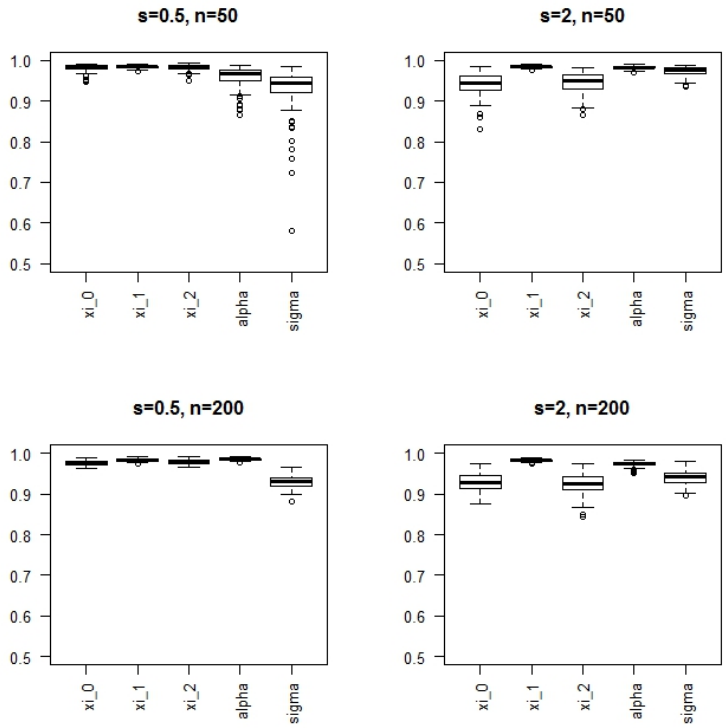


Figure 5.3: Boxplots of the agreement statistic per simulation setting for each estimated parameter.

Table 5.2: Results for the likelihood, INLA and MCMC based simulation studies. Four situations were simulated, which differed in terms of sample size ($n = 50$ and $n = 200$) and structured random effects standard deviation ($\sigma = 0.5$ and $\sigma = 2$). Other true values are $\xi_0 = 0.9$, $\xi_1 = -0.25$, $\xi_2 = -0.025$ and $\alpha = 2.5$.

n	Parameter	Mean Bias	Variance	MSE	Mean Bias	Variance	MSE
		s.d.=0.5			s.d.=2		
		INLA					
50	ξ_0	-0.0072	0.0344	0.0387	-0.0332	0.1978	0.2168
	ξ_1	-0.0129	0.0365	0.0395	0.02081	0.3617	0.3943
	ξ_2	-0.0036	0.0018	0.0018	-0.0021	0.0018	0.0017
	α	0.3194	0.9589	1.3371	0.1677	0.2765	0.3548
	σ	-0.0248	0.0092	0.0118	0.0234	0.0615	0.0636
200	ξ_0	0.0123	0.0082	0.0081	0.0134	0.0475	0.0452
	ξ_1	-0.0125	0.0086	0.0100	-0.0058	0.0868	0.0716
	ξ_2	-0.0006	0.0004	0.0004	0.0008	0.0004	0.0004
	α	0.0384	0.0966	0.0999	0.0210	0.0525	0.0510
	σ	-0.0163	0.0021	0.0023	-0.0126	0.0142	0.0156
		Likelihood					
50	ξ_0	-0.0063	0.0333	0.0385	0.0256	0.1940	0.1961
	ξ_1	-0.0119	0.0343	0.0391	-0.0166	0.3567	0.3612
	ξ_2	-0.0035	0.0018	0.0018	0.0064	0.0017	0.0019
	α	0.2003	0.6766	0.8610	0.2128	0.2756	0.3887
	σ	-0.0231	0.0090	0.0088	-0.0106	0.0601	0.0777
200	ξ_0	0.0117	0.0082	0.0080	0.0159	0.0483	0.0369
	ξ_1	-0.0122	0.0085	0.0099	0.0020	0.0882	0.0904
	ξ_2	-0.0005	0.0004	0.0004	-0.0035	0.0004	0.0004
	α	0.0222	0.0935	0.0948	0.0364	0.0530	0.0599
	σ	-0.0129	0.0021	0.0020	-0.0002	0.0146	0.0103
		MCMC					
50	ξ_0	-0.0018	0.0344	0.0387	-0.0215	0.2000	0.2267
	ξ_1	-0.0137	0.0361	0.0394	0.0835	0.3655	0.4049
	ξ_2	-0.0027	0.0018	0.0018	-0.0015	0.0018	0.0016
	α	0.4134	10.8652	3.1744	0.1474	0.2726	0.3546
	σ	-0.0221	0.0114	0.0127	0.0270	0.0679	0.0618
200	ξ_0	0.0078	0.0082	0.0079	0.0061	0.0481	0.0373
	ξ_1	-0.0082	0.0084	0.0099	0.0342	0.0829	0.0695
	ξ_2	-0.0002	0.0004	0.0004	0.0002	0.0004	0.0004
	α	0.0391	0.0978	0.1010	0.0031	0.0516	0.0525
	σ	-0.0086	0.0021	0.0020	-0.0097	0.0150	0.0164

Chapter 6

The Spatial Combined Model for Count Data

6.1 Introduction

In Chapters 4 and 5, the use of the combined model was illustrated for binomial and count data. From this point onwards, a progression will be made to spatial count data, more specifically to disease mapping, a relatively new scientific field in which disease count data are modeled in a spatial data setting. The main goal is to describe the spatial distribution of the disease with respect to the place of occurrence. Indeed, in our increasingly health-conscious and environmentally aware society, members of the public are much more likely to notice unusual aggregations of a disease in a small neighbourhood and to attribute them to some nearby industrial source of pollution. To investigate these claims, the number of cases is related to the number expected counts for a typical population, which is accomplished through disease mapping.

Typically, hierarchical Bayesian methods are used to model the overdispersion and the spatial correlation in the data. Classic random-effects based solutions to deal with overdispersion, such as the previously introduced Poisson-gamma model and the Poisson-lognormal model, have mainly been popular in the early years of disease mapping (e.g. Clayton and Kaldor, 1987). The reason for this overdispersion can be manifold, e.g., a misspecification of the model in terms of a forgotten (spatially unstructured) covariate, an excess of zero counts or many outlying counts, problems encountered frequently in disease mapping. On the other hand, focusing on the spatial autocorrelation in the data,

the conditional autoregressive model (Besag et al., 1991) has gained a lot of popularity. Due to the flexibility to apply many different weighting schemes, the model has been used extensively within the literature and has been proved to provide better solutions than the spatially unstructured counterparts. Also, its straight-forward implementation in Bayesian software such as WinBUGS (Spiegelhalter et al., 2007) has made it the most attractive among other so-called convolution models, which combine an unstructured random effects term (uncorrelated heterogeneity, UH) and a spatially-structured random effects term (correlated heterogeneity, CH) (Anselin, 1988; Cressie, 1993).

Although a convolution model provides a method to model spatially structured and unstructured variation together, the interesting conjugate feature in the Poisson-gamma framework that cannot be achieved via the Gaussian implementation has remained relatively unexplored. Wolpert and Ickstadt (1998) undertook an effort by using correlated gamma field models, however a simulation study comparing different disease mapping models (Best et al., 2005) noted a poor performance of this model. So far, extensions of the Poisson-gamma model to account for spatial heterogeneity have been limited (e.g. the gamma field models mentioned earlier). The reasons are two-fold: (1) because the gamma distribution does not easily allow for inclusion of covariate effects, and (2) because the gamma distribution does not easily extend to include spatial structure. The combined model offers a convenient method for including covariate effects and random effects. In this chapter, an extension of the combined model towards the inclusion of spatial correlation is proposed and the applicability of this model in the spatial disease mapping context will be compared with the classical used models.

The structure of this chapter is as follows: Section 6.2 gives an overview of the mostly used models, Section 6.3 focuses on prior specification, while case studies and a simulation study are conducted in Sections 6.4 and 6.5 respectively. In Section 6.6, MCMC and INLA estimation for the spatial combined model are compared and a general conclusion is given in Section 6.7.

6.2 Disease mapping models

An important feature in disease mapping is the use of offsets, the so-called expected counts, which are usually standardized for confounders such as age class and gender. Two main standardization techniques exist, being direct and indirect standardization. In direct standardization, one calculates the sex-age rates (or the rates based on the particular confounders at play) and applies it to the population, while in indirect standardization,

which is used in these analyses, it works the other way around. Here, the sex-age rates of the overall region are applied to the each individual stratum that is studied. The sum of those expected counts in each stratum provides the standardized expected counts for each area.

Disease mapping models are used to link the observed counts Y_i for spatial (lattice) location $i = 1, ..., n$ to the expected counts E_i and they mainly differ in the way they smooth away the extra variation seen in Y_i in comparison to E_i. Let ω_i denote the unknown relative risk for the ith area ($i = 1, \ldots, n$). Many models have been proposed to estimate the relative risk. A classical model assumes that Y_i simply follows a Poisson distribution with parameter $\lambda_i = E_i\omega_i$, with ω_i independent. The maximum likelihood estimator for ω_i coincides here with another important statistic of the risk in a given area, namely the standardized incidence rate

$$\widehat{\omega}_i = \mathsf{SIR}_i = Y_i/E_i,$$

(Figure 6.2). The use of SIR estimates will mostly be insufficient to model real-world dynamics because of overdispersion and the spatial dependence in the data, such that an extension of this basic statistical model is necessary. This can be done in numerous ways and in what follows, a short overview is given of some frequently used extensions, with on the one hand the Poisson-gamma and Poisson-lognormal models which model overdispersion only (the so-called uncorrelated heterogeneity, UH) and on the other hand the so-called convolution models which include terms for both the overdispersion and the spatial correlation (correlated heterogeneity, CH).

The combined model for spatial lattice data is extended in the following way from (3.4)-(3.7). By way of overview, let's assemble the different parts:

$$Y_i \sim \mathsf{Poisson}(E_i\kappa_i\theta_i), \tag{6.1}$$

$$\kappa_i = \exp\left(\xi_0 + \boldsymbol{x}_i'\boldsymbol{\xi} + \boldsymbol{b}_i\right), \tag{6.2}$$

$$\theta_i \sim \mathsf{gamma}(\alpha, \beta^*), \tag{6.3}$$

with $\omega_i = \kappa_i\theta_i$, $\boldsymbol{b}_i$ being the normal random effect terms and the rate parameter $\beta^* = 1/\beta$. Note that (only) within Section 6.2 I will apply a gamma parametrization with rate parameter β^* instead of the earlier used scale parameter β, this in order to make the presented formulations easier to interpret. A number of models can be formed by formulating different assumptions:

(1) Poisson-gamma model: The model in (6.1)-(6.3) reduces to the Poisson-gamma (PG), or negative binomial model, when one sets $\exp\left(\xi_0 + \boldsymbol{x}_i'\boldsymbol{\xi} + \boldsymbol{b}_i\right) = 1$.

As already covered in Chapter 3, a closed-form posterior distribution can be provided here and is given by a gamma distribution with parameters $Y_i + \alpha$ and $E_i + \beta^*$, respectively. As a result, the posterior mean of ω_i is a weighted average of the prior mean α/β^* and the SIR, Y_i/E_i. Because of the mathematical convenience due to the conjugacy, the Poisson-gamma model has been one of the most commonly used models in disease mapping. Because of the disadvantage that this model does not take the spatial dependence into account, together with the difficulty to include covariates in this model, the Poisson-gamma model has been criticized and shown to be inferior to more complex models such as the CAR convolution model (Lawson et al., 2000).

(2) Poisson-lognormal model: When it is assumed that $\theta_i = 1$ and $\boldsymbol{b}_i = b_{0,i} \sim N(0, \sigma_0^2)$, the Poisson-lognormal (PN) model follows, which is a GLMM with an unstructured distributed random effects term b_i and optionally covariates $\boldsymbol{x}_i$. Although in some situations the PG and the PN behave similarly (Kim et al.., 2002), the mean-variance relationship of the random-effect terms, being linear for the gamma distribution and quadratic for the lognormal distribution can cause the PN to be more conservative (less extreme in range) when estimating UH. Bayesian estimation with PN is straightforward in e.g. WinBUGS and due to the availability of powerful software packages, this model, which is easily extended with covariates, has become very popular. While this model does not yet account for spatial autocorrelation, it can be easily extended with a parameter representing CH, resulting in a so-called convolution model.

(3) Convolution model: A well-known convolution model controlling for spatial autocorrelation is the conditional autoregressive (CAR) convolution model. It is formed by assuming in (6.1)-(6.3) that $\theta_i = 1$ and that $\boldsymbol{b}_i$ consists of two normal random effects terms, namely $b_{0,i} \sim N(0, \sigma_0^2)$ to capture UH and $b_{1,i}$, an intrinsic CAR model such as introduced by Besag and Kooperberg (1995),

$$b_{1,i}|b_{1,j,i\neq j} \sim N\left(\bar{\mu}_i, \sigma_{1,i}^2\right), \tag{6.4}$$

$$\bar{\mu}_i = \frac{1}{\sum_{j=1}^N w_{ij}} \sum_{j=1}^N w_{ij} u_j, \tag{6.5}$$

$$\sigma_{1,i}^2 = \frac{\sigma_1^2}{\sum_{j=1}^N w_{ij}}, \tag{6.6}$$

which takes the heterogeneity caused by the spatial structure into account. Here, $w_{ij} = 1$ if areas i and j adjacent and 0 otherwise. Indeed, the CAR random effect is normally distributed with the mean and variance being weighted with the means and variances of adjacent areas. Although the weighting scheme presented above is

the most common one, others can be applied too. Bivand et al. (2008) provide an in-depth view on this issue. Note that the CAR convolution model presented here uses an intuitively easy to understand CH prior distribution, but more methods exist to incorporate neighbourhood dependence. Moreover, a CAR model, as presented above, is a special case of the so-called proper CAR (PCAR) models, which introduce dependence between neighbourhoods, but which also allow an additional correlation parameter. More details can be found in Stern and Cressie (1999). The CAR convolution model is known to be very robust when simulating a wide range of underlying true risk models and it is therefore widely used in spatial disease mapping (Lawson et al., 2000). Problems occur though, especially in the estimation of $b_{0,i}$ and $b_{1,i}$ separately, making it sometimes unclear whether they have been attributed the correct proportion of extra-variance. Because of this, it is important to be careful when using both uncorrelated and correlated heterogeneity terms in the same model and to avoid over-interpretation of the separate CH and UH estimates. Note that in this chapter, both the convolution model as well as the model with only the CAR distributed CH term will be used, denoted 'CAR convolution (CARCON)' and 'CAR', respectively.

(4) Spatial combined model: follows when in (6.1)-(6.3) $\boldsymbol{b}_i = b_{1,i}$ with the same specification as in (6.4)-(6.6). Indeed, this model, which stems from a very different area of statistics is closely related to the CAR convolution model. The only difference is that the uncorrelated heterogeneity is modeled via a gamma distribution instead of a lognormal distribution. There is reason to believe that this model is a valuable alternative to the commonly used CAR convolution model, since the results shown earlier along those presented by Molenberghs et al. (2010) show that the gamma distribution is able to model extra-variance very well. Again note that as a result from the strong conjugacy, the posterior distribution of ω_i given the random effect $b_{1,i}$ is

$$\omega_i | b_{1,i}, Y_i \quad \sim \quad \text{gamma}(\alpha + Y_i, \beta^* + E_i \upsilon_i)$$

with $\upsilon_i = \exp(\xi_0 + \boldsymbol{x}_i' \boldsymbol{\xi} + b_{1,i})$. Thus, the conditional mean of ω_i given the random effects $b_{1,i}$ is $(\alpha + Y_i)/(\beta^* + E_i \upsilon_i)$, and can be re-written as a weighted average of the prior mean α/β^* and the area-specific standardized incidence rate Y_i/E_i, with weights $\beta^*/(\beta^* + E_i \upsilon_i)$ and $E_i/(\beta^* + E_i \upsilon_i)$, respectively. It can also be re-written as a weighted average of the prior mean α/β^* and the ratio of the incidence rate versus spatially-structured relative risk $Y_i/(E_i \upsilon_i)$, with weights $1 - w_i$ and w_i, respectively, with $w_i = E_i \upsilon_i/(\beta^* + E_i \upsilon_i)$. While these full conditionals are not of primary interest, this relationship can give us an understanding of how smoothing in obtained is this model. The weights w_i are inversely related to the variance of Y_i/E_i. Thus, for rare diseases and small areas, there is a lot of shrinkage to the prior mean α/β^*. This is similar to the Poisson-gamma

model. When a large amount of overdispersion is present in the data (β^* small), there will be less shrinkage to the prior mean α/β^*. Now also, the weights w_i depend on the spatial structure υ_i, and thus also the amount of smoothing is spatially structured. If υ_i contains a strongly spatially structured effect, the weights (and the amount of shrinkage) will also be spatially structured. It is again important to note that to avoid overparametrization problems, again a restriction needs to be applied to α and β^*. Similar to common practice in the frailty context (Duchateau and Janssen, 2008), we assumme $\alpha = 1/\beta = \beta^*$. This standardizes the gamma random effect to mean 1.

6.3 Prior specification

It was already mentioned before that the combined model can be specified as an extension of a Poisson-gamma model or a negative binomial, the latter being a PG model in which the overdispersion effect is already integrated out. If interest lies in the two random effects, corresponding to the uncorrelated and correlated heterogeneity, Bayesian estimation using the hierarchical (PG) specification such as presented in the previous section is very appealing. In the next sections, this approach will mainly be applied. Vague priors were used for the hyperprior parameters: To ensure the range of possible values of the hyperparameters α and β in the gamma distribution of the Poisson-gamma model and the combined model (only α) to be large enough, $\alpha \sim \exp(1)$ and $1/\beta \sim$ gamma$(0.1, 1)$ were used, as also suggested by Lawson (2013). The variance parameters in the normally distributed UH term of the Poisson-lognormal model and the CAR convolution model on the one hand and the CH term of the CAR (CH) model, CAR convolution model and the combined model on the other hand had $1/\sigma_0^2 \sim$ gamma$(0.5, 2000)$ and $1/\sigma_1^2 \sim$ gamma$(0.5, 2000)$ as prior distribution, again to avoid to place restrictions on the possible values of the hyperparameter values, similar to suggestions made by Kelsall and Wakefield (1999). In a sensitivity analysis, Neyens et al. (2012) compared the gamma priors for the precision parameters with a uniform prior (U$(0, 10000)$), as suggested by Gelman (2006). Also, for parameter α in the gamma-distributed CH parameter in the Combined, a comparison was done between $\alpha \sim \exp(1)$ and $\alpha \sim$ gamma$(0.1, 1)$ as suggested by George et al. (1993). For both sensitivity analyses, the presented maps of the RR estimates which are presented later on, did not change, justifying the use of the priors presented before.

If there is no genuine interest in the uncorrelated heterogeneity, but rather interest is only in the spatially correlated heterogeneity after correction for the overdispersion in the data, estimation can proceed by first integrating out the conjugate prior distribution,

conditional on the Gaussian CAR distribution. A spatial extension, such as in (6.4)-(6.6) of the model specification given in (3.8)-(3.10) is

$$\begin{aligned} Y_i &\sim \text{NegBin}(\alpha, \kappa_i/\alpha), \\ \kappa_i &= \exp\left(\xi_0 + \boldsymbol{x}_i'\boldsymbol{\xi} + \boldsymbol{b}_i\right). \end{aligned}$$

Note that the potential of this model specification becomes more apparent now that relative risk estimation gets involved. Indeed, relative risks can now be based on the inclusion of the both random effects (PG specification) or only based on a spatial random effect (NB specification). Note that a similar model with a negative binomial specification instead of a Poisson model was proposed before by Gschössl and Czado (2008).

6.4 Case studies

Recall the male kidney cancer and prostate cancer data sets from Section 2.2.1. An in-depth data explanation is provided by Loesbergh et al. (2007). Both data sets clearly show a large amount of extra Poisson-variation (Table 2.1). Note that extremely small numbers in the lower range of these domains were sampled in a very small town, called Herstappe and high values were partly attributed by highly populated cities, such as Hasselt or Genk, may cause overdispersion. The observed counts however do not tell us much more, and although standardizing (Inskip et al., 1983) these counts for age solves a part of the problem, extra-Poisson variation is still present in the resulting expected counts (Figure 6.1). It is also very likely that part of the remaining variability can be explained by correlations through space on one hand but also by spatially uncorrelated overdispersion (e.g. caused by not standardizing for an important but still unknown factor) on the other hand. In other words, SIR estimates (Figure 6.2), may be overly simplistic and models which include random effects for both uncorrelated heterogeneity (UH) and correlated heterogeneity (CH) will probably be better suited for these data.

To illustrate the use of the combined model and to compare it with the traditional modeling options, an in-depth analysis of the kidney and prostate cancer data sets (Section 2.2.1) is presented here. As introduced above, the Poisson-gamma (PG), Poisson-lognormal (PN), CAR (CAR), the CAR convolution (CARCON) and the combined (COM) model were of primary interest and therefore applied to both data sets. Goodness-of-fit (GOF) was tested on one hand via the deviance information criterion (DIC). Note that again, when DIC differences were borderline, less complex models having lower effective number of parameters (pD) were chosen. On the other hand, the overall loss across the data was assessed by the use of the mean squared

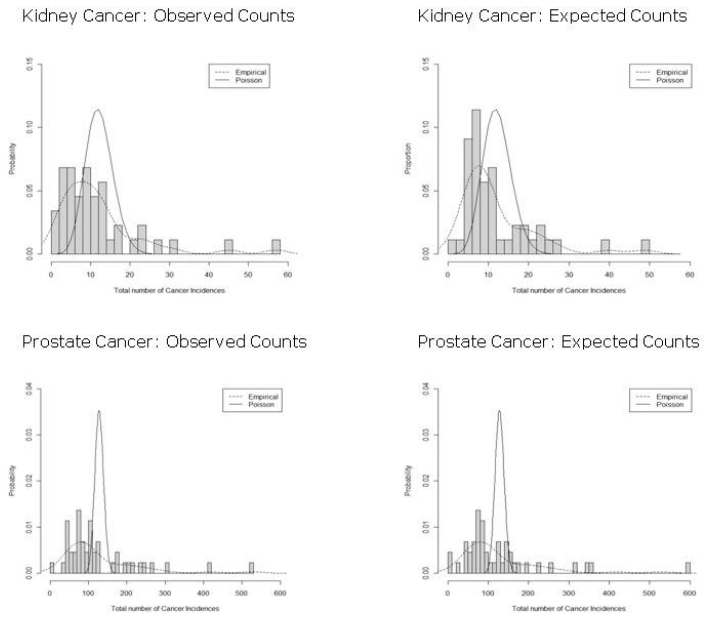


Figure 6.1: Observed and standardized expected counts for kidney and prostate cancer in the 44 municipalities of Limburg.

predictive error (MSPE), which is an average of the item-wise squared error loss, $MSPE = \sum_i \sum_j (Y_i - Y_{ij}^{pr})^2/(G \times m)$, with Y_{ij}^{pr} being the predictive data item at iteration j, m being the number of observations and G being the sampler's sample size. Note that although both measures are considered to be goodness-of-fit statistics, they do not give exactly the same information, since the DIC measures the global goodness-of-fit, while MSPE is a measure for predictive ability.

MCMC convergence was visually investigated by the use of features provided by WinBUGS, such as trace plots, history plots, etc. and it was met without any problems for all models. When looking at the fit statistics of the models of primary interest (Table 6.1, columns 2-6), one can see that in terms of DIC the CAR model is favoured for the kidney cancer data set, closely followed by the PN model, while for prostate cancer the Combined model fits best (note that DIC values differed only slightly between COM and PG, but that pD values favoured the COM model). It is noticeable too that PG and COM (the 'COM' column; the 'COM alt' column will be dealt with later) on the one hand

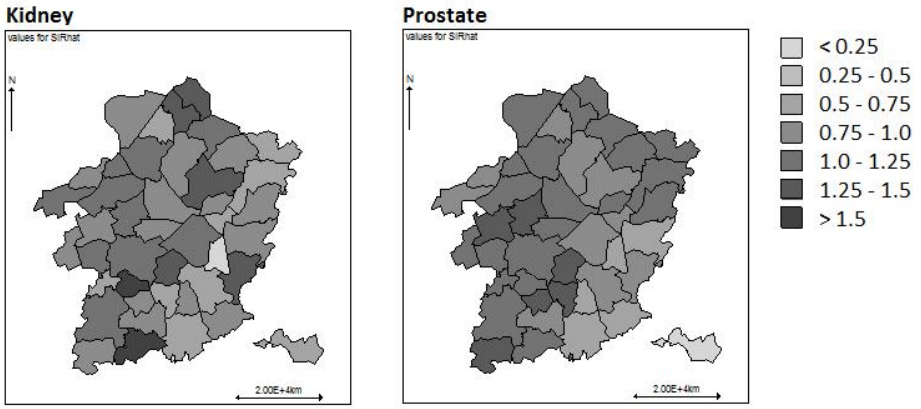


Figure 6.2: Standardized incidence rate for kidney and prostate cancer in the 44 municipalities of Limburg.

Table 6.1: DIC (pD) and MSPE values for all models (including both estimation methods for the combined model) analyzing kidney and prostate cancer data in the 44 municipalities of Limburg between 1996 and 2005.

	PG	PN	CAR		COM	COM alt	PCAR
			CH	CH + UH			
				DIC			
Kidney	235.89 (27.3)	213.55 (2.3)	**213.45 (1.7)**	214.10 (3.3)	230.74 (23.5)	231.57 (2.2)	213.23 (1.5)
Prostate	**363.98 (40.4)**	371.87 (35.8)	397.24 (30.4)	373.13 (35.8)	**364.36 (39.3)**	423.22 (2.5)	364.36 (36.1)
				MSPE			
Kidney	22.30	22.47	22.88	22.62	**21.65**	50.07	22.92
Prostate	**253.1**	258.1	284.0	258.3	**253.6**	4412.0	256.6

and PN, CAR and CARCON on the other hand have similar DIC values, leading to the impression that difference between the gamma- and lognormal UH terms is important. When looking at relative risk (RR) estimates (see Appendix), it shows that for kidney cancer for all municipalities, the credibility intervals for RR contain 1, indicating again that a CARCON or COM model is not necessary to fit the data. For prostate cancer, RR estimates do differ from 1 for several municipalities, and it is in this case that the combined model with CH and relatively large UH terms is favoured. It also has to be noticed that significance of the traditional CAR convolution model and the proposed combined model does not change amongst the models in these examples. MSPE values show somewhat different results: for both data sets, the MSPE for the combined model is small, indicating that the model has a good predictive behaviour.

Table 6.2: Parameter Estimates for all Models. σ_1^2 is the variance of the CAR random effect, $\widehat{\sigma_0^2}$ is the variance of the uncorrelated random effect at the scale of the relative risk.

	PG	PN	CAR		COM	COM
			CH	CH + UH		alt
Kidney						
ξ_0	/	-0.004699	-0.002699	-0.005883	-0.02507	-0.02392
σ_0^2	0.185726	0.00313	/	0.00316	0.118315	/
σ_1^2	/	/	0.004818	0.004909	0.004397	0.004975
Prostate						
ξ_0	/	-0.0206	-0.01366	-0.01993	0.003677	0.002367
σ_0^2	0.170858	0.05726	/	0.04920	0.111520	/
σ_1^2	/	/	0.09864	0.01298	0.004744	0.003757

Similar conclusions are drawn from the parameter estimates (Table 6.2), in which the estimated values for the intercept, the variance of the spatially structured CAR random effect, σ_1^2 and the variance of the spatially unstructured random effect, σ_0^2 are displayed. The latter comes from either the gamma distributed random effect in the PG or combined model, or from the lognormal distributed random effect in the PN and CAR convolution (CH + UH) model. First of all, for the kidney cancer data it is clear that the CAR convolution model has similar variability in the CH term as the CAR (CH) model and similar variability in the UH term as the PN, while the combined model has UH and CH term variances that are in agreement with the ones in the PG and CAR (CH) models. The prostate cancer DIC values on the other hand favoured the models with the gamma overdispersion term, especially the combined model. Parameter estimates transparently show a drop in σ_1^2 when going from the CAR based models to the combined model, while $\widehat{\sigma_0^2}$ increases.

RR maps visualize how the different models, in contrast to the SIR estimates, take the extra-variance into account. These maps show what has been seen in previous analyses too: the gamma distribution provides estimates in a larger range than the lognormal distribution, and one can conclude that some data sets will 'prefer' the larger overdispersion estimates in the COM model, while for others like kidney cancer, the lognormal UH terms in the PN and the CARCON or even the absence of an UH term (in the CAR model) will be sufficient. Figure 6.3 shows how the combined model 'combines'

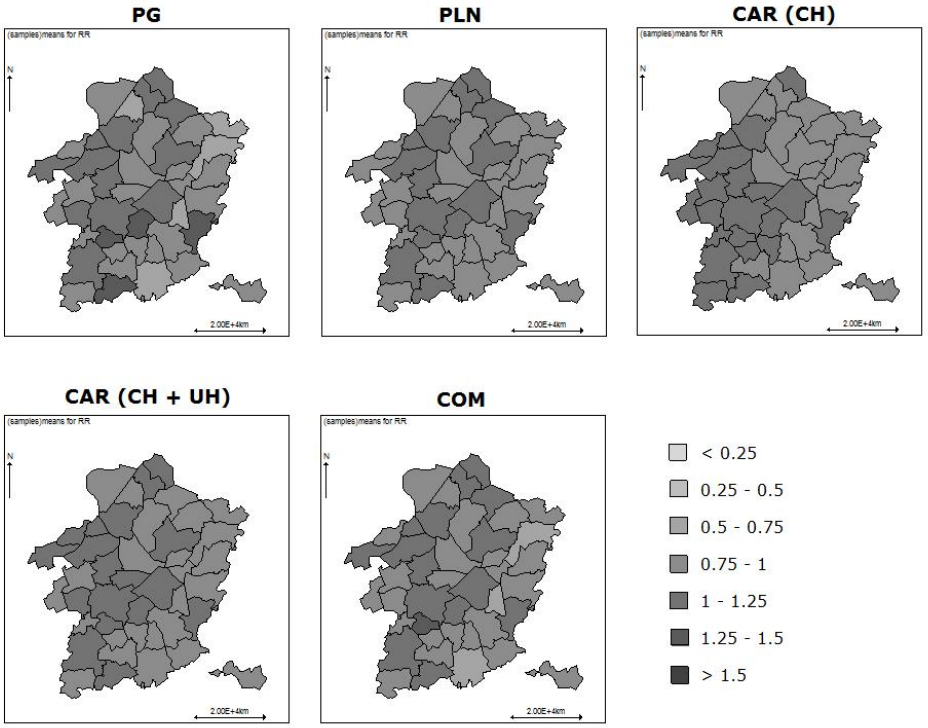


Figure 6.3: Relative risk estimates for the five models for the kidney cancer data in the 44 municipalities of Limburg.

a part of the overdispersion modeled by the PG and the spatial pattern proposed by the CAR (CH) model, while the CAR convolution model is similar to the overdispersion as proposed by the PN on the one hand and the spatial correlation represented by CAR (CH) on the other hand. The CAR convolution model with its lognormally distributed random effect seems to be more conservatively (smaller range) to model overdispersion than the combined model. We have to keep in mind though that although there are clear differences in the maps of different models, kidney cancer does not 'need' the more complex models here, as mentioned earlier. Figure 6.4 shows similar RR maps for the prostate cancer data set. Visually, the convolution model does not differ much from the other models, although in terms of DIC the combined model clearly fitted better.

In Section 3.4, it was explained that the combined model can also be fitted by first integrating out the conjugate prior distribution, conditional on the CAR distribution. Parameter estimates, DIC and MSPE values are shown in the 'COM Alt' columns in

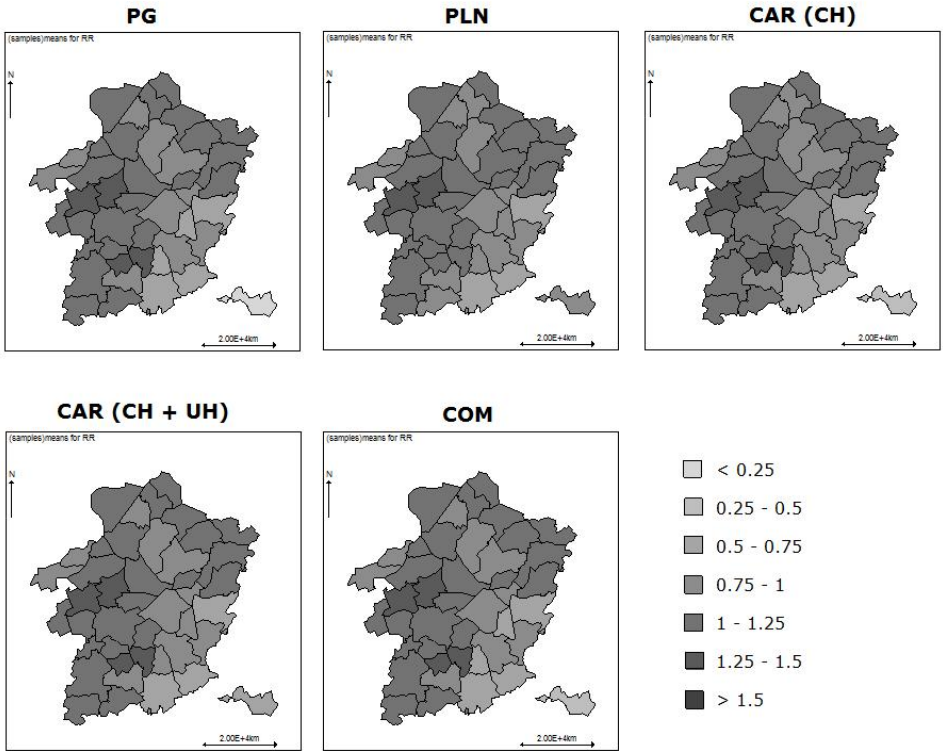


Figure 6.4: Relative risk estimates for the five models for the prostate cancer data in the 44 municipalities of Limburg.

Tables 6.1 and 6.2, and relative risk maps are shown in Figure 6.5. The parameter estimates obtained from the two estimation procedures are very similar. DIC and MSPE values are similar for the kidney data, but extremely different for the prostate data. Looking at the RR maps, it can be seen that much more smoothed estimates are obtained as compared with estimates based on the hierarchical modeling approach. This effect comes from the fact that in the hierarchical model setting the area-specific random effects related to overdispersion are taken into account in the estimation of the RR, i.e. $\widehat{\omega}_i = \widehat{\theta}_i \exp(\widehat{\xi}_0 + \widehat{b}_{1,i})$, while in the negative binomial model this random effect is not part of the estimation of the RR. Exactly the same RR estimates are obtained by using the hierarchical modeling approach but ignoring the overdispersion effect in RR estimation, i.e. $\tilde{\omega}_i = \widehat{\kappa}_i = \exp(\widehat{\xi}_0 + \widehat{b}_{1,i})$. It can be expected that $\widehat{\omega}_i$ and $\tilde{\omega}_i$ are similar when only little overdispersion is present in the data (kidney example), but very different when the overdispersion is omnipresent (prostate example). In contrast to the negative binomial estimation procedure, the hierarchical method allows the investigation of all random

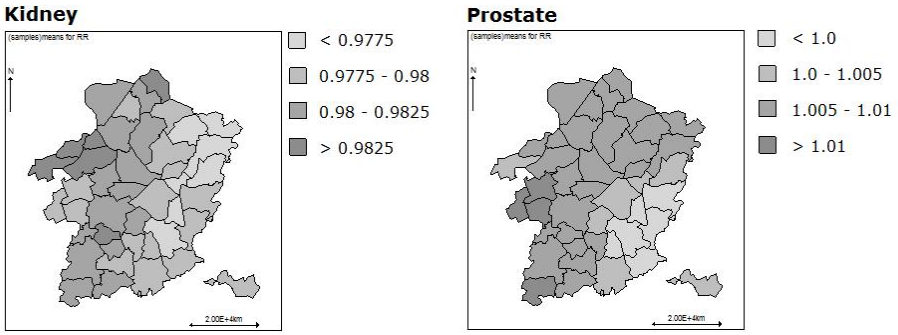


Figure 6.5: Relative risk estimates obtained by the alternative method of the combined model in the 44 municipalities of Limburg.

effects as well as estimation of $\widehat{\omega}_i$ and $\tilde{\omega}_i$.

Note that all models taking into account neighbourhood dependence were fitted via the dependence structure as indicated before. Although it was not of primary interest, it was useful to investigate how the CAR spatial structure behaves compared to the more general one in a proper CAR model (PCAR). Surprisingly, fits and predictive capabilities for the PCAR model were among, if not the best of all models, a result which encourages further research on spatial structures departing from the standard used CAR 'special case'.

6.5 Simulation study

To examine the properties of the combined model, the proposed methodology was also investigated using a simulation study. Data were simulated under three different situations: (1) the case where only uncorrelated heterogeneity is present (UH), (2) the case where only spatially correlated heterogeneity was present (CH), and (3) the case where both types of heterogeneity were present simultaneously (convolution model, CON). These three processes were simulated separately for two settings: setting A where the data contained a large amount of uncorrelated heterogeneity and only little spatially structured heterogeneity on one hand and setting B where the spatially structured heterogeneity was largely present in the data while there was only little uncorrelated heterogeneity on the other.

To achieve consistency with the data analyses, the map of Limburg was used to

simulate relative risk distributions within. The expected number of cases from the male kidney cancer data set were used. To simulate the data, the multinomial model, in line with Rodeiro and Lawson (2004), was used

$$Y_i \sim \text{Multinomial}\left(n, \frac{E_i\kappa_i}{\sum_{j=1}^{N} E_j\kappa_j}\right)$$

with $n = 44$ (total number of towns for which we had expected counts), such that the total number of cases was fixed. To introduce the three different settings in terms of included heterogeneity, the relative risks were simulated as coming from different models.

(1) Lognormal (UH) model:

$$\begin{aligned} \kappa_i &= \exp(b_{0,i}), \\ b_{0,i} &\sim \text{Normal}(0, \sigma_0^2). \end{aligned}$$

To simulate a relatively high amount of UH (setting A), we chose to use $\tau_0^2 = 0.05$, in the setting with little UH (setting B), we took $\tau_0^2 = 0.5$ (with $\tau_0^2 = 1/\sigma_0^2$).

(2) CAR (CH) model:

$$\begin{aligned} \kappa_i &= \text{exp}(b_{1,i}), \\ b_{1,i}|b_{1,j,i\neq j} &\sim N\left(\bar{\mu}_i, \sigma_{1,i}^2\right), \\ \bar{\mu}_i &= \frac{1}{\sum_{j=1}^{N} w_{ij}} \sum_{j=1}^{N} w_{ij}u_j, \\ \sigma_{1,i}^2 &= \frac{\sigma_1^2}{\sum_{j=1}^{N} w_{ij}}. \end{aligned}$$

The spatially structured heterogeneity ($b_{1,i}$) values were sampled directly from WinBUGS. By assuming a certain value for τ_1^2 ($\tau_1^2 = 1/\sigma_1^2$), one is able to control the amount of simulated spatially structured overdispersion. By setting $\tau_1^2 = 500$, only little CH was simulated (setting A), while a relatively high amount of CH (setting B) was simulated when $\tau_1^2 = 5$.

(3) CAR convolution model:

$$\kappa_i = \text{exp}(b_{0,i} + b_{1,i}),$$

with $b_{0,i}$ and $b_{1,i}$ defined as in (1) and (2). Also, exactly the same values as simulated in (1) and (2) were both included in this model.

Table 6.3: Simulation study: average MSE values for setting A (large UH, small CH) and B (small UH, large CH). The columns indicate the models from which data were simulated, while the rows indicate the fitted models.

	Setting A			Setting B		
Fitted Model	PN	CAR (CH)	CAR (CON)	PN	CAR (CH)	CAR (CON)
PG	**0.163**	0.0533	**0.191**	0.0384	0.107	0.107
PN	0.168	0.0142	0.198	0.00278	0.102	0.103
CAR (CH)	0.194	**0.0138**	0.227	**0.00276**	0.108	0.109
CARCON	0.169	0.0145	0.200	0.00306	**0.101**	**0.102**
COM	**0.163**	0.0521	**0.191**	0.0373	0.106	0.106

All simulated observed counts were analyzed with five models: the Poisson-gamma model, the Poisson-lognormal model, the CAR (CH) model, the CAR convolution (UH + CH) model and the combined model. Both settings A and B were used to simulate data from, separately 200 times and ultimately yielding 200 times 5 analyses of 3 models for each setting A and B. Mean squared error, formulated as $MSE = \sum_{n=1}^{i}((\widehat{\omega}_i - \omega^{true})^2/(n-1))$ with $i = 1, .., n$ with $n = 44$ which was averaged over the 200 simulations, was used to compare models.

Although the results presented in Table 6.3 do not show large differences in average MSE between models, they are consistent with the results seen in the previous paragraph: the combined model behaves particularly well when there is a sufficient amount of UH present in the data (setting A). In this setting, average MSE values are slightly lower for the combined and Poisson-gamma model for the cases in which UH was present in the data (PN and CARCON columns). In Setting B, in which the simulated amount of CH is large, while UH is small, average MSE values are overall smaller than those of the combined model. This again is consistent with previous observations, which state that the combined model does well when there is a large portion of uncorrelated overdispersion, but not necessarily when a map contains a lot of spatially induced extra-variance. But when there is zero or very little extra-variance present in the data, the gamma models, including the combined model, will analyze the data not as good as the normal distribution-based solutions. This simulation study shows that in several cases, the proposed method can be a good alternative to the commonly used models.

6.6 Estimation

Due to the frequent use of convolution models with CAR assumptions and the subsequent need to model complex and multiple random effects structures, disease mapping has been mainly developed in a Bayesian environment. From Chapter 5, I concluded that INLA gives fairly good results when working with count data. To investigate this for spatial data, a comparison between MCMC and INLA can be made here too, using the kidney and prostate cancer data sets. The combined model was fitted as previously,

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i\kappa_i\theta_i), \\ \kappa_i &= \exp\left(\xi_0 + b_{1,i}\right), \\ \theta_i &\sim \text{gamma}(\alpha, 1/\alpha), \end{aligned}$$

with $b_{1,i}$ being the CAR spatial random effect.

Results for the kidney cancer analysis are given in the top panel of Table 6.4. For the kidney cancer data, INLA and MCMC for estimates ξ_0 and α were very alike , with the overdispersion parameter α being relatively large and significant. Larger differences between the estimation methods are observed for the random effects standard error. This may be caused by the map of Limburg being rather small (44 municipalities), which can result in difficulties when estimating a variance parameter. The agreement statistic is very large when comparing MCMC posteriors and INLA approximate densities between the intercept and the overdispersion parameter, but the agreement is lower (71%) for the standard deviation of the spatially structured normal random effects (Figure 6.6). INLA results in a smaller standard deviation estimate σ, and will therefore produce a smoother map of the relative risks as compared to MCMC. For the prostate cancer data, results are given in the lower panel of Table 6.4. Here differences between the estimation techniques were more substantial for all three considered parameters. Parameter estimates differ consistently, while the agreement statistics remain under 75%. The overdispersion parameter has the lowest agreement (59%).

INLA and MCMC estimations differ more for the spatial data sets than the other non-spatial count data sets in Chapter 5. As mentioned earlier, this may be due to the small sample size which can affect estimation.

Table 6.4: Parameter estimates (and s.d.) for the kidney and prostate data sets. Three INLA-based and the MCMC results are provided.

	Kidney cancer data			
Model	Computation time	ξ_0 Est. (s.d.)	α Est. (s.d.)	σ_1 Est. (s.d.)
INLA (Simp. Laplace)	1.234 sec.	-0.034 (0.074)	8.441 (2.309)	0.007 (0.003)
INLA (Laplace)	1.299 sec.	-0.034 (0.074)	8.441 (2.309)	0.007 (0.004)
INLA (Gaussian)	1.929 sec.	-0.030 (0.074)	8.441 (2.309)	0.007 (0.003)
MCMC	11.887 sec.	-0.024 (0.074)	8.492 (2.331)	0.014 (0.013)
	Prostate cancer data			
Model	Computation time	ξ_0 Est. (s.d.)	α Est. (s.d.)	σ_1 Est. (s.d.)
INLA (Simp. Laplace)	1.180 sec.	-0.034 (0.050)	11.467 (2.556)	0.008 (0.003)
INLA (Laplace)	1.570 sec.	-0.032 (0.050)	11.467 (2.556)	0.008 (0.003)
INLA (Gaussian)	1.144 sec.	-0.031 (0.050)	11.467 (2.556)	0.008 (0.003)
MCMC	13.984 sec.	0.002 (0.055)	8.972 (2.112)	0.019 (0.009)

6.7 Concluding remarks

As shown in the previous sections, the spatial combined model provides an interesting alternative to the popular CAR convolution model, that sometimes suffers from spatially oversmoothing of RR maps. Compared to the lognormally distributed UH term in the CAR convolution model, the gamma distribution in the combined model is shown to allow for a wider range of relative risk estimates. The proposed model therefore provides an alternative convolution model with improved modeling capabilities when the data contain a large amount of uncorrelated heterogeneity. One of the main reasons why the CAR convolution model has known such popularity is due to the robustness and easy implementation using Bayesian software, such as WinBUGS. The combined model has these advantages too; it allows for two different specification methods to calculate the RR estimates, it can easily be extended to a model with covariates, and interpreted with respect to the strong-conjugacy properties. Unfortunately, both convolution models also share the same disadvantages, namely the possibility to suffer from identifiability problems. As a consequence of this, we chose to limit ourselves to the investigation of relative risks only, in order not to overinterpret the results by looking at the CH and UH effects separately. Finally, estimation via INLA is shown to become somewhat problematic in the spatial case, making the computationally more demanding MCMC method in WinBUGS recommended.

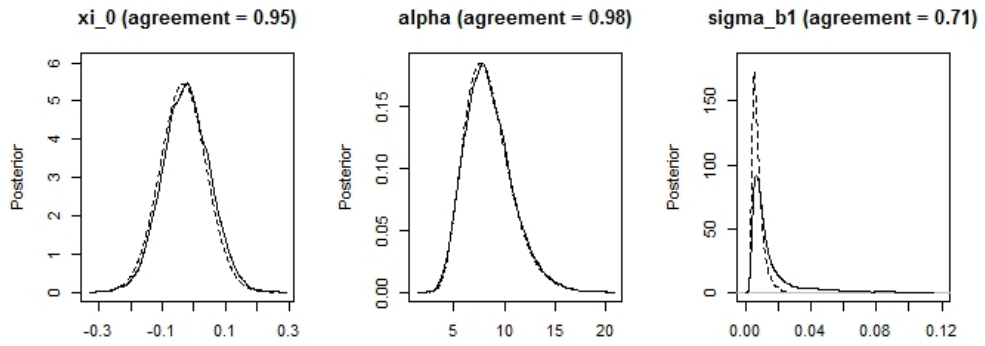


Figure 6.6: Visual representation of the posterior parameter estimates for the kidney data example. Full line = MCMC, dashed line = INLA (strategy = simplified laplace).

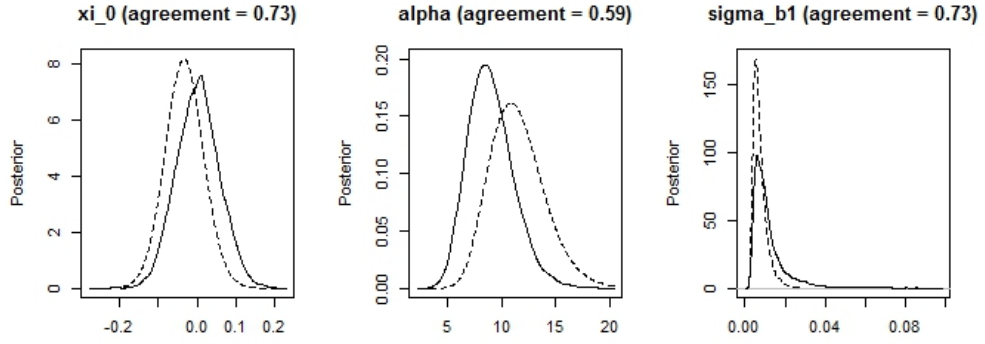


Figure 6.7: Visual representation of the posterior parameter estimates for the data data example. Full line = MCMC, dashed line = INLA (strategy = simplified laplace).

Chapter 7

The Combined Model for Excessive Zero Counts

7.1 Introduction

In the previous chapters, extensions of the Poisson model for count data were investigated. Two main reasons for doing this were one one hand the presence of a hierarchical structure in the data, e.g., due to clustering in the data, repeated measurements of the outcome, etc., causing extra-variability. On the other hand, the occurrence of overdispersion, meaning that the variability in the data is not equal to the mean as prescribed by the Poisson distribution, due to the exclusion of important covariates, has to be taken into account. A third important issue was left untouched until now: the occurrence of extra zeros beyond what a Poisson model allows for, can cause additional variability. While the first issue is often accommodated through the inclusion of normal random subject-specific effects (Engel and Keen, 1992, Breslow and Clayton, 1993, Wolfinger and O'Connell, 1993, Molenberghs and Verbeke, 2005) and overdispersion is often dealt with through an overdispersion model, such as, for count data examples, the negative-binomial model (Breslow, 1984, Lawless, 1987), where the natural parameter is assumed to follow a gamma distribution, an excessive number of zeros is regularly accounted for using so-called zero-inflated or zero-truncated models. Zero-inflated models were studied for univariate count data by Lambert (1992) and Greene (1994), with an extension towards the hierarchical setting studied in Min and Agresti (2005) and Lee et al. (2006). A zero-truncated model, the so-called hurdle model, was proposed by Mullahy (1986) and also extended to complex data settings, e.g. by Scheel et al. (2013).

While the combined model that accommodated clustering and overdispersion through two separate sets of normal and gamma random effects in a Poisson model was investigated earlier, this chapter proposes a general modeling framework in which correlation, overdispersion and an excess of zeros can appear together. An in-depth analysis of the usability of two extensions of the combined model, namely the zero-inflated combined model (ZICOM) and the hurdle (or zero-truncated) combined model (HCOM) is given. In Section 7.2, the zero-inflated and hurdle models are introduced, while estimation is considered in Section 7.3. Section 7.4 presents two case studies, while simulation study results are shown in Section 7.5. Finally, concluding remarks are given in Section 7.6.

7.2 Models for counts with an excess of zeros

While overdispersion models can deal with some extra-variance caused by a large amount of zeros, specific models for that case have been developed. Here, two of them are investigated: (1) the zero-inflated model (Lambert, 1992), assumes zeros to come from a point mass as well as a count component. (2) A hurdle model (Mullahy, 1986) on the other hand uses a binary model distinguishing between zeros and positive values, while if positive, a zero-truncated Poisson distribution is fitted.

7.2.1 Zero-inflated model

In zero-inflated count models, it is assumed that there are two processes that can generate zeros: zeros may come from both a point mass (process 1) as well as from the count component (process 2). It is assumed that for observation i at time or location j, process 1 is chosen with probability π_{ij} and process 2 with probability $1-\pi_{ij}$ (Hinde, and Demétrio, 1998a, Hinde, and Demétrio, 1998b). Process 1 generates only zeros, whereas process 2, $f_i(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij})$, generates counts from a Poisson, a negative-binomial model, a Poisson-normal GLMM, or a Poisson-normal-gamma combined model. In its most general form, the zero-inflated Poisson-normal-gamma model is given as the following mixture:

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } \pi_{ij}, \\ f_i(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{with probability } 1-\pi_{ij}, \end{cases}$$

leading to the probabilities $p(Y_{ij} = y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij})$ given by

$$p(Y_{ij} = y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij}) = \begin{cases} \pi_{ij} + (1-\pi_{ij}) f_i(0|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{if } y_{ij} = 0, \\ (1-\pi_{ij}) f_i(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{if } y_{ij} > 0. \end{cases}$$

The zero-inflation component $\pi_{ij} = \pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma} + \boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i})$ is modeled using a Bernoulli model: in the simplest case with only an intercept, but potentially containing known

regressors $\boldsymbol{x}_{2ij}$ and $\boldsymbol{z}_{2ij}$, a vector of zero-inflation coefficients $\boldsymbol{\gamma}$ to be estimated, as well as random effects $\boldsymbol{b}_{2i}$. Common link functions, such as the logit or probit, can be used. Note that $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$, and $\boldsymbol{b}_i$ from the earlier sections are now replaced by $\boldsymbol{x}_{1ij}$, $\boldsymbol{z}_{1ij}$, and $\boldsymbol{b}_{2ij}$, respectively, for the non-zero count part. The regressors in the count and zero-inflation component can either be overlapping, a subset of the regressors can be used for the zero-inflation, or entirely different regressors for the two parts can be used. In terms of the random effects to include, many options exist. A simple random-intercept model might be adequate, where $\boldsymbol{b}_{1i} = b_{1i}$, $\boldsymbol{b}_{2i} = b_{2i}$, and $\boldsymbol{z}_{1ij} = \boldsymbol{z}_{2ij} = 1$. Assuming that the random effects are normally distributed and possibly correlated with correlation parameter ρ, the variance-covariance matrix is

$$\boldsymbol{D} = \begin{pmatrix} d_1 & \rho\sqrt{d}_1\sqrt{d}_2 \\ \rho\sqrt{d}_1\sqrt{d}_2 & d_2 \end{pmatrix}.$$

The model is denoted as the zero-inflated combined model (ZICOM). Three obvious special cases are the zero-inflated Poisson-lognormal (ZIPN), the zero-inflated Poisson-gamma (ZINB), and the zero-inflated Poisson (ZIP) model. Also, all four models without zero inflation are special cases as well. The conditional mean and variance of the ZICOM are:

$$\begin{aligned} \mathsf{E}(Y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) &= \theta_{ij}\kappa_{ij}(1-\pi_{ij}), \\ \mathsf{Var}(Y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) &= \theta_{ij}\kappa_{ij}(1-\pi_{ij})[1+\theta_{ij}\kappa_{ij}(\pi_{ij}+1/\alpha)]. \end{aligned}$$

It can be seen that the conditional variance is inflated as a result of either overdispersion in the data (parameter α), or as a result of zero-inflation (parameter π_{ij}), or both.

7.2.2 Hurdle model

The hurdle model is a way of modeling count data using a two-part approach, whereby the first part is a binary model for the count value zero or positive. Within the context of the combined model, say one again defines an observation i at time or spatial location j. Given the value is positive, a count distribution f_i is truncated-at-zero and fitted for the second part. Suppose Y_{ij} is a univariate count outcome, and π_{ij} is probability of the i^{th} observation at time point or location j to be in the zero state. The hurdle model assumes Y_{ij} fulfils a distribution given by

$$p(Y_{ij} = y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij}) = \begin{cases} \pi_{ij} & \text{if } y_{ij} = 0, \\ (1-\pi_{ij})\frac{f_i(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij})}{1-f_i(0|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij})} & \text{if } y_{ij} > 0, \end{cases}$$

again with $\pi_{ij} = \pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma} + \boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i})$ defined as in Section 7.2.1. Merging ideas of the combined model of Molenberghs et al. (2010) and the hurdle model (Mullahy, 1986),

a two-part hurdle combined model is considered to deal with zero-inflated overdispersed clustered count data. While the first part models only the zero state with probability π_{ij}, the second part handles non-zero counts, which are assumed to follow a truncated-at-zero probability mass function, such as, in this case, a truncated Poisson-normal-gamma model.

7.3 Estimation

Recall that likelihood estimation of the combined model for count data was done by marginalizing analytically over the gamma random effect, with then further numerical integration over the normal random effects, via e.g. PROC NLMIXED in SAS. When the ideas in Section 5.2.1 are extended to the zero-inflated case, the partially marginalized ZICOM takes the form:

$$\begin{aligned} & f(y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\boldsymbol{b}_{2i},\boldsymbol{\gamma}) \\ &= I(y_{ij}=0)\pi_{ij} \\ &\quad +(1-\pi_{ij})\left(\begin{array}{c}\alpha_j+y_{ij}-1\\ \alpha_j-1\end{array}\right)\cdot\left(\frac{1/\alpha_j}{1+\kappa_{ij}1/\alpha_j}\right)^{y_{ij}}\cdot\left(\frac{1}{1+\kappa_{ij}1/\alpha_j}\right)^{\alpha_j}\kappa_{ij}^{y_{ij}}, \end{aligned}$$

The hurdle counterpart, the partially marginalized HCOM, becomes

$$f(y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\boldsymbol{b}_{2i},\boldsymbol{\gamma}) = I(y_{ij}=0)\pi_{ij}+(1-\pi_{ij})g_1(\boldsymbol{b}_i),$$

where

$$g_1(\boldsymbol{b}_i)=\left(\begin{array}{c}\alpha_j+y_{ij}-1\\ \alpha_j-1\end{array}\right)\cdot\left(\frac{1/\alpha_j}{1+\kappa_{ij}1/\alpha_j}\right)^{y_{ij}}\cdot\left(\frac{1}{1+\kappa_{ij}1/\alpha_j}\right)^{\alpha_j}{\kappa_{ij}}^{y_{ij}}\cdot\frac{1}{1-\left(\frac{1}{1+\kappa_{ij}^c 1/\alpha_j}\right)^{\alpha_j}},$$

with $\pi_{ij}=\pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma}+\boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i})$ in both cases. Note that in this thesis, only the logit link was investigated, but extensions to other link functions are possible, such as the probit link which was investigated by Kassahun et al. (2014b). In WinBUGS, MCMC-based sampling was performed by using the so-called zeros trick, which is used whenever a sampling distribution, here being the zero-inflated and hurdle combined model, is not included in the list of standard distributions (Spiegelhalter et al., 2007). A sample of the SAS and WinBUGS implementation code is given in the Appendix.

7.4 Case studies

In order to investigate and compare zero-inflated, hurdle and traditional models, case studies were performed in both a longitudinal and a spatial context. For the longitudinal

case, the epilepsy data, as introduced in Section 2.2.4, were used. As can be seen in Figure 2.4, zero counts represent a considerable part of the measurements (33%), while the sample average and standard deviation are 3.18 and 6.14, respectively. The investigation of a combined model approach that takes into account the excess zeros therefore seems mandatory. In the spatial framework, the mesothelioma data (Section 2.2.2) were studied. Again, a possible spatial structure should be investigated, since the presence of the disease is likely to be strongly correlated with asbestos exposure. Furthermore a sample average and standard deviation of resp. 0.32 and 1.24 and a large amount of zero counts (81%) make considering a zero-inflated or hurdle combined model useful.

7.4.1 Epilepsy data

For the epilepsy data (Section 2.2.4), let Y_{ij} represent the number of epileptic seizures that patient i experiences during week j of the follow-up period. Consider the combined model from Section 3.4, but now accounting for excess zeros, assuming that counts are generated from a combined model process with mean $\mu_{ij}^c = \kappa_{ij}\theta_{ij}$:

$$\kappa_{ij} \quad = \quad \exp\left(\xi_0 + b_{0,i} + \xi_1 Trt_i + \xi_2 Time_{ij} + \xi_3 Trt_i * Time_{ij}\right),$$

with the zero part probability (π_{ij}) modeled as $\text{logit}(\pi_{ij}) = \gamma_0 + b_{1,i} + \gamma_1 Time_{ij}$ and $b_{0,i}$ and $b_{1,i}$ being correlated with parameter ρ. Both zero-inflated and hurdle model specifications were investigated, while Poisson, Poisson-normal and negative binomial parametrizations followed in a straight-forward way. For the sake of comparison, also the non-excess-zero counterparts were fitted. Note that while originally estimation was only done in a likelihood framework (Kassahun et al., 2014a), also a Bayesian approach was investigated in this thesis. Parameters ξ_0, ξ_1, ξ_2, and ξ_3 were given normal priors with mean 0 and variance 10000, while γ_0 and γ_1 received more informative normal priors with mean 0 and variance 10, due to estimation problems when leaving those uninformative. Furthermore, $\text{U}(0, 100)$ priors were assigned to the normal random effects standard deviations $\sqrt{d_1}$ and $\sqrt{d_2}$, while $\rho \sim \text{U}(-1, 1)$. Note that for the combined and Poisson-gamma versions of the excess-zero and traditional models, a negative binomial parametrization as in (3.8)-(3.10) was applied. Again, $\alpha = 1/\beta$ and was assigned $\text{igamma}(0.01, 0.01)$. For the HNB model $\alpha \sim \exp(1)$ was used though, since the former prior specification caused estimation problems. Next to the Bayesian results, the likelihood approach is still presented, since interesting differences between both estimation methods were seen, as will be presented in the following paragraphs. A selection of codes is given in the Appendix.

While model selection in the likelihood context can be done via log-likelihood comparison, difficulties arise in the Bayesian framework. Since zero-inflated and hurdle models

are mixture models and non-excess-zero models are not, DIC values, which indicate model complexity, can not be compared as they will typically increase heavily in mixture models. A solution is provided by the conditional predictive ordinate (CPO), for individual i ($i = 1, ..., k$) on occasion j ($j = 1, ..., n_i$) defined as

$$\text{CPO}_{ij} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} [\exp(-ll_{ij})]^{-1},$$

with N the total number of observations (Geisser, 1993). From this, the marginal predictive likelihood (M) can be calculated as

$$\text{M} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \log(\text{CPO}_{ij}),$$

which can be used as a goodness-of-fit statistic (larger is better). Note however, that similar to DIC-based model selection, there are no formal tests or guidelines on how large differences between M values have to be in order to indicate improved fits.

Parameter estimates and predicted probabilities of zeros for the likelihood and Bayesian analyses are presented in Tables 7.1 and 7.2 respectively. The GOF statistics (-2log-likelihood and M) clearly show that the models for excessive zeros outperformed the traditional ones, with a substantial preference for the hurdle models. The ZINB and HNB models show an important improvement relative to the ZIP and HP models respectively, while even more improvement is gained when normal random intercepts are introduced, with the combined zero-inflated and hurdle parametrizations leading to the best fits.

None of the zero-inflated and hurdle models suggest evidence of significance in slope difference and slope ratio, except for the ZIP and HP, where significance is maintained for the slope difference (both $p = 0.0004$ in the likelihood case). However, those models, unrealistically, omit correlation and overdispersion. Furthermore, the zero-inflation and hurdle regression coefficients, which can be interpreted as model coefficients for the proportion of extra zeros, are statistically significant. It is important to note though, that while most likelihood and Bayesian results are similar, a few estimates differ substantially (e.g. $\widehat{\gamma_0}$ and $\widehat{\gamma_1}$ in the ZINB model). Furthermore, problems arose in the model fitting process. Likelihood estimation showed to be very sensitive to the choice of the initial parameters in PROC NLMIXED, which leads to the impression that estimates may come from a local maximum in the likelihood function. Next to that, uninformative priors for γ_0, γ_1 and sometimes α resulted in convergence problems. This can be expected in these models though, since the extra-variability in the data has to be allocated to a structured

random effects term, an overdispersion random effect and/or a zero process.

Finally it can be observed that both hurdle and zero-inflated models in all settings predict the probability of zeros well. When the zero part is omitted, namely when the traditional models are fitted, this prediction becomes worse, except for the COM model. Indeed, it thus seems that the combined model can deal fairly well with excessive zeros via its structured and unstructured random effects. It even seems that the inclusion of both structured and unstructured random effects without a zero process is more important that the inclusion of only one of those random effects with a zero process (e.g. by comparing -2log-likelihood between COM, HNB and HPN). It is also observed that by omitting either the overdispersion or the correlation between the normal random effects, the predicted probability of zeros are underestimated, which becomes worse when both are omitted at the same time. For example, when the ZICOM model is fitted without random effects in the zero-inflation part (not shown in Tables 7.1 and 7.2), -2log-likelihood becomes 5386.8 and the predicted probability of zeros equal to 0.3271. This implies that the inclusion of random effects in the zero-inflation part tends to have little impact on the predicted probability of zeros. However, based on likelihood comparison, model fit improves considerably. This same phenomenon is also evident in the ZIPN model fitted with random effects included only in the non-zero count part (-2log-likelihood is 5971.9, and predicted probability of zeros 0.3112).

7.4.2 Mesothelioma data

For the mesothelioma data (Section 2.2.2), let Y_i represent the number of newly diagnosed male mesothelioma cases in municipality $i = 1, ..., 308$. A model that accounts for excess zeros, overdispersion and a spatial trend seems reasonable when looking SIR estimates that show spatial trends and many zeros (Figure 7.1). A full model for excessive zero counts can be presented as a model with mean $\mu_i^c = \kappa_i \theta_i$:

$$\kappa_i \quad = \quad \exp\left(\xi_0 + b_{0,i}\right),$$

with the zero part probability (π_i) modeled as $\text{logit}(\pi_i) = \gamma_0 + b_{1,i}$. The random effects term $b_{0,i}$ introduces a spatial CAR structure, such as in (6.4)-(6.6) while $b_{1,i}$ is assumed to be unstructured and normally distributed. Again, both hurdle and zero-inflated models were fitted, with e.g. a spatial combined model following by omitting $b_{1,i}$, while a CARCON model emerges when θ_i is assumed to be equal to 1. Note that other parametrizations are possible and that the particular choice here was primarily made in order to work with a concise set of parameters, but also because of convergence problems when fitting more random effects terms simultaneously, an issue that will be

made clear in the next paragraph. Due to the spatial context, a Bayesian approach was investigated. Priors were $\xi_0 \sim \text{N}(0, \sigma^2 = 10000)$, $\gamma_0 \sim \text{N}(0, \sigma^2 = 100)$, the latter prior choice again made due to estimation problems in the uninformative case. Furthermore, a gamma$(0.05, 2000)$ was assigned as a prior for τ_0 and τ_1, the respective precisions of $b_{0,i}$ and $b_{1,i}$ similar as in Chapter 6. Again the combined and Poisson-gamma versions of all models here were fitted via the negative binomial parametrization in (3.8)-(3.10), with $\alpha = 1/\beta \sim$ igamma$(0.01, 0.01)$, except for the HNB and HCOM models, where $\alpha \sim \exp(1)$ was used. Again, a selection of codes are given in the Appendix.

When looking at the results, immediately a few interesting observations can be made (Table 7.3). In terms of the M statistic, the ZICAR and CAR models perform best. It is apparent that the hurdle options provide the worst fits, while the models without an excess-zero part fit as good or better than their excess-zero counterparts. Also noteworthy is the fact that when using uninformative priors for the zero parts, γ_0 estimates became extremely negative, making $\widehat{\pi} \approx 0$. A part of the explanation lies in the fact that although more than 80% of the observations in the data are zeros, the non-zero counts are mostly very low. Therefore, a Poisson model with a low mean will be capable of capturing a large amount of zeros. This is in line with the predicted probability of zeros that was approximately the same in a ZICOM model than in a simple Poisson model (Table 7.3). Another important part of the explanation is given by very nature of disease mapping. Since one works with one map, the information within these data becomes limited, possibly too limited to fit these complex models. Indeed, within the epilepsy context (Section 7.4.1), there were 89 entities (individuals) that were investigated through time. Here, there is only 1 entity (1 map) that is investigated through space, which makes it difficult for the estimation process to assign the correct amount of extra-variability to the corresponding random effects. This was also seen when data were generated with 80% zeros, but with 20% non-zeros with high counts (mean = 40) which were concentrated in 1 Flemish province (not shown here). In that case, the zero-inflated and hurdle models fitted better than the traditional ones, but still only when informative priors were used for the parameters in the zero part. In other words, zero-inflated and hurdle models can be good options for data analyses with many zeros, but they are difficult to use in the disease mapping context.

7.5 Simulation study

In this section, a simulation study was set up to examine the bias in estimating the regression parameters when dealing with overdispersed, longitudinal count data with excess

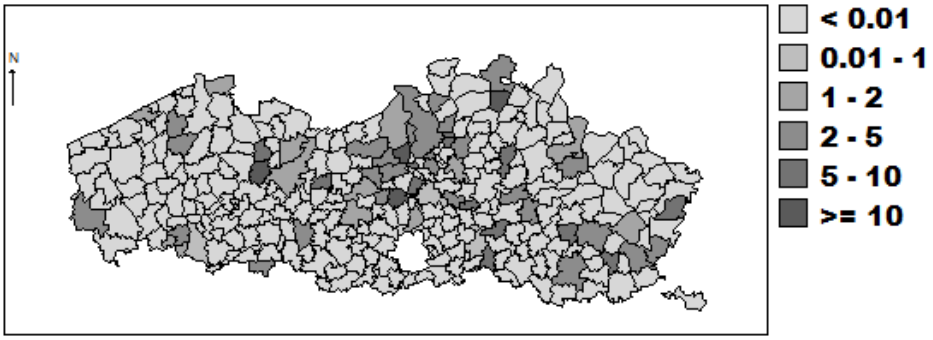


Figure 7.1: Standardized incidence rate (SIR) map of newly diagnosed male mesothelioma cases in the 308 municipalities of Flanders in 1999.

zeros. To keep the results concise, only longitudinal zero-inflated and traditional models were investigated in a likelihood setting, as was done in Kassahun et al. (2014a). Also note that only large sample scenarios were considered, in order to assess the ZI models performances in an ideal setting.

7.5.1 Simulation setting

Data were generated along a design inspired by an Ethiopian study, in which diarrhoeal disease in young children was investigated. The number of days of children's diarrhoeal illness in the two months before a visit by a medical doctor were recorded. While more information and an in-depth case study is provided by Kassahun et al. (2014a), here age in months and the status of getting medical help were used as covariates of interest.

A total of 200 data sets were randomly generated from the zero-inflated combined model for 2000 subjects with 10 measurements per subject. The response vector $\boldsymbol{y}_i$ for the i^{th} subject was generated as a correlated and overdispersed count from a negative binomial process subject to zero-inflation. That is, for each subject, $Y_{ij} \sim \text{NB}(\psi_{ij}, \theta)$, where $\theta = 1$ with $\psi_{ij} = (1 + \kappa_{ij}/\theta)^{-1}$ and where $\kappa_{ij} = \exp\{\xi_0 + b_i + \xi_1 Time_{ij} + \xi_2 Help_{ij}\}$ for $i = 1, \ldots, 2000$ and $j = 1, \ldots, 10$. Further, $Time_{ij}$ represents the time point at which the j^{th} measurement is recorded for the i^{th} subject and $Help_{ij}$ denotes whether or not the i^{th} subject is given any medical help at the j^{th} measurement occasion, generated from a Bernoulli process with $p = 0.9$. Correlation was induced via a subject-specific random intercept b_i generated from a normal distribution with mean 0 and variance 0.8. Then, zero-inflation was added by defining the final response vector $\boldsymbol{Y}_i^*$ to have

components ${Y_{ij}}^* = (1 - u_{ij})Y_{ij}$, where the u_{ij} are Bernoulli random variables with parameters π_{ij} and logit$(\pi_{ij}) = \gamma_0 + \gamma_1 Time_{ij}$.

Three different scenarios were considered for data generation: S_1: without excess zeros; S_2: with an excess of zeros of around 20%; S_3: with an excess of zeros of roughly 40%. The corresponding total zero percentages were 48%, 68%, and 88%, respectively. This was achieved, for each scenario, by appropriately choosing the zero-inflation coefficients. The true parameter values used to generate the data were $\xi = (1.12, 0.13, -1.89)^T$. Similarly, for the zero-inflation part, $\gamma = (-1, -1)^T$, $\gamma = (1, -0.25)^T$ and $\gamma = (1.8, -0.1)^T$ were used for S_1, S_2, and S_3, respectively.

7.5.2 Simulation results

The simulated data were analyzed by the ZICOM, ZINB, ZIPN, and ZIP, as well as by their non-zero-inflated counterparts. Mean, relative bias (rbias) and predicted probabilities of zero counts are summarized for the three scenarios in Tables 7.4–7.6, respectively. Parameter estimates of the ZICOM are in agreement with their true model in all scenarios. This shows that the different components (zero-inflation, overdispersion and correlation) can be well separated in practice, in settings like the ones considered here. The zero-inflated model converged for almost all simulated data sets.

Under S_1, as shown in Table 7.4, the ZICOM and the COM performed well and fairly similar in terms of relative bias, except for the intercept ξ_0 for which a larger bias is observed in the COM. The percentage of zero counts (48%) is nearly equally predicted in both cases, but severe impact starts to emerge in the non zero-inflation models when excess zero counts are present, but not accounted for (Tables 7.5 and 7.6). The predicted number of zero counts is largely underestimated in the non-zero-inflated models. When many zeros are allowed for, as in S_3, the effect is more pronounced in the intercept term and the negative binomial parameter α as compared to S_2. Moreover, the bias in the standard deviation of the random effects, for instance, in the 'true' model tends to increase in S_3, which gets substantially higher for models with neglected zero-inflation component, such as the COM and PN. The impact of omitting the overdispersion is remarkable. This can be clearly observed, for example, from the considerable increment in the relative bias of the ZIPN. When overdispersion is omitted, the zero-inflation component will try to recover part of the overdispersion. When the correlation stemming from the repeated measurements is misspecified, substantial impact appears in inferences of the ZINB, which gets even worse in the PG, as evidenced quite clearly from the larger relative bias of the intercept term. When correlation is omitted from the model, the

overdispersion term will try to recover for this misspecification. Unlike in S_1, the ZICOM significantly beats the COM, confirming the importance of accounting for the excess zeros in addition to the repeated measures nature and the overdispersion.

We conclude that failure to account for excessive zeros, overdispersion, and/or correlation has a substantial impact on bias and predicted probabilities. This was clearly shown on such key model parameters as the intercept term, the overdispersion parameter, and the variance of the random effects. All scenarios suggest that the zero-inflated combined model is the preferred one in terms of relative bias and predicted probabilities of zeros.

7.6 Concluding remarks

In this chapter, a modeling strategy for a hierarchical count data was described where excessive zeros, correlation and overdispersion can occur together and are assembled in one single model. This combined model extension to further deal with zero-inflation provides a parsimonious yet useful approach. Of course, with the considerations of not only one but multiple sets of random effects comes the obligation to reflect on the precise nature of such latent structures. As underscored by Verbeke and Molenberghs (2010), full verification of the adequacy of a random effects structure is not possible based on statistical considerations alone. Furthermore, one can question whether the data contain enough information to feed these complex random effects structures. Indeed, as seen in both case studies, but especially in the spatial context, models venture towards a border where they become intractable. If data sets are large, zero-inflated and hurdle models are good options to model data with excess zeros. In the simulation study that was conducted to investigate the impact of omitting each or a combination of zero-inflation, overdispersion and correlation, it was shown that omitting such features, while actually present, introduced considerable bias in parameter estimates and hence may lead to incorrect inferences. When data sets become small however, non-excess-zero models seem do be able to capture the excessive zeros well while estimation remains feasible. This leads to interesting avenues for further research, namely a simulation study to investigate which sample sizes allow good estimation in the zero-excess models.

Table 7.1: Epilepsy Study. Likelihood parameter estimates and standard errors for the different fitted models.

		ZICOM	HCOM	COM	ZINB	HNB	NB
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	0.9467(0.1665)	0.7707(0.1702)	0.9113(0.1755)	1.2361(0.1100)	-0.0606(0.4050)	1.2594(0.1119)
Poisson: treatment	ξ_1	-0.1106(0.2378)	0.0574(0.2250)	-0.2557(0.2500)	0.1614(0.1542)	0.4835(0.2320)	0.2156(0.1564)
Poisson: time	ξ_2	-0.0162(0.0075)	-0.0103(0.0085)	-0.0248(0.0077)	-0.0072(0.0113)	0.0079(0.0160)	-0.0126(0.0111)
Poisson: time*treatment	ξ_3	0.0101(0.0103)	0.0031(0.0119)	0.0130(0.0107)	-0.0147(0.0153)	-0.0296(0.0230)	-0.0227(0.0150)
Poisson: neg.-bin. par.	α	0.2449(0.0253)	0.2927(0.0339)	2.4640(0.2113)	1.7874(0.1004)	8.8314(4.0566)	0.5274(0.0255)
Poisson: std. dev.	$\sqrt{d_1}$	0.9974(0.0854)	1.0587(0.0952)	1.0625(0.0871)	-	-	-
Inflation: intercept	γ_0	-4.5813(0.6405)	-1.8113(0.2546)	-	-7.1064(1.3344)	-1.2545(0.1174)	-
Inflation: time	γ_1	0.0921(0.0339)	0.0517(0.0134)	-	0.2921(0.0655)	0.0588(0.0106)	-
Inflation: std. dev.	$\sqrt{d_2}$	2.5327(0.4396)	1.8086(0.2042)	-	-	-	-
R.e. correlation	ρ	-0.0961(0.1534)	-0.6403(0.0810)	-	-	-	-
Predicted prob. zeros		0.3522	0.3281	0.3206	0.3634	0.3312	0.1583
-2log-likelihood		5317.9	5285.9	5417.0	6318.9	6248.9	6326.1
		ZIPN	HPN	PN	ZIP	HP	P
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	0.9027(0.1552)	0.8149(0.1510)	0.8179(0.1677)	1.4205(0.0439)	1.4162(0.0443)	1.2662(0.0424)
Poisson: treatment	ξ_1	0.0051(0.2214)	0.1299(0.1966)	-0.1705(0.2387)	0.3404(0.0595)	0.3456(0.0597)	0.1869(0.0571)
Poisson: time	ξ_2	-0.0042(0.0047)	-0.0025(0.0048)	-0.0143(0.0044)	0.0061(0.0045)	0.0063(0.0045)	-0.0134(0.0043)
Poisson: time*treatment	ξ_3	-0.0032(0.0065)	-0.0067(0.0067)	0.0023(0.0062)	-0.0214(0.0061)	-0.0216(0.0061)	-0.0195(0.0058)
Poisson: std. dev.	$\sqrt{d_1}$	0.9713(0.0824)	1.0031(0.0878)	1.0755(0.0857)	-	-	-
Inflation: intercept	γ_0	-3.7123(0.5003)	-1.8122(0.2547)	-	-1.2879(0.1203)	-1.2545(0.1174)	-
Inflation: time	γ_1	0.0952(0.0249)	0.0517(0.0134)	-	0.0593(0.0109)	0.0588(0.0106)	-
Inflation: std. dev.	$\sqrt{d_2}$	2.2215(0.3434)	1.8100(0.2044)	-	-	-	-
R.e. correlation	ρ	-0.1541(0.1574)	0.6354(0.0795)	-	-	-	-
Predicted prob. zeros		0.3384	0.3277	0.2627	0.3316	0.3312	0.0459
-2log-likelihood		5845.1	5835.4	6271.9	9760	9758.5	11590

Table 7.2: Epilepsy Study. Bayesian parameter estimates and standard errors for the different fitted models.

		ZICOM	HCOM	COM	ZINB	HNB	NB
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	0.4944(0.0831)	0.1936(0.1124)	0.9306(0.1929)	1.2590(0.1161)	0.2618(0.2325)	1.2660(0.1046)
Poisson: treatment	ξ_1	-0.2675(0.2261)	0.0016(0.3121)	-0.2474(0.2718)	0.2183(0.1627)	0.4708(0.2297)	0.2064(0.1476)
Poisson: time	ξ_2	-0.0277(0.0089)	-0.0142(0.0134)	-0.0238(0.0069)	-0.0121(0.0117)	0.0077(0.0160)	-0.0124(0.0103)
Poisson: time*treatment	ξ_3	0.0184(0.0122)	0.0109(0.0191)	0.0114(0.0096)	-0.0230(0.0158)	-0.0283(0.0229)	-0.0218(0.0141)
Poisson: neg.-bin. par.	α	0.7744(0.0462)	1.5340(0.2036)	3.1650(0.2632)	2.1230(0.0975)	5.9750(1.3800)	0.7244(0.0309)
Poisson: std. dev.	$\sqrt{d_1}$	1.0740(0.0951)	1.2120(0.1181)	1.0810(0.0898)	-	-	-
Inflation: intercept	γ_0	-4.9720(0.9949)	-0.8955(0.1296)	-	-2.6550(2.4600)	-1.2520(0.1200)	-
Inflation: time	γ_1	0.0632(0.0939)	0.0510(0.0136)	-	-3.5740(2.0500)	0.0585(0.0109)	-
Inflation: std. dev.	$\sqrt{d_2}$	4.8680(1.0650)	1.8730(0.2190)	-	–	–	–
R.e. correlation	ρ	0.2172(0.2518)	-0.6285(0.0863)	–	–	–	–
Predicted prob. zeros		0.3864	0.3315	0.3125	0.3840	0.3314	0.1954
Marginal likelihood		-2605	-2564	-2584	-3169	-3128	-3187
		ZIPN	HPN	PN	ZIP	HP	P
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	0.4421(0.0740)	0.4084(0.0733)	0.8354(0.1626)	1.4230(0.0422)	1.4130(0.0474)	1.2620(0.0415)
Poisson: treatment	ξ_1	0.0388(0.1886)	0.1249(0.1972)	-0.1905(0.2729)	0.3367(0.0580)	0.3486(0.0654)	0.1926(0.0572)
Poisson: time	ξ_2	-0.0044(0.0046)	-0.0028(0.0049)	-0.0141(0.0042)	0.0060(0.0043)	0.0067(0.0048)	-0.0130(0.0043)
Poisson: time*treatment	ξ_3	-0.0033(0.0063)	-0.0068(0.0065)	0.0017(0.0059)	-0.0211(0.0058)	-0.0220(0.0067)	-0.0200(0.0058)
Poisson: std. dev.	$\sqrt{d_1}$	1.0070(0.0886)	1.0300(0.0923)	1.0900(0.0875)	–	–	–
Inflation: intercept	γ_0	-2.0050(0.2935)	-0.9179(0.1207)	–	-1.2730(0.1238)	-1.2410(0.1141)	–
Inflation: time	γ_1	0.0970(0.02592)	0.0523(0.0137)	–	0.0579(0.0112)	0.0575(0.0103)	–
Inflation: std. dev.	$\sqrt{d_2}$	2.5080(0.4383)	1.8790(0.2122)	–	–	–	–
R.e. correlation	ρ	−0.1541(0.1574)	-0.6160(0.0814)	–	–	–	–
Predicted prob. zeros		0.3394	0.3311	0.2690	0.3322	0.3316	0.0461
Marginal likelihood		-2732	-2717	-2981	-4883	-4882	-5797

Table 7.3: Mesothelioma study. Parameter estimates and standard errors are given for the different fitted models.

		ZICOM	HCOM	(COM)	ZICARCON	HCARCON
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	-0.1562(0.1912)	-0.1341(0.4176)	-0.1084(0.1267)	0.2159(0.1813)	0.4315(0.1503)
Poisson: neg.-bin. par.	α	5.5730(1.0010)	1.0260(0.9461)	1.8460(0.8776)	-	-
Poisson: std. dev. CH	σ_{b0}	0.0604(0.0427)	0.0796(0.0705)	0.1341(0.0666)	0.0282(0.0095)	0.0497(0.0233)
Inflation: intercept	γ_0	-10.440(5.4700)	1.4400(0.1434)	-	-3.1240(4.5530)	1.4500(0.1435)
Inflation: std. dev. UH	σ_{b1}	-	-	-	0.0474(0.0192)	0.0343(0.0243)
Predicted prob. zeros		0.8699	0.8075	0.8027	0.7863	0.8089
M		-194.2	-200.0	-176.2	-180.6	-201.0

		ZICAR	HCAR	CAR	ZIPN	HPN
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	-0.0373(0.1706)	0.4330(0.1554)	-0.0840(0.1118)	0.2355(0.1818)	0.4283(0.1532)
Poisson: std. dev. CH	σ_{b0}	0.4548(0.1303)	0.0705(0.0575)	0.4643(0.1647)	-	-
Inflation: intercept	γ_0	-6.8540(6.2030)	1.4410(0.1462)	-	-2.9150(4.5320)	1.4450(0.1430)
Inflation: std. dev. UH	σ_{b1}	-	-	-	0.1084(0.0736)	0.0388(0.0189)
Predicted prob. zeros		0.7847	0.8076	0.7756	0.7873	0.8081
M		-169.1	-201.1	-170.3	-180.4	-201.1

		ZINB	HPG	PG	ZIP	HP	P
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	-0.1582(0.1919)	-0.1268(0.4191)	-0.1024(0.1244)	0.2446(0.1791)	0.4319(0.1500)	0.0155(0.1010)
Poisson: neg.-bin. par.	α	5.5690(1.0040)	1.0060(0.9380)	1.7600(0.8304)	-	-	-
Inflation: intercept	γ_0	-10.420(5.2520)	1.4430(0.1457)	-	-2.4790(3.8350)	1.4520(0.1441)	-
Predicted prob. zeros		0.8700	0.8079	0.8033	0.7881	0.8094	0.765
M		-194.2	-199.9	-177.4	-180.4	-201.1	-182.5

Table 7.4: Simulation study under scenario S_1. Mean, standard error, and relative bias of the parameter estimates in ZICOM, ZINB, ZIPN, ZIP, and its non-zero-inflated counterparts.

			ZICOM		COM		ZINB		NB	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Poisson: intercept	ξ_0	1.12	1.068(0.003)	0.046	0.991(0.004)	0.115	1.277(0.003)	0.139	2.404(0.005)	1.147
Poisson: time	ξ_1	0.13	0.125(0.001)	0.040	0.136(0.001)	0.046	0.125(0.001)	0.040	0.133(0.001)	0.026
Poisson: help	ξ_2	-1.89	$-1.794(0.002)$	0.051	$-1.796(0.002)$	0.049	$-1.705(0.002)$	0.098	$-1.708(0.002)$	0.096
Poisson: neg.-bin. par.	α	1.00	0.953(0.002)	0.047	0.995(0.002)	0.005	1.774(0.003)	0.774	0.552(0.001)	0.448
Poisson: std. dev.	$\sqrt{d}$	0.80	0.780(0.001)	0.025	0.779(0.001)	0.026	–	–	–	–
Inflation: intercept	γ_0	-1.00	$-0.856(0.099)$	0.104	–	–	$-0.265(0.123)$	0.725	–	–
Inflation: time	γ_1	-1.00	$-1.049(0.098)$	0.049	–	–	$-1.698(0.122)$	0.687	–	–
Predicted prob. zeros		0.48	0.493		0.481		0.359		0.291	
Frequency of convergence			199		200		200		200	
			ZIPN		PN		ZIP		P	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Poisson: intercept	ξ_0	1.12	1.216(0.003)	0.086	0.892(0.003)	0.204	1.661(0.003)	0.483	1.250(0.003)	0.116
Poisson: time	ξ_1	0.13	0.101(0.001)	0.225	0.127(0.001)	0.026	0.089(0.001)	0.318	0.124(0.001)	0.043
Poisson: help	ξ_2	-1.89	$-1.467(0.002)$	0.224	$-1.693(0.002)$	0.104	$-1.275(0.002)$	0.326	$-1.682(0.002)$	0.109
Poisson: std. dev.	$\sqrt{d}$	0.80	0.796(0.001)	0.005	0.861(0.001)	0.076	–	–	–	–
Inflation: intercept	γ_0	-1.00	$-0.386(0.005)$	0.614	–	–	0.247(0.003)	1.247	–	–
Inflation: time	γ_1	$-.00$	$-0.094(0.001)$	0.906	–	–	$-0.094(0.001)$	0.906	–	–
Predicted prob. zeros		0.48	0.473		0.365		0.483		0.255	
Frequency of convergence			200		200		200		200	

Table 7.5: Simulation study under scenario S_2. Mean, standard error, and relative bias of the parameter estimates in ZICOM, ZINB, ZIPN, ZIP, and its non-zero-inflated counterparts.

			ZICOM		COM		ZINB		NB	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Poisson: intercept	ξ_0	1.12	1.079(0.004)	0.037	1.833(0.005)	0.637	1.089(0.005)	0.027	2.796(0.006)	1.497
Poisson: time	ξ_1	0.13	0.123(0.001)	0.052	0.239(0.001)	0.839	0.125(0.001)	0.040	0.225(0.001)	0.730
Poisson: help	ξ_2	-1.89	-1.766(0.003)	0.066	-1.776(0.003)	0.060	-1.671(0.003)	0.116	-1.703(0.003)	0.099
Poisson: neg.-bin. par.	α	1.00	0.908(0.004)	0.093	0.372(0.001)	0.628	2.379(0.008)	1.379	0.266(0.001)	0.734
Poisson: std. dev.	$\sqrt{d}$	0.80	0.772(0.002)	0.035	0.754(0.002)	0.058	–	–	–	–
Inflation: intercept	γ_0	1.00	1.056(0.005)	0.056	–	–	0.993(0.006)	0.003	–	–
Inflation: time	γ_1	-0.25	-0.246(0.001)	0.014	–	–	-0.354(0.001)	0.416	–	–
Predicted prob. zeros		0.68	0.696		0.398		0.549		0.367	
Frequency of convergence			200		200		200		200	
			ZIPN		PN		ZIP		P	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Poisson: intercept	ξ_0	1.12	1.183(0.004)	0.056	-0.235(0.004)	1.210	1.666(0.004)	0.488	0.215(0.004)	0.808
Poisson: time	ξ_1	0.13	0.099(0.001)	0.235	0.212(0.001)	0.633	0.087(0.001)	0.329	0.210(0.001)	0.613
Poisson: help	ξ_2	-1.89	-1.444(0.003)	0.236	-1.679(0.003)	0.112	-1.261(0.002)	0.420	-1.664(0.003)	0.120
Poisson: std. dev.	$\sqrt{d}$	0.80	0.834(0.001)	0.042	0.976(0.001)	0.220	–	–	–	–
Inflation: intercept	γ_0	1.00	1.473(0.004)	0.473	–	–	1.816(0.003)	0.816	–	–
Inflation: time	γ_1	-0.25	-0.209(0.001)	0.163	–	–	-0.202(0.001)	0.193	–	–
Predicted prob. zeros		0.68	0.677		0.520		0.682		0.422	
Frequency of convergence			200		200		200		200	

Table 7.6: Simulation study under scenario S_3. Mean, standard error, and relative bias of the parameter estimates in ZICOM, ZINB, ZIPN, ZIP, and its non-zero-inflated counterparts.

			ZICOM		COM		ZINB		NB	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Poisson: intercept	ξ_0	1.12	1.076(0.007)	0.039	4.005(0.009)	2.828	0.980(0.009)	0.125	4.494(0.008)	3.012
Poisson: time	ξ_1	0.13	0.125(0.001)	0.042	0.216(0.001)	0.658	0.121(0.001)	0.070	0.202(0.001)	0.554
Poisson: help	ξ_2	-1.89	$-1.757(0.005)$	0.070	$-1.765(0.005)$	0.067	$-1.676(0.005)$	0.113	$-1.701(0.005)$	0.100
Poisson: neg.-bin. par.	α	1.00	0.887(0.007)	0.112	0.088(0.001)	0.912	3.041(0.034)	2.041	0.076(0.001)	0.924
Poisson: std. dev.	$\sqrt{d}$	0.80	0.765(0.003)	0.043	0.609(0.004)	0.239	–	–	–	–
Inflation: intercept	γ_0	1.80	1.862(0.006)	0.034	–	–	1.487(0.009)	0.174	–	–
Inflation: time	γ_1	-0.10	$-0.102(0.001)$	0.017	–	–	$-0.118(0.001)$	0.177	–	–
Predicted prob. zeros		0.88	0.884		0.590		0.807		0.604	
Frequency of convergence			200		200		200		200	
			ZIPN		PN		ZIP		P	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Poisson: intercept	ξ_0	1.12	1.051(0.008)	0.061	$-1.515(0.007)$	2.353	1.660(0.006)	0.482	$-0.631(0.007)$	1.563
Poisson: time	ξ_1	0.13	0.104(0.001)	0.203	0.195(0.001)	0.502	0.088(0.001)	0.323	0.193(0.001)	0.487
Poisson: help	ξ_2	-1.89	$-1.473(0.005)$	0.221	$-1.669(0.005)$	0.117	$-1.257(0.004)$	0.335	$-1.661(0.005)$	0.121
Poisson: std. dev.	$\sqrt{d}$	0.80	0.941(0.002)	0.176	1.416(0.002)	0.769	–	–	–	–
Inflation: intercept	γ_0	1.80	2.205(0.005)	0.225	–	–	2.629(0.004)	0.382	–	–
Inflation: time	γ_1	-0.10	$-0.112(0.001)$	0.122	–	–	$-0.127(0.001)$	0.271	–	–
Predicted prob. zeros		0.88	0.876		0.756		0.877		0.675	
Frequency of convergence			200		200		200		200	

Chapter 8

The Spatial Bivariate Combined Model for Count Data

8.1 Introduction

While Chapter 6 and 7 dealt with implementing the combined model into the spatial context, only the univariate data setting was investigated. In practice however, counts per area are available for different diseases or for different population groups, and interest may be in the spatial distribution of both diseases or population groups, as well as in the correlation between the spatial distributions. Existing spatial modeling frameworks for multivariate data are based on the extension of the CAR convolution model (Section 6.2), assuming a Poisson distribution for the counts conditional on the spatial process, and assuming that the spatial process is the sum of a Gaussian Markov random field plus an additional unstructured Gaussian variation, such as proposed by Besag et al. (1991), and illustrated in e.g. Clayton and Bernardinelli (1992), Rue and Held (2005) and Lawson (2013) amongst others. Recall that the convolution model allows to account for both the overdispersion, also called uncorrelated heterogeneity, via the unstructured variation, as well as for the spatial correlation, also called correlated heterogeneity, via the Markov random field, explaining the wide use and applicability of the convolution model. Also recall that, as proposed by Neyens et al. (2012), the combined model assumes that the spatial process is the sum of a Gaussian Markov random field, to account for the spatial heterogeneity, plus a gamma-distributed unstructured heterogeneity term.

Extensions of the convolution model towards two or more diseases have been given

by e.g. Knorr-Held and Best (2001), Gelfand and Vounatsou (2003) and Lawson (2013, chapter 10). Knorr-Held and Best (2001) propose the use of a shared component model, assuming a shared Gaussian Markov random field for both diseases. Gelfand and Vounatsou (2003) propose a multivariate conditional autoregressive model, introducing correlation in the spatial component. Lawson (2013) discusses introducing correlation in both the aspatial component, and in the spatially structured component. This has been used also by Kramer and Williamson (2013), showing the flexibility of the latter model and the possibility to quantify the correlation between the spatial processes. In this chapter, an alternative extension of the latter model is proposed, with correlation between the aspatial components introduced via gamma-distributed random effects. This forms an alternative method to the commonly assumed Gaussian convolution model, providing the practitioner with more tools to efficiently model the bivariate disease distributions.

In this chapter, Section 8.2 gives an explanation on existing methods to model spatial counts bivariately and proposes a new bivariate extension of the combined model. The data application considers the study of asthma and COPD in Georgia (USA), and the study of bladder cancer in males and females in Limburg (Belgium) and is covered in Section 8.3. A conclusion is provided in Section 8.4.

8.2 Bivariate disease mapping

Throughout the previous chapters, different extensions of the combined model were investigated. These were all based on data coming from one population, although working towards the inclusion of two populations seems interesting. Indeed, when looking at the spatial data setting covered in Chapters 6 and 7, the univariate disease map gave a number of answers to scientific questions, but issues concerning interactions between the two diseases remain unsolved and require a bivariate approach. In what follows, a general modeling framework for two diseases will be given before focusing on the bivariate extension of the combined model.

8.2.1 Bivariate convolution model

To put the definitions of the proposed bivariate combined model into perspective, it is important to keep in mind that basically two strategies exist in the bivariate setting, namely models with common random effects and models with correlated random effects. The first type of models assumes that specific spatial or non-spatial extra-variance terms are shared between the models, while both models may also have a set of separate terms. As an extension of the convolution model, one could consider the following model (for an

overview of models, see Lawson, 2013):

$$\begin{aligned}\kappa_{1i} &= \exp\left(\xi_{0,1} + b_{0,i} + b_{1,1i}\right),\\ \kappa_{2i} &= \exp\left(\xi_{0,2} + b_{0,i} + b_{1,2i}\right),\end{aligned}$$

with $b_{1,1i}$ and $b_{1,2i}$ defined univariately as mentioned in Section 6.2 and $b_{0,i} \sim N(0, \sigma_0^2)$. Alternatively, one can choose to take the spatial random effects as shared ($b_{1,1i} = b_{1,2i} = b_{1,i}$) while the UH term may be taken as disease-specific ($b_{0,li}, l = 1, 2$). When the diseases are not equally common, it is proposed to take into account a scaling component (Knorr-Held and Best, 2001). By doing so, the amount of overdispersion explained by the random effects is not equal any more. As an example, a typical model is:

$$\begin{aligned}\theta_{1i} &= \exp\left(\xi_{0,1} + \delta b_{0,i} + b_{1,1i}\right),\\ \theta_{2i} &= \exp\left(\xi_{0,2} + b_{0,i}/\delta + b_{1,2i}\right),\end{aligned}$$

with δ the scaling component. On the other hand, it is not difficult to understand that models with shared terms are somehow too restrictive and indeed, intuitively two related diseases are seen as diseases that act alike, not the same. This can be modeled by assuming that both random effects are correlated. In the disease mapping context, such models are not yet widely available, although the multivariate CAR convolution model has been developed (Stern and Cressie, 2000), which in the bivariate framework uses the MCAR specification, which models spatial correlation on a multivariate scale (Gelfand and Vounatsou, 2001), while it uses $(b_{0,1i}, b_{0,2i}) \sim MVN(0, \Sigma)$ for the aspatial extra-variance. As an alternative to this model, we propose a bivariate extension of the combined model, given in the next section.

8.2.2 Bivariate combined model

In the bivariate combined model, we assume that

$$\begin{aligned}Y_{i1} &\sim \text{Poisson}(E_{i1}\kappa_{i1}\theta_{i1}),\\ Y_{i2} &\sim \text{Poisson}(E_{i2}\kappa_{i2}\theta_{i2}),\end{aligned}$$

with the relative risks modeled as

$$\begin{aligned}\kappa_{i1} &= \exp\left(\xi_{0,1} + b_{1,1i}\right),\\ \kappa_{i2} &= \exp\left(\xi_{0,2} + b_{1,2i}\right).\end{aligned}$$

The random effects $b_{1,1i}$ and $b_{1,2i}$ are conditional autoregressive random effects, and are specified either univariately as in Section 6.2 or multivariately, using the MCAR specification (Gelfand and Vounatsou, 2001; Stern and Cressie, 2000). The gamma-distributed

overdispersion random effects g_{i1} and g_{i2} can also be modeled either uni- or bivariately, with the bivariate distribution defined as follows:

$$\begin{aligned}\theta_{i1} &= \frac{1}{k_0+k_1}(\gamma_{i0}+\gamma_{i1}) \sim \Gamma(k_0+k_1, \frac{1}{k_0+k_1}) \\ \theta_{i2} &= \frac{1}{k_0+k_2}(\gamma_{i0}+\gamma_{i2}) \sim \Gamma(k_0+k_2, \frac{1}{k_0+k_1}).\end{aligned}$$

where k_0, k_1 and k_2 are three real positive variables, similar as in Hens et al. (2009). The terms γ_0, γ_1, and γ_2 are assumed as independent gamma-distributed random variables. The common term is assumed to be $\gamma_0 \sim \Gamma(k_0, 1)$, and the two disease- or population-specific terms are $\gamma_1 \sim \Gamma(k_1, 1)$ and $\gamma_2 \sim \Gamma(k_2, 1)$. Consequently, the aspatial heterogeneity terms are associated, with the Pearson product-moment correlation coefficient equal to $\rho = \frac{k_0}{\sqrt{(k_0+k_1)(k_0+k_2)}}$. This is clearly less restrictive than the assumption of a perfect correlation when assuming a shared aspatial effect $g_{i1} = g_{i2}$. Note however that the correlation coefficient for the correlated gamma model is bounded such that $0 \leq \rho \leq \min(\frac{\sigma_{g1}}{\sigma_{g2}}, \frac{\sigma_{g2}}{\sigma_{g1}})$. Further, note that by using this formulation, g_1 and g_2 remain gamma-distributed, such that the conjugacy between the Poisson and gamma distributions still yields a closed-form distribution for each disease.

8.3 Data application

In this section, we compare the different models within this modeling framework. The asthma and COPD data from Georgia (Section 2.2.3) and the male and female bladder cancer data from Limburg (Section 2.2.1) were used. Both data sets contain two outcomes that are interesting to model simultaneously: For the first data, asthma and COPD counts were collected in Georgia. It is likely that the occurrences of these outcomes are correlated, since they are both respiratory diseases and therefore can be caused by the same agents. The bladder cancer data on the other hand consists out of counts for males and females. A similar risk could be expected for these two sub-populations, and therefore it seems interesting to model of the spatial distribution of males and females simultaneously.

From Figure 2.2 in Section 2.2.3, it was noticeable that observed counts for asthma and COPD were very high in and around Atlanta. This was probably due to Georgia's largest population sizes being found in those regions, since SIR estimates seem to be highest in the southeastern region (Figure 8.1). Biased interpretations can also emerge when only looking at the observed male and female bladder cancer counts in Limburg (Figure 2.1). This cancer type is diagnosed relatively frequently, more in males than females, and it is clear that for both genders, the southwestern towns have increased SIR's (Figure 8.2). As was mentioned before, SIR estimates do not suffice to capture

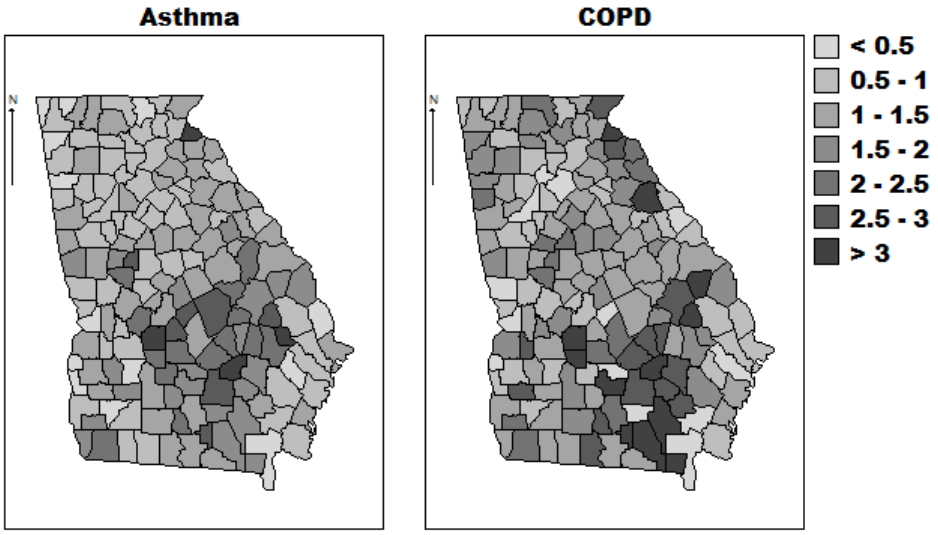


Figure 8.1: Standardized Incidence Rates ($\text{SIR}_i = Y_i/E_i$) per county for asthma (left panel) and COPD (right panel) in Georgia (USA).

spatial or non-spatial extra-variance in the data, which point towards the need of models, more precisely those that model both populations simultaneously to take possible inter-population correlations into account.

Several models will be considered for both case studies and are listed in Table 8.1. It can be seen that there is a large set of possible models, amongst which the best model can be chosen. Included are univariate models fitted separately on the two outcomes, joint models with a correlated random effect and joint convolution models (two random effects). In order to evaluate goodness-of-fit, DIC and MSPE were used again (for an in-depth explanation, consult Sections 5.2 and 6.4.2). Furthermore, the empirically-based Pearson correlations between both diseases/population groups were calculated for the estimated spatial random effects, non-spatial random effects and relative risks, i.e. $r_{x_1,x_2} = \frac{cov(x_1,x_2)}{sd(x_1)sd(x_2)}$ with $x = b_{0,i}, b_{1,i}, g_i$ or κ_i.

8.3.1 Asthma and COPD in Georgia

Table 8.2 gives an overview of the model fits. For the Georgia data, Model 3 (d) did not converge, even after considerable extra iterations. The best model according to DIC was achieved by model 3 (c), with 2 univariate CAR random effects and a bivariate UH term.

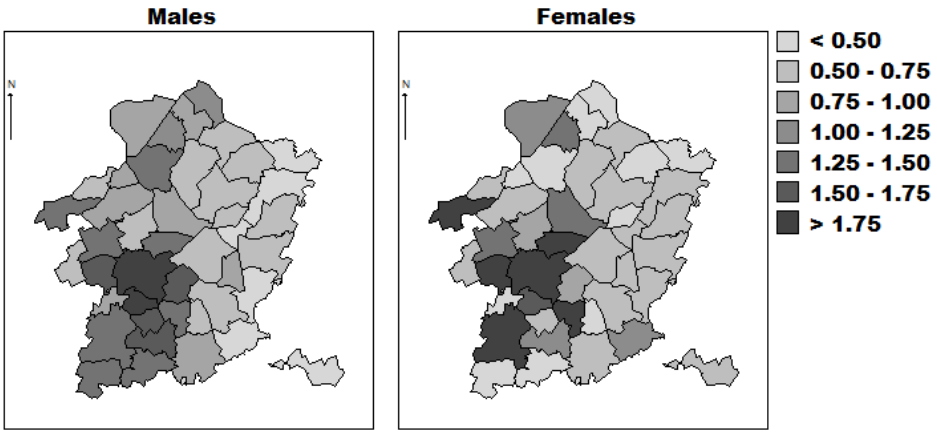


Figure 8.2: Standardized Incidence Rates for bladder cancer ($\text{SIR}_i = Y_i/E_i$) per municipality for males (left panel) and females (right panel).

It is also apparent that the models with gamma bivariate terms had consistently smaller DIC values than those with normal bivariate UH random effects, e.g. when comparing bivariate model 2 (b) and 2 (c) or the convolution bivariate models 3 (a) and 3 (c). Further, it can be observed that the univariate models do not seem to do much worse than bivariate models, e.g. univariate model 1 (a) had a lower DIC value than many other bivariate models, such as the convolution model 3 (b). In terms of MSPE, different things can be seen: Model 2 (c) and 2 (b), the models with respectively only a gamma and a bivariate normal random effects term, had the lowest error values followed by the univariate gamma model. Although differences between gamma and normal bivariate UH terms were small, it can be stated that also in terms of MSPE, the bivariate gamma term performed equally or better than the bivariate normal term. Lastly, it is interesting to point out that of all models, model 2 (a), with only a MCAR term, did worse in terms of both DIC and MSPE. This may be due to the combination of the need for a UH random effects term and the non-necessity of a MCAR term as spatial random effect.

The last column in Table 8.2 shows the empirical-based correlation estimates for the relative risks. These estimates summarize the correlation between COPD and asthma. From all models, it can be seen that the correlation between the two diseases is estimated around 0.5. At first sight, this is unexpected, since the first two models (1 (a) and 1 (b)) model the two diseases independently. This indicates that these models are indeed insufficient in understanding the source of association between

Table 8.1: An overview of the fitted case study models for the Georgia and Limburg bladder cancer data.

Model	Family	Random Effects	
		Spatial	Non-Spatial
1(a)	Univariate models on two outcomes	/	disease-specific univariate gamma
1(b)	Bivariate models with correlated random effects	disease-specific UCAR	/
2(a)	Bivariate models with correlated random effects	MCAR	/
2(b)	Bivariate models with correlated random effects	/	bivariate normal
2(c)	Bivariate models with correlated random effects	/	bivariate gamma
3(a)	Bivariate convolution models	disease-specific UCAR	bivariate normal
3(b)	Bivariate convolution models	MCAR	bivariate normal
3(c)	Bivariate convolution models	disease-specific UCAR	bivariate gamma
3(d)	Bivariate convolution models	MCAR	bivariate gamma

the diseases. In addition, when modeling the association between the two diseases, correlations increase to around 0.6, indicating that a misspecification of the models might underestimate the correlation. Among the bivariate models, the correlation of the RRs remains relatively alike. However, in order to understand the source of association, also the empirical correlations of the random effect terms were calculated. Looking at the best fitting models (models 3 (c) based on DIC and models 2 (b) and 2 (c) based on MSPE), a significant correlation is observed for the non-spatial heterogeneity terms and a non-significant correlation for the spatial component. This consistently indicates that the correlation between two diseases does not have an environmental cause.

When looking at the relative risk maps (Figure 8.3), based on model 3 (c), we indeed see that both diseases are correlated, with a shared increase in relative risk in the central southeastern part of Georgia, while there are also disease-specific patterns scattered around the map. When investigating Figure 8.4, which shows maps of γ_0, γ_1 and γ_2, the shared and disease-specific parts of the bivariate gamma-distributed random effects, we see that γ_0 represents the shared relative risk increase in the southeast, while γ_1 and γ_2 take the disease-specific relative risk changes throughout the map into account. Furthermore, γ_0, γ_1 and γ_2 only give information about the UH in the data, while relative risks also include the CH, which is shown in Figure 8.5. Indeed, also the CH maps show higher values in the southeastern part of the map, which indeed suggests that in this case extra-variance, which was not taken into account by the spatial random effect, was modeled by γ_0, γ_1 and γ_2.

Table 8.2: Model fits and empirically based correlations (with 95% credible intervals) between random effects and relative risks.

Model	Random Effects		Model Fit			EB RE Corr		EB
	Spat.	Non-Spat.	DIC	pD	MSPE	Spatial	Non-spatial	RR corr
Georgia Data								
1(a)	/	univ. gam.	2376.7	281.0	**330.7**	/	0.5002 [0.4299; 0.5683]	0.5002 [0.4299; 0.5683]
1(b)	UCAR	/	2399.2	275.1	347.9	0.47 [0.4047; 0.4709]	/	0.5186 [0.4512; 0.5839]
2(a)	MCAR	/	2452.4	297.8	402.3	0.5731 [0.506; 0.636]	/	0.6101 [0.5429; 0.6732]
2(b)	/	biv. norm.	2380.5	273.3	**330.2**	/	0.5646 [0.4949; 0.6294]	0.5935 [0.5229; 0.6599]
2(c)	/	biv. gam.	2359.8	265.4	**330.1**	/	0.6257 [0.559; 0.6879]	0.6257 [0.559; 0.6879]
3(a)	UCAR	biv. norm.	2375.7	271.5	336.2	0.145 [-0.3594; 0.5042]	0.6221 [0.4932; 0.7346]	0.5836 [0.5136; 0.6506]
3(b)	MCAR	biv. norm.	2393.0	280.2	342.6	0.6289 [0.3105; 0.8144]	0.4278 [0.06492; 0.7373]	0.5889 [0.5193; 0.6545]
3(c)	UCAR	biv. gam.	**2346.1**	254.3	335.4	0.13 [-0.4128; 0.543]	0.6736 [0.5786; 0.7636]	0.6124 [0.5458; 0.675]
3(d)	MCAR	biv. gam.	/	/	/	/	/	/
Limburg Data								
1(a)	/	univ. gam.	465.6	55.3	57.7	/	0.3525 [0.125; 0.5621]	0.3525 [0.125; 0.5621]
1(b)	UCAR	/	479.9	43.8	68.2	0.5264 [0.2859; 0.7164]	/	0.5913 [0.3515; 0.7572]
2(a)	MCAR	/	459.3	43.7	63.15	0.7305 [0.478; 0.8868]	/	0.7375 [0.496; 0.8881]
2(b)	/	biv. norm.	459.8	51.1	57.11	/	0.5897 [0.3102; 0.7889]	0.6102 [0.3541; 0.7983]
2(c)	/	biv. gam.	**450.5**	39.8	**56.94**	/	0.8367 [0.6385; 0.9509]	0.8367 [0.6385; 0.9509]
3(a)	UCAR	biv. norm.	460.6	52.0	57.44	0.01876 [-0.4529; 0.4948]	0.5744 [0.2711; 0.5903]	0.595 [0.3311; 0.788]
3(b)	MCAR	biv. norm.	463.4	57.2	**56.7**	0.5227 [-0.09472; 0.8513]	0.3091 [-0.1761; 0.6829]	0.5283 [0.2604; 0.7354]
3(c)	UCAR	biv. gam.	454.2	42.0	57.27	0.03011 [-0.4401; 0.501]	0.8236 [0.586; 0.9585]	0.8065 [0.5666; 0.952]
3(d)	MCAR	biv. gam.	453.7	45.1	57.58	0.5549 [-0.02187; 0.8565]	0.7362 [0.2832; 0.9449]	0.6675 [0.396; 0.8531]

8.3.2 Bladder cancer in Limburg

The analysis of the Limburg data tells a different story. Results are summarized in the bottom panel of Table 8.2. When looking at the DIC (and pD) values, the univariate models resulted in poorer fits than the multivariate models. Again, the convolution models, which combine CH with UH terms, did not have better fits than the other bivariate models (models 2 (a) - 2 (c)). In fact, the model with the lowest DIC was model 2 (c), which only has a bivariate gamma term. Furthermore, it again is striking that the bivariate gamma term provides better fits than the bivariate normal one, e.g. when comparing DIC values from model 2 (b) with 2 (c), or the convolution models 3 (a) with 3 (c) and 3 (b) with 3 (d), the models with a bivariate gamma term consistently turn out as the better ones. When looking at the MSPE statistics, it seems that models without UH terms (Models 1 (b) - 2 (a)) do worse than the others. The best models in terms of MSPE are models 3 (b) (MCAR CH random effects and a bivariate random effect) and 2 (c) (only a bivariate gamma term). When looking at the DIC and MSPE investigations together, model 2 (c) would then constitute as the best model.

When looking at the RR correlation estimates, we see that model 1 (a) has a substantially smaller estimate than the other models. Furthermore, RR correlation estimates were less stable among the models when compared with those given by the Georgia data analyses. When focusing on the random effects correlation estimates, again relatively high values were estimated for the univariate models. Again all correlation

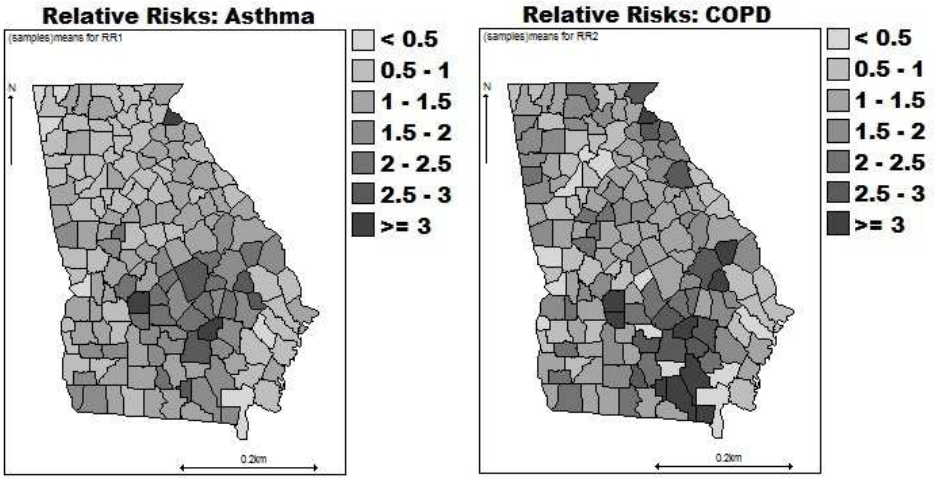


Figure 8.3: RR maps for asthma and COPD counts in the counties of Georgia.

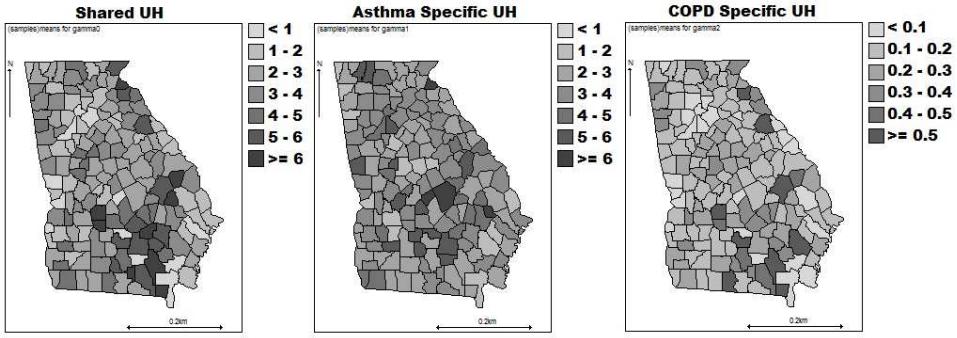


Figure 8.4: Maps of γ_0, γ_1 and γ_2 for asthma and COPD counts in the counties of Georgia.

estimates are positive, but not all of them differ significantly from zero. When looking at the convolution models in particular, the spatial random effects correlations are never significant and for model 3 (b) also the non-spatial random effects are not significantly correlated between males and females. However, RR correlation estimates do show significant positive correlations for all models.

When looking at the RR estimates (Figure 8.6), based on model 2 (c), it is clear that there is a high correlation between the male and female cases with elevated risks in the southwestern parts of Limburg. This high correlation is also visible in Figure 8.7 (b), where the shared UH (γ_0) has large values, while the disease specific UH terms (γ_1 and

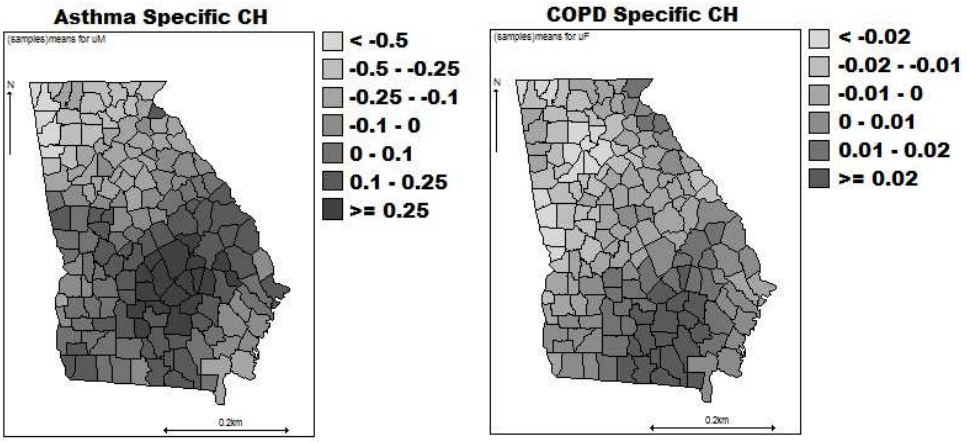


Figure 8.5: Disease-specific CH maps for asthma and COPD counts in the counties of Georgia.

γ_2) do not contribute that much.

8.4 Concluding remarks

In this chapter, a novel method for the bivariate analysis of spatial disease counts was proposed, and an investigation of the source of correlation between diseases or population was presented. As indicated by the modeling results, the bivariate combined model, which introduces a bivariate gamma distributed random effects term to capture aspatial extra-variance, fitted better than a model with the same spatial random effects but with a multivariate normal random effect for UH. As data sets differ, the best way to analyze them will differ too. Therefore, the bivariate combined model proposes an interesting option when working with two possibly related data sets existing out of counts. Although prior and initial values choices for the shape parameters in γ_0, γ_1 and γ_2 have to be assigned thoughtfully to avoid iterating difficulties and malconvergence, it has been shown that the bivariate combined model presents an interesting new piece in the spatial statistician's toolbox.

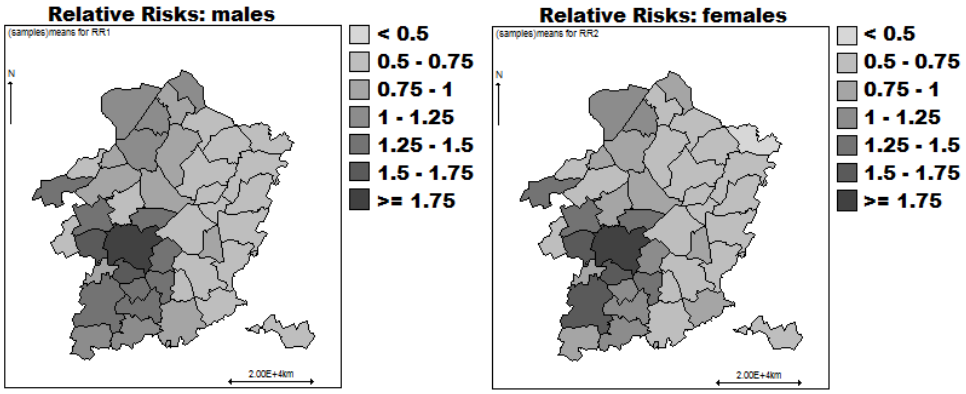


Figure 8.6: RR maps for male and female bladder cancer counts in the municipalities of Limburg.

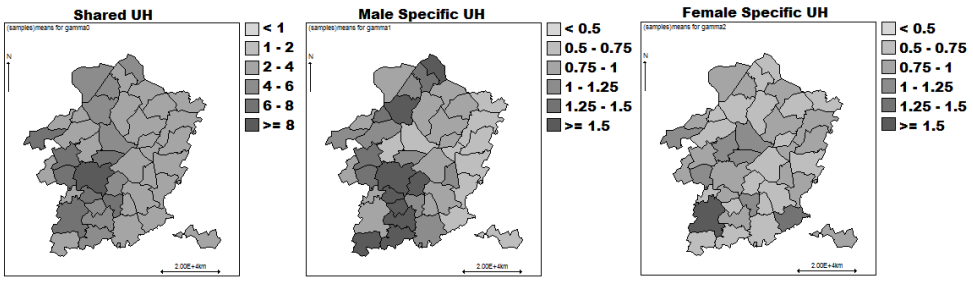


Figure 8.7: Maps of γ_0, γ_1 and γ_2 for male and female bladder cancer counts in the municipalities of Limburg.

Chapter 9

General discussion and conclusion

This thesis was built around the combined model, which is a GLMM that specifically models uncorrelated and correlated extra-variance by combining overdispersion models that exploit the conjugacy characteristic and normal random effects that take the structural aspects in the data into account. I focused on binomial and count data types. This thesis has shown that the combined model is very useful in most cases. A number of remarks and avenues for future research can be noted.

In Chapters 4 and 5, I closely compared different estimation techniques. The reason here lies in the fact that the combined model was developed in a likelihood framework and used partial marginalization along the exponential model family's conjugacy characteristic to allow for both a normal and a conjugate random effects term. Not surprisingly, the combined model provided very good results when using it in PROC NLMIXED (Molenberghs et al., 2010). However, recent statistical science has developed tools to adequately work with multi-hierarchical data, with MCMC being the most popular one. Indeed, it feels natural to further develop the combined model, which provides a way to efficiently model extra-variance caused by different sources in the data, in a Bayesian setting. I have been closely involved in these developments (Neyens et al., 2012, Neyens et al., 2015a, Neyens et al., 2015c), while I also have been involved with research that specifically compared the model's behaviour between the likelihood and MCMC framework (Neyens et al., 2015b, Kassahun et al., 2012, Kassahun et al., 2015). Without going into detail, it is safe to say that the combined model extends well into

the MCMC territory. For both binary and count data, in longitudinal and spatial data settings, MCMC provides ways to deal with multiple hierarchies within the context of the combined model, presuming the data set is sufficiently large. In Neyens et al. (2015b), which is mainly covered in Chapter 5, a comparison between the likelihood, MCMC and INLA estimation techniques was undertaken. INLA is an approximate Bayesian method that uses Laplace approximations to work around MCMC's long computing times, which is a well-known issue and MCMC's most important drawback. Although INLA works extremely fast and mostly provides good estimates, the reported problems which have to do with random effects parameter estimation, make its usefulness doubtful in a number of cases. Further research is needed though, mainly because of the following reason: Bayesian methodology promised from the beginning that the possibility to build models with many complexities, would revolutionize modern statistical science. It has partly done so, but in order to receive good results from MCMC, one has to wait very long, mostly to find out that still a few parameters have not been converged. The paradox here is that inferential validity rests on the convergence of a Markov chain to its equilibrium distribution, which in fact can never be perfectly achieved and therefore is difficult to verify empirically. To fully exploit the possibilities that Bayesian statistical methodology offers, its use should be made more attractive to the average end user. I strongly believe that approximation techniques are the best way forward. Although INLA still remains to be improved, not only in terms of parameter estimation, but also in its modeling flexibility, it seems to be an important avenue for future research, among other approximation techniques (e.g Ghebretinsae et al., 2012a).

The combined model has been a source of research for many statisticians and not all of the research I was involved with, has become part of this thesis. In Chapter 4, I focused on the combined model for the logit case, based on Kassahun et al. (2012). As was explained in Chapter 3, the use of the combined model is not as natural in the logit case as in the Poisson case, on which the rest of this thesis is built. It however gave very good modeling opportunities. The combined model has been extended towards the joint modeling case too, research in which I was closely involved too (Kassahun et al., 2015a), but in which others also did important work (Njagi et al., 2013). Next to investigations by Efendi et al. (2013), I was also part of research towards the implementation of the combined model in the context of marginalized models for count data (Kassahun et al., 2014b), in which it also worked very well. For count data, I have focused mainly on spatial data, when modeling one spatial count outcome (Neyens et al., 2010), or when investigating two counts simultaneously (Neyens et al., 2015a). In the univariate case, the combined model proved to be a good tool and a worthy counterpart for the

popular CAR convolution model. In the bivariate case, the bivariate gamma random effects were shown to be superior to the frequently used bivariate log-normal random effects. Finally, I have been involved with the extensions of the combined model towards the setting with many zeros, in the longitudinal (Kassahun et al. 2014a, Kassahun et al, 2015a, Kassahun et al, 2015b) and the spatial setting (Neyens et al., 2015c). While the longitudinal cases all considered large sample cases and have shown the extensive capabilities of zero-excess models, the spatial models, as they were based on smaller data sets needed more assumptions in order to make the models estimable. Keeping this in mind, the fully multivariate and spatio-temporal case becomes an interesting research avenue. With the arrival of supercomputers, it has become tempting to model many variables simultaneously through time and space. Although the statistical complexities that arise are not to be underestimated, many have been investigating those cases (e.g. Cressie and Wilke, 2011) and it has shown to be an important field in future statistical science. I believe that this path provides the way forward and extensions of the successful combined model into those territories will be obligate.

Bibliography

Agresti, A. (2002) *Categorical Data Analysis* (2nd ed.). New York: John Wiley & Sons.

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Dordrecht (NE): Kluwer Academic Publishers.

Aregay, M., Shkedy, Z., and Molenberghs, G. (2013) A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion: A simulation study. *Computational Statistics & Data Analysis*, **57**, 233–245.

Bergmann, K. E., Bergmann, R. L., Von Kries, R., Bohm, O., Richter, R., Dudenhausen, J. W., and Wahn, W. (2003) Early determinants of childhood overweight and adiposity in a birth cohort study: role of breast-feeding. *International Journal of Obesity*, **27**, 162–172.

Besag J. and Kooperberg, C. (1995) On conditional and intrinsic autoregressions. *Biometrika*, **82**, 733–746.

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**, 1–59.

Best N. S., Richardson S., and Thompson A. (2005) A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, **14**, 35–59.

Bethlehem, J. (2009). *Survey Methods: A Statistical Perspective*. Hoboken: Wiley.

Bivand R. S., Pebesma E. J., and Gómez-Rubio V. (2008) *Applied Spatial Data Analysis with R*. New York: Springer.

Bliddal, S., Feldt-Rasmussen, U., Boas, M., Faber, J., Juul, A., Larsen, T., and Hansen Precht, D. (2011) Gestational age-specific reference ranges from different laboratories misclassify pregnant women's thyroid status: comparison of two longitudinal prospective cohort studies. *European Journal of Endocrinology*, **170**, 329-339.

Booth, J. G., Casella, G., Friedl, H., and Hobert, J. P. (2003) Negative binomial loglinear mixed models. *Statistical Modelling*, **3**, 179–181.

Breslow, N. E. (1984) Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38–44.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brooks, S. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Computing Science and Statistics*, **7**, 434–455.

Clayton, D. G. and Bernardinelli, C. (1992) Bayesian methods for mapping disease risk. In Elliott, P., Cuzick, J., English, D., and Stern R. (Eds.), *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. Oxford University Press: Oxford.

Clayton, D. G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman & Hall, p. 370.

Cressie, N. A. C. (1993) *Statistics for Spatial Data* (rev. ed.) New York: Wiley.

Cressie, N. A. C. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. New York: Wiley.

Diggle, P. J. (1981) Binary mosaics and the spatial pattern of Heather. *Biometrics*, **37**, 531–539.

Duchateau, L. and Janssen, P. (2008) *The Frailty Model*. New York: Springer.

Efendi, A., Molenberghs, G., Njagi, E. N., and Dendale, P. (2013) A joint model for longitudinal continuous and time-to-event outcomes with direct marginal interpretation. *Biometrical Journal*, **55**, 572–588.

Elliott, P., Wakefield, J., Best, N., and Briggs, D. (2000) *Spatial Epidemiology*. New York: Oxford University Press Inc.

Engel, B. and Keen, A. (1992) A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**, 1–22.

Faes, C., Ormerod, J. T., and Wand, M. P. (2011) Variational Bayesian inference for parametric and nonparametric regression with missing data. *Journal of the American Statistical Association*, **106**, no. 495.

Faught, E., Wilder, B. J., Ramsay, R. E., Reife, R. A., Kramer, L. D., Pledger, G. W., and Karim, R. M. (1996) Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages. *Neurology*, **46**, 1684–1690.

Ferrís-i-Tortajada, J., Berbel-Tornero, O., Garcia-i-Castell, J., López-Andreu, J. A., Sobrino-Najul, E., and Ortega-García, J. A. (2011) Non-dietary environmental risk factors in prostate cancer. *Actas Urológicas Españolas (English Edition)*, **35**, 289–295.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2009) *Longitudinal Data Analysis*, Boca Raton: Chapman & Hall.

Geisser, S. (1993) *Predictive Inference: An Introduction*. New York: Chapman & Hall.

Gelfand, A. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, **4**, 11–25.

Gelman, A. E. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

Gelman, A. E., Carlin, J., Stern, H., and Rubin, D. B. (2004) *Bayesian Data Analysis*, Second Edition. Boca Raton: Chapman & Hall.

Gelman, A. E. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

George, E. I., Makov, U. E., and Smith, A. F. M. (1993) Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, **20**, 147–156.

Ghebretinsae, A. H., Faes, C., and Molenberghs, G. (2012a) Gaussian variational approximation for overdispersed generalized linear mixed models. *Proceedings of the Joint Statistical Meetings*, **C2**.

Ghebretinsae, A. H., Faes, C., Molenberghs, G., Geys, H., and Van de Leede, B.-J. (2012b) Joint modeling of hierarchically clustered and overdispersed non-gaussian continuous outcomes for comet assay data. *Pharmaceutical Statistics*, **11**, 449–455.

Gilks, W. and Richardson, S., and Spiegelhalter, D. (1996) *Markov Chain Monte Carlo in Practice*. Boca Raton: Chapman & Hall.

Gillman, M. W., Rifas-Shiman, S. L., Berkey, C. S., Frazier, A. L., Rockett, H. R., Camargo Jr, C. A., Field, A. E., and Colditz, G. A. (2006) Breast-feeding and overweight in adolescence: within-family analysis. *Epidemiology*, **17**, 112–114.

Goeyvaerts, N., Potter, G., Van Kerckhove, K., Willem, L., Beutels, P. and Hens, N. (2014) Within-household contact networks: implications for epidemic modeling. *Technical Report*.

Goldstein, H. (2002) *Multilevel Statistical Models.* Third Edition. Oxford: Oxford University Press.

Greene, W. (1994) Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC- 94-10, Department of Economics, New York University. *Working Paper*, 9–10.

Gschössl, S. and Czado, C. (2008) Modeling count data with overdispersion and spatial effects. *Statistical Papers*, **49**, 531–552.

Hens, N., Wienke, A., Aerts, M., and Molenberghs, G. (2009) The correlated and shared gamma frailty model for bivariate current status data. An illustration for cross-sectional serological data. *Statistics in Medicine*, **28**, 2785–2800.

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

Hinde, J. and Demétrio, C. G. B. (1998) Overdispersion: models and estimation. *São Paulo: XIII Sinape*.

Inskip, H., Beral, V., Fraser, P., and Haskey, P. (1983) Methods for age-adjustments of rates. *Statistics in Medicine*, **2**, 483–493.

Kalema, G. and Molenberghs, G. (2014) Pseudo-likelihood methodology for hierarchical count data. *Communications in Statistics: Theory and Methods*, **43**, 4790–4805.

Karlis, D. and Ntzoufras, I. (2003) Analysis of sports data by using bivariate Poisson models. *The Statistician*, **52**, 381–393.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2012) Modeling overdispersed longitudinal binary data using a combined beta and normal random-effects model. *Archives of Public Health*, **70**: 7.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2014a). A zero-inflated overdispersed and hierarchical Poisson model. *Statistical Modelling*, **14**, 439–456, DOI:10.1177/1471082X14524676.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2014b) Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. *Statistics in Medicine*, **33**, 4402-–4419, DOI: 10.1002/sim.6237.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2015a). A joint model for hierarchical continuous and zero-inflated overdispersed count data. *Journal of Statistical Computation and Simulation*,**84**, 552–571, DOI: 10.1080/00949655.2013.829058.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2015b) The zero-inflated and hurdle combined model for count data in a Bayesian context. *Work in progress*.

Kazemi, I., Mahdiyeh, Z. Mansourian, M., and Park, J. J. (2013) Bayesian analysis of multivariate mixed models for a prospective cohort study of skew-elliptical distributions. *Biometrical Journal*, **55**, 495–508.

Kelsall, J. and Wakefield, J. (1999) Modelling spatial variation in disease risk. *Technical Report, Imperial College, London*.

Kim, H., Sun, D., and Tsutakawa, R. K. (2002) Lognormal vs. gamma: extra variations. *Biometrical Journal*, **44**, 305–323.

Kleinman, J. (1973) Proportions with extraneous variance: single and independent samples. *Journal of the American Statistical Association*, **68**, 46–54.

Knorr-Held, L. and Best, N. G. (2001) A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society*, **164**, 73–85.

Kramer, M. R. and Williamson, R. (2013) Multivariate Bayesian spatial model of preterm birth and cardiovascular disease among Georgia women: Evidence for life course social determinants of health. *Spatial and Spatio-temporal Epidemiology*, **6**, 25–35.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lawless, J. (1987) Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 209–225.

Lawson, A. B. (2013) *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Second Edition*. Boca Rotan: Chapman & Hall.

Lawson, A. B., Biggeri, A. B., Boehning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P., and Divino, F. (2000) Disease mapping models: an empirical evaluation. *Statistics in Medicine*, **19**, 2217–2241.

Lee, A. H., Wang, K., Scott, J., Yau, K. K. W. and McLachlan, G. J. (2006) Multilevel zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research*, **15**, 47–61.

Lee, Y., Nelder, J., and Pawitan Y. (2006) *Generalized Linear Models With Random Effect*. London: Chapman & Hall, p. 178.

Loesbergh, D., Cloes, E., Op de Beeck, L., Rummens, J. L., Vanden Brande, J., Faes, C., Bruckers, L., Molenberghs, G., Dhollander, D., Kellen, E., Hensen, K., Lathouwers, D., Meekers, E., and Buntinx, F. (2007) *Ten Years of Cancer in the Belgian Province of Limburg*. Hasselt, Leuven: Limburgs Kankerregister.

Macro, I. (2008) *Nutrition of Young Children and Women Ethiopia 2005*. Maryland: Macro International.

Matheron, G. (1963) Principles of geostatistics. *Economic Geology*, **58**, 1246–1266.

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

Mellemgaard, A., Engholm, G., McLaughlin, J. K., and Olsen, J. H. (1994) Occupational risk factors for renal-cell carcinoma in Denmark. *Scandinavian Journal of Work, Environment and Health*, **20**, 160–165.

Min, Y. and Agresti, A. (2005) Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1–19.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. New York: Springer.

Molenberghs, G., Verbeke, G., and Demétrio, G. (2007) An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, **13**, 513–531.

Molenberghs, G., Verbeke, G., Demétrio, G., and Vieira, A. (2010) A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical science*, in press.

Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series B*, **135**, 370–384.

Neyens, T., Faes, C., and Molenberghs, G. (2012) A generalized Poisson-gamma model for spatially overdispersed data. *Spatial and Spatio-temporal Epidemiology*, **3**, 185–194.

Neyens, T., Lawson, A. B., Kirby, R. S., and Faes, C. (2015a) The bivariate combined model for spatial data analysis. *Statistics in Medicine*, in revision.

Neyens, T., Faes, C., and Molenberghs, G. (2015b) Hierarchical Bayesian Inference using Integrated Nested Laplace Approximation for the Combined Model: a simulation study. *Computational Statistics and Data Analysis*, submitted.

Neyens, T., Faes, C., and Molenberghs, G. (2015c) The zero-inflated and hurdle combined model for spatial data analysis. *Work in Progress*.

Njagi, E. N., Molenberghs, G., Verbeke, G., Kenward, M. G., Dendale, P., and Willekens, K. (2013) A flexible joint-modelling framework for longitudinal and time-to-event data with overdispersion. *Statistical Methods in Medical Research*, Published online before print July 18, 2013, doi: 10.1177/0962280213495994.

Owen, C. G., Martin, R. M., Whincup, P. H., Smith, G. D., and Cook, D. G. (2005) Effect of infant feeding on the risk of obesity across the life course: a quantitative review of published evidence. *Official Journal of The Americal Academy of Pediatrics*, **115**, 1367–1377.

Paul, M., Riebler, A., Bachmann, L. M., Rue, H., and Held, L. (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Statistics in Medicine*, **29**: 1325–1339.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed Effects Models in S and S-Plus*. New York: Springer-Verlag.

Poisson S.-D. (1937) *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, Précédées des Règles Générales du Calcul de Probabilités*. Paris: Bachelier.

Riebler A., Held L., Rue H. and Bopp, M. (2011) Gender-specific differences and the impact of family integration on time trends in age-stratified Swiss suicide rates. *Journal of the Royal Statistical Society*, **175**, 473–490.

Ripley, B. D. (1987) *Stochastic Simulation*. New York: Wiley.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall-CRC Press.

Rue, H., Martino, S., and Chopin, N. (2009) A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical science*, **25**, 325-347.

Rushworth, A., Lee, D., and Mitchell, R. (2014) A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-temporal Epidemiology*, DOI: 10.1016/j.sste.2014.05.001.

Scheel I., Ferkingstad E., Frigessi A., Haug O., Hinnerichsen M., and Meze-Hausken E. (2013) A Bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society, Series C*, **62**, 85–100.

Schrödle B. and Held L. (2010) Spatio-temporal disease mapping using INLA. *Environmetrics*, **22**, 725-734.

Serra, J. (1980) The Boolean model and random sets. *Computer Graphics and Image Processing*, **12**, 99-126.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, **10**, 257–261.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling*. London: Chapman & Hall/CRC.

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

Spiegelhalter, D. Thomas, A., Best, N., and Lunn, D. (2007) *WinBUGS manual*, Cambridge, UK: MRC Biostatistics Unit. version 1.4.3.

Stern, H. S. and Cressie, N. A. C. (2000) Posterior predictive model checks for disease mapping models *Statistics in Medicine*, **19**, 2377–2397.

Stevens, AS., Pirotte, N., Plusquin, M., Willems, M., Neyens, T., Artois, T., and Smeets, K. (2015) Toxicity profiles and solvent-toxicant interference in the planarian *Schmidtea mediterranea* after dimethylsulfoxide (DMSO) exposure. *Journal of Applied Toxicology*, **35**, 319–326.

Taylor, B. M. and Diggle, P. J. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, **84**: 2266–2284, DOI:10.1080/00949655.2013.788653

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*, New York: Springer.

Verbeke, G. and Molenberghs, G. (2010) Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statistical Modelling*, **10**, 391–419.

Vidal Rodeiro, C. L. and Lawson, A. B. (2004) An evaluation of the edge effects in disease map modelling. *Computational Statistics and Data Analysis*, **49**, 45–62.

Wang, Y., Du, Q., Ren, F., Liang, S., Lin, D. N., Tian, Q., Chen, Y., and Li, J. J. (2014) Spatio-temporal variation and prediction of ischemic heart disease hospitalizations in Shenzhen, China. *International Journal Of Environmental Research And Public Health*, **11**: 4799–4824.

WHO, O. (2009) *World Health Statistics*, Switzerland: WHO Press.

Wolfinger, R. and O'Connell, M. (1993) Generalized linear mixed models. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Wolpert, R. L. and Ickstadt, K. (1998) Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**, 251–267.

Appendix A

Appendix

A.1 Chapter 4

WinBUGS combined model codes for the epilepsy data set.

```
{for(j in 1 : N) {

# Specifying the likelihood:
            Y[j] ~ dnegbin(p[j],alpha)
            p[j] <- alpha/(alpha+mu[j])}
            log(mu[j]) <- ksi0 + b0[ id[j] ] + ksi1*trt[j]
                  + ksi2*studyweek[j]

# Normal random effects:
for (l in 1:89) {
b0[l]  ~ dnorm(0.0, tau.b0)}

# Other priors:
ksi0 ~ dnorm(0.0,0.0)
ksi1  ~ dnorm(0.0,0.001)
ksi2  ~ dnorm(0.0,0.001)
tau.b0 ~ dgamma(1,1.0E-5)
sigma.b0 <- sqrt(1/tau.b0)
        alpha~ dgamma(1,1)}
```

INLA combined model codes for epilepsy data set.

```
formula = Y ~ Trt + Time + f(id,model="iid")

inla(formula,data=data,family="nbinomial",verbose=TRUE,
control.inla = list(strategy="simplified.laplace"),
control.compute=list(dic=TRUE))
```

SAS combined model codes for epilepsy data set.

```
proc nlmixed data=epilepsy qpoints=50;
title 'Poisson-normal Model';
parms ksi0=0.5 ksi1=1 ksi2=0.1 tau.b0=1 alpha=5;
eta = ksi0 + ksi1*trt + b0 + ksi2*studyweek ;
kappa = exp(eta);
beta=alpha**-1;
loglik=lgamma(alpha+Y)-lgamma(alpha)+Y*log(beta)
-(Y+alpha)*log(1+beta*kappa)+Y*eta;
model Y ~ general(loglik);
random b0 ~ normal(0,tau.b0**-1) subject = id;
estimate 'sigma.b0' sqrt(tau.b0**-1);
run;
```

WinBUGS combined model codes for the Flemish contact data set.

```
{for(j in 1 : N) {

# Specifying the likelihood:
                Y[j] ~ dnegbin(p[j],alpha)
    p[j] <- alpha/(alpha+mu[j])
log(mu[j]) <- ksi0 + b1[hh_id[j]] +b2[zipcode[j]] +
ksi1*Sex[j] + ksi2*Time[j]}

# Normal random effects:
        for (l in 1:336) {
b1[l]  ~ dnorm(0.0, tau.b1)}
        for (k in 1:211) {
b2[k]  ~ dnorm(0.0, tau.b2)}

# Other priors:
ksi0 ~ dnorm(0.0,0.0)
ksi1 ~ dnorm(0.0,0.001)
ksi2 ~ dnorm(0.0,0.001)
tau.b1 ~ dgamma(1,5.0E-5)
sigma.b1 <- sqrt(1/tau.b1)
        tau.b2 ~ dgamma(1,5.0E-5)
        sigma.b2 <- sqrt(1/tau.b2)
        alpha ~ dgamma(1,1)}
```

INLA combined model codes for the Flemish contact data set.

```
formula = Y ~ Sex + Time + f(zipid,model="iid") + f(hhid,model="iid")

inla(formula,data=data,family="nbinomial",verbose=TRUE,
control.inla = list(strategy="simplified.laplace"),
control.compute=list(dic=TRUE))
```

INLA combined model codes for the Jimma infant growth study.

```
formula = Y ~ TIME + SEX + RUR + URB + BF + f(ID,model="iid")

inla(formula,data=data,family="betabinomial",verbose=TRUE,
control.inla = list(strategy"simplified.laplace"),
control.compute=list(dic=TRUE))
```

INLA combined model codes for the Jimma longitudinal family survey of youth.

```
formula = Y ~ AGE + URB + SURB + WORK + SEX + ROUND + f(ID,model="iid")

inla(formula,data=data,family="betabinomial",verbose=TRUE,
control.inla = list(strategy = "simplified.laplace"),
control.compute=list(dic=TRUE))
```

A.2 Chapter 5

SAS combined model codes for the Jimma infant growth study.

```
proc nlmixed data = infant noad qpoints = 10;
title 'Combined Model-Jimma infants with const = beta/
alpha';
parms ksi0 = -3.23 ksi1 = 0.0602 ksi2 = 0.0402
ksi3 = -0.8369 ksi4 = -0.552
ksi5 =1.7266 ksi6 = -0.003 ksi7 = -0.0262
ksi8 = -0.0184 ksi9 = -0.1584
sd0 = 1.3662 sd1 = 0.2576 const = 0.0944;
eta = ksi0 + b0+ (ksi1 + b1)
*time + ksi2*sex + ksi3*(place = 1) + ksi4*
(place = 2)
+ksi5*(Bf ) + ksi6*(sex)*time + ksi7*time*
(place = 1) + ksi8*time*(place = 2)
+ ksi9*time*(BF);
kappa = exp(eta);
ll = -log(1 + const) +Y*eta - Y*log
(1 + kappa)
+ (1-Y)*log((1-kappa/(1 + kappa)) + const);
model Y ~ general(ll);
random b0 b1 ~ normal([0,0],[sd0**2,0,sd1**2])
subject = id;
run;
```

WinBUGS combined model codes for the Jimma infant growth study.

```
model { for (i in 1:49112) {

# Specifying the likelihood:
          Y[i] ~ dbern(p[i])
          p[i]<-kappa[i]*theta[i]
          logit(kappa[i])<- xi0 + (b0[ID[i]] + xi1)*TIME[i]
          + xi2*SEX[i] + xi3*RUR[i] + xi4*URB[i] + xi5*BF[i]
          + xi6*SEX[i]*TIME[i] + xi7*RUR[i]*TIME[i]
          + xi8*URB[i]*TIME[i] + xi9*BF[i]*TIME[i] + b1[ID[i]]

# Overdispersion random effects:
          theta[i] ~ dbeta(alpha,beta)}

# Normal random effect:
          for (j in 1:7969) {
          b0[j] ~ dnorm(0.0,tau.b0)
          b1[j] ~ dnorm(0.0,tau.b1)}

# Other priors:
    alpha ~ dunif(3,5)
    beta ~ dunif(1.1,1.5)
    c<-beta/alpha
    xi0 ~ dnorm(0.0,1.0E-6)
    xi1 ~ dnorm(0.0,1.0E-6)
    xi2 ~ dnorm(0.0,1.0E-6)
    xi3 ~ dnorm(0.0,1.0E-6)
    xi4 ~ dnorm(0.0,1.0E-6)
    xi5 ~ dnorm(0.0,1.0E-6)
    xi6 ~ dnorm(0.0,1.0E-6)
    xi7 ~ dnorm(0.0,1.0E-6)
    xi8 ~ dnorm(0.0,1.0E-6)
    xi9 ~ dnorm(0.0,1.0E-6)
    tau.b0 ~ dgamma(0.001,0.001)
    tau.b1 ~ dgamma(0.001,0.001)
    sd0<-sqrt(1/tau.b0)
    sd1<-sqrt(1/tau.b1)}
```

SAS combined model codes for the Jimma longitudinal family survey of youth.

```
proc nlmixed data = ado noad qpoints = 10 ;
title 'Combined Model-Jimma youth with const = beta/
alpha';
parms ksi0 =1.1652 ksi1 = 0.04351 ksi2 = 1.0911
ksi3 = 1.1051
ksi4 = -1.2249 ksi5 = 0.1471 ksi6 = 0.3903
const = 0.05 sd = 0.5;
eta = Beta_0 + ksi1*age + ksi2*(typplace = 1)
+ ksi3*(typplace = 2) + ksi4*currwork + ksi5*sex
+ksi6*round + b0;
kappa = exp(eta);
ll = -log(1 + const) + Y*eta - Y*log
(1 + kappa)
+ (1-Y)*log((1-kappa/(1 + kappa)) + const);
model Y ~ general(ll);
random b0 ~ normal(0,sd*sd) subject = id ;
run;
```

WinBUGS combined model codes for the Jimma longitudinal family survey of youth.

```
{for (i in 1:3815) {

# Specifying the likelihood:
           Y[i] ~ dbern(p[i])
           p[i]<-theta[i]*kappa[i]
           logit(kappa[i])<- xi0 + xi1*AGE[i] + xi2*URB[i]
           + xi3*SURB[i] + xi4*WORK[i] + xi5*SEX[i]
           + xi6*TIME[i] + b0[ID[i]]

# Overdispersion random effects:
            theta[i] ~ dbeta(alpha,beta)}

# Normal random effects:
           for (j in 1:1956) {
           b0[j] ~ dnorm(0,tau.b0)}

# Other priors:
       alpha ~ dunif(110,210)
       beta ~ dunif(1.1,2.2)
       c<-beta/alpha
       xi0 ~ dflat()
       xi1 ~ dnorm(0.0,0.00001)
       xi2 ~ dnorm(0.0,0.00001)
       xi3 ~ dnorm(0.0,0.00001)
       xi4 ~ dnorm(0.0,0.00001)
       xi5 ~ dnorm(0.0,0.00001)
       xi6 ~ dnorm(0.0,0.00001)
       tau.b0 ~ dgamma(1,0.00001)
       sd0<-1/sqrt(tau.b0)}
```

A.3 Chapter 6

WinBUGS combined model codes for the kidney and prostate cancer data sets.

```
 model{ for (i in 1 :N) {

# Specifying the likelihood:
      Y[i]        ~ dpois(mu[i])
      log(mu[i]) <- log(E[i]) + ksi0 + b[i] + log(theta[i])
      RR[i] <- exp(ksi0 + b[i] + log(theta[i]))

# Overdispersion random effect:
  theta[i] ~ dgamma(alpha,alpha)}

# CAR random effects:
b[1:N] ~ car.normal(adj[], weights[], num[], tau.b)
for(k in 1:sumNumNeigh) {
      weights[k] <- 1}

# Other priors:
   ksi0 ~dflat()
   mean <- exp(ksi0)
   tau.b ~dgamma(0.5, 0.0005)
   sigma.b <- 1 / tau.b
   alpha ~ dexp(1)}
```

INLA combined model codes for kidney cancer data set.

```
inla(Y~f(spatialgrid$locationID, model="besag", graph="spatialgrid.adj"
,adjust.for.con.comp = FALSE),
control.inla = list(strategy = "simplified.laplace"),
family="nbinomial", E=kidney$E, data=kidney,
control.fixed=list(compute=TRUE),control.compute=list(config e=TRUE))
```

A.4 Chapter 7

SAS ZICOM model codes for the epilepsy data set.

```
proc nlmixed data=epil.epilepsy qpoints=20;
parms beta0= 0.8511 beta2=-0.01048 beta1=-0.0346 beta3=0.00248
alpha=0.2937 d11=1.0810 rho=0 d22=3.19 gamma0=-1.78 gamma1=0.052;
eta = beta0 + beta1*Trt + beta2*Time + beta3*Time*Trt + b0;
lambda = exp(eta);
eta_prob = gamma0+gamma1*Time+b1;
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*lambda);
if Y=0 then
ll = log(p_0 + (1-p_0)*(p**m));
else ll = log(1-p_0) + log(gamma(m + Y)) - log(gamma(Y + 1))
- log(gamma(m)) + m*log(p) + Y*log(1-p);
random b0 b1 ~ normal([0,0], [d11**2,rho*d11*d22,d22**2]) subject = id;
model Y ~ general(ll);
predict p_0+(1-p_0)*(1/(1+lambda/m))**m out=ZIPNG;
run;
```

SAS HCOM model codes for the epilepsy data set.

```
proc nlmixed data=epil.epilepsy qpoints=20;
title 'HPNG';
parms beta0= 0.8511 beta2=-0.01048 beta1=-0.0346 beta3=0.00248
alpha=0.2937 d11=1.0810 rho=0 d22=3.19 gamma0=-1.78 gamma1=0.052;
eta = beta0 + beta1*Trt + beta2*Time + beta3*Time*Trt + b0;
lambda = exp(eta);
eta_prob = a0+a1*Time+b1 ;
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*lambda);
if Y=0 then ll = log(p_0);
else ll = log(1-p_0) + log(gamma(m + Y)) - log(gamma(Y + 1))
- log(gamma(m)) + Y*log(alpha*lambda)-
(Y+m)*log(1/p)-log(1-(1/p)**(-m));
model Y ~ general(ll);
random b0 b1 ~ normal([0,0],[d11**2,rho*d11*d22,d22**2]) subject=id;
predict p_0 out=HPNG;
run;
```

WinBUGS ZICOM model codes for the epilepsy data set.

```
model{for (i in 1 :N) {

#Zeros trick:
         ze[i]<-0
         ze[i]~dpois(phi[i])
         phi[i]<- -ll[i]+10000

#Count part:
         mu[i]<-theta[i]
         log(theta[i]) <- beta0 + beta1*Trt[i] + beta2*Time[i] +
                          beta3*Trt[i]*Time[i] + b[id[i],0]

#Group membership part:
         logit(q0[i]) <- gamma0 + gamma1*week[i] + b[id[i],1]
         p0[i]<-max(0.001,min(0.999,q0[i]))

#Log-Likelihood:
         zero[i]<-equals(Y[i],0)
         p[i]<-1/(1+theta[i]*alpha)
         ll[i] <- zero[i]*log(p0[i]+ (1-p0[i])*pow(p[i],1/alpha))
                  +(1-zero[i])*(log(1-p0[i])+loggam(1/alpha+Y[i])
                  - loggam(Y[i] + 1) - loggam(1/alpha)
                  + 1/alpha*log(p[i]) +Y[i]*log(1-p[i]))
         pzero[i]<-p0[i]+ (1-p0[i])*pow(p[i],1/alpha)

#M-statistic:
         CPinv[i]<-exp(-ll[i])
         cpo[i]<-pow(CPinv[i],-1)
         Lcpo[i]<-log(cpo[i])}
         MargL<-sum(Lcpo[])

#Predicted zeros:
         predzero<-mean(pzero[])

#Priors:
         gamma0~dnorm(0, 0.1)
         gamma1~dnorm(0, 0.1)
         beta0~dnorm(0, 0.0001)
         beta1~dnorm(0, 0.0001)
         beta2~dnorm(0, 0.0001)
         beta3~dnorm(0, 0.0001)
         alpha<- 1/invalpha
         invalpha ~ dgamma(0.01, 100)
         sig1~dunif(0, 100)
         sig2~dunif(0, 100)
```

```
rho~dunif(-1, 1)
Sigma.B[1, 1] <- pow(sig1, 2)
Sigma.B[2, 2] <- pow(sig2, 2)
Sigma.B[1, 2] <- rho*sig1*sig2
Sigma.B[2, 1] <- Sigma.B[1, 2]
covariance <- Sigma.B[1, 2]
Tau.B[1:2, 1:2] <- inverse(Sigma.B[,])

for (i in 1:89) {
B.hat[i, 1] <- beta0
B.hat[i, 2] <- gamma0
b[i, 1:2]~dmnorm(B.hat[i, ], Tau.B[,])
b1[i] <- b[i, 1] # random intercept 1
b2[i] <- b[i, 2] # random intercept 2}}
```

WinBUGS HCOM model codes for the epilepsy data set.

```
model{for (i in 1 :N) {

#Zeros trick:
         ze[i]<-0
         ze[i]~dpois(phi[i])
         phi[i]<- -ll[i]+10000

#Count part:
         mu[i]<-theta[i]
         log(theta[i]) <- beta0 + beta1*Trt[i] + beta2*Time[i] +
                          beta3*Trt[i]*Time[i] + b[id[i],0]

#Group membership part:
         logit(q0[i]) <- gamma0 + gamma1*week[i] + b[id[i],1]
         p0[i]<-max(0.001,min(0.999,q0[i]))

#Log-Likelihood:
         zero[i]<-equals(Y[i],0)
         p[i]<-1/(1+theta[i]*alpha)
         ll[i] <- zero[i]*log(p0[i])+(1-zero[i])*(log(1-p0[i])
                  + loggam(1/alpha+Y[i]) - loggam(Y[i]+1)
                  - loggam(1/alpha)+Y[i]*log(alpha*theta[i])
                  -(Y[i]+1/alpha)*log(1/p[i])
                  -log(1-pow((1/p[i]),(-1/alpha))))
         pzero[i]<-p0[i]

#M-statistic:
         CPinv[i]<-exp(-ll[i])
         cpo[i]<-pow(CPinv[i],-1)
         Lcpo[i]<-log(cpo[i])}
         MargL<-sum(Lcpo[])

#Predicted zeros:
         predzero<-mean(pzero[])

#Priors:
         gamma0~dnorm(0, 0.1)
         gamma1~dnorm(0, 0.1)
         beta0~dnorm(0, 0.0001)
         beta1~dnorm(0, 0.0001)
         beta2~dnorm(0, 0.0001)
         beta3~dnorm(0, 0.0001)
         alpha<- 1/invalpha
         invalpha ~ dgamma(0.01, 100)
         sig1~dunif(0, 100)
```

```
sig2~dunif(0, 100)
rho~dunif(-1, 1)
Sigma.B[1, 1] <- pow(sig1, 2)
Sigma.B[2, 2] <- pow(sig2, 2)
Sigma.B[1, 2] <- rho*sig1*sig2
Sigma.B[2, 1] <- Sigma.B[1, 2]
covariance <- Sigma.B[1, 2]
Tau.B[1:2, 1:2] <- inverse(Sigma.B[,])

for (i in 1:89) {
B.hat[i, 1] <- beta0
B.hat[i, 2] <- gamma0
b[i, 1:2]~dmnorm(B.hat[i, ], Tau.B[,])
b1[i] <- b[i, 1] # random intercept 1
b2[i] <- b[i, 2] # random intercept 2}}
```

WinBUGS ZICOM model codes for the mesothelioma data set.

```
model{for (i in 1 :N) {

#Zeros trick:
         ze[i]<-0
         ze[i]~dpois(phi[i])
         phi[i]<- -ll[i]+10000

#Count part:
    zero[i] <- equals(Y[i], 0)
    mu[i]<-(EE[i]+0.0001)*theta[i]
    log(theta[i]) <- beta0 + b0[i]

#Group membership part:
        logit(p0[i]) <- gamma0

#Log-Likelihood:
        p[i]<-1/(1+mu[i]*alpha)
        ll[i] <- zero[i]*log(p0[i]+ (1-p0[i])*pow(p[i],1/alpha))
                 +(1-zero[i])*(log(1-p0[i])+loggam(1/alpha+Y[i])
                 - loggam(Y[i] + 1) - loggam(1/alpha)
                 + 1/alpha*log(p[i]) +Y[i]*log(1-p[i]))
        pzero[i]<-p0[i]+ (1-p0[i])*pow(p[i],1/alpha)

#M-statistic:
         CPinv[i]<-exp(-ll[i])
         cpo[i]<-pow(CPinv[i],-1)
         Lcpo[i]<-log(cpo[i])}
         MargL<-sum(Lcpo[])

#Predicted zeros:
         predzero<-mean(pzero[])

#Priors:
b0[1:N] ~ car.normal(adj[], weights[], num[], tau.b0)
for(k in 1:sumNumNeigh) {weights[k] <- 1}
    alpha<- 1/invalpha
        invalpha ~ dgamma(0.01, 100)
        beta0 ~ dnorm(0, 0.0001)
        gamma0 ~ dnorm(0, 0.01)
        tau.b0 ~ dgamma(0.5, 0.0005)  # prior on precision
        sig.b0<-sqrt(1/tau.b0)}}
```

WinBUGS HCOM model codes for the mesothelioma data set.

```
model{for (i in 1 :N) {

#Zeros trick:
         ze[i]<-0
         ze[i]~dpois(phi[i])
         phi[i]<- -ll[i]+10000

#Count part:
    zero[i] <- equals(Y[i], 0)
    mu[i]<-(EE[i]+0.0001)*theta[i]
    log(theta[i]) <- beta0 + b0[i]

#Group membership part:
        logit(p0[i]) <- gamma0

#Log-Likelihood:
        p[i]<-1/(1+mu[i]*alpha)
        ll[i] <- zero[i]*log(p0[i])+(1-zero[i])*(log(1-p0[i])
                 + loggam(1/alpha+Y[i]) - loggam(Y[i]+1)
                 - loggam(1/alpha)+Y[i]*log(alpha*mu[i])
                 -(Y[i]+1/alpha)*log(1/p[i])
                 - log(1-pow((1/p[i]),(-1/alpha))))
        pzero[i]<-p0[i]

#M-statistic:
         CPinv[i]<-exp(-ll[i])
         cpo[i]<-pow(CPinv[i],-1)
         Lcpo[i]<-log(cpo[i])}
         MargL<-sum(Lcpo[])

#Predicted zeros:
         predzero<-mean(pzero[])

#Priors:
b0[1:N] ~ car.normal(adj[], weights[], num[], tau.b0)
for(k in 1:sumNumNeigh) {weights[k] <- 1}
    alpha ~ dexp(1)
        beta0 ~ dnorm(0, 0.0001)
        gamma0 ~ dnorm(0, 0.01)
        tau.b0 ~ dgamma(0.5, 0.0005)  # prior on precision
        sig.b0<-sqrt(1/tau.b0)}}
```

A.5 Chapter 8

WinBUGS combined (Bivariate gamma and MCAR random effects) model codes.

```
model{ for(i in 1:N){
        Y[i,1]<-Y1[i]
        Y[i,2]<-Y2[i]
        E[i,1]<-E1[i]
        E[i,2]<-E2[i]}

for (i in 1:N) {for (k in 1:2) {

# Specifying the likelihood:
           Y[i, k] ~ dpois(mu[i, k])
           log(mu[i, k]) <- log(E[i, k]) + ksi0[k] + b0[k, i]
                            + log(g[i,k])}
           RR1[i] <- exp(ksi0[1] + b0[1,i]+ log(g[i,1]))
           RR2[i] <- exp(ksi0[2] + b0[2,i]+ log(g[i,2]))
           RRprod[i]<-RR1[i] *RR2[i]
           b01[i]<-b0[1,i]
           b02[i]<-b0[2,i]

# Bivariate overdispersion random effects:
           g[i,1] <-var1*(gamma0[i] + gamma1[i])
           g[i,2] <-var2*(gamma0[i] + gamma2[i])
           g1[i]<-g[i,1]
           g2[i]<-g[i,2]
           logg1[i]<-log(g[i,1] )
           logg2[i]<-log(g[i,2] )
           loggamma0[i]<-log(gamma0[i])
           loggamma1[i]<-log(gamma1[i])
           loggamma2[i]<-log(gamma2[i])
           gamma0[i]~dgamma(k0,1)
           gamma1[i]~dgamma(k1,1)
           gamma2[i]~dgamma(k2,1)


# Specifying MSPE:
           ypred1[i]~dpois(mu[i, 1])
           pres1[i]<-Y1[i]-ypred1[i]
           spe1[i]<-pow(pres1[i],2)
           ypred2[i]~dpois(mu[i, 2])
           pres2[i]<-Y2[i]-ypred2[i]
           spe2[i]<-pow(pres2[i],2)}
           mspe1<-mean(spe1[])
           mspe2<-mean(spe2[])
           mspe<-mspe1+mspe2
```

```
# MVN random effects:
           b0[1:2, 1:N] ~ mv.car(adj[], weights[], num[], omega[ , ])
           for (i in 1:sumNumNeigh) {weights[i] <- 1}

# Other priors:
     for (k in 1:2) {
     ksi0[k] ~ dflat()}
     k0~dexp(1)
     k1~dexp(1)
     k2~dexp(1)

# Empirical correlations:
     var1<-1/(k0+k1)
     var2<-1/(k0+k2)
     mu1<-mean(RR1[])
     mu2<-mean(RR2[])
     sd1<-sd(RR1[])
     sd2<-sd(RR2[])
     mu12<-inprod(RR1[],RR2[])/N
     CRR12<-(mu12-mu1*mu2)/(sd1*sd2)

     uh_mu1<-mean(g1[])
     uh_mu2<-mean(g2[])
     uh_sd1<-sd(g1[])
     uh_sd2<-sd(g2[])
     uh_mu12<-inprod(g1[],g2[])/N
     uh_CRR12<-(uh_mu12-uh_mu1*uh_mu2)/(uh_sd1*uh_sd2)

     ch_mu1<-mean(b01[])
     ch_mu2<-mean(b02[])
     ch_sd1<-sd(b01[])
     ch_sd2<-sd(b02[])
     ch_mu12<-inprod(b01[],b02[])/N
     ch_CRR12<-(ch_mu12-ch_mu1*ch_mu2)/(ch_sd1*ch_sd2)}}
```

Table A.1: RR estimates and standard deviations for the kidney cancer data set MCMC results. There were no significant differences from '1'.

	PG		PLN		CAR (UH)		CAR CON		COM	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
RR[1]	1.462	0.347	1.015	0.07546	1.01	0.06021	1.025	0.08737	1.36	0.3076
RR[2]	0.8396	0.2959	0.9931	0.06959	0.992	0.05588	0.9858	0.07748	0.8565	0.2626
RR[3]	1.053	0.1866	1.002	0.06548	1.005	0.05509	1.006	0.07361	1.042	0.1782
RR[4]	0.7842	0.1785	0.9841	0.06736	0.9937	0.0537	0.9788	0.07403	0.7996	0.1697
RR[5]	1.091	0.2919	1.002	0.06913	0.9963	0.05798	0.9986	0.08009	1.059	0.261
RR[6]	0.8817	0.2631	0.9934	0.06836	0.9998	0.05826	0.9928	0.07868	0.8878	0.2385
RR[7]	0.8934	0.2442	0.9933	0.06801	0.9899	0.06145	0.9842	0.0806	0.8971	0.2266
RR[8]	1.256	0.2807	1.009	0.07127	1.003	0.05613	1.013	0.08008	1.203	0.2552
RR[9]	0.9511	0.2359	0.9953	0.06746	0.9904	0.06193	0.987	0.07999	0.9439	0.2183
RR[10]	1.115	0.1579	1.01	0.06497	1.0	0.04993	1.011	0.06918	1.103	0.1521
RR[11]	0.9314	0.2922	0.9952	0.06869	1.004	0.07315	0.9992	0.09057	0.9273	0.2609
RR[12]	0.8888	0.2794	0.9942	0.06892	1.0	0.07065	0.9949	0.08888	0.8946	0.2504
RR[13]	0.9566	0.2867	0.9961	0.06884	1.005	0.06755	1.001	0.08677	0.9506	0.257
RR[14]	1.179	0.2781	1.006	0.06985	1.005	0.07143	1.011	0.09025	1.139	0.2546
RR[15]	1.152	0.1465	1.016	0.06551	1.007	0.05129	1.022	0.06999	1.139	0.1419
RR[16]	1.118	0.299	1.002	0.06919	1.0	0.05653	1.003	0.0783	1.081	0.2663
RR[17]	1.271	0.3511	1.006	0.07171	1.004	0.06269	1.01	0.08547	1.193	0.3088
RR[18]	1.008	0.2791	0.9982	0.06839	1.004	0.06079	1.002	0.08081	0.9942	0.2518
RR[19]	0.9966	0.4502	0.9977	0.07073	0.9945	0.08658	0.9923	0.104	0.9685	0.3599
RR[20]	1.157	0.2171	1.008	0.06815	1.007	0.05556	1.015	0.07628	1.129	0.2031
RR[21]	0.9777	0.2935	0.9969	0.06881	0.9965	0.05697	0.9935	0.07859	0.9644	0.2598
RR[22]	0.9858	0.2094	0.9968	0.06621	0.9997	0.05143	0.9972	0.07197	0.9774	0.1955
RR[23]	0.7269	0.2404	0.9878	0.06947	0.9862	0.06566	0.9753	0.08393	0.761	0.223
RR[24]	0.8112	0.2828	0.9918	0.06865	1.0	0.05419	0.9925	0.07685	0.836	0.2548
RR[25]	1.259	0.2373	1.013	0.07092	1.001	0.0586	1.015	0.08109	1.218	0.2221
RR[26]	1.142	0.2772	1.004	0.069	1.007	0.06546	1.01	0.0845	1.107	0.2521
RR[27]	0.9059	0.1871	0.991	0.06552	0.995	0.06319	0.9872	0.07885	0.906	0.1794
RR[28]	0.8961	0.2457	0.9936	0.06844	1.003	0.05664	0.9964	0.07747	0.9004	0.2275
RR[29]	0.6875	0.1825	0.9808	0.06917	0.9868	0.05781	0.9688	0.07792	0.7166	0.1762
RR[30]	0.9678	0.1862	0.9954	0.06536	0.9947	0.0553	0.991	0.07349	0.9632	0.1766
RR[31]	1.184	0.3063	1.005	0.07032	0.9958	0.05542	1.001	0.07913	1.131	0.2712
RR[32]	1.176	0.2771	1.005	0.06968	0.9998	0.05993	1.007	0.08111	1.137	0.252
RR[33]	0.8447	0.2958	0.9932	0.0699	1.005	0.06017	0.9978	0.08081	0.8628	0.2642
RR[34]	0.8775	0.2897	0.9942	0.06948	0.9937	0.05582	0.9885	0.0778	0.8862	0.2585
RR[35]	0.7285	0.2285	0.9866	0.06932	0.9938	0.06132	0.9817	0.08091	0.7647	0.2172
RR[36]	0.9868	0.2616	0.997	0.06818	0.9973	0.05552	0.9957	0.07726	0.9737	0.2389
RR[37]	0.9148	0.2348	0.9936	0.06713	0.9944	0.0589	0.9889	0.07774	0.9151	0.2164
RR[38]	1.095	0.1832	1.005	0.06536	1.006	0.05757	1.011	0.07536	1.079	0.1732
RR[39]	1.057	0.2564	0.9999	0.06792	1.007	0.07218	1.008	0.0893	1.036	0.2347
RR[40]	0.6996	0.1602	0.977	0.0695	0.992	0.05523	0.9702	0.0758	0.7216	0.1542
RR[41]	0.8604	0.3244	0.9943	0.06988	0.9982	0.04339	0.9928	0.07002	0.8745	0.2788
RR[42]	0.9456	0.3118	0.9961	0.0691	1.004	0.05808	0.9999	0.08027	0.9414	0.2756
RR[43]	0.9903	0.2395	0.9974	0.06741	1.003	0.05616	1.001	0.07674	0.9812	0.2199
RR[44]	0.6664	0.271	0.9883	0.0701	0.9934	0.05636	0.9825	0.07823	0.7254	0.2455
Average	0.9857	0.2578	0.9977	0.0686	0.9991	0.0595	0.9973	0.0799	0.9760	0.2339

Table A.2: RR estimates and standard deviations for the prostate cancer data set MCMC results. Significant differences from '1' are denoted by *.

	PG		PLN		CAR (UH)		CAR CON		COM	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
RR[1]	1.107	0.1163	1.087	0.1081	1.12	0.09665	1.09	0.1074	1.103	0.1144
RR[2]	0.9621*	0.1349	0.9646*	0.1194	0.9357*	0.09229	0.9601*	0.1178	0.9609*	0.1313
RR[3]	1.175*	0.06626	1.166*	0.06448	1.147*	0.06336	1.165*	0.06428	1.174*	0.06631
RR[4]	0.7918	0.06158	0.8047	0.05946	0.7853	0.05331	0.8025	0.05914	0.7936	0.06123
RR[5]	1.03	0.109	1.022	0.1012	1.031	0.08753	1.023	0.1007	1.028	0.1078
RR[6]	0.8692	0.09884	0.8844	0.09186	0.9187	0.08332	0.887	0.0917	0.8716	0.09771
RR[7]	1.164	0.1043	1.142	0.09769	1.142	0.09207	1.143	0.0971	1.159	0.1022
RR[8]	1.243*	0.105	1.215*	0.09963	1.14	0.09037	1.208*	0.09947	1.235*	0.1034
RR[9]	1.157	0.09543	1.139	0.09074	1.109	0.08834	1.135	0.09065	1.152	0.09523
RR[10]	0.8228*	0.04409	0.8283*	0.04334	0.8412*	0.04326	0.8289*	0.04327	0.8236*	0.04417
RR[11]	1.243	0.1334	1.201	0.1228	1.214	0.1228	1.203	0.1233	1.233	0.1315
RR[12]	1.022	0.1147	1.014	0.1059	1.06	0.1087	1.021	0.1069	1.02	0.1134
RR[13]	0.9557	0.1119	0.9582	0.1024	0.9601	0.097	0.9589	0.1025	0.9556	0.11
RR[14]	1.082	0.09548	1.069	0.09046	1.069	0.09094	1.071	0.09043	1.079	0.09473
RR[15]	1.119*	0.04581	1.115*	0.04534	1.115*	0.04759	1.115*	0.04515	1.118*	0.04621
RR[16]	1.025	0.1115	1.017	0.1023	1.013	0.08681	1.017	0.1018	1.023	0.1094
RR[17]	1.017	0.1269	1.008	0.114	0.9952	0.1002	1.008	0.1132	1.015	0.1235
RR[18]	1.104	0.1108	1.085	0.1033	1.117	0.097	1.089	0.1032	1.099	0.1098
RR[19]	1.065	0.4091	1.018	0.2361	0.7604	0.2375	0.9847	0.2469	1.027	0.3461
RR[20]	1.291*	0.07958	1.273*	0.07764	1.262*	0.07586	1.272*	0.07726	1.287*	0.07974
RR[21]	0.7395*	0.1012	0.7839*	0.09423	0.7821*	0.07856	0.7815*	0.09283	0.747*	0.1004
RR[22]	1.015	0.07552	1.012	0.07243	1.016	0.06556	1.013	0.07204	1.015	0.07486
RR[23]	1.121	0.1168	1.099	0.1084	1.119	0.1035	1.102	0.108	1.116	0.1154
RR[24]	1.327*	0.156	1.262*	0.1415	1.136	0.1065	1.251	0.1404	1.309*	0.1515
RR[25]	0.8289*	0.06685	0.8399*	0.06434	0.7963*	0.06005	0.8354*	0.06446	0.8301*	0.06662
RR[26]	1.035	0.0954	1.027	0.08978	1.032	0.08692	1.028	0.08973	1.033	0.09471
RR[27]	1.05	0.06943	1.044	0.06739	1.035	0.06757	1.044	0.06739	1.049	0.06902
RR[28]	1.436*	0.1154	1.391*	0.1107	1.359*	0.102	1.389*	0.1097	1.424*	0.1141
RR[29]	1.21*	0.0847	1.192*	0.08171	1.17*	0.07662	1.191*	0.08141	1.205*	0.08449
RR[30]	0.6972*	0.0538	0.7164*	0.05246	0.7285*	0.05089	0.7164*	0.05235	0.6996*	0.05361
RR[31]	0.9731	0.1097	0.9717	0.1009	1.004	0.08663	0.9755	0.1001	0.9721	0.1077
RR[32]	1.004	0.09459	0.9996	0.08861	0.9953	0.08057	1.001	0.08824	1.003	0.09324
RR[33]	1.033	0.1375	1.021	0.1227	1.073	0.1113	1.027	0.1227	1.03	0.1345
RR[34]	1.124	0.1402	1.094	0.126	1.059	0.1036	1.091	0.1245	1.116	0.1368
RR[35]	0.8587	0.09339	0.8737	0.08692	0.8926	0.08104	0.8764	0.08705	0.8613	0.09198
RR[36]	0.9035	0.09575	0.9122	0.08904	0.9328	0.0773	0.9147	0.08834	0.9038	0.09405
RR[37]	0.6138*	0.06883	0.6595*	0.06725	0.6486*	0.06025	0.6571*	0.06683	0.6209*	0.06865
RR[38]	1.168*	0.06145	1.159*	0.06047	1.151*	0.06065	1.159*	0.06026	1.166*	0.0619
RR[39]	0.7978*	0.0807	0.8182*	0.07653	0.827*	0.07747	0.8196*	0.07694	0.801*	0.08036
RR[40]	0.6565*	0.051	0.6785*	0.05045	0.7003*	0.04764	0.6794*	0.05032	0.6595*	0.0513
RR[41]	0.2418*	0.07968	0.4965*	0.09751	0.9866	0.01644	0.53*	0.1185	0.2864*	0.08607
RR[42]	1.336*	0.1579	1.268*	0.1441	1.229*	0.1203	1.264*	0.1426	1.317*	0.1544
RR[43]	1.028	0.08917	1.021	0.08457	1.029	0.07958	1.023	0.08418	1.026	0.08863
RR[44]	0.7104*	0.1238	0.7816	0.1111	0.7583*	0.08461	0.7761	0.1091	0.723*	0.1219
Average	10.035	0.1044	10.030	0.0947	10.045	0.0850	10.029	0.0950	10.016	0.1019

Samenvatting

In zijn boek over spatiale statistische analyse dat nog steeds tot de belangrijkste werken behoort binnen deze wetenschappelijke niche, definieert Cressie statistiek als de wetenschap van de onzekerheid (Cressie, 1993) en geeft daarmee een allesomvattende beschrijving aan iets dat voor velen een zwarte doos blijft binnen wetenschappelijk onderzoek. Statistiek tracht variatie in observaties te verklaren aan de hand van gezamenlijke kenmerken en maakt daarvoor gebruik van kanstheorie. Statistisch modelleren gaat een stap verder door de realiteit weer te geven in een wiskundige formule die een observatie verbindt met één of meerdere factoren en daarmee probeert alle geobserveerde variatie te verklaren. In theorie zou, wanneer je alle verklarende variabelen vooraf kent, alle variatie in de zogenaamde uitkomstvariabele kunnen worden uitgelegd a.d.h.v. deze variabelen, maar dit is zelden het geval. Daarbij komt dat statistisch modelleren meestal gebeurt binnen de context van een wiskundige verdeling die specifiek is voor het type data dat de uitkomstvariabele betreft. Die verdelingen leggen typisch ook bepaalde beperkingen op aan de data, zoals wat betreft de mate van variatie die de uitkomstvariabele mag vertonen. In de praktijk zien we dat aan deze variatieassumptie in heel wat gevallen niet wordt voldaan, wat statistische analyses via het desbetreffende type model en de onderliggende verdeling onbetrouwbaar maakt. De oorzaak hiervan is niet altijd duidelijk, maar heeft meestal te maken met twee belangrijke problemen: (1) er zit een structuur in de data, m.a.w. de data zijn niet willekeurig verzameld, of (2) er zijn onbekende verklarende factoren die een invloed hebben op de uitkomstvariabele.

In deze verhandeling werd een oplossing gezocht voor scenario's waarin een te grote mate van variabiliteit voorkomt en waarbij dus de assumptie omtrent variatie wordt geschonden. Er werd specifiek gekeken naar gegroepeerde binaire data en data verkregen door tellingen, waarbij gegroepeerde binaire data via de binomiale verdeling worden

geanalyseerd, terwijl bij tellingen de Poisson-verdeling typisch wordt gebruikt. Centraal in dit onderzoek stond het zogenaamde *combined model* (Molenberghs et al., 2010). Dit is een model dat ontwikkeld werd voor meerdere datatypes en dat letterlijk twee soorten modellen, nl. overdispersiemodellen en GLMM's, combineert: (1) een overdispersiemodel tracht via een dispersieparameter met een specifieke achterliggende (conjugate) verdeling extravariatie toe te laten, terwijl (2) een GLMM typisch een normaal verdeeld random effect gebruikt om structurele aspecten in de data in rekening te brengen. Het combined model voegt op een slimme manier beide modellen samen tot één model met enerzijds een conjugaat random effect en anderzijds één of meerdere normaal verdeelde random effecten. Op deze manier kan extravariatie, veroorzaakt door respectievelijk onbekende verklarende variabelen en structuur in de data, toegelaten worden.

Binnen de context van het combined model keek ik in eerste instantie naar het gebruik ervan binnen verschillende statistische schattingsmethodes. Het combined model werd namelijk ontwikkeld binnen de likelihood-methodologie. De Bayesiaanse schattingsmethode daarentegen heeft als groot voordeel in staat te zijn complexere datastructuren te analyseren. In Hoofdstuk 4 en 5 toonde ik aan dat parameterschattingen nagenoeg hetzelfde zijn wanneer in de likelihood- of in de Bayesiaanse setting wordt gewerkt. Echter, een belangrijk nadeel van de Bayesiaanse statistiek is dat de meest gebruikte methode, MCMC, gebaseerd is op Gibbs sampling, een methode die gebruik maakt van simulaties die erg lang kan duren. Enkele snelle methodes werden ontwikkeld, maar zijn door hun benaderende aard meestal minder goed wat betreft de kwaliteit van parameterschatten. Ik onderzocht in Hoofdstuk 5 (en deels in Hoofdstuk 6) ook de werking van zo'n methode, met name INLA, maar besloot dat vooral parameterschattingen van de variantieschatters voor de random effecten matig tot slecht waren.

Verder deed ik onderzoek naar het gebruik van het combined model in een aantal settings. Zo onderzocht ik zijn toegevoegde waarde in longitudinale data analyse voor binaire data (Hoofdstuk 4) en vergeleek ik de resultaten met resultaten verkregen uit de traditionele modellen: het combined model deed het in de meeste gevallen beter. Dit kan verklaard worden door het feit dat het combined model een goed onderscheid kan maken tussen extravariatie komende van structuur in de data enerzijds en niet-gestructureerde variabiliteit anderzijds. Dit heeft als gevolg dat de effecten van de verklarende variabelen ook correcter worden geschat. Wanneer ik het combined model toepaste op teldata, was dit meestal binnen de spatiale setting, m.a.w. de setting waarin de datastructuur plaatsgebonden is. In Hoofdstuk 6 werkte ik het combined model uit voor spatiale teldata en vergeleek deze met traditionele modellen binnen die niche. Een belangrijke conclusie

was dat het combined model erg goede resultaten opleverde, vooral wanneer er veel ongestructureerde variabiliteit aanwezig was. Juist in deze situatie gaf het combined model betere resultaten dan het populaire en slechts licht verschillende CAR-convolutiemodel.

Naast het voorkomen van onbekende verklarende factoren en een datastructuur, kan ook het disproportioneel voorkomen van nul-observaties in teldata leiden tot schendingen van de variabiliteitassumpties. Daarom onderzocht ik extensies van het combined model in deze setting voor longitudinale en spatiale data. Vermits deze modellen een grote hoeveelheid informatie trachten te destilleren uit data die dat niet altijd hebben, bleken deze modellen erg goed te werken bij de analyse van grote datasets. Wanneer datasets kleiner werden, onstonden er problemen tijdens het schattingsproces. Een interesante onderzoekspiste is daarom ook het kwantificeren van de grens in termen van steekproefgroottes waarop een model voor extra nullen nuttig wordt.

Ook onderzocht ik in Hoofdstuk 8 de mogelijkheid tot uitbreiding van het combined model in de spatiale setting wanneer men twee i.p.v. één uitkomstvariabele wil modelleren, waarbij een bivariate gamma-verdeelde dispersiefactor kan worden gebruikt. Hoewel de complexiteit van deze analyses soms computationele problemen met zich meebrengt, geeft dit zogenaamde bivariate combined model erg goede resultaten die in veel gevallen superieur zijn t.o.v. de bestaande methodes. Het is ook in deze richting dat veel wetenschappelijke mogelijkheden liggen: het simultaan analyseren van meerdere uitkomstvariabelen wanneer rekening wordt gehouden met zowel een spatiale als een longitudinale structuur. Een uitbreiding van het combined model naar deze setting lijkt logisch vermits het via de vooruitgang in computationele technieken mogelijk wordt om erg complexe datastructuren te analyseren.