Master's thesis

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

Marte Marie Saghagen

A Comparison of Model-Based and Design-Based Methods for Spatial Modelling Using Complex Survey Data

A Case Study for Neonatal Mortality in Kenya

Master's thesis in Applied Physics and Mathematics
Supervisor: Geir-Arne Fuglstad
June 2019

NTNU
Norwegian University of Science and Technology

Marte Marie Saghagen

A Comparison of Model-Based and Design-Based Methods for Spatial Modelling Using Complex Survey Data

A Case Study for Neonatal Mortality in Kenya

Master's thesis in Applied Physics and Mathematics
Supervisor: Geir-Arne Fuglstad
June 2019

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

Abstract

In recent years, considerable progress has been made in the estimation of subnational child mortality rates with sparse survey data. The widely used design-based methods are easy to implement and produce estimates that are consistent with the survey design. However, design-based methods require sufficiently large sample sizes to obtain accurate estimates in each geographic unit of interest. Design-based methods are therefore not applicable in areas where there is not enough sampled data. Also, estimation of rates on coarse scales may hide fine-scale trends. Alternative methods are necessary for fine-scale space-time estimation, such as model-based methods.

This thesis provides a comparison of design-based and model-based approaches to the estimation of neonatal mortality rates (NMR) at the county level in Kenya for 2009-2014 with complex survey data. The dataset used is a demographic and health survey (DHS) conducted in Kenya in 2014. The design-based approach, a spatially smoothing design-based approach and model-based approaches are tested. In addition, a new approach combining design-based and model-based ideas is explored. Design-based rates are estimated with the package survey in R. The other methods are formulated as Bayesian hierarchical models, and inference is conducted using integrated nested Laplace approximations (INLA) with the package R−INLA in R. The methods are evaluated on a set of scoring rules and on computational time through a simulation study and on real survey data.

The model-based methods were the best performing methods in the simulation study. The results of the methods used on real survey data show that model-based approaches are superior at handling more sparse data. In addition, model-based methods are able to obtain estimates on finer spatial scales, where design-based methods are not applicable.

Sammendrag

I de siste årene har det blitt gjore store framskritt i estimering av barnedødelighet på subnasjonalt nivå med survey data. De mest brukte metodene er design-baserte, som er enkle å implementere og produserer estimater som er konsistente med surveydesignet. Problemet med design-baserte metoder er at de krever tilstrekkelig store samples for å oppnå nøyaktige estimater i hver geografiske enhet av interesse. Design-baserte metoder er derfor ikke andvendbare i områder hvor det ikke er nok innsamlet data. I tillegg kan estimering av dødelighet på grove skalaer i tid og rom gjemme trender på finere skala. Derfor er det nødvendig med alternative metoder for estimering på fin skala i tid og rom.

Denne masteroppgaven gir en sammenligning av design-baserte og modell-baserte fremgangsmåter for estimering av neonatale dødelighetsrater på fylkenivå i Kenya i perioden 2009-2014 med komplekse survey data. Datasettet som er brukt er en demographic and health survey (DHS) utført i Kenya i 2014. En design-basert metode, en romlig glattende design-basert metode og modell-baserte fremgangsmåter er testet. I tillegg er en ny foreslått tilnærming utforsket, som kombinerer design-baserte og modell-baserte ideer. Design-baserte rater er estimert med pakken survey i R. De andre metodene er formulert som Bayesianske hierarkiske modeller og inferens er utført med integrated nested Laplace approximations (INLA) med pakken R−INLA i R. Metodene er evaluert ved hjelp av et sett av mål og på gjennomsnittlig kjøretid i et simuleringsstudie og på ekte survey data.

De modell-baserte metodene var metodene som oppnådde de beste resultatene i simuleringsstudiet. Resultatene av metodene brukt på ekte survey data viser at modell-baserte metoder er bedre enn design-baserte metoder på å håndtere lite data. I tillegg er det også mulig å oppnå estimater på finere romlig skala der design-baserte metoder ikke kan brukes.

Preface

This master's thesis was written during the last semester of my Master of Technology degree at the Norwegian University of Science and Technology (NTNU). This final assignment marks the end of the five-year study programme "Applied Physics and Mathematics", with specialization in "Industrial Mathematics".

Working on this thesis has been exciting and rewarding, and has given me both practical and theoretical challenges. It has allowed me to utilize the knowledge I have gained during the last five years and develop a deeper understanding of several fields in statistics. The code used for the simulations and the generation of the results is written in R. I used the package survey (Lumley (2010)) for obtaining design-based rates and the package INLA (Rue et al. (2009)) for Bayesian inference.

I want to give special thanks to my supervisor Geir-Arne Fuglstad for his help and guidance. I would also like to thank my family for all the support, and my friends for five incredible years at NTNU.

Marte Marie Saghagen
Trondheim, June 2019

Table of Contents

Chapter 1

Introduction

In the year 2000, the United Nations Millennium Declaration was adopted by world leaders (UN (2019b)). The nations committed to eight Millennium Development Goals (MDGs) which contain measurable targets to reduce poverty, hunger and the spread of diseases such as HIV/AIDS and malaria. The post-2015 Development Agenda was a UN-led process to define the new goals that would succeed the MDGs. The new set of goals are called the Sustainable Development Goals (SDGs) consisting of 17 goals building on the MDGs. Two important indicators of the health of a nation are the under-5 mortality rate (U5MR) and the neonatal mortality rate (NMR). U5MR is defined in UNICEF (2018b) as "the probability of dying between birth and exactly 5 years of age, expressed per 1000 live births". NMR is defined in UNICEF (2018a) as "the probability of dying during the first 28 days of life, expressed per 1000 live births". Goal 3.2 of the SDGs is "by 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births" (UN (2018)).

Estimating mortality rates is complicated. Many of the countries that are falling behind when it comes to improving mortality rates are developing countries where vital registration systems are often limited. When there does not exist a complete system for registrations of every birth and death in the country, it is not possible to obtain exact mortality rates. The alternative is to estimate mortality rates based on data arising from surveys and censuses.

This thesis will focus on subnational estimation of NMRs while accounting for complex survey designs. Quantifying the NMR is important for the countries when interventions are to be done. As stated by UNICEF (2018a), "children face the highest risk of dying in their first month of life", making accurate estimation of the NMR an important field of research. Also, stated by UNICEF (2018a), "a child born in sub-Saharan Africa or in South Asia is nine times more likely to die in the first month than a child born in a high-income country". It is not only essential to obtain accurate national estimates of the NMR, but also sub-nationally. Information about the subnational variation can make it easier for the countries to allocate resources in a productive way.

The structure of this thesis is as follows. Chapter 1 outlines the current practice within the field, the scope of this thesis and the application. In Chapter 2, the relevant background theory will be explained. The methods to be used in this thesis are thoroughly described in Chapter 3. Further, Chapter 4 presents a simulation study to assess the performance of the methods. Chapter 5 presents the results of the methods applied to a real survey conducted in Kenya in 2014. Finally, findings are discussed in Chapter 6.

1.1 Current practice

When the commitment was made to realize the 2030 Agenda for Sustainable Development (SDGs), the member states recognized that the targets of the SDGs should be "met for all nations and people, and all segments of society, leaving no one behind" (UN (2019a)). However, the disaggregated data needed to address subgroups of the population is sparse. For the countries, it is essential to monitor progress towards the SDGs on the geographical level where policies are decided and interventions can be made. Typically, this is not on the national level, but on subnational levels such as counties or constituencies. Methods that can reveal subnational variation and create accurate estimates for small areas or subgroups are therefore a necessity.

The UN Inter-Agency Group for Child Mortality Estimation (IGME (2019)) is responsible for developing methods for child mortality estimation and reporting the progress towards child survival goals. The method used by IGME for estimating child mortality globally is the Bayesian B-spline bias-reduction model proposed by Alkema and New (2014). The model produces national estimates and cannot produce estimates at subnational levels.

Great progress has been made in producing subnational estimates of child mortality. Mercer et al. (2015) proposed a discrete space-time smoothing model that accounts for the survey design to obtain subnational estimates of the U5MR. The model uses the design-based approach where the goal of the analysis is to estimate features of the fixed population, not generalizing the findings to other populations. After that, the estimates are smoothed in space and time. However, design-based methods require sufficiently large sample sizes to obtain estimates in each geographic unit of interest. This method will therefore not be applicable in areas where there is not enough sampled data. This type of approach is still highly relevant and recently used by Zehang et al. (2019) to estimate the U5MR for subnational areas in 35 countries in Africa. However, they do not produce other mortality indicators such as mortality in the neonatal period, which is the period of life where children face the highest risk of dying. Also, they use a coarse scale such as provinces in Kenya, which may hide fine-scale trends.

In addition to discrete approaches to estimation with survey data, continuous models have also been developed. Golding et al. (2017) use a Bayesian model-based approach that is continuous in space and yearly in time to produce estimates of U5MR and NMR at a resolution of 5×5 km grid cells for 46 African countries. The results were also aggregated to obtain estimates on national, and subnational administrative levels 1 and 2. There are several problems with this approach. It is computationally expensive to estimate in fine scale for both space and time. Also, it is not straightforward how to aggregate the grid cells up to subnational and national level when the survey design has to be accounted for.

Spatial and spatio-temporal modelling of complex survey data differs from standard spatial or spatio-temporal statistics. Data collected through surveys is correlated and dependent which present additional challenges when it comes to modelling. When working with survey data it is necessary to account for the survey design, which includes stratification and clustering. Stratification ensures coverage of all desirable subgroups, such as the inclusion of residents from each subnational area. Clustering saves time and resources by only sampling from restricted areas, called clusters, yielding correlated observations. When using standard spatio-temporal models, stratification and clustering that introduce dependence between the observations are not accounted for. The assumption of independence between the locations and what is measured do not hold for survey data. In addition, standard models do not account for the clustering of the observations resulting in variance estimates that are too low.

1.2 Approach

This thesis will estimate subnational NMR with complex survey data. Survey data is an important data source in the process of assessing the progress towards the SDGs. In much of the developing world, vital registration is often absent or deficient. However, information about the populations' health can be obtained through surveys such as the Demographic and Health surveys conducted by the DHS program. The DHS surveys are used to asses progress towards the SDGs and collect important population characteristics. The surveys are conducted in over 90 countries every fifth year such that the results of the surveys can be compared over time, and the survey design is chosen in such a way that the estimates are representative for the full population. The DHS surveys collect the complete reproductive history of interviewed women, including the date of birth of each child, if the child is deceased and the date of death.

The first approach explored is design-based, where county NMRs are directly estimated for the fixed population based on the survey design. The second approach is a smoothed design-based approach greatly inspired by the methods outlined in Mercer et al. (2015). The design-based estimates with uncertainty, are spatially smoothed, in the hope of obtaining more accurate estimates with less variance. The third approach is model-based, where a binomial distribution is specified to the random process that generates the occurrences of neonatal deaths. The approach incorporates both spatially smoothing effects, a fixed effect for the stratification and unstructured effects for the levels of clustering in the survey design. Lastly, a new approach that combines the smoothed design-based method and the model-based method is explored. The approach is a joint model with two likelihoods. With this approach, the smoothed design-based approach is used in counties where there are more than zero observed neonatal deaths, and the model-based approach is used in counties where there is not enough data to produce design-based estimates. Combining the two approaches makes it possible to follow design-based ideas in areas where this is possible and take advantage of modelling in the remaining areas. Two levels of complexity are considered for the model-based approach and the combined approach. First, unstructured effects at the first level of clustering and second, unstructured effects at both levels of clustering.

The design-based approach is the benchmark when it comes to estimating health indi-

cators from surveys as the design of the survey is fully accounted for. However, design-based methods require large sample sizes from each subgroup one wishes to obtain estimates for. Each new cluster that is added to a survey costs time and money, and the budget is limited. When using design-based approaches, it is not possible to monitor the progress of small areas. It is simply not doable to conduct surveys large enough to produce yearly estimates at the administrative levels where policies and interventions are decided. Methods that spatio-temporally smooth the design-based estimates can reduce the variance of the design-based estimates by borrowing strength in space and time. However, smoothing design-based methods cannot be used on smaller areas or with sparse data.

An alternative to the descriptive, design-based approach is modelling the problem with a more traditional model-based approach. Modelling makes it possible to obtain estimates at arbitrarily small scales. In addition, model-based approaches make it possible to obtain information about underlying structures in the data, such as how large the spatial variation is, the significance of urban or rural areas, and intra-cluster correlation. However, several questions arise when using model-based approaches. Where design-based estimates are asymptotically unbiased, model-based estimates may produce bias. Fine-scale information about the population is missing, such as the location of every child born. Therefore, ad-hoc procedures must be used to create estimates for finer spatial units. However, model-based approaches can produce more accurate estimates and less total variance. Since resources are limited, model-based methods are favorable if they are sufficiently accurate.

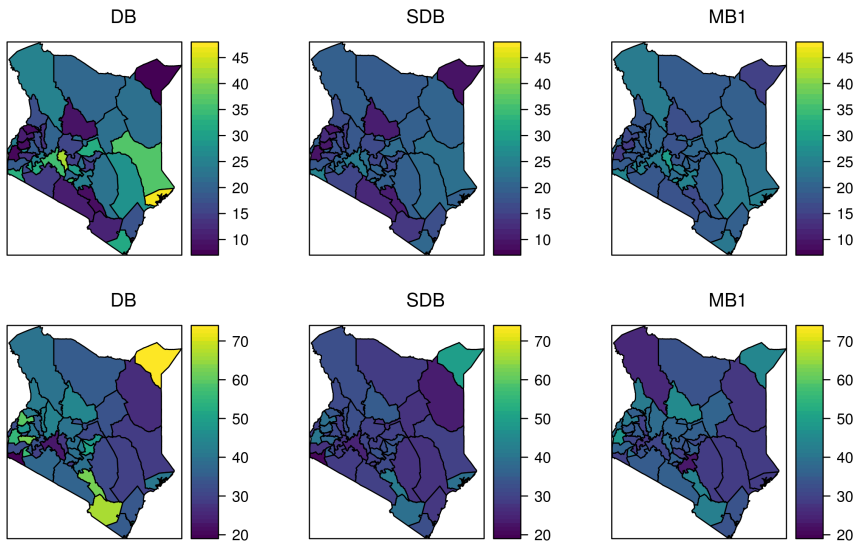


Figure 1.1: Maps of Kenya with estimated county NMRs and relative standard deviation (RSD). The results are presented as maps with estimated county NMRs, r_i, per 1000 live births for counties $i = 1, ..., 47$ in the first row and the relative standard deviation (RSD) $\frac{\sigma_i}{r_i} \cdot 100$ in the second row, where r_i and σ_i is the estimated rate and estimated standard deviation of the rate in county i respectively. First column: Design-based (DB). Second column: Smoothed design-based (SDB). Third column: Model-based 1 (MB1).

Design-based estimates may have high variance and considerable variability, as shown in the first column of Figure 1.1. Since the estimates are calculated directly as rates between the number of neonatal deaths and births, it is not possible to obtain any information about possible structures in the data. Therefore, one cannot know if the noise in the design-based estimates is random or structured variability between the areas. The SDB method and the MB1 method may considerably reduce variance and noise as shown in the second and thirds column of Figure 1.1. However, it is difficult to decide which of the three methods that provide the most correct estimates of the NMR. The DB method is mostly used and accepted within the field of survey statistics. In addition, design-based estimates are asymptotically unbiased. If the design-based approach provides estimates with low variance it is the preferred method. Since the SDB approach is based on the design-based estimates there is no reason for not to trust the approach if the design-based estimates have low variance. However, the DB method cannot be used on finer scales as this requires larger sample sizes. On the contrary, model-based approaches will always be applicable and provide estimates with low variances. The downside to the MB1 method is that it is not based on survey theory and one has to trust that the chosen model is correct.

When estimating mortality rates at subnational level with the design-based method, there is no assumption of any spatial relationship between the areas. However, results by among others Mercer et al. (2015) and Zehang et al. (2019) show that including spatial- and temporal dependency produce more precise estimates. In this thesis, spatially smoothing effects are incorporated with the commonly used spatial model by Besag, York and Mollié (BYM) outlined in Besag et al. (1991). In the BYM model one assumes that the spatial effect in one area is dependent on the neighboring areas. The spatial effect is divided into a structured and an unstructured part, making it possible to find out how much of the spatial variation that is systematic. Continuous spatial models can also be used. However, it is not necessary for comparing the different proposed approaches in this thesis.

As the methods explored in this thesis contain several model components, such as spatial effects and unstructured cluster effects, hierarchical Bayesian models are appropriate. Markov chain Monte Carlo (MCMC) methods are common approaches to estimate posterior densities from Bayesian models. MCMC can obtain arbitrarily accurate results, depending on the number of steps made. The software stan (Carpenter et al. (2017)) performs full Bayesian statistical inference with MCMC sampling. Stan is general and can be used to fit the models in this thesis, however, this approach is slow. Another approach, which is beneficial computationally, is used instead. The approach is called integrated nested Laplace approximations (INLA, Rue et al. (2009)). INLA can be used on a subclass of Bayesian hierarchical models called latent Gaussian models (LGMs). For LGMs, Gaussian distributions are assigned to the latent field which is not directly observed but inferred from the observed data. The models that are developed in this thesis fit within the class of LGMs. The INLA approach is implemented in the package R−INLA, which can be accessed at `http://www.rinla.org` (Rue (2019)).

The methods are assessed on their predictive performance through a simulation study with different cases of spatial dependence: no spatial effect, unstructured effect and structured effect. Within each case, several scenarios with different fixed and random effects are carried out such that the methods are tested on a wide range of situations. The methods are evaluated on mean-absolute error (MAE), root-mean-squared error (RMSE), continu-

ous ranked probability score (CPRS, Gneiting and Raftery (2007)), mean bias error (MBE) and time-complexity.

1.3 Neonatal mortality in Kenya

The methods are applied to the case of estimating NMRs in the counties of Kenya. Kenya is used as an example as this is a country where health indicators are estimated through surveys. Even though Kenya is chosen to demonstrate the methods used, the methods are also applicable to other low- and medium-income countries. Also, one would expect that the methods can be used on other health indicators, such as the U5MR. A DHS conducted in Kenya in 2014 is used. This particular survey is used as this is the most newly conducted DHS in Kenya and can produce reliable estimates not only on national and regional levels, as previous surveys, but also at the county level because of the increased number of sampled clusters.

Kenya is situated in East Africa, bordering Ethiopia (north), Somalia (northeast), Tanzania (south), Uganda (west), and South Sudan (northwest). A new constitution was approved in 2010, and consequently, 47 counties emerged in 2013. The country had previously been dived into eight provinces subdivided into 46 districts. A map of Kenya is presented in Figure 1.2. The size of Kenya is 582,646 square kilometers, of which 98% is dry land area. In 2009, a census was conducted to enumerate Kenya's population (KNBS (2012)). The census enumerated the population of Kenya to 38.6 million, of which 26.1 million lived in rural areas and 12.5 million lived in urban areas.

Figure 1.2: Map of Kenya and its 47 counties, taken from Kenya National Bureau of Statistics et al. (2015).

The survey consists of an extensive questionnaire reporting among others interviewed women's reproductive history, sanitary conditions and other health indicators. For more information about the DHS program, see DHS (2019). As outlined in Kenya National Bureau of Statistics et al. (2015), the sample in the 2014 Kenya DHS was drawn from a master sampling frame based on information obtained in the 2009 Kenya Population and Housing Census. The 2009 census established 96251 enumeration areas (EAs). Each EA is solely categorized as either an urban or a rural area. The sample is selected with a complex design that consists of stratification and two levels of clustering where the units were selected proportional to size. The population is stratified on counties and urban and rural areas within the counties. Kenya has 47 counties where Nairobi and Mombasa consist of solely urban areas, yielding a total of 92 strata. In the first stage of clustering, 1600 clusters were selected from the EAs and 25 households were selected within the EAs in the second stage of clustering. In this thesis, the EAs is called clusters, where each cluster is either urban or rural. The officially reported national NMR from the survey, estimated by IGME (2019) with the B-spline model proposed by Alkema and New (2014) is 22.301 per 1000 children.

The approaches explored in this thesis are tested through a simulation study. The purpose of conducting the study is to evaluate the set of proposed approaches within a controlled framework. As one is dealing with survey data, true values of the NMR do not exist, making it difficult to evaluate the approaches on the real survey data. The goal of the simulation study is to compare the methods and find the best performing method.

The best performing method from the simulation study is compared to the design-based and smoothed design-based methods on the real survey data from the 2014 DHS in Kenya. It is desired to explore and quantify how compatible the estimates obtained from the best method are to the established and commonly used methods. Can model-based or combined approaches be used instead of the design-based methods? A related question is how sensitive the estimates are to the choice of model and assumptions. What effects are appropriate to include in the modelling? The final goal is to obtain methods that can estimate rates on a fine scale both in space and time, but this is outside the scope of this thesis. To achieve this, it may be necessary to balance accuracy and computational time. Is it possible to obtain sufficiently accurate estimates on fine scales within tolerable running time?

Earlier surveys conducted in Kenya have around 400 clusters, such as the surveys conducted in 2003 and 2008/2009, while the 2014 DHS had 1600 clusters. Are 1600 clusters necessary or could time and money be saved by using fewer clusters? Can model-based or combined methods obtain sufficiently accurate estimates with smaller sample sizes? The 2014 DHS is designed to produce reliable estimates at the county level and national level. Are there fine-scale features hidden by producing estimates at the county level? This is explored with a finer spatial scale model of constituencies within the counties.

Chapter 2

Background Theory

In this section, the relevant background theory needed to understand the methods used in this thesis is presented.

2.1 Complex survey methodology

Samples obtained from surveys differ from samples considered in other fields of statistics. In most areas of statistics, the sampled units are assumed to be a subsample from an infinitely large population. Usually, the population is assumed to belong to a parametric family, and the goal is to learn something about the underlying parameters. On the other hand, data arising from surveys is sampled from a finite population of units. Thus, there is a finite number of possible combinations of units that can be selected. The units in the sample are randomly selected according to a predefined design, often called the survey design or sampling method. The theory of complex survey methodology presented here is based on the book by Lohr (2010).

A sample is commonly denoted by S, where S is a subset consisting of n units from a finite population consisting of N units. Each unit in the population has a measurable value $y_i, i = 1, \ldots, N$. The y_i's are fixed, unknown quantities, unless the corresponding units appear in the sample S. Each sample S has a probability greater than zero of being selected from the population, and they sum to 1. In addition, each unit in the population has a probability greater than zero of being included in the selected sample S, called the inclusion probability, denoted π_i for unit i in the population. The inverse of the inclusion probabilities are called weights and denoted $w_i = 1/\pi_i$, $i = 1, \ldots, N$.

The most common population quantity of interest is the population total. The population total, denoted t, is the sum of a quantity y for each unit in the population, $t = \sum_{i=1}^{N} y_i$. A population total of interest can, for example, be the total number of neonatal deaths in Kenya. From the population total, t, one can also obtain the population mean, $\bar{y} = t/N$. The sample variance for the full population is $s_N^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$. Often, in the developing world, these quantities are not available and must be estimated from a sample

of the population. The goal of sampling is to get sufficiently accurate estimates of quantities that describe the full population. The choice of the sampling method depends on the goal of the research. Here, three methods are outlined: simple random sampling, stratified sampling and cluster sampling.

The simplest sampling method is called simple random sampling (SRS). Here, n units are selected to be in the sample completely at random. For an SRS without replacement, the estimators for the population total and mean of a measured quantity y, are $T_y = \frac{N}{n}\sum_{i\in S} y_i$ and $\bar{Y} = \frac{1}{n}\sum_{i\in S} y_i$. The estimated variance of the estimator of the population total of y, $\hat{V}(T_y)$ is

$$\hat{V}(T_y) = N^2\Big(1 - \frac{n}{N}\Big)\frac{s^2}{n}.$$

Here, $(1 - \frac{n}{N})$ is the finite population correction (fpc) and s^2 is the sample variance of y. The fpc is a correction of the variance that accounts for the finite population. When the population size N is very large compared to n, the total population can be regarded as infinite, and the fpc becomes close to 1. On the other hand, if n and N are of equal magnitude, the fpc corrects the overestimated variance.

Simple random sampling does not ensure that subgroups of the population with specific properties are included and is seldom used in practice. For example, for the 2014 DHS conducted in Kenya, it is desired to obtain separate estimates for urban and rural areas. Simple random sampling could lead to only units in urban areas being sampled, and estimates for rural areas are then unattainable. If the goal of the research is to get estimates of subgroups of the population, such as geographical divisions, stratified sampling should be used. In stratified sampling, the population is partitioned into subgroups called strata. Strata are mutually exclusive groups such that each unit in the population belongs to only one stratum. The units are sampled independently from each stratum such that units from each stratum are guaranteed to be included in the final sample. In the 2014 DHS in Kenya, stratification was used to ensure that units within each county and urban and rural areas within the counties were sampled.

A population of N units are divided into H strata, with N_h units in stratum $h, h = 1, \ldots, H$. In stratification of the simplest form, an SRS is taken independently within each stratum, where the selected sample from stratum h is denoted S_h, with n_h units. The probability of including unit i from stratum h in the sample is $\pi_{hi} = n_h/N_h, i = 1, \ldots, N_h$. The estimator of the total of a quantity within each stratum is $T_h = \frac{N_h}{n_h}\sum_{i\in S_h} y_{hi}$. The estimated totals and variances of the totals for each stratum are summed to obtain the estimated population total

$$T_y = \sum_{h=1}^{H} T_h. \tag{2.1}$$

Further, the estimate of the variance of the estimated total in Equation (2.1) is

$$\hat{V}(T_y) = \sum_{h=1}^{H}\Big(1 - \frac{n_h}{N_h}\Big)N_h^2\frac{s_h^2}{n_h}. \tag{2.2}$$

Here, $s_h^2 = \sum_{i \in S_h} \frac{(y_{hi} - \hat{\bar{y}}_h)^2}{n_h - 1}$ is the sample variance for stratum h where $\hat{\bar{y}}_h$ is the estimate of the mean $\bar{y}_h$ in stratum h. Equation (2.2) shows that the variance of the estimated population total for a stratified sample conducted with SRS is the sum of the variance from each stratum.

If the goal of the research is to produce estimates with limited resources, cluster sampling can be used. In cluster sampling, the population is divided into subgroups, called clusters or primary sampling units (psu). Clusters are geographical units such as villages or households. A unit in the population is only included in the sample if the cluster it belongs to is included. One goal of cluster sampling is to limit the number of geographical areas to sample from, which again can reduce the cost of the sampling. However, units in the same cluster tend to be more similar than units selected at random. It is expected that the NMR is more alike for units in the same village or household due to, for example, access to health clinics. When several units from the same cluster are sampled, similar information is often repeated instead of obtaining new information as one would if the units were less alike. However, in large surveys such as the 2014 DHS in Kenya, it is necessary to cluster because of time and limited budget.

Cluster sampling can be one-staged or multi-staged. In one-stage cluster sampling, all units within a cluster are included in the sample if that particular cluster is sampled. The number of clusters, or primary population units (ppu) in the population, is denoted N and the number of primary sampling units (psu) is denoted n. In cluster sampling of the simplest form, an SRS of n clusters is taken such that each unit in the population has an equal probability of being included in the sample. The total of a quantity in a sampled cluster S_i is $t_i = \sum_{j \in S_i} y_{ij}$, where y_{ij} is the measured quantity of unit j in cluster i. From this, an unbiased estimator of the population total is

$$T_y = \frac{N}{n} \sum_{i \in S} t_i. \tag{2.3}$$

The estimated variance of the estimator of the population total in Equation (2.3) is

$$\hat{V}(T_y) = N^2 \Big(1 - \frac{n}{N}\Big) \frac{s_t^2}{n}. \tag{2.4}$$

The sample variance, s_t^2, around the mean cluster total for N clusters is

$$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \Big(t_i - \frac{\hat{T}_y}{N}\Big)^2,$$

where $\hat{T}_y$ is the computed value of the estimator T_y. Sometimes it is costly or time-consuming to measure all units within each cluster. Two-stage clustering or many-stage clustering can, in this case, be included in the survey design. This method is often used in practice in real and complex surveys where, for example, the geographical areas or the population is large. Instead of including all units in the sampled clusters, only a sub-sample of the units are included in the final sample, such as in the 2014 DHS in Kenya where a subset of the households in each cluster is selected. The number of secondary population units (spu) in cluster i is denoted M_i, and the number of sampled secondary units (ssu) is

denoted m_i. After n clusters are randomly sampled, an SRS is taken of $m_i, i = 1, \ldots, n$ secondary sampling units from the n clusters. The estimator of the total of a quantity for cluster i is $T_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij}$. The estimator of the population total is

$$T_y = \frac{N}{n} \sum_{i \in S} \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij}. \tag{2.5}$$

The estimator of the variance of the population total in Equation (2.5) is the variance for one-stage clustering in Equation (2.4) with an added term to account for the variance that arises from estimating the cluster totals instead of measuring them directly as in one-stage clustering. The resulting variance estimate, as derived in Section 6.6 in Lohr (2010), is

$$\hat{V}(T_y) = N^2 \Big(1 - \frac{n}{N}\Big) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in S} \Big(1 - \frac{m_i}{M_i}\Big) M_i^2 \frac{s_i^2}{m_i}.$$

The sample variance around the estimated cluster mean for cluster i is

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j \in S_i} (y_{ij} - \hat{\bar{y}}_i)^2,$$

where $\hat{\bar{y}}_i$ is the estimate of the mean $\bar{y}_i$ in cluster i.

In this thesis, the population quantity that will be estimated from survey data is the NMR. The estimated NMR, denoted $\hat{r}$, between the estimated total number of children deceased d, $\hat{t}_d$, and the estimated total number of children born b, $\hat{t}_b$, is defined as

$$\hat{r} = \frac{\hat{t}_d}{\hat{t}_b}, \tag{2.6}$$

where the totals are estimated according to the survey design. The estimated variance, $\hat{V}(\hat{r})$, of the rate estimator in Equation (2.6), is given by

$$\hat{V}(\hat{r}) = \Big(1 - \frac{n}{N}\Big) \frac{s_e^2}{n\bar{b}^2}, \tag{2.7}$$

where s_e^2 is the sample variance of the residuals $e_i = d_i - \hat{r} b_i$ for the $i = 1, \ldots, n$ observations in the sample and $\bar{b}$ is the mean of the quantity b in the sample, which is the number of children born. It is desired to evaluate the estimated rates $\hat{r}$ on the logit scale since the rates cannot be assumed to follow a simple, symmetric distribution when expressed as values between 0 and 1, which can be assumed on the logit scale. The logit of the estimated rate is

$$\text{logit}(\hat{r}) = \log\Big(\frac{\hat{r}}{1 - \hat{r}}\Big), \tag{2.8}$$

where the estimated rates are found using Equation (2.6). The rates are, by transforming to the logit scale, mapped from $[0, 1]$ to $[-\infty, \infty]$. One can assume that the logit rates are asymptotic normally distributed, as this has been shown by Mercer et al. (2014) to perform well on estimation with complex survey data

$$\text{logit}(\hat{r}) \sim N(\text{E}[\text{logit}(\hat{r})], \hat{V}(\text{logit}(\hat{r}))). \tag{2.9}$$

The estimated variance, $\hat{V}(\text{logit}(\hat{r}))$, is found via the delta method as outlined in Section 5.5 in Casella and Berger (2001). If $\text{E}(\hat{r}) = r$, the first-order approximation of $\text{logit}(\hat{r})$ is

$$\text{logit}(\hat{r}) = \text{logit}(r) + \text{logit}'(r)(\hat{r} - r),$$

where $\text{logit}'(r)$ is the first derivative of $\text{logit}(r)$. If $\text{logit}(\hat{r})$ is an estimator of $\text{logit}(r)$, the variance of $\text{logit}(\hat{r})$ can be approximated as

$$\hat{V}(\text{logit}(\hat{r})) \approx [\text{logit}'(r)]^2 \hat{V}(\hat{r}), \tag{2.10}$$

where the estimated variance of the rate, $\hat{V}(\hat{r})$ is found from Equation (2.7).

After the estimated rate on the logit scale is found using Equation (2.8) following the survey design, and the variance of the estimate is found via the delta method using Equation (2.10), all considerations necessary to handle the design of the survey are taken. Also, since estimates in different counties are independent through Equation (2.9), they can be treated as spatially referenced observations with a Gaussian likelihood.

2.2 Spatial modelling

Most of the data that are collected have space and time coordinates. Whether this information is relevant or not depends on the kind of study performed. In most experimental studies, information about where the study is performed is not essential when analyzing the outcome of the study. All the information that is relevant for the outcome is in the explanatory variables, and the experiments are independent.

Studies in, for example, epidemiology or environmental sciences are observational, not experimental. The observed outcomes cannot be changed or replicated. In this case, records of the place and time of events can be relevant. One example is survey data, introduced in Section 2.1. The geographical locations of the clusters and households in a sample influence the observed outcomes. Spatial modelling can play an important part in explaining spatial variation in the measured quantities obtained from surveys. It is reasonable to assume that the NMR is more similar in clusters in the same county than for clusters in different counties. Data that is geographically referenced is called spatial data. For spatial data, one should account for the spatial dependence in the model. If the spatial relation between the data is dismissed in the modelling, it may result in biases in the estimates and loss of efficiency.

Spatial data is defined in Blangiardo and Cameletti (2015) as realizations of a process indexed by space

$$Y(\boldsymbol{s}) \equiv \{y(\boldsymbol{s}), \boldsymbol{s} \in D\}, \tag{2.11}$$

where D is a fixed subset of $\mathbb{R}^d$. Banerjee et al. (2004) classify spatial data into three types: point-referenced data, areal data and, point-pattern data. For point-referenced data (or geostatistical data), the observed value $Y(\boldsymbol{s})$ in Equation (2.11), is a random vector at sites $\boldsymbol{s} \in \mathbb{R}^d$. The sites, $\boldsymbol{s}$, varies continuously over a domain D, where D is a fixed subset of $\mathbb{R}^d$ that is a d-dimensional rectangle. For areal data, the domain D of regular or irregular shape is divided into areal units with well-defined boundaries between them. The values, $Y(\boldsymbol{s})$, consist of a single aggregated measured value per areal unit. Lastly, point

pattern data have a random domain D where its index set contains the sites of random events.

Both areal data and point-referenced data are considered in the methods in this thesis. The design-based estimates are spatially smoothed with the smoothed design-based approach. Here, areal data $Y(\boldsymbol{s})$ are the NMRs from the design-based method for each county. The areas are denoted by $s_i, i = 1, ..., n_c$, where n_c is the number of counties in the domain D. For the model-based approach, point-referenced data are considered. Here, $Y(\boldsymbol{s})$ are random outcomes at specific locations with GPS coordinates. The data are represented as a set of observations $y(s_1), \ldots, y(s_{n_{cl}}), y(s_{n_h})$, where $s_1, \ldots, s_{n_{cl}}, s_{n_h}$ are the point-reference locations of the observations and the number of locations is n_{cl} for clusters and n_h for households. Thus, the observed values are the outcomes, i.e. the number of neonatal deaths in each cluster or household.

The spatial correlation between the areas can be accounted for using the class of spatial models called intrinsic conditional auto-regressive models (ICAR) introduced by Besag (Besag et al. (1991)). Even though intrinsic models are improper with precision matrices not of full rank, Besag and Kooperberg (1995) point out several advantages over the standard auto-regressive models. "They often avoid difficulties in parameter estimation, without apparent loss, or exhibit appealing invariances" (Besag and Kooperberg (1995)).

Given a set of observed values from n areas, spatial correlation between a pair of areas s_i and s_j can be modelled conditionally with a spatial random variable, $\phi = (\phi_1, \ldots, \phi_n)^T$. The association between the areas or locations is dependent on the neighbourhood arrangement of the areas. The neighbourhood arrangement is quantified in an $n \times n$ adjacency matrix denoted W where the entries w_{ij} quantifies the spatial association between areas s_i and s_j. The adjacency matrix is binary such that $w_{ii} = 0, w_{ij} = 1$ if areas i and j are neighbours and $w_{ij} = 0$ otherwise.

The full conditional distribution for value ϕ_i for area i in the ICAR model is

$$p(\phi_i|\phi_j, j \neq i, \tau_i^{-1}) = N\left(\frac{\sum_{i\sim j}\phi_i}{d_i}, \frac{1}{d_i\tau_i}\right), \tag{2.12}$$

where the conditional mean of ϕ_i is the average of the spatial effects over it's neighbours, d_i is the number of neighbours of area i, the precision parameter is τ_i, which is the inverse of the variance σ_i^2 and $i \sim j$ indicates that areas i and j are neighbours. The conditional precision is proportional to the number of neighbours such that the conditional variance will be smaller if an area has many neighbours. The joint distribution of ϕ is given by

$$\phi \sim N(0, [\tau_\phi(D - W)]^{-1}, \tag{2.13}$$

where D is an $n \times n$ diagonal matrix with the number of neighbours d on the diagonal. The joint distribution in Equation (2.13) can be rewritten in pairwise difference form as

$$p(\phi|\tau_\phi) \propto \exp\Big(-\frac{\tau_\phi}{2}\sum_{i\sim j}(\phi_i - \phi_j)^2\Big) = \exp\Big(-\frac{\tau_\phi}{2}\phi^T\boldsymbol{Q}\phi\Big)$$

where $\boldsymbol{Q} = D - W$ is the precision matrix with entries

$$Q_{ij} = \begin{cases} d_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

This is an example of an intrinsic Gaussian Markov random field (GMRF). A GMRF is a Markov random field (MRF) following a multivariate normal distribution. Following Rue and Held (2005), an MRF satisfies the following conditional independence assumption: The pairs of entries of an MRF are conditionally independent given the remaining elements and vice versa. This results in computational benefits as the precision matrix, $\boldsymbol{Q}$, is sparse with zeros for pairs of conditionally independent values. The non-zero entries of the precision matrix $\boldsymbol{Q}$ are given by the neighbourhood structure, where the entries of the precision matrix are only non-zero for the neighbouring areas. The specification intrinsic means that the precision matrix is not of full rank. In addition, the joint distribution is non-identifiable as adding any constant to all of the elements of ϕ leaves the joint distribution unchanged. This issue is solved by adding a sum-to-zero constraint on the spatial variables: $\sum_i \phi_i = 0$.

In the ICAR model in Equation (2.12) the spatial relationship between the areas is considered to be structured. Hence, unstructured noise within each area is modelled as structured spatial correlation, giving misleading estimates of the spatial effect. Therefore, the Besag-York-Mollié (BYM) model (Besag et al. (1991)) decomposes the spatial effect, denoted υ, into a structured, ϕ, and an unstructured spatial component, γ, such that $\upsilon = \phi + \gamma$. The structured spatial component is the ICAR model with joint distribution as in Equation (2.13). The second component γ is independent random noise with joint distribution $\gamma \sim N(0, 1/\tau_\gamma \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix. The resulting covariance matrix of υ is

$$\text{Var}(\upsilon|\tau_\phi, \tau_\gamma) = 1/(\tau_\phi \boldsymbol{Q}) + 1/(\tau_\gamma \boldsymbol{I}).$$

The BYM model is used to model spatial effects both for areal data in the smoothed method and for point-referenced data in the model-based approach. For the smoothed design-based method, there is one spatial effect υ_i for each areal data point $i = 1, \ldots, n_c$ for all counties. For the model-based method, the spatial model is similar since the same spatial effect υ_i is assigned to all the point-referenced clusters within the same county i.

Models with spatial effects can be constructed within the Bayesian framework by expanding the concept of hierarchical models to account for similarities based on the neighbourhood structure. The next section presents the theory of Bayesian statistics and outline Bayesian hierarchical models.

2.3 Bayesian hierarchical modelling

The classical approach to statistical analysis is the frequentist approach. As stated in Bolstad (2007), "in frequentist statistics, it is assumed that the data is distributed according to a model with fixed and unknown parameters". Inference about the parameters is based

on likelihoods calculated from the distribution the data is assumed to follow. These likelihoods are based on all possible samples that can be drawn from the assumed distribution. They are not conditional on the sample collected.

Here is where the difference between the Bayesian and frequentist approach appears. In the Bayesian approach, it is assumed that the parameters of the model are random. Inference about the parameters of interest is based on the posterior distribution of the parameters given the collected data. The theory of Bayesian statistics and hierarchical models presented here is based on the book by Gelman et al. (2014).

Inference about the parameter of interest, x, given the observed value, y, is based on the posterior distribution $p(x|y)$. The parameter x can, for example, be the fixed effect of the urban and rural clusters and the response y can be the number of observed neonatal deaths in a cluster or household. To obtain the posterior distribution, one needs the joint probability distribution. The joint probability distribution can be written as a product of the prior distribution of x, denoted $p(x)$, and the conditional distribution, denoted $p(y|x)$. The joint distribution becomes

$$p(x, y) = p(x)p(y|x).$$

The distribution of the fixed effect x, $p(x)$, is the knowledge one has about the parameter before the data is observed. The core of the Bayesian approach to inference about the parameter x is Bayes' rule. Conditioning on the outcome y, Bayes' rule yields the posterior distribution of x

$$p(x|y) = \frac{p(x)p(y|x)}{p(y)}. \tag{2.15}$$

Since $p(y)$ is not dependent on x, Equation (2.15) can be written in unnormalized form as

$$p(x|y) \propto p(x)p(y|x).$$

Many situations where it is desired to perform statistical analysis involve several parameters that are related in some way. For example, for spatial data as introduced in Section 2.2, it is desired to incorporate the spatial dependence of the data in the model. Also, when working with clustered samples as outlined in Section 2.1, the dependence between observations in the same cluster can be included. One needs a way to construct such complex models with several effects in a simple and interpretive way. This type of data can be modeled hierarchically.

A Bayesian hierarchical model can be split into stages. Assume a set of residents $i = 1, \ldots, n$ is sampled in a survey, where resident i has measured quantity y_i. Then it can be assumed that the distribution of y_i is conditional on a set of latent variables, and the distribution of the latent variables is conditional on a set of parameters. In a three-stage model, those parameters are called hyperparameters with prior distributions. The latent field is denoted $\boldsymbol{x}$ and the hyperparameters are denoted $\boldsymbol{\theta}$. Three-stage hierarchical models can be defined by the following stages.

$$\text{First stage: } \boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta} \sim \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$$
$$\text{Second stage: } \boldsymbol{x}|\boldsymbol{\theta} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta})$$
$$\text{Third stage: } \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$$

The models developed in this thesis belongs to a class of models called latent Gaussian Models (LGMs). LGMs are a class of hierarchical models, where the prior distribution of every component of the latent field must be Gaussian conditional on the parameters. LGMs are a subclass of structure additive regression models, for example, formulated in Chapter 9 of Farhmeir et al. (2013). For structure additive models, the response y_i is assumed to belong to an exponential family, where the mean of y_i, μ_i, is linked to the predictor η_i through a link function, such that $g(\mu_i) = \eta_i$.

Models with two kinds of distributions on the observational level are considered. For the smoothed design-based method (SDB), the design-based NMRs on the logit scale, denoted y_i, are assumed to follow a Gaussian distribution with mean η_i and estimated design-based variance σ_i^2. For the model-based methods (MB1 and MB2), the observed values are the number of neonatal deaths in each cluster or household linked to the predictor with a logit link with a binomial likelihood. The number of trials is denoted b_i for the number of births in either each cluster or household. Following a similar notation as in Rue et al. (2009), the SDB Gaussian model is

$$y_i \sim N(\eta_i, \sigma_i^2), \quad i = 1, \ldots, n_c,$$

where n_c is the number of counties. The MB binomial model is

$$y_i \sim \text{Binomial}(p_i, b_i), \quad \text{logit}(p_i) = \eta_i, \quad i = 1, \cdots, n_{cl}, n_h,$$

where b_i is the number of births in cluster or household i, p_i is the probability of death and n_{cl} and n_h is the number clusters and households in county i. The structured additive predictor is

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i.$$

The predictor accounts for effects of different covariates in an additive manner. In the Gaussian model, $i = 1, \ldots, n_c$ represents observations from each county. In the binomial model, $i = 1, \ldots, n_{cl}, n_h$ represents the observations from each cluster or household. The $\{\beta_k\}$s are the linear effects of the covariates $z_k, k = 1, \ldots, n_\beta$. In this context, one fixed effect separating urban and rural clusters in the binomial model and none in the Gaussian model. The intercept is α and ε_is are unstructured error terms.

The functions $\{f^{(j)}(\cdot)\}$, called model components by Rue et al. (2017), are unknown functions of the covariates $u_{ji}, j = 1, \ldots, n_f$. The functions $\{f^{(j)}(\cdot)\}$ can take many different forms. Here, model components are added to incorporate spatial effects and random effects. Spatial dependence follows the BYM model outlined in Section 2.2. Then, one of the $f^{(j)}(\cdot)$ terms is $f(u_i) = f_i$ for counties $i = 1, \ldots, n_c$ with spatial effect u_i. In the binomial model, all clusters and households in a county are assigned the same spatial effect. For the correlation between observations in the clusters and households, the variables f_i are independent Gaussian with zero mean for clusters or households $i = 1, \ldots, n_{cl}, n_h$. For LGMs, Gaussian distributions are assigned to the latent field, $\boldsymbol{x}$ consisting of α, $\{f^{(j)}(\cdot)\}$, $\{\beta_k\}$ and $\{\varepsilon\}$. The vector of hyperparameters, $\boldsymbol{\theta}$, are not necessarily Gaussian. LGMs can be used on a wide range of applications and can model a sum of various components such as both spatial and temporal dependencies, random effects

and linear and smooth effects of some covariates. For a list of examples of applications of LGMs, see section 1.2 in Rue et al. (2009).

Once a Bayesian model is defined through a likelihood, possible model components and fixed effects and parameters, the posterior probabilities of the parameters can be computed. However, it is only in some cases that the posteriors can be computed exactly. In most situations, the posteriors have to be approximated. There are many methods available, and some of them are presented in the next section.

2.4 Inference with Bayesian hierarchical models

The most common approach to inference of Bayesian hierarchical models is Markov chain Monte Carlo (MCMC) sampling. The theory of Markov chain simulations is found in, for example, Chapter 11 of Gelman et al. (2014). MCMC is a method based on drawing values of the parameter θ from approximate distributions and correcting the draws by a acceptance/rejection rule to approve the approximation of the target posterior distribution $p(\theta|y)$. The draws form a Markov which is defined as a sequence of random variables $\theta_1, \theta_2, \ldots$, where for any n, the distribution of θ_n given all previous θ's only depend on the last value, θ_{n-1}. The more steps of the Markov chain that are taken, the closer the distribution of the samples is to the desired posterior distribution.

The favorable feature of MCMC methods is that the error can be made arbitrarily small. MCMC methods guarantee asymptotically exact recovery of the posterior distribution as the number of posterior samples grows. However, this may be time-consuming or require great computational power. For approximating posteriors of LGMs, MCMC methods tend to perform poorly as discussed in 1.4 in Rue et al. (2009). The components of the latent field $\boldsymbol{x}$ are strongly dependent on each other, the same is true for the latent field and the hyperparameters $\boldsymbol{\theta}$, requiring modifications to the algorithm. Rue, Martino and Chopin propose in Rue et al. (2009) an alternative, deterministic approach to approximating posteriors for LGMs called integrated nested Laplace Approximations (INLA). Rue et al. (2009) argue that the INLA approach outperforms MCMC algorithms for a given computational cost. According to Rue et al. (2009), for fine-scale spatial models where the number of observations can be thousands, approximations to the posterior marginals are computed with INLA within minutes, compared to maybe hours with MCMC approaches.

Rue et al. (2017) makes the following three assumptions that are required for the approximations to be accurate and computationally feasible. The number of hyperparameters $\boldsymbol{\theta}$ is small, not exceeding around 20. Also, the distribution of the latent field $\boldsymbol{x}|\boldsymbol{\theta}$ is Gaussian and is a GMRF when $\dim(\boldsymbol{x})$ is high. Lastly, the observations $\boldsymbol{y}$ are mutually conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\theta}$. These assumptions are required both for computational reasons and to ensure that the approximations are accurate with a high degree of certainty.

The INLA method is based on Laplace approximations. Following Blangiardo and Cameletti (2015), Laplace approximation is a technique used to approximate integrals of the form

$$\int f(x)dx = \int \exp(\log(f(x))dx, \tag{2.16}$$

where $f(x)$ is the density of a random variable X. The second order Taylor series expansion of the term $\log f(x)$ around $x = x_0$ is

$$\log f(x) \approx \log f(x_0) + (x - x_0)\frac{\partial \log f(x)}{\partial x}\bigg|_{x=x_0} + \frac{(x - x_0)^2}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\bigg|_{x=x_0}.$$

x_0 is set equal to the mode $x^* = \text{argmax}_x \log f(x)$, such that the derivative of $\log f(x)$ evaluated at $x_0 = x^*$ is zero. The integral in Equation (2.16) can then be approximated as

$$\int f(x)dx \approx \exp(\log f(x^*)) \int \exp\bigg(- \frac{(x - x^*)^2}{2\sigma^{2*}}\bigg),$$

where the integrand has the form of the density of a Gaussian distribution setting $\sigma^{2*} = -1/\frac{\partial^2 \log f(x)}{\partial x^2}|_{x=x^*}$.

The posteriors of interest are the variance of the spatial effect and the cluster effects, and the size of the fixed effect of urban and rural stratification. The posterior marginals can be written as

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \tag{2.17}$$

$$\pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}, \tag{2.18}$$

where $\boldsymbol{\theta}_{-j}$ is all θ's except θ_j. The key feature of the INLA approach is to construct nested approximations to Equation (2.17) and (2.18) of the form

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \int \tilde{\pi}(x_i|\boldsymbol{\theta}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

$$\tilde{\pi}(\theta_j|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j},$$

where $\tilde{\pi}(\cdot|\cdot)$ is an approximated density of its arguments. Approximations to $\pi(x_i|\boldsymbol{y})$ are computed by approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$, and using numerical integration to integrate out $\boldsymbol{\theta}$. The integration is made possible due to the small dimension of $\boldsymbol{\theta}$. The approach is based on the following approximation $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ of the marginal posterior of $\boldsymbol{\theta}$

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\bigg|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta}),} \tag{2.19}$$

where $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation obtained by performing Laplace approximation and $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the mode of the full conditional for $\boldsymbol{x}$, for a given $\boldsymbol{\theta}$. For the posterior marginals of the latent field, the density of $x_i|\boldsymbol{\theta}, \boldsymbol{y}$ can be approximated with the Gaussian marginal derived from $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. This approximation can be integrated numerically with respect to $\boldsymbol{\theta}$. One obtains the following expression of approximations of the marginals for the latent field

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta_k}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta_k}|\boldsymbol{y})\Delta_k, \tag{2.20}$$

where the sum is over the $\boldsymbol{\theta}$s with area weights Δ_k.

The first step of INLA is to compute the approximation to the posterior marginal of $\boldsymbol{\theta}$ in Equation (2.19). The main use of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is to integrate out the uncertainty with respect to $\boldsymbol{\theta}$ when approximating the posterior marginal of x_i in Equation (2.20). Firstly, the mode of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$, denoted $\boldsymbol{\theta}^*$, is located by maximizing $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$ with respect to $\boldsymbol{\theta}$, by the use of a quasi-Newton method which creates an approximation to the second derivatives of $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$. Thereafter, the negative Hessian matrix $\boldsymbol{H} > 0$ of $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$ is computed at the mode $\boldsymbol{\theta}^*$, using finite differences. Let $\boldsymbol{\Sigma} = \boldsymbol{H}^{-1}$, which is the covariance matrix of $\boldsymbol{\theta}$ if the density is Gaussian. To correct for scale and rotation, and simplify numerical integration, a reparametrization is done such that $\boldsymbol{\theta}$ is defined via the standardized variables $\boldsymbol{z}$

$$\boldsymbol{\theta} = \boldsymbol{\theta}^* + \boldsymbol{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{z}.$$

Here, $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ is the eigendecomposition of $\boldsymbol{\Sigma}$. If $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is Gaussian, then $\boldsymbol{z}$ is standard Gaussian. The third step is to explore $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$, using the $\boldsymbol{z}$-parametrization, to locate the majority of the probability mass. A grid is constructed and the $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$ is computed for all points on the grid, locating where $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$ is considered significant, which is used in the numerical integration in Equation (2.20). Finally, the points computed in the grid are used to construct an interpolant to $\log(\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}))$, and to compute the marginals $\tilde{\pi}(\theta_j|\boldsymbol{y})$.

Now that a set of weighted points $\boldsymbol{\theta}_k$ are obtained, the next step is to obtain approximations for the posterior marginal of the latent field x, $\tilde{\pi}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$. There are three different approximations: Gaussian, Laplace and simplified Laplace. As argued in 3.2 in Rue et al. (2009), Laplace approximations are preferred in general. However, the simplified Laplace has much smaller cost that can compensate for the slight loss in accuracy. All three approaches are outlined in 3.2 in Rue et al. (2009). The main benefit of the INLA method is computational, as the method can produce accurate approximations in seconds or minutes. In addition, a wide range of different LGMs can easily be implemented with the same general code by changing the likelihood, model components, and priors.

A set of measures and scoring rules must be defined to assess the performance of the methods. The next section outlines the scoring rules that will is used in this thesis.

2.5 Model assessment

For validating the accuracy of the proposed methods, one needs to select one or more forms of accuracy measures. In this section, four scoring rules for assessing the predictive performance of the methods are presented. The chosen scoring rules are the mean absolute error (MAE), root-mean-squared error (RMSE), continuous rank probability score (CRPS) and the mean bias error (MBE). The scoring rules are averaged over all n counties, $i = 1, \ldots, n$, where the true neonatal mortality rate and the predicted neonatal mortality rate in coutny i are denoted r_i and $\hat{r}_i$, respectively. Also, the average running time of the methods is considered. The methods are evaluated on running time as it is desired to obtain methods that are scalable to fine spatio-temporal scales. Also, it is desired to assess if it is worth choosing more complex methods when it comes to accuracy vs. running time.

The first scoring rule considered is the mean absolute error (MAE). The MAE is the average of the absolute difference between the true rate, r_i, and the predicted rate, $\hat{r}_i$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|r_i - \hat{r}_i|.$$

The MAE is a linear score which means that all the individual errors are weighted equally when averaging over the counties. The root-mean-squared error (RMSE) is the square root of the average squared difference between the true rate, r_i, and the predicted rate, $\hat{r}_i$. The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2}.$$

Since the errors are squared before they are averaged, RMSE penalizes large errors stronger than MAE. Therefore, RMSE is more informative when large errors are particularly undesirable. The RMSE incorporates both the variance of the estimator and its bias.

As explained in Gneiting and Raftery (2007), the continuous ranked probability score (CRPS) measures the difference between the predicted and occurred cumulative distributions. The CRPS handles not only the uncertainty in the predictions but also the uncertainty in the observations. Thus, the CRPS will favor predictions that are further away from the true rate, but with uncertainty that reflects this, compared to predictions closer but with an uncertainty that is too small. The CRPS is a generalized version of the MAE and can be used on probabilistic forecasts, as apposed to more straightforward scoring rules such as MAE and RMSE which are not directly applicable on probabilistic forecasts.

Let R be a random variable and let F be the cumulative distribution function of R, such that $F(r) = \text{P}[R \leq r]$. Let $\hat{r}$ be the observations. The crps is defined as

$$\text{crps}(F, \hat{r}) = -\int_{-\infty}^{\infty}(F(r) - \mathbb{1}[\hat{r} \leq r])^2 dr,$$

where $\mathbb{1}$ is the heaviside step function that attains the value of 1 if the argument is positive or zero and the value of 0 otherwise. If the predictive distribution is Gaussian with mean μ and variance σ^2, then the crps is

$$\text{crps}(N(\mu, \sigma^2), \hat{r}) = \sigma\left[\frac{1}{\sqrt{\pi}} - 2\phi\Big(\frac{\hat{r}-\mu}{\sigma}\Big) - \frac{\hat{r}-\mu}{\sigma}\left(2\Phi\Big(\frac{\hat{r}-\mu}{\sigma}\Big) - 1\right)\right],$$

where ϕ and Φ are the probability density function and the cumulative distribution function of a standard Gaussian variable, respectively. The CRPS is expressed in the same unit as the observations and is in practice averaged over the observations. Let CRPS denote the average of the crps defined as

$$\text{CRPS} = \frac{1}{n}\sum_{i=1}^{n}\text{crps}(F_i, \hat{r}_i).$$

The final measure is the mean bias error (MBE). MBE captures the average bias in the prediction and is calculated as

$$\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_i - r_i).$$

The bias of a predicted rate shows the tendency of a model to over- or underestimate the rate. The MBE is often used to decide if any steps need to be taken to correct the bias in the model and are usually not used as a measure of the model error as high individual errors in prediction can still produce a low MBE. The MBE, MAE and RMSE are related by the following inequalities: MBE $\leq$ MAE $\leq$ RMSE.

Chapter 3

Method

Four different methods for estimating NMRs at the county level are explored. The design-based method and the smoothed design-based method are described in Section 3.1 and in Section 3.2, respectively. The next method considered follows the model-based approach and is presented in Section 3.3. The last method is a combination of the smoothed design-based method and the model-based methods. The combined methods are outlined in Section 3.4. Finally, details of the implementation of the methods in INLA is described in Section 3.5.

3.1 Design-based method

Neonatal mortality rates are estimated for Kenya's 47 counties. As mentioned in Section 1.3, the 2014 DHS in Kenya is a stratified two-stage cluster sample. The 47 counties crossed with an urban/rural indicator for the EAs within each county, are the strata. There is a total of 92 strata, as the counties Nairobi and Mombasa are both solely urban. In the first stage of clustering, 1612 clusters (EAs) was selected, 995 rural clusters and 617 urban clusters. A map of Kenya with the 47 counties and cluster locations is presented in Figure 3.1. At the second stage of the sampling procedure, 40300 households were sampled from the selected clusters.

The dataset from the 2014 Kenya DHS can be accessed through the DHS program (DHS (2019)). Each row in the dataset represents a child, and the columns in the data set are the reported values of the variables from the survey questionnaire. The variables are coded according to the DHS recode manual. For the 2014 Kenya DHS, the recode manual can be found in DHS (2013). The relevant variables from the dataset are presented in Table 3.1.

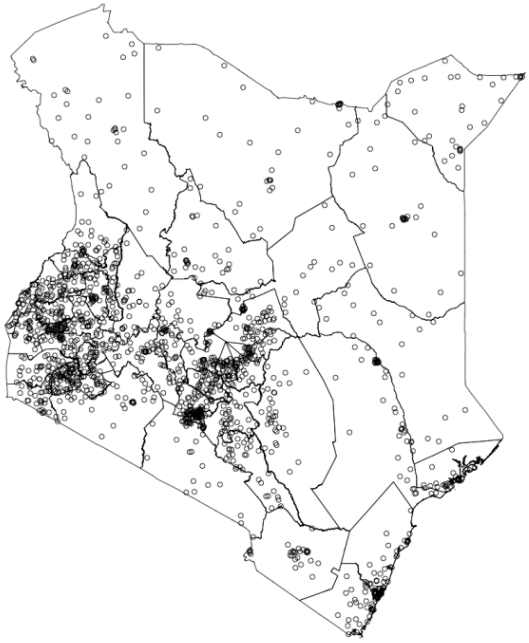


Figure 3.1: Map of Kenya with the 47 counties and cluster locations.

v002	Household ID
v005	Sample weight
v008	Century month code of date of interview
v021	Cluster ID
v023	Strata ID (county and urban/rural stratum)
v102	Urban/rural identifier
b3	Century month code for the date of birth of the child
b6	Age at death of the child (NA if the child is alive)
scounty	County ID

Table 3.1: Relevant variables from the DHS dataset

The full dataset consists of 83591 observations of 1135 variables. The first step when working with the data set is to prepare the data for the analysis. The health indicator of interest in this thesis is the NMR in the years 2009-2014. The observations and variables that are not of interest are removed to give a better overview and speed up the computation time. First, all variables except the variables presented in Table 3.1 are removed. Secondly, the children that are born less than 30 days before the interview are removed based on the value of the variables v008 and b3. Also, children that are born outside the period of interest (2009-2014) are removed based on the value of the variable b3. Lastly, a new variable is created, flagging the children that died in the neonatal period, based on the variable b6. The

variable has value 1 if a child died in the neonatal period, and 0 otherwise. The IDs for the counties, strata, clusters and households are the values of the variables scounty, v023, v021 and v002, respectively. Finally, the urban/rural clusters identifier is in the variable v102.

Estimates of the neonatal mortality rate for each county and strata are found with the help of the sampling weights defined in Section 2.1. The weight of each child in the dataset is given in the variable v005 and are based on the stratum, cluster and household of each child. The estimates are calculated as the rate between the estimated total number of neonatal deaths and the estimated total number of children born for each county, see Equation (2.6).

The rates are estimated with the use of functions from the survey package in R. Documentation and examples of the package can be found in Lumley (2010). First, a design object is created using the function svydesign from the survey package. The function requires the arguments id, strata , weights and data. The argument id specifies the cluster ids from the first level to the last level of clustering. The argument is a combination of the cluster ID and the household ID as the design is two-stage clustered. The argument strata specify the strata, in this case, which county the stratum belongs to and if it is urban or rural. The weights are specified in weights and data is the dataset. How the survey design is created is found in Listing 3.1.

Listing 3.1: Creating a complex survey design.

```
design = svydesign(id=~v021+v002, strata=~v023, weights=~v005,
    data=data)
```

Here, v021 and v002 are the cluster ID's and household ID's respectevily, v023 is the strata ID's, v005 is the weights and data is the dataset.

After the design object is created, the rates can be estimated with the function svyratio from the survey package. The function requires a numerator, a denominator and a design object. The numerator is a vector of zeros and ones where a row in the vector represents a child with value zero if the child survived the neonatal period and value one if the child died in the neonatal period. The denominator is an equally long vector with ones representing births. The design object made with svydesign is the final argument where the observations belonging to the specific county are extracted from the object. The function estimates the population total of children born and children that died in the neonatal period in the specific county according to the survey design, and thereafter the ratio between the totals. The results of the svyratio call are the estimated rate and the estimated variance of the rate. Code for estimating the rate and the variance of the rate for a county with ID $= i$ is presented in Listing 3.2.

Listing 3.2: Estimating design-based county rates.

```
design_county = design[design$variables$scounty==i]
rate = svyratio(~neonatal_mortality, ~ones, design_county)
```

Here, design_county is the design object for the county with county ID specified in scounty $= i$. The nominator is neonatal_mortality and denominator is ones. The result of the svyratio function is stored in the object rate. As explained in Section 2.1, it is desired to estimate the rates on the logit scale. The logit of the rates are found with Equation

(2.8). The corresponding variances of the logit estimates are found with the delta method in Equation (2.10).

The design-based method requires the number of observed neonatal deaths to be larger than zero for each county. If non of the observations in a county died in the neonatal mortality period, it will result in estimated rates and variances of zero, which cannot be used for estimation of the NMR. As a consequence, the design-based method only works on survey data that have enough samples in each subnational area one wishes to obtain estimates.

The results of this method are design-based estimates of the NMR on the logit scale and the variance of the estimates for each county. Let y_i and σ_i^2 denote the estimated NMR and estimated variances for counties $i = 1, \ldots, n_c$, where n_c is the number of counties. The results are used in the next section in the smoothed design-based approach.

3.2 Smoothed design-based method

The answers collected through surveys from residents in the same county may be more similar than answers from residents in different counties. Therefore, it can also be assumed that counties that share a common border have more similar NMRs than counties further apart. If this assumption is correct, one can improve the design-based estimates by spatially smoothing them over the counties, reducing the variance in the estimates.

The smoothed design-based method (SDB) is constructed with the design-based estimates for the counties from Section 3.1. The design-based NMRs are represented as areal data which are smoothed with the BYM model described in Section 2.2. The SDB method is fitted with the package INLA in R, which estimates posteriors with integrated nested Laplace approximations as described in Section 2.4.

For each county i the model is

$$y_i|\eta_i \sim N(\eta_i, \sigma_i^2), \quad \eta_i = \mu + \phi_i + \gamma_i, \quad i = 1, \ldots, n_c.$$

The outcomes $y_i, i = 1, \ldots, n_c$, are the design-based NMRs on the logit scale, which is conditioned on the true, unobserved mortality η_i that is desired to estimate and σ_i^2 is the estimated design-based variance of the estimated rate y_i. The linear predictor η_i contains an intercept μ and the spatial random variables in the BYM model outlined in Section 2.2: $\phi_i \sim \text{ICAR}(\sigma_\phi^2)$ and $\gamma_i \sim_{\text{iid}} N(0, \sigma_\gamma^2)$. The distribution of the intercept and the priors of the hyperparameters are

$$\begin{aligned} \mu &\sim N(0, \infty), \\ \log\tau_\phi &\sim \text{logGamma}(0.5, 0.0005), \\ \log\tau_\gamma &\sim \text{logGamma}(0.5, 0.0005), \end{aligned}$$

where the intercept is a fixed effect modelled with a normal prior with zero mean and large variance, $\tau_\phi = 1/\sigma_\phi^2$ and $\tau_\gamma = 1/\sigma_\gamma^2$. INLA assigns by default non-informative priors of logGamma$(1, 0.0005)$ for the precision of the random effects ϕ and γ. The priors are changed as the default prior has a prior 95% CI of $[0.01, 0.14]$ for σ_ϕ and σ_γ which results in unrealistically small effects. The new prior of has a prior 95% CI of $[0.1, 1.1]$ which is more sensible.

The neighbourhood structure needs to be specified in a graphical structure which assigns the set of neighbours for each county. The graphical structure is created in an ASCII file with $n_c + 1$ rows. The first row contains the total number of counties, n_c. The next n_c rows specify the county ID, the number of neighbours and the IDs for the neighbours. An example of a row in the file is in Listing 3.3.

Listing 3.3: A row in the graph file specifying the neighbourhood structure.

```
1 8 5 12 20 31 37 43 44 47
```

Here, the ID for the county is 1, the county has 8 neighbours and their county IDs are 5, 12, 20, 31, 37, 43, 44 and 47. After defining the neighbourhood structure, the formula for the model is specified in R with the code in Listing 3.4.

Listing 3.4: Formula specifying SDB to be used in INLA.

```
formula = y ~ 1
    + f(ID_County, model="bym", graph="Kenyaadm.graph",
    constr=TRUE, scale.model=TRUE, hyper=c(0.5,5e-04))
```

Here, y is the response and f is the model component for the spatial effects at the county level. ID_County represents the IDs for the counties and through the graph option the name of the object containing the neighborhood structure is included. The type of spatial model to be used is defined by the parameter model, which is set to BYM. The priors of the hyperparameters are defined in hyper. The argument scale.model is a boolean variable indicating whether to scale the model so that the generalized variance is 1. Since the ICAR model is intrinsic, it does not have a proper marginal variance. Therefore, the BYM model is scaled to be able to interpret the precision for the spatial effects τ_ϕ and τ_γ. constr is a boolean variable indicating whether to set a sum to zero constraint on the spatial term which is set to true, following the discussion in Section 2.2.

As for the design-based method in Section 3.1, the smoothed design-based estimates can only produce estimates of the NMR in the counties where there are more than zero observed neonatal deaths.

3.3 Model-based method

The traditional approach to spatial modelling is model-based. Model-based methods do not account directly for the survey design, but indirectly by specifying a probability model for the random process that generates the data. In model-based approaches, it is assumed that there is an underlying infinite population and that the response, given a set of covariates and model components are randomly drawn from it. The estimated model coefficients are the coefficients that best describe the underlying trend for the infinite population. To the extent that the model represents the process that generated the data, it is possible to draw conclusions that can be generalized to other situations where the same process operates. The true NMRs cannot be determined, as the model can only ever be an approximation.

Two different model-based models are proposed. The first model (MB1) contains iid effects on the first stage of clustering, while the second model (MB2) is an extension of MB1 with additional iid effects for the second stage of clustering. The most correct model

is MB2, while MB1 is an approximation to MB2. The iid effects are added to resemble the dependence that exists between individuals in the same cluster and household. Here, the observations are point-referenced with specific locations for the clusters or households. The observed values are modelled with a Binomial model where the response is assumed to be outcomes of a set of Bernoulli trials.

MB1

Model MB1 becomes

$$y_{ij} \sim \text{Binomial}(p_{ij}, b_{ij}), \quad \text{logit}(p_{ij}) = \eta_{ij}, \quad i = 1, \ldots, n_c, \;\; j = 1, \ldots, n_{icl},$$

where the response y_{ij} is the number of children in cluster j in county i who died in the neonatal period, n_c is the number of counties and n_{icl} is the number of clusters in county i. The response y_{ij} is linked to an additive predictor η_{ij} through a logit link and binomial likelihood where the number of trials b_{ij} is the number of births in cluster j in county i and p_{ij} is the probability of death for children in cluster j in county i. The additive predictor is

$$\eta_{ij} = \mu + \mathbb{1}[cl(j) = 1]\beta + \phi_i + \gamma_i + \nu_j, \quad i = 1, \ldots, n_c, \;\; j = 1, \ldots, n_{icl}.$$

The intercept is μ and the fixed stratum effect is β, where the indicator function $\mathbb{1} = 1$ if cluster j is urban and 0 if cluster j is rural. The spatial random variables are defined with the BYM model described in Section 2.2: $\phi_i \sim \text{ICAR}(\sigma^2_\phi)$ and $\gamma_i \sim_{\text{iid}} N(0, \sigma^2_\gamma)$. The same spatial random variables are assigned to clusters within the same county. The final term is the independent random noise at the cluster level, $\nu_j \sim_{\text{iid}} N(0, \sigma^2_\nu)$. The distribution of μ and β and the priors of the hyperparameters are

$$\begin{aligned}
\mu &\sim N(0, \infty), \\
\beta &\sim N(0, 1000), \\
\log\tau_\phi &\sim \text{logGamma}(0.5, 0.0005), \\
\log\tau_\gamma &\sim \text{logGamma}(0.5, 0.0005), \\
\log\tau_\nu &\sim \text{logGamma}(0.5, 0.0005),
\end{aligned}$$

where $\tau_\phi = 1/\sigma^2_\phi$, $\tau_\gamma = 1/\sigma^2_\gamma$ and $\tau_\nu = 1/\sigma^2_\nu$. The formula for the model is specified in R with the code in Listing 3.5.

Listing 3.5: Formula specifying MB1 to be used in INLA.

```
formula = y ~ 1 + urban
    + f(ID_County, model="bym", graph="Kenyaadm.graph",
    constr=TRUE, scale.model=TRUE, hyper=c(0.5,5e-04))
    + f(ID_Cluster, model="iid", hyper=c(0.5,5e-04))
```

Here, y is the response and urban is the vector of ones and zeros for urban or rural cluster respectively. The model components f are a spatial effect modelled with the BYM model and independent random noise at the levels of clustering. The independent random noise is defined in model="iid". The IDs for the counties and clusters are in the vectors ID_County and ID_Cluster.

MB2

MB2 is an extension of MB1 where an additional model component, independent random noise at the household level, is added. The model becomes

$$y_{ijk} \sim \text{Binomial}(p_{ijk}, b_{ijk}), \quad \text{logit}(p_{ijk}) = \eta_{ijk},$$
$$i = 1, \ldots, n_c, \;\; j = 1, \ldots, n_{icl}, \;\; k = 1, \ldots, n_{ijh},$$

where the response y_{ijk} is the number of children in household k in cluster j in county i that died in the neonatal period and n_{ijh} is the number of households in cluster j in county i. The response y_{ijk} is linked to an additive predictor η_{ijk} through a logit link and binomial likelihood where b_{ijk} is the number of births in household k in cluster j in county i and p_{ijk} is the probabilities of neonatal death. The additive predictor becomes

$$\eta_{ijk} = \mu + \mathbb{1}[cl(j) = 1]\beta + \phi_i + \gamma_i + \nu_j + \varepsilon_k,$$
$$i = 1, \ldots, n_c, \;\; j = 1, \ldots, n_{icl}, \;\; k = 1, \ldots, n_{ijh},$$

where the terms are the same as for MB1 with an additional independent random noise effect $\varepsilon_k \sim_{\text{iid}} N(0, \sigma^2_\varepsilon)$ at the household level. The distribution of μ and β and the priors of the hyperparameters are

$$\begin{aligned}
\mu &\sim N(0, \infty), \\
\beta &\sim N(0, 1000), \\
\log\tau_\phi &\sim \text{logGamma}(0.5, 0.0005), \\
\log\tau_\gamma &\sim \text{logGamma}(0.5, 0.0005), \\
\log\tau_\nu &\sim \text{logGamma}(0.5, 0.0005), \\
\log\tau_\varepsilon &\sim \text{logGamma}(0.5, 0.0005),
\end{aligned}$$

where $\tau_\phi = 1/\sigma^2_\phi$, $\tau_\gamma = 1/\sigma^2_\gamma$, $\tau_\nu = 1/\sigma^2_\nu$ and $\tau_\varepsilon = 1/\sigma^2_\varepsilon$. The formula for the model is specified in R with the code in Listing 3.6.

Listing 3.6: Formula specifying MB2 to be used in INLA.

```
formula = y ~ 1 + urban
    + f(ID_County, model="bym", graph="Kenyaadm.graph",
    constr=TRUE, scale.model=TRUE, hyper=c(0.5,5e-04))
    + f(ID_Cluster, model="iid", hyper=c(0.5,5e-04))
    + f(ID_Household, model="iid", hyper=c(0.5,5e-04))
```

where the formula is the same as for MB1 with an additional model component at the household level where ID_Household is the vector of household IDs.

3.4 Combining design-based and model-based methods

The last method is a combination of the smoothed design-based approach outlined in Section 3.2 and the model-based methods in Section 3.3. Two models are considered. The first model contains iid effects at the cluster level (CM1), and the second model contains

iid effects at the cluster level and household level (CM2). Each of the models, CM1 and CM2, are split into two parts. The first part, SDB, is used on the strata where there is enough data to obtain design-based estimates. Here, the observations are areal data with the design-based estimates of the logit NMR for each stratum with fixed, estimated variances. The model is constructed with a Gaussian model as explained in Section 3.2. Let the response for the first part be denoted y_1.

The second part is used on the strata where there are zero observed neonatal deaths. Here, the observations are point-referenced with specific locations for the clusters or households. The observed values are modelled with a Binomial model where the response is outcomes of a set of Bernoulli trials, as outlined in Section 3.3. Let the response for the second part be denoted y_2. The combined response becomes $y = [y_1, y_2]$, constructed with the joint modelling of a Gaussian distribution for y_1 and binomial distribution for y_2.

For the strata where there are more than zero cases of neonatal mortality, rates on the logit scale can be estimated with the design-based method outlined in Section 3.1. The estimated design based rates are used as the response, $y_{1i}, i \in s_1$, where s_1 consists of the strata where the design-based rates can be estimated. For each stratum $i \in s_1$, the model becomes

$$y_{1i}|\eta_{1i} \sim N(\eta_{1i}, \sigma_i^2) \quad \eta_{1i} = \mu + \mathbb{1}[s(i) = 1]\beta + \phi_i + \gamma_i, \quad i = 1, \ldots, n_{s_1},$$

The responses $y_{1i}, i = 1, \ldots, n_{s_1}$ are the stratum rates on the logit scale, estimated with the design-based method and n_{s_1} is the number of s_1 strata. The results of this model are the smoothed design-based estimates, as in Section 3.2, with an indicator function which is 1 if stratum i is rural and 0 if urban. Thus, when there is enough data to generate design-based estimates in all strata, the combined methods reduce to SDB.

The second part of the model uses the model-based approach on the strata where there are zero occurrences of neonatal mortality in the sample. The model-based approach makes it possible to obtain estimated rates for the strata where there is not enough data to obtain design-based estimates. It is desired to explore these combined approaches in addition to purely model-based methods since the joint modelling makes it possible to use design-based estimates where they are attainable.

CM1

For CM1, the response, y_{2ij}, is zero for each child born in cluster j in the zero strata, $i \in s_2$, where s_2 is the set of strata where there are zero observed neonatal deaths in the sample. The model CM1 becomes, for s_2,

$$y_{2ij} \sim \text{Binomial}(p_{ij}, b_{ij}), \quad \text{logit}(p_{ij}) = \eta_{2ij}, \quad i = 1, \ldots, n_{s_2}, \;\; j = 1, \ldots, n_{icl},$$

where n_{s_2} is the number of s_2 strata. The response y_{2ij} is linked to an additive predictor η_{ij} through a logit link, where b_{ij} is the number of births in cluster j in stratum $i \in s_2$ and p_{ij} is the probability of death for children in cluster j in stratum i. The additive predictor is

$$\eta_{2ij} = \mu + \mathbb{1}[cl(j) = 1]\beta + \phi_i + \gamma_i + \nu_j, \quad i = 1, \ldots, n_{s_2}, \;\; j = 1, \ldots, n_{icl}.$$

The predictor and the prior distribution of the parameters are the same as in MB1. The formula for the joint model with response $y = [y_1, y_2]$ is specified in R with the code in Listing 3.7.

Listing 3.7: Formula specifying CM1 to be used in INLA.

```
formula = y ~ 1 + urban
    + f(ID_County, model="bym", graph="Kenyaadm.graph",
    constr=TRUE, scale.model=TRUE, hyper=c(0.5,5e-04))
    + f(ID_Cluster, model="iid", fixed=TRUE,
    initial=log(initial_cluster))
```

The hyperparameters of the iid effects on the cluster and household level are fixed by fixed =TRUE. This is done because it is assumed that there is not enough data in the strata where the model-based approach is used for estimation to capture the cluster- and household effects. Therefore, a fully model-based model (MB2) is fitted to the full dataset with INLA and the estimated 0.5 quantile of the precision of the cluster and household effects are extracted from the model. The value for the cluster effect to be used in Listing 3.7 is found with the code in Listing 3.8.

Listing 3.8: Extracting the hyperparameter of the cluster effect from MB2.

```
initial_cluster = MB2$summary.hyperpar$`0.5quant`[3]
```

The object summary.hyperpar is a matrix containing the mean, SD and quantiles of the hyperparameters of the model. The log of the estimated precisions are used as initial, fixed hyperparameters for the cluster- and household effects in the joint models. The values are inserted in the argument initial =log(initial _cluster) for the clusters, where initial _cluster is the estimated the precision of the cluster effect.

CM2

CM2 is an extension of CM1 where an additional model component, iid effects at the household level, is added. The model becomes, for s_2,

$$y_{2ijk} \sim \text{Binomial}(p_{ijk}, b_{ijk}), \quad \text{logit}(p_{ijk}) = \eta_{2ijk},$$
$$i = 1, \ldots, n_{s_2}, \;\; j = 1, \ldots, n_{icl}, \;\; k = 1, \ldots, n_{ijh}.$$

The response y_{2ijk} is linked to an additive predictor η_{ijk} through a logit link, where b_{ijk} is the number of births in household k in cluster j in stratum $i \in s_2$ and p_{ijk} is the probability of death. The additive predictor is

$$\eta_{2ijk} = \mu + \mathbb{1}[cl(j) = 1]\beta + \phi_i + \gamma_i + \nu_j + \varepsilon_k,$$
$$i = 1, \ldots, n_{s_2}, \;\; j = 1, \ldots, n_{cl}, \;\; k = 1, \ldots, n_h.$$

The predictor and the prior distribution of the parameters are the same as in MB2. The formula for the joint model with response $y = [y_1, y_2]$ is specified in R with the code in Listing 3.9.

Listing 3.9: Formula specifying CM2 to be used in INLA.

```
formula = y ~ 1 + urban
    + f(ID_County, model="bym", graph="Kenyaadm.graph",
    constr=TRUE, scale.model=TRUE, hyper=c(0.5,5e-04))
    + f(ID_Cluster, model="iid", fixed=TRUE,
        initial=log(initial_cluster))
    + f(ID_Household, model="iid", fixed=TRUE,
        initial=log(initial_house))
```

3.5 Details for implementation in INLA

The method of integrated nested Laplace approximations (INLA), described in Section 2.4, is used to implement the methods outlined in Sections 3.2, 3.3 and 3.4. R-INLA can be downloaded by following the steps provided here: `http://www.rinla.org` (Rue (2019)). The function inla in the INLA package performs a full Bayesian analysis of additive models using integrated nested Laplace approximations. The function returns an object of class inla with among others, summary of the hyperparameters, linear predictors, fixed effects and random effects.

Smoothed design-based method

The smoothed design-based method (SDB) in Section 3.2 is defined with the function inla as in the code in Listing 3.10.

Listing 3.10: Fitting SDB in INLA.

```
SDB = inla(formula,
           data=list(y=DB_rate, ID_County=ID_county),
           scale=1/DB_var,
           control.family=list(initial=0,fixed=TRUE),
           control.predictor=list(compute=TRUE))
```

Here, formula, is the formula provided in Listing 3.4 in Section 3.2. The data are provided in data, which is a list containing the variables in the model. For the SDB model, data consist of the response y which is the estimated design-based rates on logit scale, and the IDs for the counties. The argument scale is set to the inverse of the estimated design-based variances. control.family is used to control the hyperparameters of the model. Here, the initial value for the hyperparameters are set to zero and the argument fixed is set to true. This is done to fix the variance of the noise of the observations since the variance of the estimates already is estimated with the design-based method. The correct variance is obtained when this is combined with scale. The last argument in the SDB model is control.predictor which is used to control variables. Here, compute is set to true, which indicates that the marginals for the linear predictor are computed. As the model includes an intercept, the argument is set to true to obtain the combination of the intercept and the spatial effect.

The estimated county rates and their standard deviation are found from the linear predictor of the fitted INLA object as presented in the Code in Listing 3.11.

Listing 3.11: Extracting estimated county rates and SD from the INLA object.

```
rates = SDB$summary.linear.predictor$`0.5quant`
SD = SDB$summary.linear.predictor$sd
```

The rates are on logit scale, which is as desired for further analysis.

Model-based methods

The model-based methods, MB1 and MB2, outlined in Section 3.3 are defined with the code in Listing 3.12.

Listing 3.12: Fitting MB1 in INLA

```
MB1 = inla(formula, family="binomial", Ntrials=births,
           data=list("y"=deaths, "urban"=urban,
           "ID_County"=ID_County, "ID_Cluster"=ID_Cluster),
           control.compute=list(config=TRUE),
           control.predictor=list(compute=TRUE))
```

The argument formula is defined for MB1 and MB2 in Listings 3.5 and 3.5 in Section 3.3. The response y is a vector with the number of neonatal deaths in each cluster (MB1) or household (MB2). The argument family is a string indicating the likelihood family. The default is Gaussian with identity link and for the model-based methods the family is set to binomial with logit link. The argument Ntrials takes a vector containing the number of trials for the binomial likelihood. In this case, the number of births in each cluster (MB1) or household (MB2). For MB2, there is an additional ID identifier in the list data for the household IDs called ID_Household.

The estimated rates are found by adding new observations to the vectors, and predicting the posterior of the latent field of the observations. Two new observations are added for each county, one from the urban stratum and one from the rural stratum. This results in $47 * 2 = 94$ new observations. The new observations are added at the end of each of the vectors y, urban, ID_County, ID_Cluster, ID_Household and Ntrials. The response y and the number of trials Ntrials for the new observations are unknown and set to NA. The IDs for strata, counties, cluster and households for the new observations are defined as in Listing 3.13.

Listing 3.13: New observations for prediction.

```
new_urban = [0,1,0,1,...,0,1,0,1]
new_ID_County = [1,1,2,2,...,46,46,47,47]
new_ID_Cluster = [Max_Cluster+1,..., Max_Cluster+94]
new_ID_household = [Max_Household+1,..., Max_Household+94]
```

Here, Max_Cluster and Max_Household are the largest cluster and household ID's in the sample. Thus, the predictions are assigned to new clusters and households. Once the new observations are created and the models are fitted with the function inla, predicted rates for the new observations can be approximated. This is done with the function inla.posterior.sample. The function generates samples from the approximated posterior of the inla-object. The samples are generated as in Listing 3.14.

Listing 3.14: Generating samples from the fitted MB1.

```
MBsample = inla.posterior.sample(100,MB1,seed=123)
sample = matrix(data=NA,nrow=100,ncol=nr_counties*2)
for (i in 1:100){
  sample[i,]=MBsample[[i]]$latent[Max_Cluster+(1:nr_counties*2)
      ,]
}
```

Here, nr_counties, is the number of counties in the sample, which is 47 for Kenya. The value Max_Cluster is the largest cluster ID for MB1 and replaced by Max_Household for MB2. The samples are found in the final elements of MBsample[[i]]$latent for each sample $i, i = 1, ..., 100$. The estimated rates for each of the 94 new observations for the 100 samples are saved in the matrix called sample. First, the estimates are transformed from the logit to the probability scale. Afterwards, the estimates are weighted with the number of residents in urban and rural areas in each county to produce aggregated rates for each county for each sample. The number of residents is taken from Kenya National Bureau of Statistics et al. (2015). The final estimated county rates are found by averaging over all 100 samples and transforming them back to the logit scale. Lastly, the standard deviation of the estimates are found from the 100 samples.

Combining design-based and model-based methods

There is a priori information about the offset between the officially reported national NMR and the estimated rates from the CM methods. Adding cluster and households effect increases the mean of the mortality rates. Also, the same spatial effect is used for both parts of the model, on county level for SDB with areal data and on cluster level for MB with point referenced data. For the first part, the rates are modelled directly. For the second part, the rates are modelled indirectly through a binomial likelihood and with added cluster- and household effects. Therefore, it is not only the spatial effects that control the estimates. An offset is needed to correct for the skewness in the estimated rates that arises from the model-based approach. The skewness results in estimates of the NMR that are too high. The value for the offset is found by using the function optimize in R to minimize the squared difference between the expected rate with added cluster- and household effects and the estimated national rate of approximately 22 per 1000 live births as reported in Kenya National Bureau of Statistics et al. (2015). The combined methods, CM1 and CM2, are defined with the code in Listing 3.15.

Listing 3.15: Fitting CM1 in INLA.

```
CM1 = inla(formula, family=c("gaussian", "binomial"),
           Ntrials=Ntrials,
           data=list("y"=deaths, "urban"=urban,
           "ID_County"=ID_County, "ID_Cluster"=ID_Cluster),
           scale=scale, offset=offset,
           control.family=list(list(initial=0,fixed=TRUE),list())
               ,
           control.compute=list(config=TRUE),
           control.predictor=list(compute=TRUE))
```

The two likelihoods are defined with family=c("gaussian", "binomial"). The arguments in data, Ntrials, scale and offset are vectors split into three parts. The first elements of the vectors belong to the strata where the design-based rates can be obtained. The vectors for the first part are defined in the following way. The response y is the design-based estimates for each stratum in s_1. The stratum indicator urban is 0 for rural stratum and 1 for urban stratum. The argument ID_County is the county ID for the strata. The argument ID_Cluster is set to NA for all strata as the cluster effect is not included in the SDB method. The same is true for ID_Household in CM2. The elements in the first part of Ntrials are set to NA as this argument belongs to the binomial part. The elements in the first part of scale are set to the precision of the design-based estimates as described above. The elements in the first part of offset are set to zero as the offset is only needed in the binomial part. The likelihood for the first part is Gaussian.

The second part of the vectors belong to the stratum for the design-based method produce estimates of 0. This is the strata in s_2 where there are zero observed neonatal deaths in the sample. Here, the response y is set to NA as this is the rates that are unattainable with the design-based method and are to be predicted. urban is again set to 0 for rural strata and 1 for urban strata. The county IDs are again the county ID for the strata with zero deaths. Again, the IDs for the household and clusters are set to NA. The elements in the second part of Ntrials are again set to NA. The parameter scale is set to 0 as the design-based variances are unattainable. Lastly, the parameter offset are again set to 0. The likelihood is again Gaussian for the second part.

The third and last elements of the vectors are similar to the vectors in the model-based methods. The response y and Ntrials are the number of observed neonatal deaths and the number of births in each cluster or household in the strata with zero deaths s_2. As there are zero observed deaths, y is a vector of only zeros. The argument urban is 0 for clusters or households in rural strata and 1 for clusters or households in urban strata. The argument ID_County is the county ID for which the clusters or households belong to. The IDs for the clusters and households is the elements in the third part of ID_Cluster and ID_Household. The parameter scale is set to 0 and offset is the offset values found as described above. The likelihood for this last part is binomial.

Estimates of the strata rates and their standard deviation are approximated from generated samples of the posterior of the linear predictor according to the areal data model. The samples are generated as in Listing 3.16.

Listing 3.16: Generating samples from the fitted CM1.

```
CMsample = inla.posterior.sample(100,CM1, seed=123)
sample = matrix(data=NA, nrow=100, ncol=nr_strata)
for (i in 1:100){
  sample[i,]=CMsample[[i]]$latent[1:nr_strata,]
}
```

Here, nr_strata $=s_1+s_2$, is the total number of strata in the sample, which is 92 for Kenya. The strata rates and their standard deviation estimated with the smoothed design-based method are the first elements in the posterior sample (s_1), while the rest of the elements are the estimates predicted with the help of the model-based method (s_2). The estimated rates are found in the output CMsample[[i]]$latent for each $i = 1, ..., 100$ sample.

The estimated rates for each stratum for the 100 samples are saved in the matrix called sample. First, the estimates are transformed from the logit to the probability scale. Afterwards, the estimates are weighted with the number of residents in urban and rural areas in each cluster to produce aggregated rates for each county for each sample. Thus, a final aggregated county rate may either consist of SDB rates or MB rates for both strata, or one SDB rate and one MB rate. The final estimated county rates are found by averaging over all 100 samples and transforming them back to the logit scale for further analysis. Lastly, the standard deviation of the estimates are found from the 100 samples.

Chapter 4

Simulation Study

The methods outlined in Chapter 3 are tested with a simulation study. The estimates are evaluated according to the scoring rules outlined in Section 2.5 and average running time. The design of the simulation study is described in Section 4.1 and Sections 4.2, 4.3 and 4.4 present the results. Finally, all findings are discussed in Section 4.5.

4.1 Design of study

The purpose of conducting the simulation study is to evaluate the set of proposed methods within a controlled framework. When working with real survey data, the truth is not known, as that would require a census of the population. Since there does not exist exact answers, it is difficult to evaluate and compare the methods on the real data. When conducting a simulation study, one can simulate rates such that the estimated rates obtained with the methods can be evaluated against 'true' simulated rates.

There are several goals of the study. First, find the best method based on the set of scoring rules outlined in Section 2.5 and running time. Secondly, investigate what kinds of conditions that make it challenging for the methods to predict mortality rates. Lastly, explore if the proposed combined methods are better than design-based and model-based approaches.

In the simulations, the survey design is retained. The stratified, two-stage cluster design with the true weights and number of births in each household are kept as in the real data set and the neonatal deaths are simulated. Therefore, this study will only be valid for this specific sample, as the clusters and households are retained. The general performance of the methods is not tested, only the performance on this data set. However, one can expect similar results for other surveys.

Occurrences of neonatal deaths are simulated using specific 'true' rates. The number of simulations N for each scenario is pre-decided to a number large enough such that there are at least 100 simulations where the design-based method can obtain estimates of the NMR for each county ($N > 100$). A set of different scenarios are considered to evaluate the methods both on simple scenarios, and on more complex ones to see how the methods

handle effects such as the effect of rural or urban areas and correlation within clusters and households.

The number of simulated neonatal deaths y_{ijk} in household k in cluster j in county i are generated by a binomial distribution

$$\begin{aligned} y_{ijk} &\sim \text{Binomial}(p_{ijk}, b_{ijk}), \quad \text{logit}(p_{ijk}) = \eta_{ijk}, \\ i &= 1, \ldots, n_c, \;\; j = 1, \ldots, n_{icl}, \;\; k = 1, \ldots, n_{ijh} \end{aligned} \tag{4.1}$$

The number of reported births in household k in cluster j in county i from the real survey data is denoted b_{ijk} and η_{ijk} is the link function on the logit scale. The link function depends on the type of simulated scenarios. The simulated deaths are used to obtain neonatal mortality rates at the county level with the methods outlined in Section 3.

Cases

Three cases of spatial dependencies are considered, where the added spatial effect is denoted u. In the first case, scenarios without spatial effects are considered i.e. $u = 0$. The main purpose of these simulations is a sanity check, as this case should work for all approaches. The results of the simulations with no spatial effect are presented in Section 4.2.

In the second case, an unstructured spatial effect at the county level is added. The spatial effects are denoted u_i for counties $i = 1, ..., n_c$. This case can be appropriate in a scenario where counties make interventions independently of each other, such that the variation in the mortality rates are random. The unstructured spatial effect is normally distributed with mean 0 and standard deviation $\sigma_u = 0.25, u_i \sim N(0, \sigma_u^2)$. The value for σ_u was found by fitting a fully model-based model (MB2) to the real data and estimating the standard deviation of the total spatial effect. The results of the simulations with unstructured spatial effects are presented in Section 4.3.

In the third and last case, a structured spatial effect at the county level is added. Structured spatial variation can arises, for example, if the mortality rates are affected by phenomenons that are dependent and vary in space in a structured way. The spatial effects u_i are sampled from the precision matrix $\boldsymbol{Q}$ of the neighbourhood structure of the counties with the code in Listing 4.1.

Listing 4.1: Generating structured spatial effects.

```
Q = INLA:::inla.pc.bym.Q("Kenyaadm1.graph")
sigma = 0.25
Q_tmp = inla.scale.model(Q, constr=list(A=matrix(1,nrow=1,ncol
    =47),e=matrix(0,nrow=1,ncol=1)))+Diagonal(47, 1e-9)
Q_new = Q_tmp/sigma^2
u = inla.qsample(n,Q_new, constr=list(A=matrix(1,nrow=1,ncol=47),
    e=matrix(0,nrow=1,ncol=1)),seed=123)
```

INLA:::inla.pc.bym.Q constructs a sparse symmetric neighbourhood matrix, where the non-zero pattern is defined by the graph. The number of neighbours for each county is on the diagonal, and the entries for the neighbours are set to -1 as given in Equation (2.14) in Section 2.2. The function inla.scale.model scales the neighbourhood matrix such that the

geometric mean of the marginal variances is one. The neighbourhood matrix is divided by a variance $\sigma_u^2 = 0.25^2$, yielding the precision matrix $\boldsymbol{Q}$. Lastly the spatial effects u_i are generated from $N(0, Q^{-1})$ with a sum-to-zero constraint with the function inla.qsample. The results of the simulations with structured spatial effects are presented in Section 4.4.

Scenarios

For each case, four different scenarios are considered. In the first scenario, occurrences of neonatal mortality are simulated with the same rate for all counties, strata, clusters, and households. The second scenario is simulated with a stratum effect, resulting in separate rates for urban and rural clusters. In the third scenario, occurrences of neonatal mortality are simulated with iid effects at the cluster- and household levels, resulting in different rates for each household. In the fourth and last scenario, occurrences of neonatal mortality are simulated with both a stratum effect and iid effects at the cluster- and household elevels. Again, resulting in different rates for each household, also separating households in urban and rural clusters.

The first scenario is simulated with a baseline rate of μ for all counties, clusters, and households and a spatial effect u depending on the case. The resulting link function η_{ijk} for household k in cluster j in county i is

$$\eta_{ijk} = \mu + u_i, \quad i = 1, \ldots, n_c, \;\; j = 1, \ldots, n_{icl}, \;\; k = 1, \ldots, n_{ijh}, \tag{4.2}$$

where u_i is the spatial effect for county i. For the simulations with no spatial effects, $u_i = 0$ for all counties, $i = 1, ..., n_c$, where n_c is the number of counties. County rates are estimated with the methods in Section 3 from the simulated occurrences of neonatal deaths obtained by Equation (4.1) with the link function in Equation (4.2). The estimated county rates $\hat{r}_i$ on logit scale are compared to the true rate on logit scale for each county i

$$r_i = \mu + u_i, \quad i = 1, \ldots, n_c. \tag{4.3}$$

The second scenario contains a stratum effect on the urban and rural clusters. The effect is one for urban clusters and zero for rural clusters. The values of the stratum effect are motivated by the higher estimated NMR of 26 per 1000 births for the urban sample, compared to 21 per 1000 births for the rural sample as reported on pages 363 and 365 in Kenya National Bureau of Statistics et al. (2015). With the stratum effect, the rate for each county becomes a weighted sum of the rate for the urban and rural strata within the county. The link function for cluster j in county i is

$$\eta_{ijk} = \mu + u_i + \mathbb{1}[cl(j) = 1], \quad i = 1, \ldots, n_c, \;\; j = 1, \ldots, n_{icl}, \;\; k = 1, \ldots, n_{ijh}, \tag{4.4}$$

where the spatial effect is denoted u_i and the indicator function is 1 if household k is in an urban cluster j and 0 if household k is in a rural cluster j. County rates are estimated with the methods in Section 3 from the simulated occurrences of neonatal deaths obtained by Equation (4.1) with the link function in Equation (4.4). The estimated county rates $\hat{r}_i$ are compared to the true probability for each county i

$$r_i = r_{iu}\frac{m_{iu}}{m_i} + r_{ir}\frac{m_{ir}}{m_i}, \quad i = 1, \ldots, n_c, \tag{4.5}$$

where $m_{iu} + m_{ir} = m_i$ is the number of residents in urban and rural clusters respectively in county i, as reported in the Kenya 2009 Population and Housing census, KNBS (2012). The stratum effect results in the following urban rate r_{iu} and rural rate r_{ir}

$$r_{iu} = \text{expit}(\text{logit}(\mu) + 1 + u_i)$$
$$r_{ir} = \text{expit}(\text{logit}(\mu) + u_i)$$

where μ is again the baseline rate and $\text{expit}(\cdot)$ is the inverse of $\text{logit}(\cdot)$.

The third scenario contains iid effects at the cluster- and household level. The link function for household k in cluster j county i is

$$\eta_{ijk} = \mu + u_i + \nu_j + \varepsilon_k, \quad i = 1, \ldots, n_c, \;\; j = 1, \ldots, n_{icl}, \;\; k = 1, \ldots, n_{ijh}, \tag{4.6}$$

with baseline rate μ, spatial effect u_i and iid cluster- and household effects $\nu_j \sim_{iid} N(0, \sigma_\nu^2)$ and $\varepsilon_k \sim_{iid} N(0, \sigma_\varepsilon^2)$. The variance of the cluster and household effects, σ_ν^2 and σ_ε^2 are chosen by fitting the model-based model MB2 on the real data and using the 0.5 quantile of the precision of the hyperparameters as the inverse of the variances. The standard deviation of the effects are $\sigma_\nu = 0.046$ and $\sigma_\varepsilon = 0.052$.

County rates are estimated with the methods in Section 3 from the simulated occurrences of neonatal deaths obtained by Equation (4.1) with the link function in Equation (4.6). The estimated county rates $\hat{r}_i$ are compared to the true rates

$$r_i = \text{E}[\text{expit}(\text{logit}(\mu) + u_i + \nu_j + \varepsilon_k)], \quad i = 1, \ldots, n_c, \tag{4.7}$$

where the expectation is taken over a large sample under the assumption of an infinite number of clusters and households.

The fourth and last scenario considered contains a stratum effect and iid effects at the cluster- and household level. This scenario is the one that is expected to resemble to real situation the most. It is reasonable to believe that the mortality rate vary between urban and rural areas, are similar for children within the same cluster and even more similar for children within the same households. The resulting link function for household k in cluster j county i is

$$\begin{gathered} \eta_{ijk} = \mu + u_i + \mathbb{1}[cl(j) = 1] + \nu_j + \varepsilon_k, \\ i = 1, \ldots, n_c, \;\; j = 1, \ldots, n_{icl}, \;\; k = 1, \ldots, n_{ijh}, \end{gathered} \tag{4.8}$$

with baseline rate μ, spatial effect u_i and an urban or rural cluster indicator. Again, cluster and household effects ν_j and ε_k are iid. County rates are estimated with the methods in Chapter 3 with the simulated occurrences of neonatal deaths obtained by Equation (4.1) with the link function in Equation (4.4). The estimated county rates $\hat{r}_i$ are compared to the true probability for each county i

$$r_i = r_{iu}\frac{m_{iu}}{m_i} + r_{ir}\frac{m_{ir}}{m_i}, \quad i = 1, \ldots, n_c. \tag{4.9}$$

The true, simulated rates for each stratum type, urban and rural, are generated by assuming an infinite number of clusters and households in the sample

$$r_{iu} = \text{E}[\text{expit}(\text{logit}(\mu) + u_i + 1 + \nu_j + \varepsilon_k)],$$
$$r_{ir} = \text{E}[\text{expit}(\text{logit}(\mu) + u_i + \nu_j + \varepsilon_k)].$$

Again, the urban and rural rates are separated by increasing the urban rate as in scenario 2.

For scenarios 2 and 4 with stratum effect, the baseline rate μ is corrected for the increased expectation in the rates arising from the added stratum effect. The baseline rate, μ, is decreased such that on average, one obtains simulated rates of $r_i \approx \mu_0 = 0.022$ when the stratum rates are aggregated to a combined county rate. Hhere, μ_0 is the estimated neonatal mortality rate from the full sample as reported on page 361 in Kenya National Bureau of Statistics et al. (2015).

Methods

The methods described in Chapter 3 are used to estimate neonatal mortality rates on the logit scale for each of the three spatial cases and the four scenarios. The first method is the design-based method (DB) described in Section 3.1. The second method is the smoothed design-based approach (SDB) outlined Section 3.2. The third approach is the model-based methods MB1 and MB2 described in Section 3.3. The final method is the combined methods CM1 and CM2 where design-based- and model-based approaches are combined as described in Section 3.4.

Due to the randomness of the simulations, some simulations will result in zero simulated neonatal deaths in one or more counties. Then, as mentioned in Sections 3.1 and 3.2, DB and SDB cannot be used to obtain estimates of the rates. Therefore, the results from these two models are only presented for the subset of the simulations where all counties have occurrences of neonatal deaths and design-based estimates can be obtained. The number of these simulations is denoted n, which is a subset of the total number of simulations denoted N. The two other methods, MB and CM, can obtain estimates of the rates also when some of the counties have zero neonatal deaths. Thus, the results from these models are presented for all N simulations.

Model assessment

The performance of each method is evaluated with the measures presented in Section 2.5: MAE, RMSE, CRPS, MBE, and on average running time. The estimated county rates on the logit scale $\hat{r}$ from each method are evaluated against the 'true' simulated county rates r, and the results are presented as tables and boxplots. The tables display the mean value score of MAE, RMSE, CRPS and MBE on the logit scale, averaged over all the simulations, and the average running time in seconds for one run of the simulations for each of the methods. The first four rows of the tables present the mean value of the scores and average running time in seconds for DB, SDB, MB1 and MB2 on the subset of n simulations where design-based estimates can be obtained. The next four rows present the mean value of the scores and the average running time in seconds for CM1, CM2, MB1 and MB2 on all N simulations. The best scores are colored yellow.

The figures present boxplots of the scores of MAE, RMSE, CRPS and MBE on the logit scale averaged over all counties for each simulation. The first four boxplots of each figure present the estimates from DB, SDB, MB1 and MB2 on the subset of n simulations where design-based estimates can be obtained. The next four boxplots of each figure present the estimates from CM1, CM2, MB1 and MB2 on all N simulations.

4.2 Case 1: No spatial effect

The four scenarios of case 1, (1.1, 1.2, 1.3 and 1.4) are simulated without a spatial effect. The spatial effects u_i are set to zero for all four scenarios.

Case 1.1

For case 1.1, no spatial effects or other effects are included. Neonatal deaths are simulated with the link function in Equation (4.2). In total $N = 110$ simulations were conducted and $n = 105$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.3). The mean value of the scores for case 1.1 are reported in Table 4.1. Box-plots of the scores for case 1.1 are found in Figure 4.1.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.3256	0.4341	0.2449	-0.0760	14.6
SDB	0.1997	0.2919	0.1567	-0.1740	17.4
MB1	0.0436	0.0496	0.0359	0.0012	20.5
MB2	0.0451	0.0515	0.0385	0.0014	274.9
CM1	0.3206	0.4505	0.2562	-0.2889	72.1
CM2	0.3197	0.4497	0.2553	-0.2875	146.6
MB1	0.0450	0.0509	0.0365	-0.0013	20.5
MB2	0.0463	0.0527	0.0391	-0.0011	274.6

Table 4.1: Case 1.1: No effects. Mean value of the scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

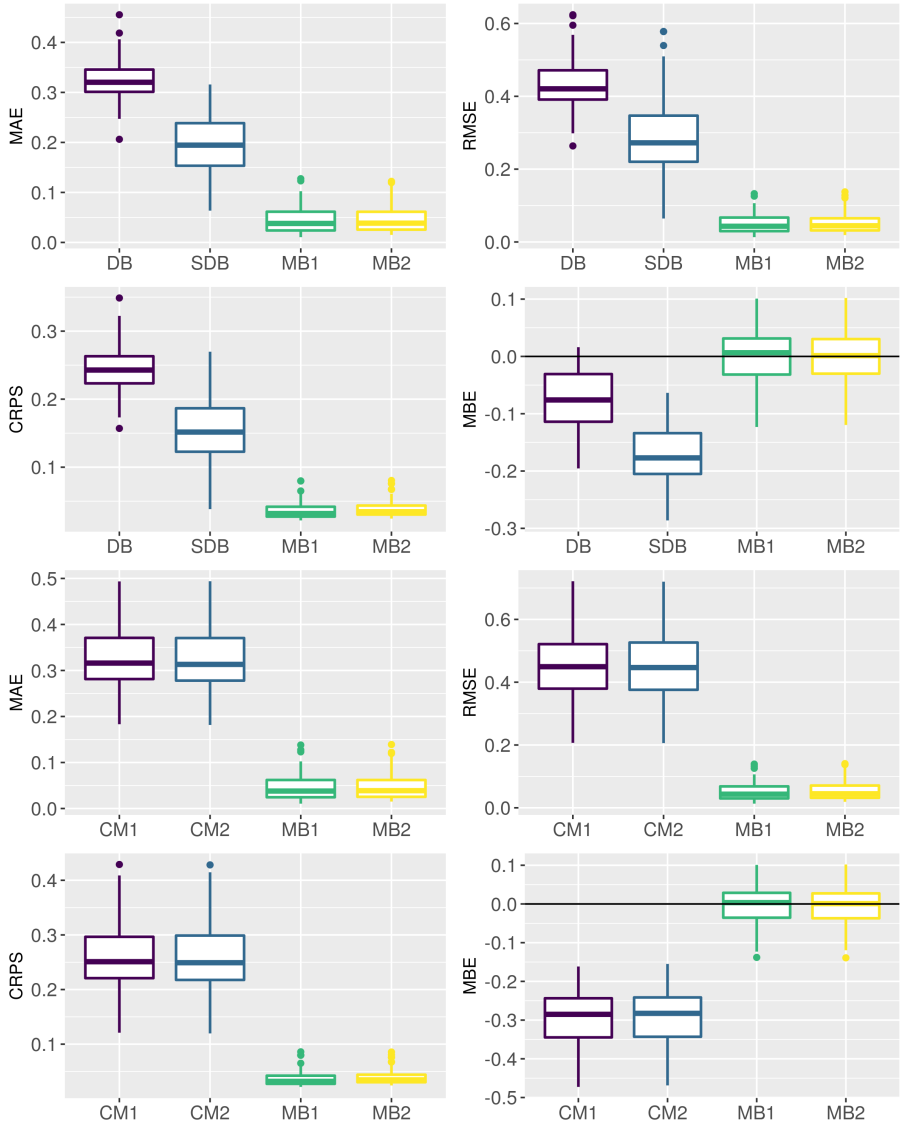


Figure 4.1: Case 1.1: No effects. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

For case 1.1, one can see from Table 4.1 that MB1 has the lowest average score of MAE, RMSE and CRPS on the subset of n simulations. In addition, MB1 has the smallest MBE. DB has the lowest running time but is outperformed by SDB on MAE, RMSE and CRPS. However, the SDB introduces some bias to the estimates. The results in Figure 4.1 show that the model-based methods, MB1 and MB2, perform better than the other methods on all four scoring rules. There is no clear difference in performance between MB1 and MB2 indicating that the household effect included in MB2 is excessive. Also, Figure 4.1 show that SDB has the largest spread in scores and that CM1 and CM2 are clearly outperformed by MB1 and MB2 on all N simulations. The best method for case 1.1 is MB1.

Case 1.2

Case 1.2 is simulated with a fixed stratum effect and no spatial effect. The link function in Equation (4.4) is used. In total $N = 130$ simulations were conducted and $n = 109$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.5). The mean value of the scores for case 1.2 are reported in Table 4.2. Box-plots of the scores for case 1.2 are found in Figure 4.2.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.5282	0.6921	0.4377	-0.4376	14.8
SDB	0.5173	0.6588	0.4267	-0.5127	17.7
MB1	0.3304	0.3863	0.2731	-0.3303	21.1
MB2	0.3309	0.3938	0.2691	-0.3306	252.7
CM1	0.7133	0.8434	0.5926	-0.7077	74.7
CM2	0.7127	0.8425	0.5917	-0.7070	181.8
MB1	0.3361	0.3913	0.2783	-0.3359	20.7
MB2	0.3367	0.3988	0.2744	-0.3364	254.3

Table 4.2: Case 1.2: Stratum effect and no spatial effect. Mean value of the scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

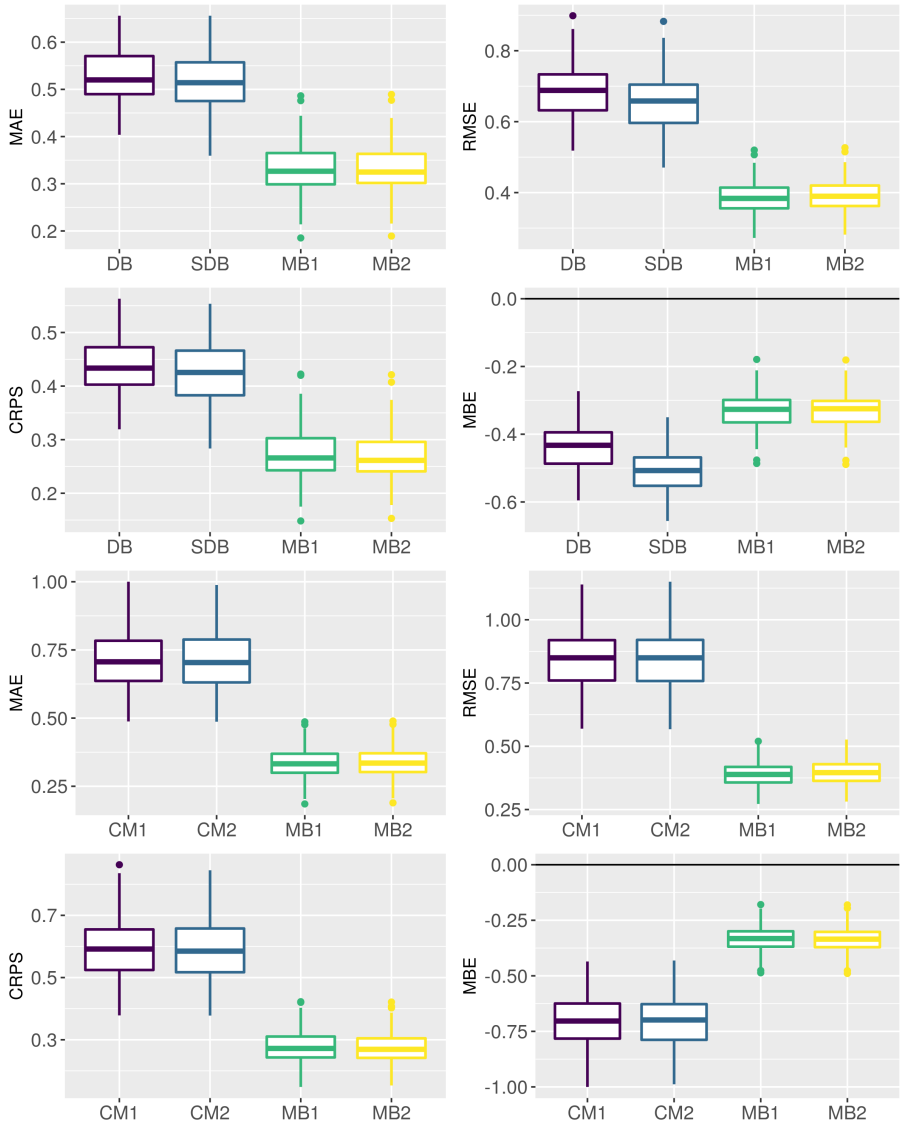


Figure 4.2: Case 1.2: Stratum effect and no spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

In Table 4.2, one can observe that the added stratum effect causes a large increase in bias for all methods. The negative MBE indicates that all methods predict lower rates than the simulated rate. MB1 has the lowest average score of MAE, RMSE and smallest MBE. MB2 has the best CRPS score but the longest average running time. Again, the model-based methods outperform the combined methods. CM2 has better results than CM1 on all scoring rules but the household effect results in longer running time. From Figure 4.2 one can see that SDB does not improve the DB estimates as much as in case 1.1. Comparing the scores for CM1 and CM2, one can see that CM2 has slightly larger first and third quartiles than CM1. Considering the scores of MAE, RMSE and MBE is MB1 the preferable method, while MB2 is the preferable method considering the CRPS score.

Case 1.3

Case 1.3 is simulated with iid cluster- and household effects and no spatial effect. The link function is in Equation (4.6). In total $N = 150$ simulations were conducted and $n = 140$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.7). The mean value of the scores for case 1.3 are reported in Table 4.3. Box-plots of the scores for case 1.3 are found in Figure 4.3.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.3379	0.4449	0.2556	-0.0996	11.3
SDB	0.2193	0.3069	0.1719	-0.1951	12.8
MB1	0.0733	0.0775	0.0531	-0.0191	6.9
MB2	0.0733	0.0782	0.0535	-0.0188	161.6
CM1	0.3446	0.4766	0.2761	-0.3078	30.6
CM2	0.3423	0.4748	0.2745	-0.3055	83.4
MB1	0.0736	0.0781	0.0531	-0.0194	6.9
MB2	0.0735	0.0786	0.0535	-0.0192	161.6

Table 4.3: Case 1.3: Cluster- and household effects and no spatial effect. Mean value of the scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

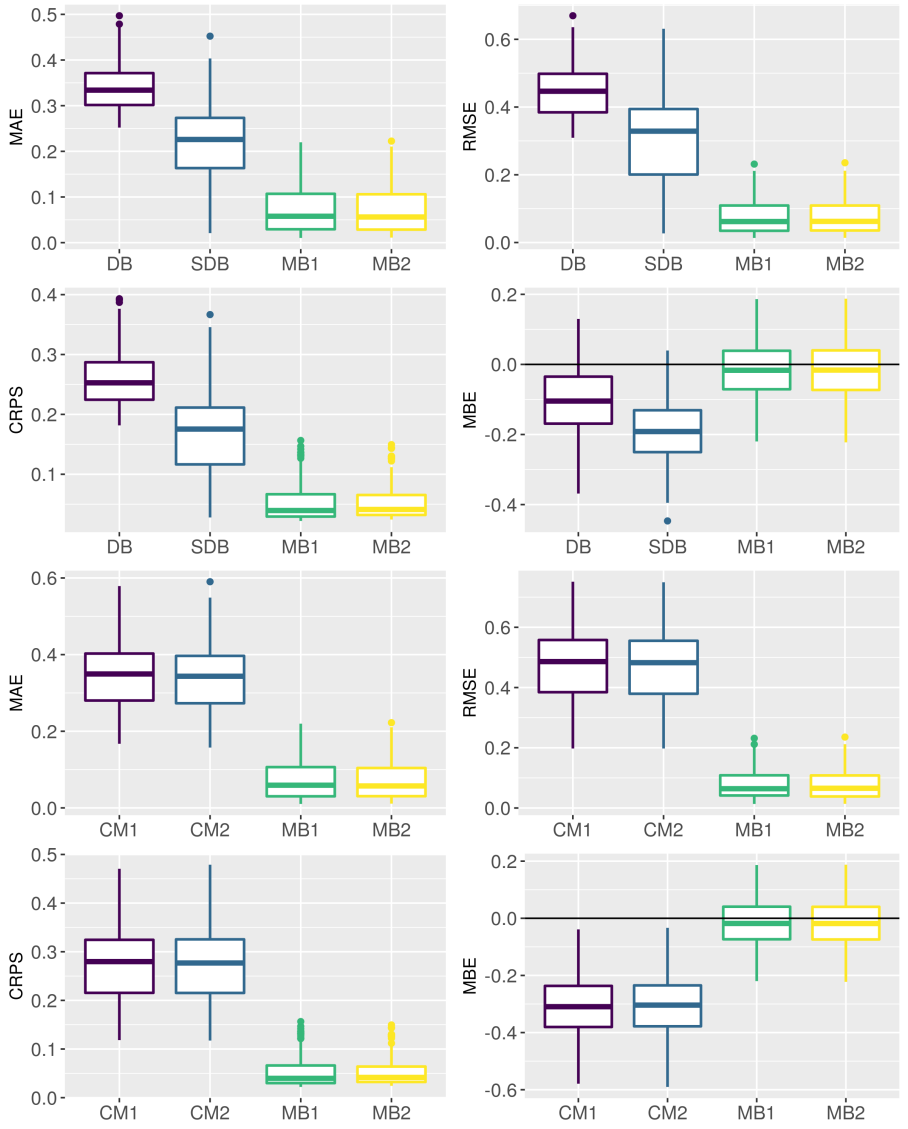


Figure 4.3: Case 1.3: Cluster- and household effects and no spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

The results from case 1.3 indicate that all methods perform better on data simulated with cluster- and household effects compared to case 1.2 simulated with a stratum effect. From Table 4.3 one can see that all methods have a tendency to underestimate the rates, from the small negative MBE. Again, the model-based methods perform best on all scoring rules, with smaller MBE for MB2, but, in general, very similar results for MB1 and MB2. One can also observe that CM1 and CM2 perform poorer than DB and SDB, indicating that the methods are poor at predicting the rates for the simulations where zero deaths occurs. From the box-plots in Figure 4.3, it is clear that the model-based methods perform best when cluster and household effects are considered. SDB performs better than DB but with a larger spread in the scores. For case 1.3 are MB1 and MB2 best, with very similar results except for the average running time, where MB1 is a much faster method.

Case 1.4

Case 1.4 is simulated with a fixed stratum effect, cluster- and household effects and no spatial effect. The link function in Equation (4.8) is used. In total $N = 200$ simulations were conducted and $n = 161$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.9). The mean value of the scores for scenario 4 are reported in Table 4.4. Box-plots of the scores for case 1.4 are found in Figure 4.4.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.5445	0.7055	0.4541	-0.4556	7.1
SDB	0.5364	0.6728	0.4449	-0.5322	7.9
MB1	0.3469	0.4037	0.2916	-0.3461	4.9
MB2	0.3470	0.4108	0.2873	-0.3458	152.5
CM1	0.7318	0.8572	0.6107	-0.7269	31.3
CM2	0.7324	0.8576	0.6110	-0.7279	68.5
MB1	0.3504	0.4067	0.2942	-0.3494	4.9
MB2	0.3506	0.4139	0.2902	-0.3494	151.9

Table 4.4: Case 1.4: Stratum effect, cluster- and household effects and no spatial effect. Mean value of the scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

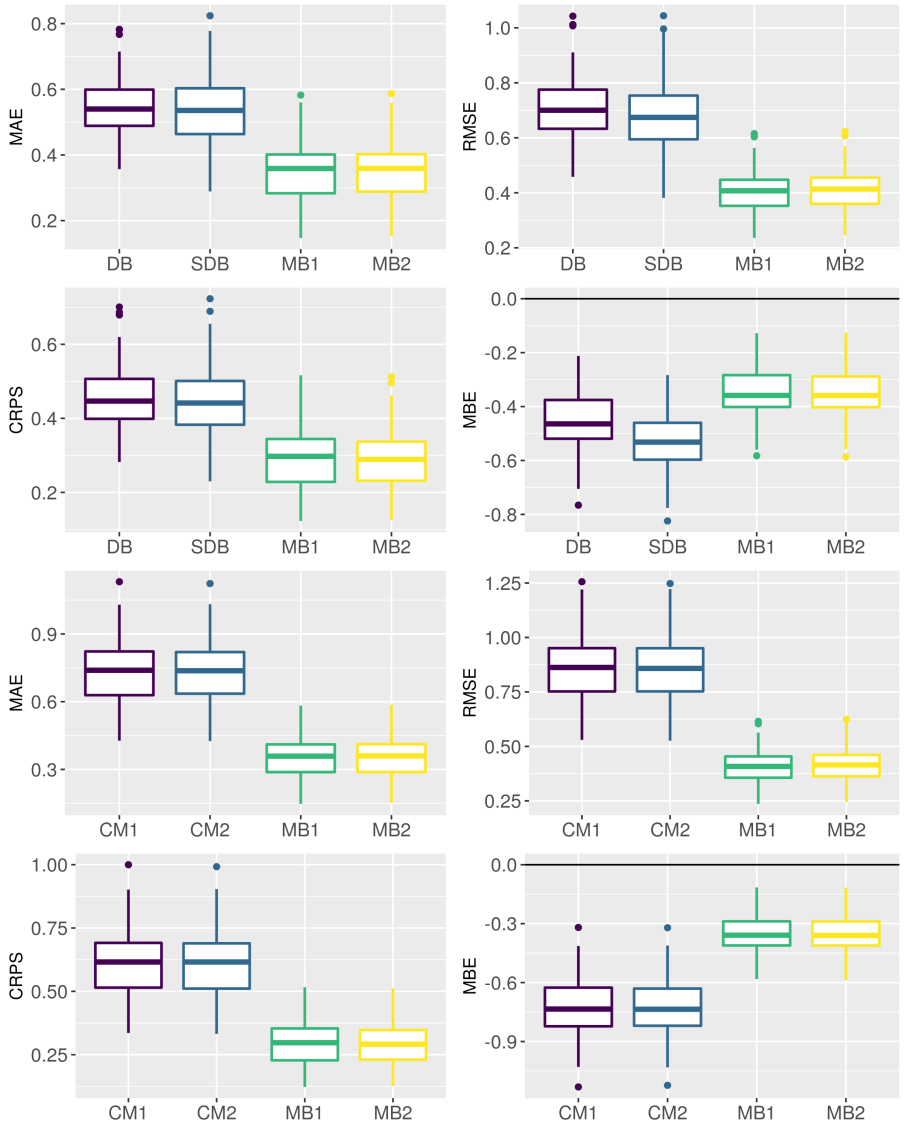


Figure 4.4: Case 1.4: Stratum effect, and cluster- and household effects and no spatial effect. Boxplots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

Table 4.4 shows that the methods again have large negative MBE due to the stratum effect. MB1 has the best scores for MAE and RMSE, and MB2 has the best CRPS score. MB1 has the best running time. From Figure 4.4 one can again, as in case 1.2, see that SDB only slightly improves the DB estimates when the stratum effect is included. The model-based methods are better than the combined methods on all the scoring rules. MB1 and MB2 are the best methods for case 1.4.

	Measure			
Method	MAE	RMSE	CRPS	Time (s)
DB	0.4340	0.5692	0.3481	11.9
SDB	0.3682	0.4826	0.3000	13.9
MB1	0.1986	0.2293	0.1634	13.4
MB2	0.1991	0.2336	0.1621	210.4
CM1	0.5276	0.6569	0.4339	52.2
CM2	0.5268	0.6561	0.4332	120.1
MB1	0.2013	0.2318	0.1655	13.3
MB2	0.2018	0.2360	0.1643	210.6

Table 4.5: Average scores of MAE, RMSE and CRPS and running time across all scenarios in case 1 (1.1, 1.2, 1.3 and 1.4). The first four rows display the average scores for methods DB, SDB, MB1 and MB2 for the subset of n simulations. The next four rows display the average scores for methods CM1, CM2, MB1 and MB2 for all N simulations.

Table 4.5 presents the average scores of MAE, RMSE and CRPS for DB, SDB, CM1, CM2, MB1 and MB2 across all four scenarios in case 1, and average running time in seconds. MB1 has overall the best score for MAE and RMSE. DB has the lowest running time and MB2 has the best CRPS score. As there are no spatial effects included in the scenarios, it is not evident why SDB performs better than DB. However, SDB is still believed to reduce variance which can explain these results. MB1 the preferable method across all 4 scenarios.

4.3 Case 2: Unstructured spatial effect

The four scenarios of case 2 (2.1, 2.2, 2.3 and 2.4) are simulated with unstructured spatial effects $u_i \sim N(0, \sigma_u^2)$ for counties $i = 1, \ldots, n_c$.

Case 2.1

Case 2.1 is simulated with an unstructured spatial effect and constructed with the link function in Equation (4.2). In total $N = 110$ simulations were conducted and $n = 103$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.3). The mean value of the scores for case 2.1 are reported in Table 4.6. Box-plots of the scores for case 2.1 are found in Figure 4.5.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.4278	0.5475	0.3230	-0.0763	16.6
SDB	0.3291	0.4330	0.2631	-0.1640	18.6
MB1	0.2640	0.3210	0.2180	0.0064	11.9
MB2	0.2639	0.3209	0.2135	0.0064	254.9
CM1	0.4939	0.6395	0.4781	-0.2124	50.2
CM2	0.4939	0.6395	0.4781	-0.2122	122.5
MB1	0.2676	0.3242	0.2211	-0.0098	11.9
MB2	0.2674	0.3242	0.2165	-0.0099	255.8

Table 4.6: Case 2.1: Unstructured spatial effect. Mean value of scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

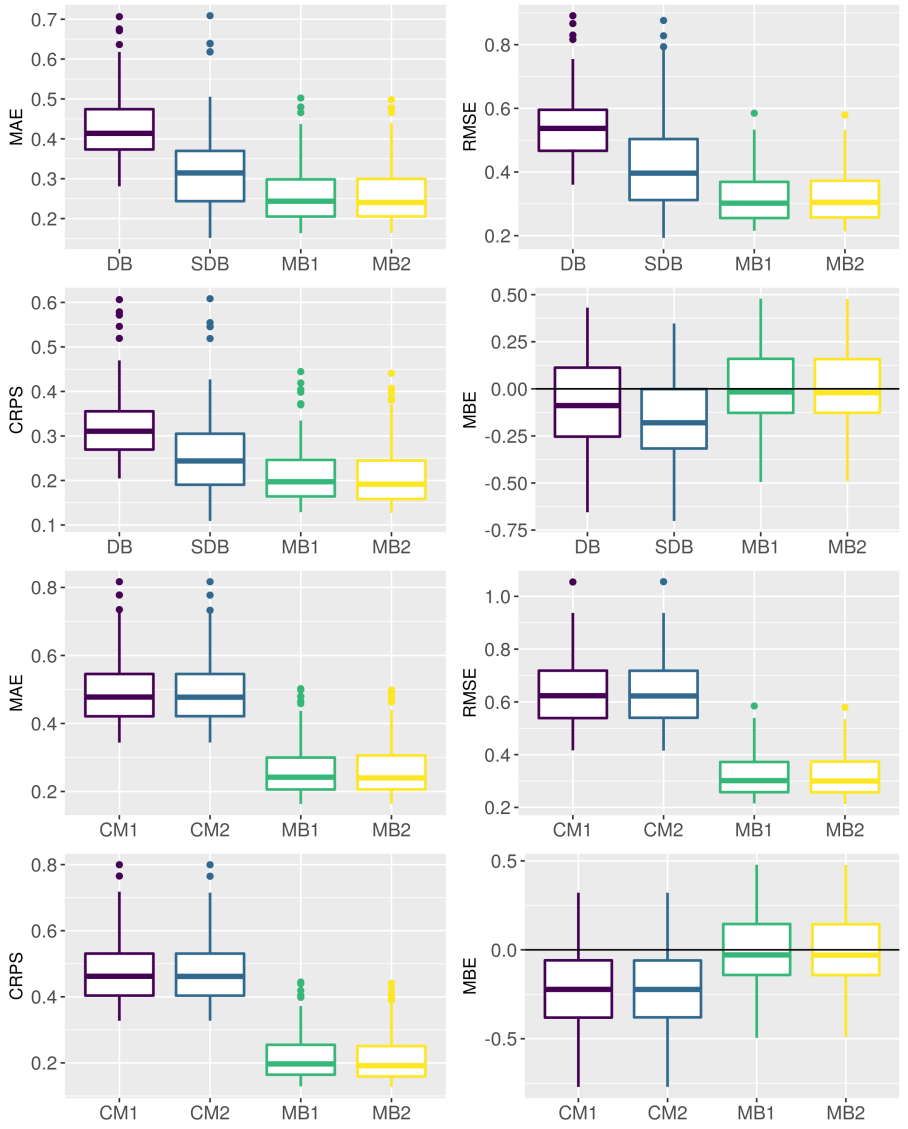


Figure 4.5: Case 2.1: Unstructured spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

In Table 4.6, SDB has better average scores of MAE, RMSE and CRPS than DB. However, DB has smaller MBE compared to SDB. The lowest scores of MAE, RMSE, CRPS and MBE across all methods is with method MB2, and MB1 has the shortest running time. CM1 and CM2 have the poorest results for all scoring rules. Figure 4.5 shows that all methods produce several high outliers and that the model-based methods are superior.

Case 2.2

Case 2.2 is simulated with a fixed stratum effect and an unstructured spatial effect. The link function in Equation (4.4) is used. In total $N = 150$ simulations were conducted and $n = 111$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.5). The mean value of the scores for case 2.2 are reported in Table 4.7. Box-plots of the scores for case 2.2 are found in Figure 4.6.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.5855	0.7460	0.4891	-0.4113	14.3
SDB	0.5467	0.6963	0.4655	-0.4894	16.1
MB1	0.4000	0.4795	0.3482	-0.3045	10.8
MB2	0.4023	0.4855	0.3456	-0.3043	244.5
CM1	0.7675	0.9067	0.6540	-0.7394	49.7
CM2	0.7667	0.9057	0.6529	-0.7382	128.3
MB1	0.4438	0.5238	0.3897	-0.3691	10.5
MB2	0.4457	0.5296	0.3864	-0.3689	242.3

Table 4.7: Case 2.2: Stratum effect and unstructured spatial effect. Mean value of scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

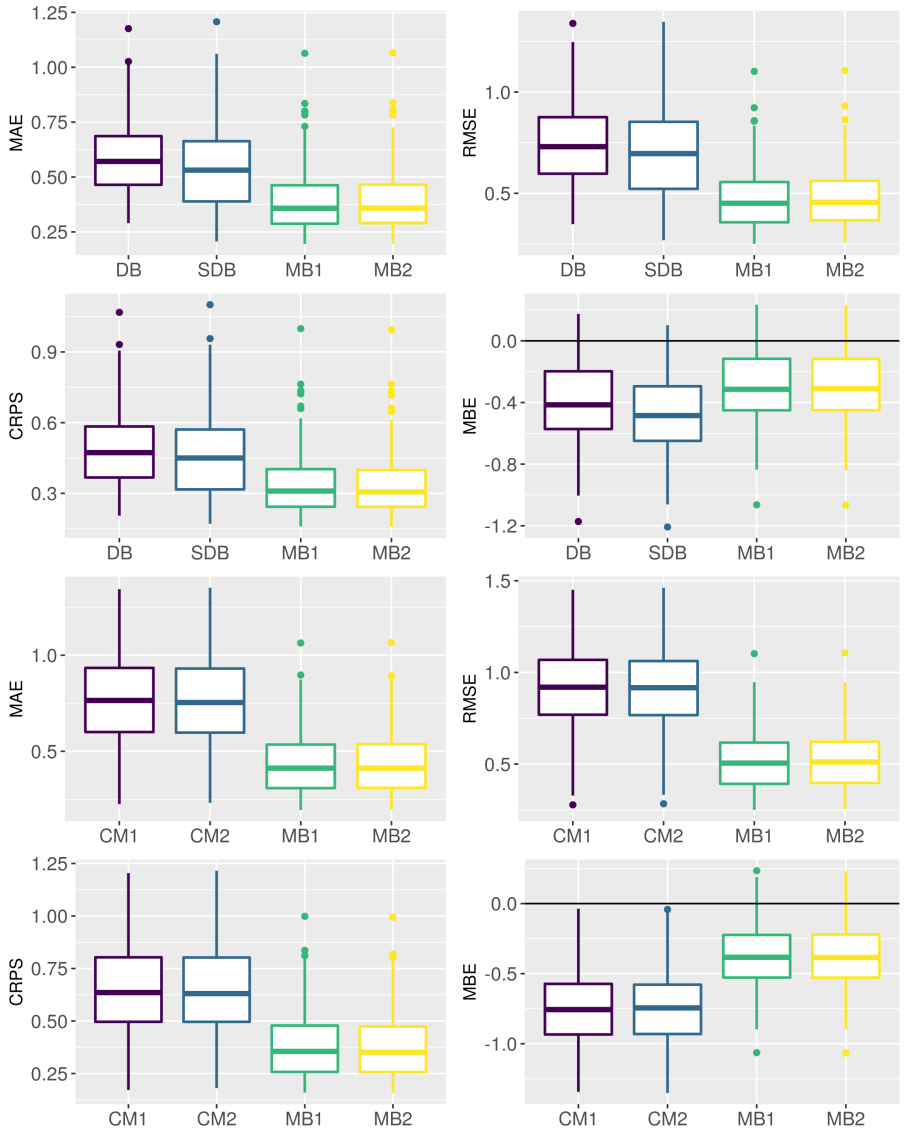


Figure 4.6: Case 2.2: Stratum effect and unstructured spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

In Table 4.7, MB1 has the lowest average score of MAE, RMSE, and shortest running time. MB2 has lowest CRPS score and smallest bias. SDB has better results on the scoring rules compared to DB, except for the bias. CM1 and CM2 perform considerably weaker than MB1 and MB2 on all measures. Again, Figure 4.6 shows that MB1 and MB2 perform the best on case 2.2, while SDB only slightly improves the DB estimates with the added stratum effect. The household effects in CM2 and MB2 do not improve the overall predictive performance of the methods. MB1 and MB2 have the best results on case 2.2.

Case 2.3

Case 2.3 is simulated with iid cluster- and household effects and an unstructured spatial effect. The link function in Equation (4.6) is used. In total $N = 150$ simulations were conducted and $n = 138$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.7). The mean value of the scores for case 2.3 are reported in Table 4.8. Box-plots of the scores for case 2.3 are found in Figure 4.7.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.4566	0.5810	0.3484	-0.0821	15.5
SDB	0.3636	0.4739	0.2967	-0.1720	17.6
MB1	0.2947	0.3510	0.2483	-0.0023	11.5
MB2	0.2949	0.3514	0.2435	-0.0022	249.6
CM1	0.4537	0.5858	0.3742	-0.3164	49.8
CM2	0.4526	0.5848	0.3732	-0.3146	113.9
MB1	0.3038	0.3604	0.2566	-0.0304	11.2
MB2	0.3039	0.3608	0.2519	-0.0301	247.5

Table 4.8: Case 2.3: Cluster- and household effects and unstructured spatial effect. Mean value of scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

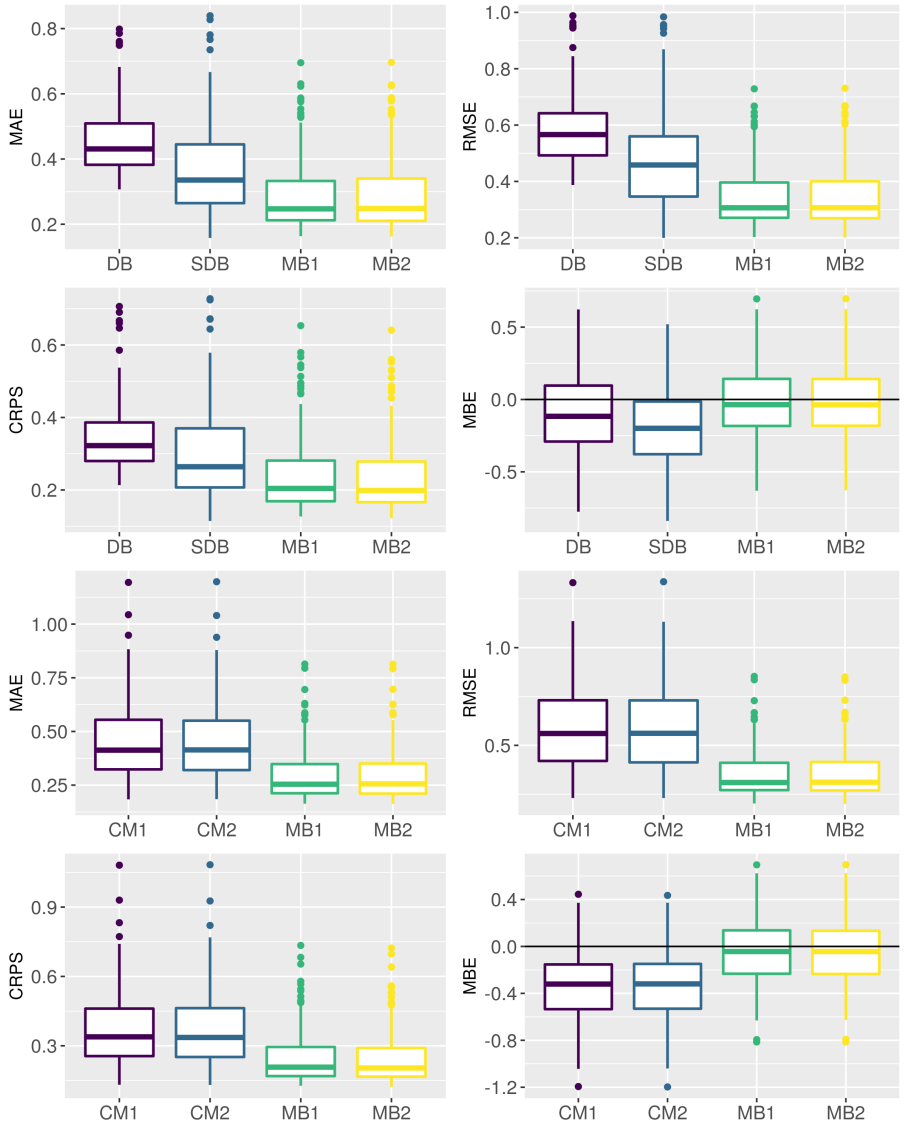


Figure 4.7: Case 2.3: Cluster- and household effects and unstructured spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

In Table 4.8, the best average scores of MAE and RMSE are with MB1, which also has the shortest running time. From the box-plot in Figure 4.7 one can see that also with cluster and household effects are MB1 and MB2 better than DB and SDB. The predictive performance of SDB relative to DB is better than in case 2.2. Again, from Figure 4.7 one can see that the household effect in CM2 and MB2 do not improve the estimates, and the model-based methods are preferable over the combined methods on all scoring rules.

Case 2.4

Case 2.4 is simulated with a fixed stratum effect, iid cluster- and household effects and an unstructured spatial effect. The link function in Equation (4.8) is used. In total $N = 150$ simulations were conducted and $n = 110$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.9). The mean value of the scores for case 2.4 are reported in Table 4.9. Box-plots of the scores for case 2.4 are found in Figure 4.8.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.5911	0.7553	0.4897	-0.3871	15.3
SDB	0.5446	0.6991	0.4623	-0.4682	17.2
MB1	0.3980	0.4774	0.3455	-0.2754	10.2
MB2	0.4009	0.4838	0.3439	-0.2751	233.4
CM1	0.7771	0.9147	0.6653	-0.7394	50.1
CM2	0.7758	0.9139	0.6640	-0.7384	129.9
MB1	0.4586	0.5370	0.4026	-0.3655	10.1
MB2	0.4611	0.5430	0.4006	-0.3652	238.9

Table 4.9: Case 2.4: Stratum effect, cluster- and household effects and unstructured spatial effect. Mean value of scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

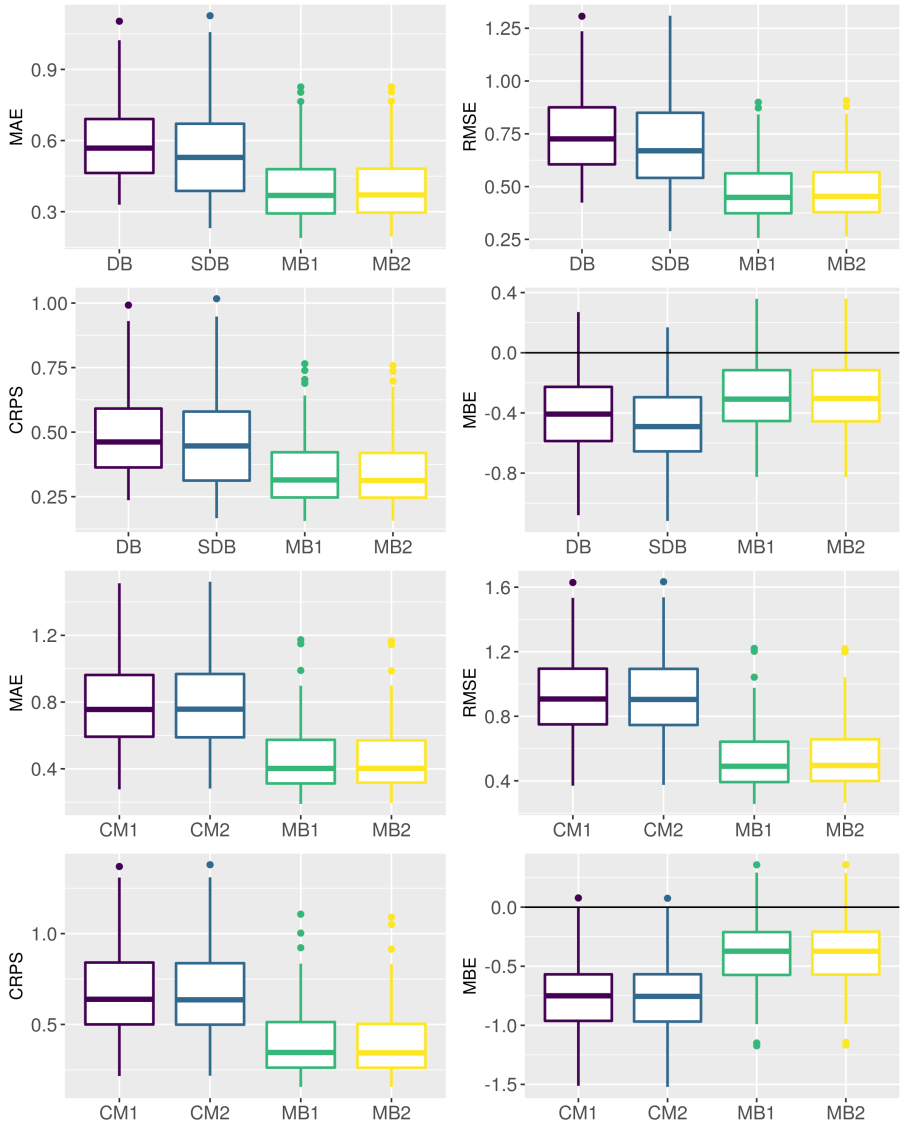


Figure 4.8: Case 2.4: Stratum effect, cluster- and household effects and unstructured spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

The results in Table 4.9 show that MB1 has the lowest average score of MAE, RMSE and running time, while MB2 has the lowest average CRPS score and smallest MBE. CM1 and CM2 have the poorest results on all scores. The box-plots in Figure 4.8 show that MB1 and MB2 again perform best on all four scoring rules and that SDB outperforms DB, however with large spread in the results. Again, as in case 2.2, all methods have large negative MBE due to the added stratum effect.

	Measure			
Method	MAE	RMSE	CRPS	Time (s)
DB	0.5152	0.6574	0.4125	15.4
SDB	0.4460	0.5756	0.3719	17.4
MB1	0.3392	0.4072	0.2900	11.1
MB2	0.3405	0.4104	0.2866	245.6
CM1	0.6231	0.7616	0.5429	49.9
CM2	0.6222	0.7610	0.5420	123.6
MB1	0.3684	0.4364	0.3175	10.9
MB2	0.3695	0.4394	0.3138	246.1

Table 4.10: Average score of MAE, RMSE and CRPS and running time across all scenarios in case 2 (2.1, 2.2, 2.3 and 2.4). The first four rows display the average scores for methods DB, SDB, MB1 and MB2 for the subset of n simulations. The next four rows display the average scores for methods CM1, CM2, MB1 and MB2 for all N simulations.

Table 4.10 presents the average scoring rules for DB, SDB, CM1, CM2, MB1 and MB2 across all four scenarios in case 2. MB1 has the best overall scores for MAE and RMSE and the best running time. MB2 has the best CRPS score. SDB has a better overall score for MAE, RMSE and CRPS compared to DB, and the methods are in the same range in terms of running time. The combined methods CM1 and CM2 perform the weakest out of all the methods on average. CM2 has a slightly better score for MAE, RMSE and CRPS compared to CM1. As expected, all methods perform poorer with unstructured effects, compared to case 1 with no spatial effect in Section 4.2. This is reasonable as the county rates vary more due to the added randomness. Again, SDB improves on the DB estimates but introduces bias. The superior method is MB1, when considering the overall performance.

4.4 Case 3: Structured spatial effect

The four scenarios of case 3 (3.1, 3.2, 3.3 and 3.4) are simulated with structured spatial effects $u_i \sim N(0, Q^{-1})$ generated with the function inla.qsample.

Case 3.1

Case 3.1 is simulated with a structured spatial effect and constructed with the link function in Equation (4.2). In total $N = 120$ simulations were conducted and $n = 114$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.3). The mean value of the scores for case 3.1 are reported in Table 4.11. Box-plots of the scores for case 3.1 are found in Figure 4.9.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.4212	0.5389	0.3182	-0.0785	14.4
SDB	0.3191	0.4213	0.2533	-0.1693	16.8
MB1	0.2468	0.3070	0.2025	0.0039	13.8
MB2	0.2467	0.3072	0.1988	0.0038	272.9
CM1	0.4062	0.5390	0.3306	-0.2927	60.1
CM2	0.4067	0.5390	0.3307	-0.2925	146.7
MB1	0.2472	0.3076	0.2028	-0.0027	13.8
MB2	0.2472	0.3079	0.1992	-0.0028	273.1

Table 4.11: Case 3.1: Structured spatial effect. Mean value of the scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

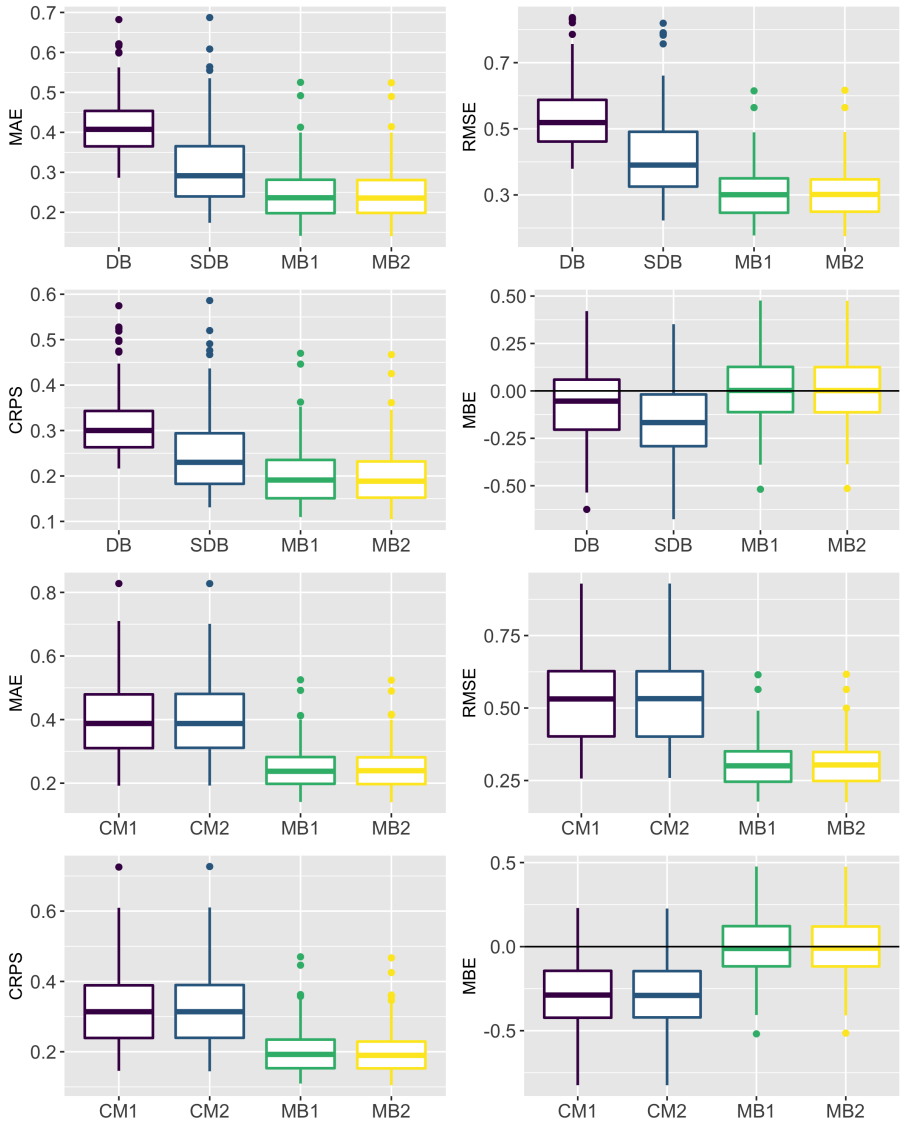


Figure 4.9: Case 3.1: Structured spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

The results in Table 4.11 show that MB2 has the best scores on MAE, CRPS and MBE, while MB1 has the best RMSE score. Both DB, CM1 and CM2 perform poorly, especially when evaluated on MAE and RMSE. From Figure 4.9, one can see that SDB, CM1 and CM2 have a large spread in the scoring results, while the results for MB1 and MB2 are consistent. Again, MB1 and MB2 are the preferable methods. MB2 has the best scores but increased running time.

Case 3.2

Case 3.2 is simulated with a fixed strata effect and a structured spatial effect. The link function in Equation (4.4) is used. In total $N = 200$ simulations were conducted and $n = 160$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.5). The mean value of the scores for case 3.2 are reported in Table 4.12. Box-plots of the scores for case 3.2 are found in Figure 4.10.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.5811	0.7437	0.4813	-0.4059	13.1
SDB	0.5373	0.6920	0.4543	-0.4872	15.4
MB1	0.3795	0.4616	0.3272	-0.2951	13.7
MB2	0.3820	0.4677	0.3246	-0.2950	262.8
CM1	0.7240	0.8706	0.6145	-0.6931	62.1
CM2	0.7241	0.8703	0.6144	-0.6932	180.7
MB1	0.4050	0.4874	0.3510	-0.3305	13.9
MB2	0.4072	0.4933	0.3485	-0.3303	264.4

Table 4.12: Case 3.2: Stratum effect and structured spatial effect. Mean value of the scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

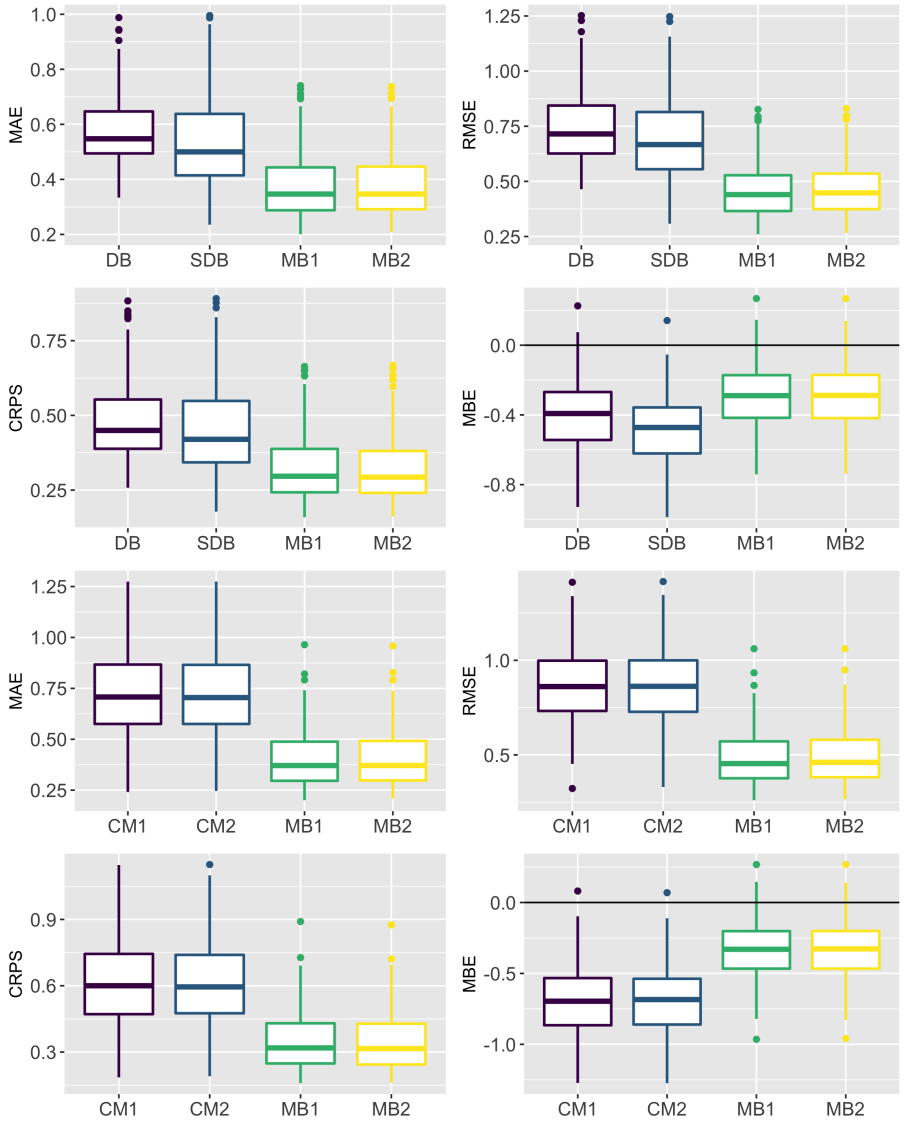


Figure 4.10: Case 3.2: Stratum effect and structured spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

For case 3.2, the results in Table 4.12 shows that MB1 and MB2 again have the best scores on all measures. Again, the added stratum effect causes large negative MBE for all methods. Figure 4.10 shows that several methods have high outliers in the scores, especially DB and SDB, while the most uniform scores are obtained with MB1 and MB2. The best performing methods for case 3.2 are MB1 and MB2.

Case 3.3

Case 3.3 is simulated with iid cluster- and household effects and a structured spatial effect. The link function in Equation (4.6) is used. In total $N = 150$ simulations were conducted and $n = 140$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.7). The mean value of the scores for case 3.3 are reported in Table 4.13. Box-plots of the scores for case 3.3 are found in Figure 4.11.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.4228	0.5422	0.3196	-0.0844	12.2
SDB	0.3228	0.4277	0.2581	-0.1746	14.1
MB1	0.2560	0.3168	0.2117	-0.0058	12.5
MB2	0.2563	0.3172	0.2076	-0.0053	227.1
CM1	0.4094	0.5396	0.3328	-0.2928	51.6
CM2	0.4093	0.5396	0.3327	-0.2918	126.7
MB1	0.2573	0.3181	0.2125	-0.0111	12.4
MB2	0.2576	0.3185	0.2086	-0.0106	228.7

Table 4.13: Case 3.3: Cluster- and household effects and structured spatial effect. Mean value of the scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

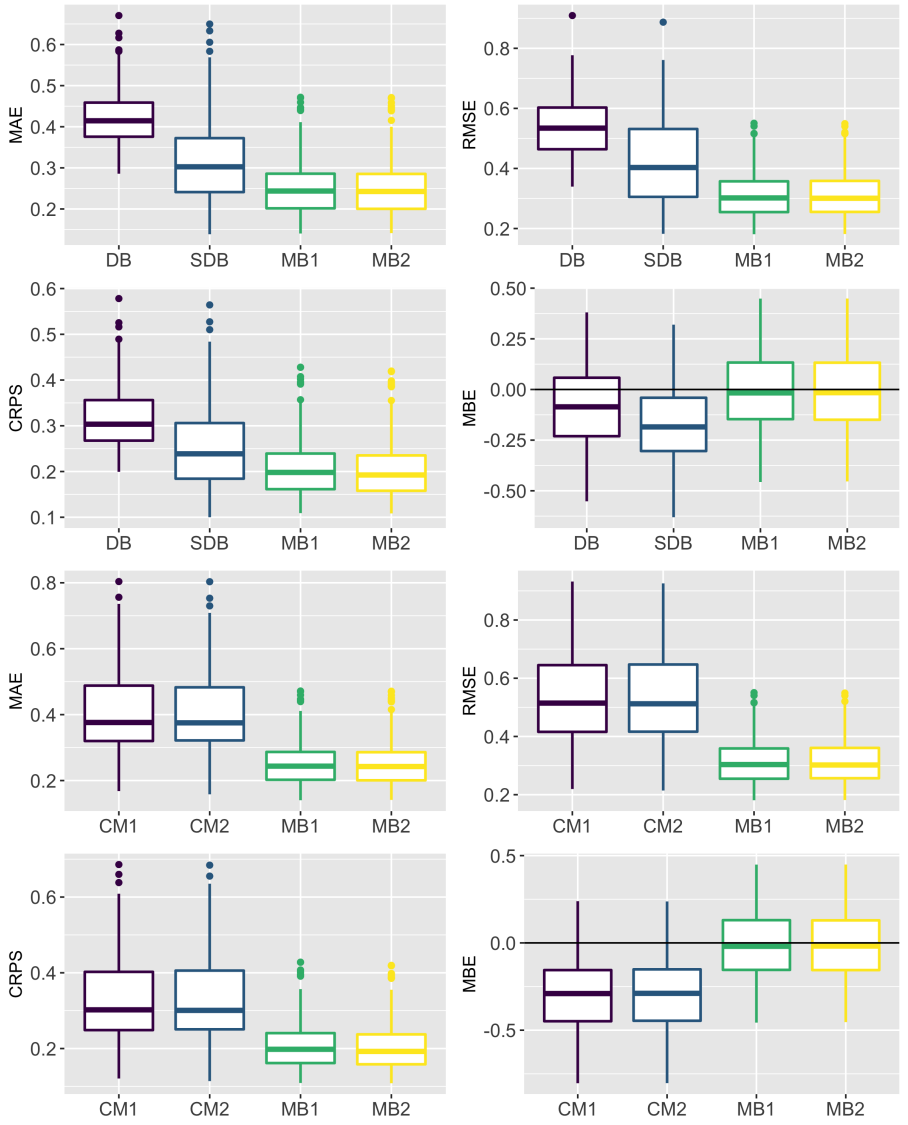


Figure 4.11: Case 3.3: Cluster- and household effects and structured spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

From Table 4.13, one can see that MB1 has the best scores for MAE and RMSE, while MB2 has the best scores for CRPS and MBE. CM1 and CM2 have large negative MBE compared to the other methods. DB results in the poorest scores for MAE and RMSE. The results in Figure 4.11 show that SDB, CM1 and CM2 have large spreads in the results. However, in general, the DB estimates are improved by SDB. Again, MB1 and MB2 are the preferred methods.

Case 3.4

Case 3.4 is simulated with a fixed stratum effect, iid cluster- and household effects and a structured spatial effect. The link function in Equation (4.8) is used. In total $N = 200$ simulations were conducted and $n = 150$ of them generated enough neonatal deaths to obtain design-based rates for each county. The county rates estimated with the methods are compared to the true county rates found with Equation (4.9). The mean value of the scores for case 3.4 are reported in Table 4.14. Box-plots of the scores for case 3.4 are found in Figure 4.12.

	Measure				
Method	MAE	RMSE	CRPS	MBE	Time (s)
DB	0.5933	0.7549	0.4945	-0.4270	8.5
SDB	0.5535	0.7053	0.4698	-0.5077	9.9
MB1	0.3922	0.4747	0.3410	-0.3121	9.8
MB2	0.3947	0.4808	0.3382	-0.3120	202.5
CM1	0.7426	0.8804	0.6257	-0.7140	36.8
CM2	0.7417	0.8800	0.6250	-0.7134	121.6
MB1	0.4096	0.4929	0.3563	-0.3374	9.9
MB2	0.4122	0.4993	0.3537	-0.3376	203.8

Table 4.14: Case 3.4: Stratum effect, cluster- and household effects and structured spatial effect. Mean value of the scores of MAE, RMSE, CRPS and MBE, and mean running time in seconds averaged over the simulations. The first four rows present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations and the next four rows present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

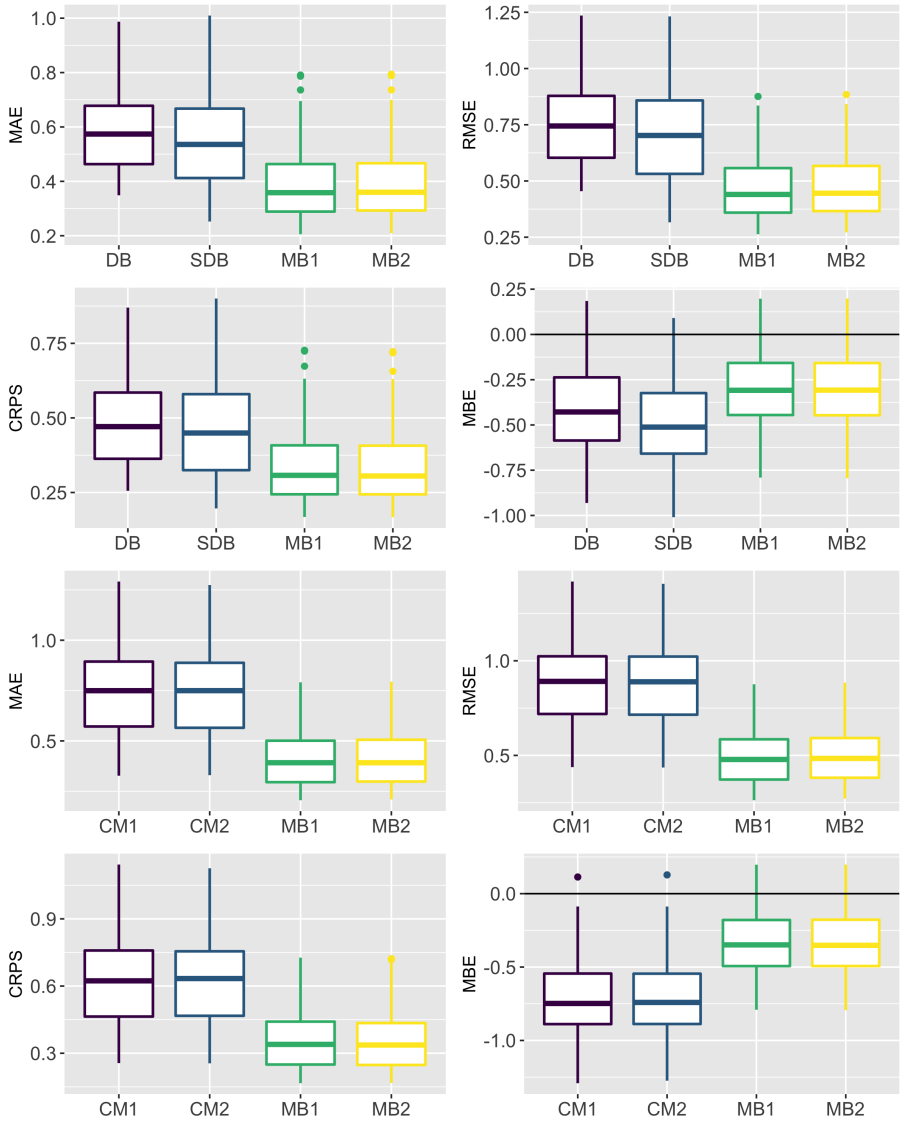


Figure 4.12: Case 3.4: Fixed effect, cluster- and household effects and structured spatial effect. Box-plots of the scores of MAE, RMSE, CRPS and MBE. The first four plots present the results for methods DB, SDB, MB1 and MB2 on the subset of n simulations. The next four plots present the results for methods CM1, CM2, MB1 and MB2 on all N simulations.

The results in Table 4.14 show MB1 has the best scores for MAE and RMSE, while MB2 has the best CRPS and MBE scores. Again, the results from DB are improved by SDB, except for the introduced MBE. CM1 and CM2 have the poorest average scores for all measures. Figure 4.12 show that the scores for DB are only slightly improved with SDB and that MB1 and MB2 provides the most consistent results with little spread.

	Measure			
Method	MAE	RMSE	CRPS	Time (s)
DB	0.5046	0.6449	0.4034	12.1
SDB	0.4332	0.5616	0.3588	14.1
MB1	0.3186	0.3900	0.2706	12.5
MB2	0.3199	0.3932	0.2673	241.3
CM1	0.5706	0.7074	0.4759	52.7
CM2	0.5705	0.7072	0.4757	143.9
MB1	0.3298	0.4015	0.2807	12.5
MB2	0.3311	0.4048	0.2775	242.5

Table 4.15: Average score of MAE, RMSE and CRPS and running time across all scenarios in case 3 (3.1, 3.2, 3.3 and 3.4). The first four rows display the average scores for methods DB, SDB, MB1 and MB2 for the subset of n simulations. The next four rows display the average scores for methods CM1, CM2, MB1 and MB2 for all N simulations.

Table 4.15 presents the average scores for DB, SDB, CM1, CM2, MB1 and MB2 across all four scenarios. MB1 has the overall best average score for MAE and RMSE, while MB2 has the overall best average score for CRPS. The worst performing methods across all scenarios are CM1 and CM2. The preferable methods are MB1 and MB2, however, MB1 is superior considering the average running time.

4.5 Discussion

All methods perform the best when there is no spatial effect, in Section 4.2. For SDB, CM and MB, there is more strength for estimating the intercept and cluster variance as there are no spatial effects to estimate. Thus, this may result in a better estimation of the intercept. Out of the three spatial cases considered, one can assume that the simulations with the unstructured and structured spatial effects are the most realistic. The spatial effects make it more difficult to obtain accurate estimates for the spatial methods, SDB, CM and MB, while DB provides similar results across all three cases. It may be more difficult for the methods to learn the spatial effect as this depends on the prior. Interesting future work is to investigate the effect of the priors on prediction more thoroughly.

It is reasonable to believe that the scenarios with both stratum effects and iid effects at

the cluster- and household level are the most realistic. This scenario is also the one that is the most difficult to obtain accurate rates from. As expected, it is generally easiest to estimate accurate rates when the simulated deaths are generated by a uniform rate in each county. The results show that the fixed stratum effect makes it challenging to estimate the county rates. The difference in urban and rural rates and the lowered baseline rate μ results in more stochastic variation as there are less simulated deaths in rural areas making prediction in rural areas more difficult. Also, since the stratum rates are weighted with the number of residents in urban and rural areas from the 2009 census, where the majority live in rural areas, the gathered county rates will be most influenced by the estimated rural rates.

The results show that the stratum effect gives bias for all methods. The reason for this may be how the estimates from urban and rural areas are weighted to obtain the aggregated county rates. Here, the rates are weighted with the number of residents in urban and rural areas in each county from the 2009 census, which may not give the same weighting as the survey weights. A solution that may correct the bias is to weight the stratum rates with the number of births in urban and rural areas in the time period 2009-2014. However, this information is not available. Also, since only one sample is considered in this study the survey weights will not provide the exact ratio of the number of births in urban vs. rural areas. There is only over repeated sampling that the weights will give the exact ratio.

DB is the method that performs the poorest across most of the 12 simulations performed. The method estimates the rates directly without accounting for the added effects. Therefore, DB estimates variable rates which result in poor scores. This may indicate that there is too much noise in the estimates and that a smoothing method such as SDB or MB is preferable. The SDB method is able to improve the design-based estimates but introduces some bias. In addition, SDB is outperformed by both MB1 and MB2 on all scoring rules on all cases. However, one advantage of SDB is that the method respects the survey design by using the design-based estimates, and it is intuitive and computationally fast.

Based on the results of the simulations, it seems clear that it is unnecessary to include iid household effects in MB2 and CM2. The effects do not improve precision and, as expected, the increased complexity of the methods increase running time considerably. The non-increased performance may be due to the methods not being able to fully capture the intra-household correlation. Further exploration into the priors of the cluster and household effects can be done to see if this affects the performance of CM2 and MB2. It is somewhat unexpected that CM1 and CM2, in general, perform unsatisfactorily. However, as MB outperforms SDB, there is no reason to believe that CM are the preferred methods.

One goal of the study is to assess whether it is possible to obtain methods that can estimate NMRs on fine scales in space and time. If this is desired, CM1 and MB1 should be preferred over CM2 and MB2 as the running times for the more complex models are already extensive for estimation at the county level. Also, if the goal is to obtain estimates on finer spatial scales, DB and SDB can only be used if the sample size is increased, which undesirable due to limited resources. The same is true if it is desired to reduce the sample size. The methods CM and MB are preferred for both cases as it would be impossible to obtain reliable estimates with DB and SDB on smaller spatial scales or with smaller sample sizes.

It is important to point out that the results obtained from this simulation study are

only valid for the specific sample drawn. However, one can expect similar results for all possible samples. If one wishes to obtain general results, it is necessary to draw new samples of clusters and households. As information about the non-sampled enumeration areas is unattainable, it is not possible to perform a complete simulation study.

The simulation study concludes that the preferable method is MB1. It performs well across all scenarios on all scoring rules and is computationally fast. In the next chapter, the performance of MB1 on the real data is explored and compared to DB and SDB.

Chapter 5

Estimating neonatal mortality rates in Kenya 2009-2014

In this chapter, subnational neonatal mortality rates are estimated with the best performing method from the simulation study, i.e. MB1. The estimates obtained from MB1 are compared to estimates obtained with the design-based methods DB and SDB. Even though MB1 in general performed better in the simulation study, it is desired to compare its predictive performance against established and commonly used methods.

5.1 Full sample

Figure 1.1 in Chapter 1 presents maps of estimated NMRs at the county level and the relative standard deviation (RSD) from methods DB, SDB and MB1. The RSD is calculated as the estimated standard deviation divided by the estimated rate and provides a measure of how large the variance of the estimates are. The rates are obtained from the full sample of approximately 1600 clusters. DB produced the most varying rates across the counties, while MB1 produces more similar rates. The smoothing of the rates is anticipated as the modelling reduces noise in the data and causes attenuation. One can observe that high estimated rates produced by DB are smoothed out with SDB. However, it may seem like this is not the case for low estimated rates. SDB also reduce the relative standard deviation as expected.

In Figure 5.1, scatter plots of the rates from MB1 and SDB against the rates from DB for the full sample are presented. The left plot is on the logit scale, and the right plot is on the probability scale. MB1 smooths out high and low values and produce less spatial variation compared to DB, as expected. One can see that SDB produces rates more similar to DB, than MB1.

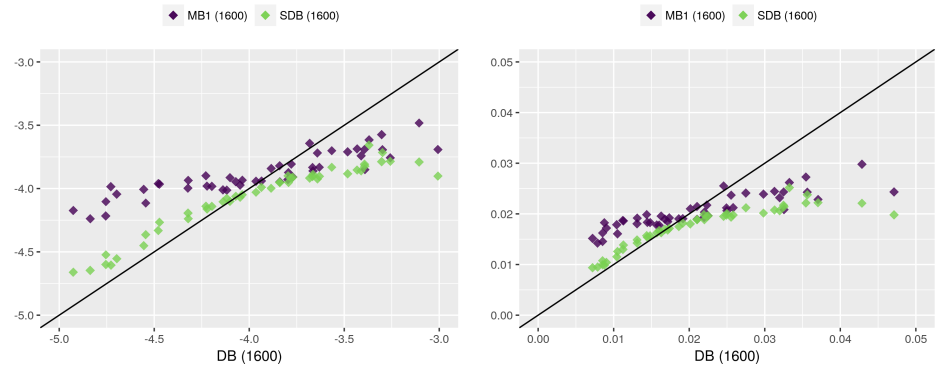


Figure 5.1: Scatter plots of the estimated rates from MB1 and SDB against DB rates for the full sample (1600 clusters). Left: Logit scale. Right: Probability scale.

	Method	
Parameter	MB1	SDB
$\exp\beta$	1.05[0.85, 1.29]	
σ_ν	0.06[0.02, 0.63]	
σ_υ	0.21[0.10, 0.35]	0.33[0.23, 0.46]
ξ	0.42[0.03, 0.94]	0.32[0.02, 0.89]

Table 5.1: Median, 0.025 and 0.975 quantiles of the hyperparameters for MB1 and SDB. β is the fixed strata effect, σ_ν is the standard deviation for the random cluster effects, σ_υ is the standard deviation of the total spatial effect ($\phi + \gamma$) and ξ is the proportion of variance explained by the structured spatial effect, ϕ.

Table 5.1 presents the hyperparameters for models MB1 and SDB. The cluster effect and total spatial effect are expressed as standard deviation ($\sigma = 1/\sqrt{\tau}$). ξ is the proportion of the spatial variability that is structured, where the total spatial variability is the sum of the structured and unstructured effects: $\upsilon = \phi + \gamma$. β is expressed as $\exp\beta$. One can observe that β is insignificant as the 0.025 and 0.975 quantiles cover 1. The finding agrees with the officially reported urban and rural sample rates of 26 and 21 per 1000 live births reported in Kenya National Bureau of Statistics et al. (2015), which are quite similar. For both MB1 and SDB the median of the proportion of structured variability is smaller than 0.5, 0.42 and 0.32 respectively, indicating that the spatial variability in the county rates consists of more random noise than structured variation. The standard deviation of the iid cluster effect is small with median 0.06, suggesting little variation between the clusters in the same stratum. SDB reports larger spatial effects, while MB1 reports larger ξ, however with a very a large credibility interval of $[0.03, 0.094]$.

To evaluate the difference in the rates for MB1 and SDB compared to the design-

based rates, median and 0.025 and 0.975 quantiles are computed from samples using the estimated rates. For MB1, n samples are obtained by the function `inla.posterior.sample` as explained in Section 3.5. For DB and SDB, n samples of the estimated rates are generated from $N(\mu_i, \sigma_i^2)$ for county i, where the mean, μ_i, is the estimated rate in county i on logit scale and the variance, σ_i^2, is the variance of the rate. 0.025 and 0.975 quantiles and the median of the difference between the samples from DB and MB, and DB and SDB for each county are computed and reported in Figure 5.2.

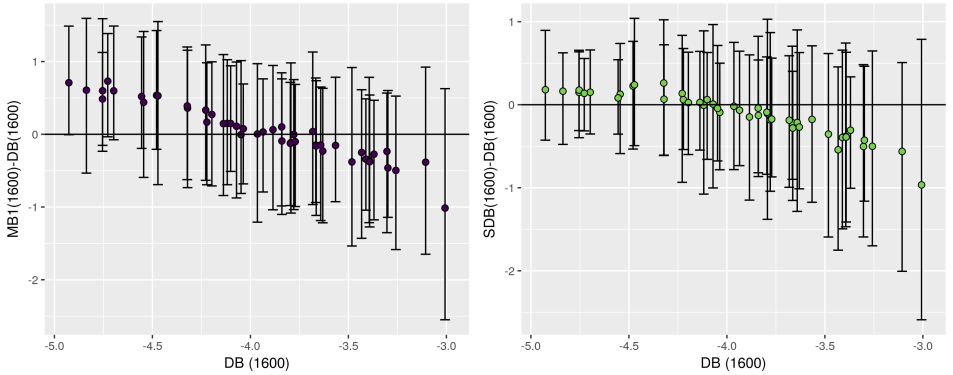


Figure 5.2: Median, 0.025 and 0.975 quantiles of the difference between samples from MB1 and DB, and SDB and DB for the full sample. Left plot: MB1-DB. Right plot: SDB-DB.

From Figure 5.2, one can see that the rates produced by SDB are more compatible with the DB rates. The result is consistent with the findings in Figure 5.1, as MB1 produces the most homogeneous rates that are dissimilar for the counties where DB produce very high or low rates. However, most of the rates from MB are not significantly different from the DB rates.

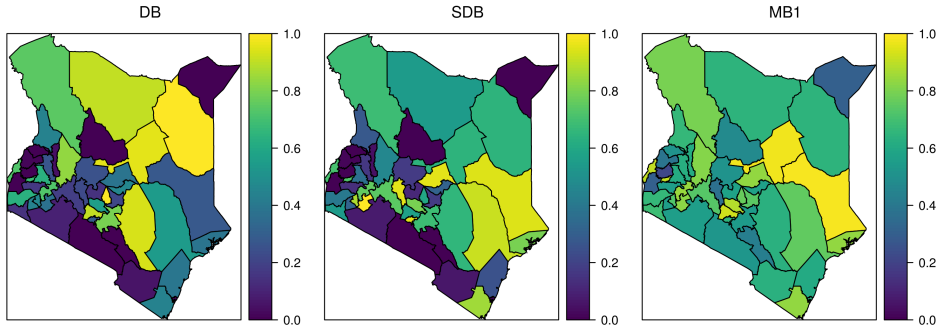


Figure 5.3: Maps with two-sided p-values for the county rates from DB, SDB and MB1 for the full sample (1600 clusters). Left plot: DB. Middle plot: SDB. Right plot: MB1.

Figure 5.3 presents maps with computed two-sided p-values assuming normal distri-

bution for the estimated county rates from DB, SDB and MB1 for the full sample. The null hypothesis is no significant difference between the national NMR of 22 per 1000 live births as reported in Kenya National Bureau of Statistics et al. (2015) and the county rates from the three methods. Small p-values indicate that the county rate is significantly different from the national rate. The results show that the rates from MB1 are the most compatible with the officially reported national rate, while many county rates for DB and SDB are significantly different from the national estimate. It is questionable whether the apparent spatial heterogeneity estimated by DB and SDB is real or if the NMR is more similar in each county then the methods indicate.

5.2 Reducing the sample size

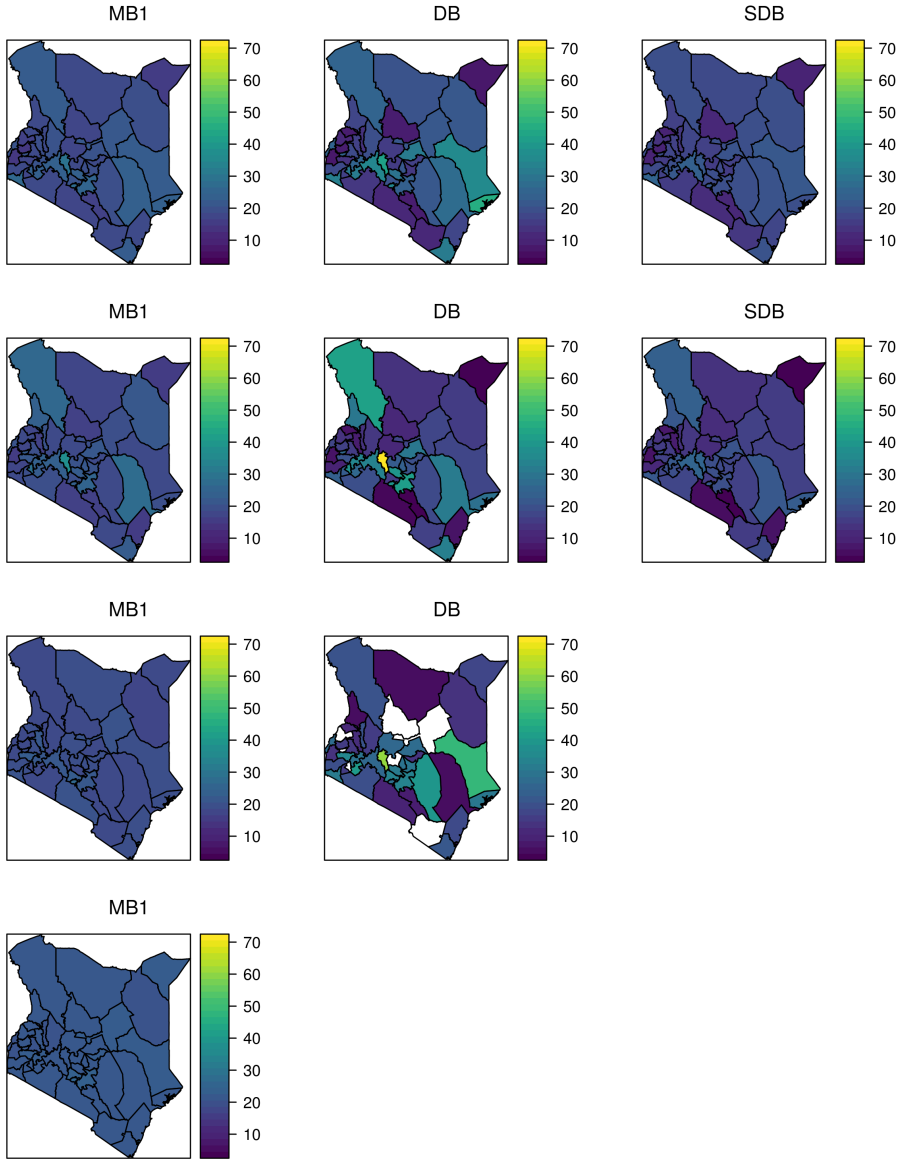


Figure 5.4: Maps of the estimated NMRs at the county level from samples of 1600, 800, 400 and 200 clusters. Left column: MB1. Middle column: DB. Right column: SDB. Top row: 1600 clusters. Second row: 800 clusters. Third row: 400 clusters. Bottom row: 200 clusters.

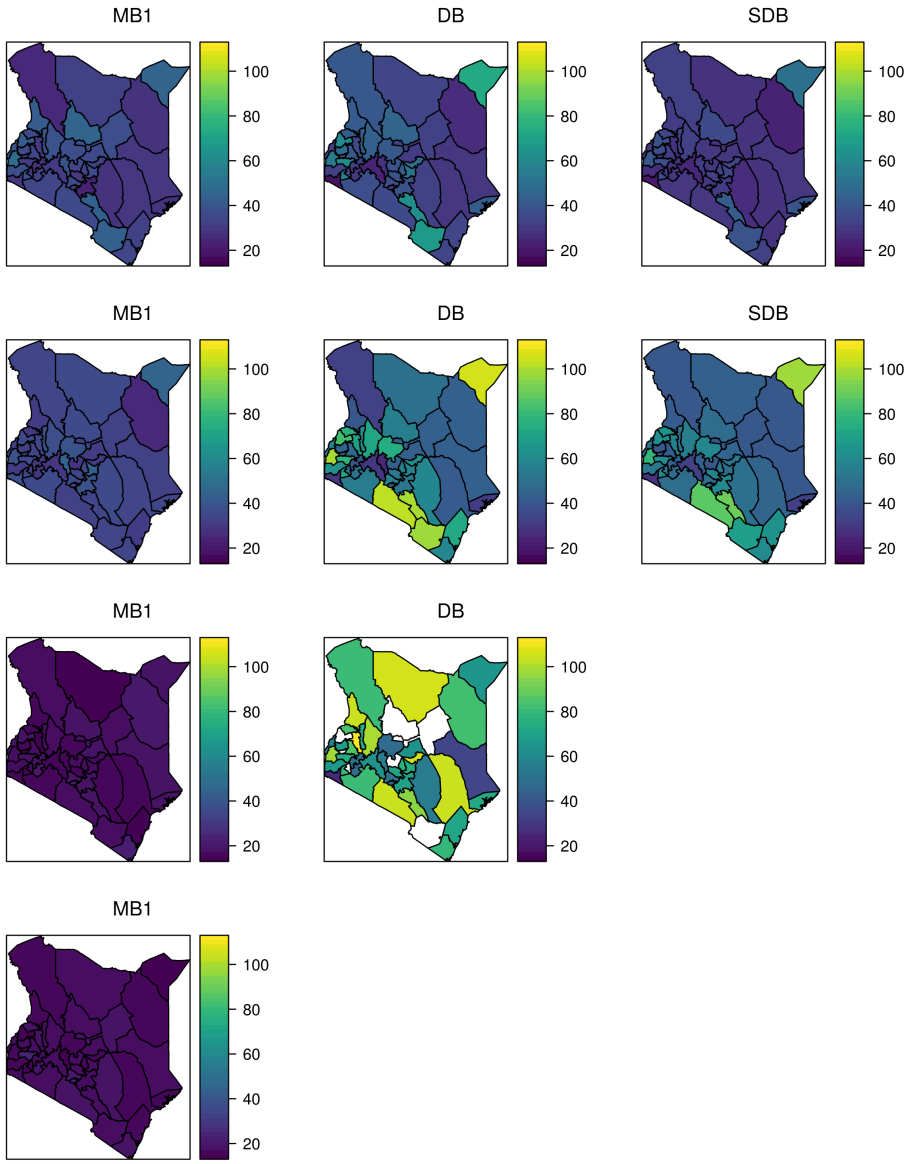


Figure 5.5: Maps of the RSD at the county level from samples of 1600, 800, 400 and 200 clusters. Left column: MB1. Middle column: DB. Right column: SDB. Top row: 1600 clusters. Second row: 800 clusters. Third row: 400 clusters. Bottom row: 200 clusters.

In Figure 5.4, maps of estimated rates at the county level from samples of 1600, 800, 400 and 200 clusters for MB1, DB and SDB are presented. DB produces outliers for several counties with 800 clusters, which are smoothed by SDB. The two methods produce similar rates excluding the outliers. Also, MB1 and SDB produce similar rates, with more

uniform rates from MB1. For a sample with 400 clusters, there is not enough data to produce design-based rates for all counties. The areas coloured white for DB in Figure 5.4 are counties where rates are unavailable. Therefore, SDB rates cannot be obtained as explained in Section 3.2. For 400 clusters, DB estimates both very high and low rates for some counties. The rates from MB1 is rather uniform, indicating that MB1 tend to smooth the rates too much. For a sample with 200 clusters, it is not possible to obtain any design-based rates as several strata contain zero or only one cluster. The rates obtained by MB1 are very similar to the results for 400 clusters.

In Figure 5.5, maps of the relative standard deviation of the rates at the county level from samples of 1600, 800, 400 and 200 clusters for MB1, DB and SDB are presented. For 1600 clusters, all three methods have similar RSD, with DB having the highest. When the sample size is reduced to 800 clusters, DB and SDB have larger RSD in many of the counties compared to MB1. The smoothing of SDB reduces the variability in the rates from DB. For 400 clusters the RSD is reduced for MB1, while the RSD is increased for DB. Some of the counties have RSD > 100, which means that the standard deviation of the rate is larger than the rate itself. Therefore, the rates are highly uncertain. Also, the RSD is set to zero for the counties with zero rates, presented with white for these counties. For both 400 and 200 clusters are the uncertainties small for MB1. It appears that MB1 is unable to capture the spatial structures for the small sample sizes of 400 and 200, yielding very low values for the RSD. This can be a result of the choice of priors.

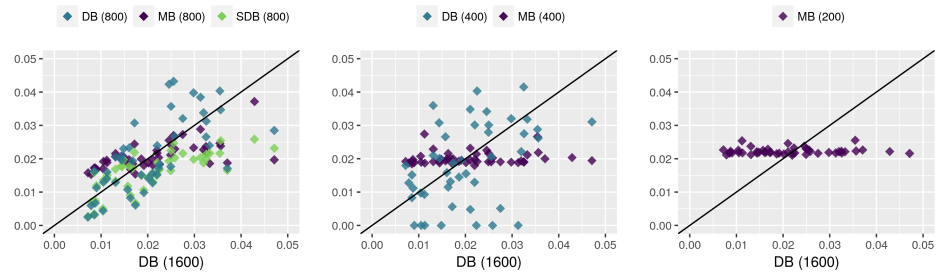


Figure 5.6: Scatter plots of the rates from DB, SDB and MB1 against rates from DB for the full sample (1600 clusters) on probability scale. Left: 800 clusters. Middle: 400 clusters. Right: 200 clusters.

Figure 5.6 shows scatter plots of the rates from DB, SDB and MB1 for reduced sample sizes (800, 400 and 200 clusters) against the rates from DB for the full sample. It seems like the rates from SDB for 800 clusters are the most compatible with the design-based rates for the full sample. With 400 and 200 clusters the MB1 rates are very uniform, indicating that the spatial variation is difficult to capture with small sample sizes. The rates from DB for 400 and 1600 clusters are compatible, with a large amount of noise. The DB rates for 400 clusters are invalid if there is in reality little spatial variation.

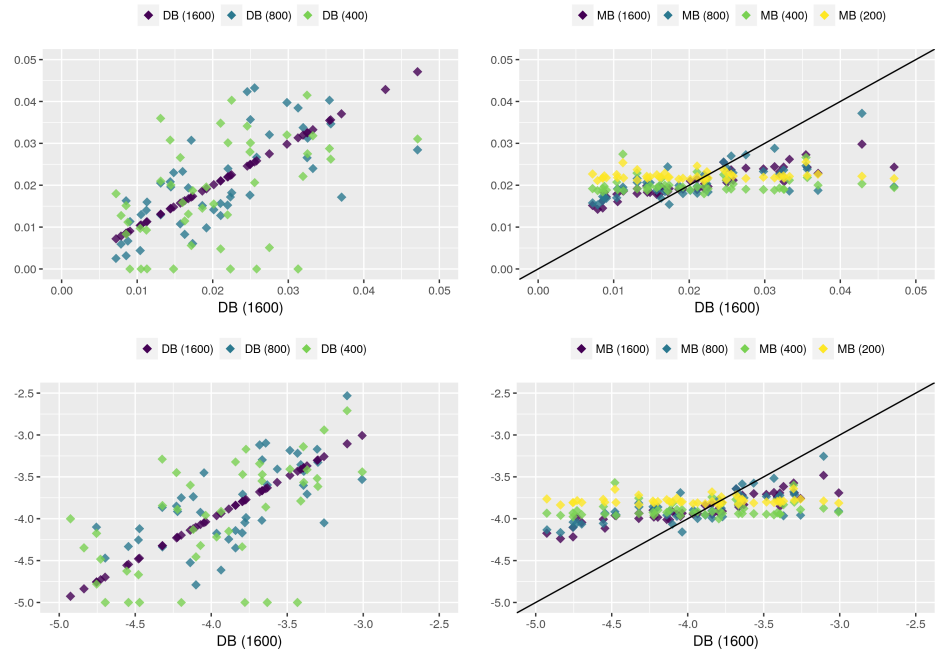


Figure 5.7: Scatter plots of the rates from DB and MB1 against rates from DB for the full sample (1600 clusters). Top left: Rates from DB for 1600, 800 and 400 clusters on probability scale. Bottom left: Rates from DB for 1600, 800 and 400 clusters on logit scale. Top right: Rates from MB for 1600, 800, 400 and 200 clusters on probability scale. Bottom right: Rates from MB for 1600, 800, 400 and 200 clusters on logit scale.

Figure 5.7 shows scatter plots of the rates from DB and MB1 against rates from DB for the full sample (1600 clusters) on logit and probability scale. For 800 clusters, DB estimates are rather compatible with the full sample estimates, while the MB1 estimates are more and more uniform as the sample is reduced, which is not compatible with the noisy estimates obtained by DB for the full sample.

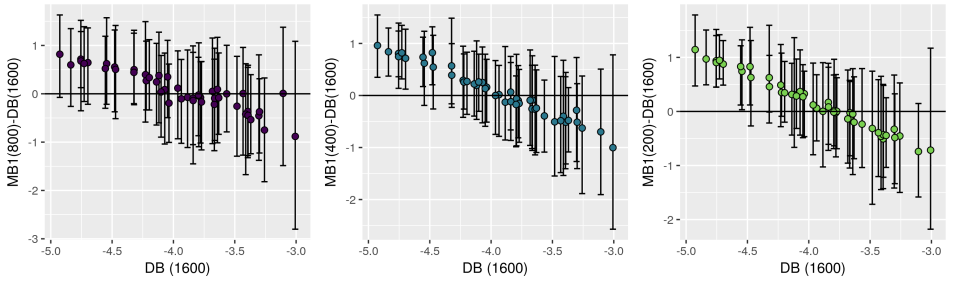


Figure 5.8: Median, 0.025 and 0.975 quantiles of the difference between samples from MB1 and DB. Left plot: MB1 (800) - DB (1600). Middle plot: MB1 (400) - DB (1600). Right plot: MB1 (200) - DB (1600).

Figure 5.8 presents plots of 0.025 and 0.975 quantiles and the median of the difference between samples of MB1 and DB for each county. The samples of the estimates are from MB1 for 800, 400 and 200 clusters and for the full sample from DB. The MB1 samples for 800 clusters are compatible with the 1600 clusters DB samples. With 400 and 200 clusters several county rates are outside the credible region. The results can be explained by the very uniform estimates with low uncertainty produced by MB1.

	Number of clusters			
Parameter	1600	800	400	200
$\exp\beta$	1.05[0.85, 1.29]	1.00[0.75, 1.33]	1.54[1.04, 2.27]	1.23[0.71, 2.09]
σ_ν	0.06[0.02, 0.64]	0.06[0.02, 0.51]	0.05[0.02, 0.35]	0.05[0.02, 0.30]
σ_υ	0.21[0.10, 0.35]	0.14[0.03, 0.36]	0.04[0.01, 0.20]	0.05[0.01, 0.26]
ξ	0.42[0.03, 0.94]	0.25[0.01, 0.89]	0.29[0.01, 0.95]	0.29[0.01, 0.93]

Table 5.2: Median, 0.025 and 0.975 quantiles of the hyperparameters for MB1 for samples of 1600, 800, 400 and 200 clusters. β is the fixed strata effect, σ_ν is the standard deviation for the random cluster effects, σ_υ is the standard deviation of the total spatial effect ($\phi + \gamma$) and ξ is the proportion of variance explained by the structured spatial effect, ϕ.

Table 5.2 presents the median, 0.025 and 0.975 quantiles of the hyperparameters for MB1. The fixed strata effect, β, is insignificant for samples of 1600, 800 and 200. For 400 clusters, it seems like the effect is significant. However, this is likely by chance as the sample is reduced at random or a spurious signal caused by the low amount of data. The iid cluster effect remains somewhat constant as the sample is reduced. The same can not be said for the total spatial effect which is drastically reduced from a median value of 0.14 for 800 clusters to a median value of 0.04 for 400 clusters. The small spatial effects for smaller samples may contribute to the very low RSD as presented in Figure 5.5. The proportion of variability explained by the structured spatial effect ξ is rather constant with large credible intervals.

	Number of clusters			
Method	1600	800	400	200
DB	0.0000	0.3883	0.4178	X
SDB	0.2172	0.3831	X	X
MB1	0.2891	0.3213	0.4072	0.4181

Table 5.3: MAE between the estimated rates for DB, SDB and MB1 from all sample sizes and the estimated rates from DB for 1600 clusters on logit scale. The best scores are colored yellow.

Table 5.3 presents the MAE between the rates from DB, SDB and MB1 for all sample sizes and DB rates from the full sample. SDB has the smallest MAE of 0.2172 for 1600, yielding the most compatible rates with DB for the full sample. For 800 clusters MB1 rates are more compatible with MAE $= 0.3213$ and the same yields for 400 clusters with MAE $= 0.4072$. The score for MB1 with 400 clusters is almost as good as the DB and SDB scores with 800 clusters, indicating again that MB1 is better at handling sparse data, compared to DB and SDB.

	Number of clusters			
Method	1600	800	400	200
DB	0.0000	0.4689	0.5633	X
SDB	0.2848	0.4508	X	X
MB1	0.3588	0.4061	0.4925	0.5182

Table 5.4: RMSE between the estimated rates for DB, SDB and MB1 from all sample sizes and the estimated rates from DB for 1600 clusters on logit scale. The best scores are colored yellow.

Table 5.4 presents the RMSE between the rates from DB, SDB and MB1 for all cluster samples and DB rates from the full sample. The difference in RMSE scores for 400 clusters is larger than the MAE scores as RMSE penalizes outliers, which are very much present in the DB rates for 400 clusters with RMSE $= 0.4925$. The score is better for SDB than for MB1 when considering the full sample, but as the sample is reduced to 800, MB1 gives more compatible estimates.

5.3 Finer spatial scale

It is useful to have a tool for detecting which areas within the counties are worse off, such that the counties' policy makers can decide where interventions need to be made. It may be possible that finer-scale features are hidden by producing estimates at the county level. Therefore, neonatal mortality rates for the administrative level 2 (constituencies) within the counties are estimated with MB1 for the full sample. The result of MB1 for administrative level 1 (counties) and level 2 (constituencies) are presented in Figure 5.9.

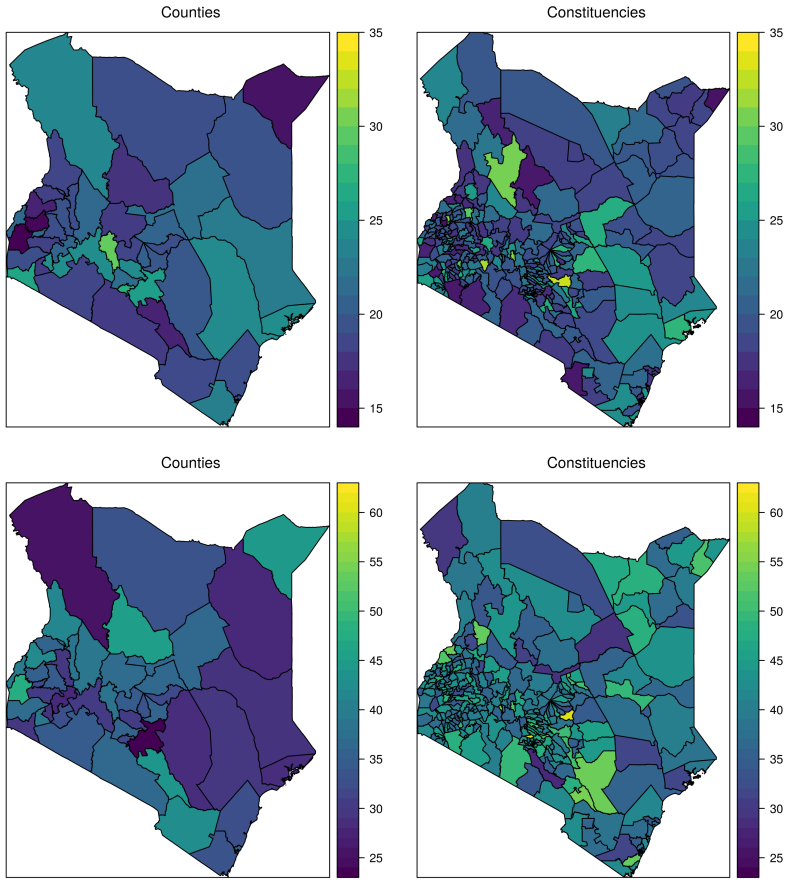


Figure 5.9: Estimated NMRs and RSD obtained by MB1 at administrative level 1 and 2 (counties and constituencies). Top row: NMRs. Bottom row: RSD. Left: Counties. Right: Constituencies.

Based on the results presented in Figure 5.9, it seems like there exist differences in the neonatal mortality rates within the counties. For example, in Turkana, the most north-western county in Kenya, there is one area with a relatively high mortality rate (light green) and several areas with relatively low mortality rates (dark blue). The median and 0.025 and 0.975 quantiles of the difference between posterior samples of the estimated rates from

the areas are calculated with inla . posterior .sample to investigate whether the difference in rates is significant. The quantiles and median are $0.7024[-0.3736, 1.9560]$. Since the prediction interval contains 0, it can not be concluded that the difference in rates within Turkana is significant.

Nevertheless, having methods that can obtain within-county estimates are important. Estimates of the NMR for the different areas within the counties can be useful for classifying counties according to the degree of within-county spatial heterogeneity. Quantifying the differences within the counties can be used to evaluate the state of each county and make it easier to decide where interventions should be made.

Model-based methods such as MB1 makes it possible to produce estimates on finer spatial scales as illustrated in Figure 5.9 for constituencies within the counties. Such investigation into the spatial heterogeneity within the counties cannot be obtained by the design-based methods DB and SDB with the current sample size of 1600 clusters. Obtaining sufficiently accurate estimates of finer spatial scale NMR with DB and SDB is not doable due to the considerable financial costs required to increase the size of the survey.

Chapter 6

Discussion

The overall goal of this thesis is to compare design-based and model-based methods for estimation of subnational neonatal mortality rates. Monitoring the progress towards the SDGs is challenging for several reasons. It is not straightforward to obtain fine-scale estimates with sparse survey data. The more accepted methods that respect the survey design, such as the smoothed method introduced by Mercer et al. (2015), cannot obtain accurate estimates with sparse data, as was demonstrated in Section 5.2. Model-based approaches make this possible, but bias is introduced and there is skepticism within the field to model-based methods. The survey design for the 2014 DHS in Kenya is constructed such that one can obtain reliable estimates at national and county level with design-based methods, but not for more fine-scale divisions. Finer scale estimation would require too much resources and money which is not feasible. In addition, one is dealing with dependent data which has to be considered when conducting inference.

Another source of data that could have been used is census data. Here, every resident in the population is asked a smaller set of questions. The advantage of census data is that the whole population is asked, removing uncertainty. However, as of now, the appropriate questions are not asked if the goal is to obtain child mortality rates. Every woman is only asked how many children they have born and how many of them died before the age of five. Such vague information is much more difficult to handle compared to the detailed responses obtained from, for example, DHS. An idea might be to combine the data sources to explore if this can improve estimation.

In this thesis, model-based and combined approaches is proposed and a simulation study is conducted to asses the performance of the proposed method compared to design-based approaches on a set of spatial cases and scenarios. The study is extensive and required considerable computing power. The conclusion of the simulation study is that the preferred method is MB1. The method generally performs satisfactorily across all spatial cases and scenarios and has a short running time. The more complex model-based method, MB2, also has satisfactory results on the study. However, due to the highly increased computational time combined with no definite improvement in accuracy, it is not the desired method to be used on a finer spatial or spatio-temporal scale. It can be argued that SDB

can also be chosen as the preferred method if it is desired to chose a method that follows the design-based approach. However, it is not possible to use SDB to obtain estimates on finer spatial scales, or for surveys with smaller sample sizes.

Several variations of the simulation study can be explored. It might be interesting to investigate more scenarios, such as only cluster correlation and a combination of unstructured and structured spatial effects. As discussed in Section 4.5, the stratum effect results in increased bias that may be due to how the estimated stratum rates are weighted to obtain gathered county rates. Alternative methods for the weighting can be explored to remove the bias. Also, further investigation into the effect of the priors can be done to see if this changes the outcome of the methods. If the study was to be extended, new populations should be simulated, where also new clusters and households are drawn. This is not done as the non-sampled clusters and households are unavailable. Thus, the results of the study only apply for the 2014 DHS in Kenya, but it can easily be extended. However, one should be careful to generalize the conclusions that are drawn from this particular study.

The methods are assessed with the scoring rules MAE, RMSE, CRPS and MBE. The scores give different conclusions. MB1 is generally the preferred method when considering RMSE and MAE. However, considering the CRPS MB2 has the best score in many scenarios. When it comes to the MBE, both model-based methods are preferable. However, conclusions should not be drawn only based on the bias as it does not reveal other aspects of the distribution. On the contrary, CRPS takes into account both the uncertainty in the true rates and the estimates, crediting estimates with uncertainty reflecting the distance to the true rates. Therefore, one can argue that CRPS is the most informative score and that MB2 is the best method. However, the difference in the CRPS score for MB1 and MB2 is small, typically at the 3rd or 4th decimal place. Also, MB1 is more scalable to finer spatial-temporal scales when considering the computational cost.

The methods MB1, DB and SDB are used to estimate subnational NMRs for Kenya with the 2014 DHS data. The estimated rates for MB1 and SDB are both compatible with the DB rates for the full sample. As expected, SDB reduces the variance and noise of the DB rates. Most of the spatial variability in MB1 was unstructured, which can indicate that the varying rates reported by DB are mostly noise and not systematic variation. Reducing the sample size increases uncertainty and noise in the DB and SDB estimates. For 400 and 200 clusters several estimated county rates for MB1 are not compatible with the DB rates for the full sample, as the MB1 rates become more uniform with decreased uncertainty. The spatial effect is decreased with the decreasing sample size, indicating that the method is unable to capture the spatial structure.

One of the goals of this thesis is to obtain accurate estimates when the sample size is reduced. Surveys conducted in other countries often consist of only 400 clusters, and the same is true for surveys conducted in Kenya prior to the one performed in 2014. It is clear from the results in Section 5.2 that the design-based method is highly dependent on the sample size and is only applicable for the full sample of 1600 clusters for this particular survey. The uncertainty in the estimates with 800 clusters is high and some estimates are unattainable with 400 clusters as the sample has zero observed neonatal deaths for several counties. The smoothed design-based method removes noise and reduces uncertainty in the design-based estimates, but is also dependent on obtaining reliable design-based rates and uncertainty estimates.

The results in Section 5.2 show that the model-based approach MB1 is superior at handling more sparse data, compared to DB and SDB. However, as the number of clusters is reduced to 400 and 200, the method is not able to capture the spatial effects, which results in unreasonably low variance and uniform estimates. This may be a result of the choice of priors, which has not been investigated in depth in this thesis but would be an interesting direction for future work.

Another goal of this thesis is to investigate how the methods perform on even finer spatial scales. Obtaining estimates for the constituencies within the counties is not possible with the design-based method as the data is too sparse. On the other hand, the model-based approach can obtain fine-scale estimates with sparse data, as presented in Section 5.3. However, the results show that the uncertainty in the estimates is increased and it is not straightforward to know if the difference in the estimates within the counties is noise or structured spatial variation. Nonetheless, model-based methods can give continuous models which are necessary to achieve the goal of obtaining fine-scale estimates in both space and time.

There are several things to highlight in the comparison of design-based and model-based methods. Results of the simulation study in Section 4 show that the model-based approaches perform the best and, in particular, MB1 is computationally fast. When it comes to limited resources, model-based methods can handle smaller sample sizes than the design-based method, making it possible to save time and money. In addition, model-based methods can obtain estimates on finer scales and gives information about fixed and random effects, such as the effect of urban or rural areas and the spatial structures are quantified. However, the choice of priors is not insignificant and should be studied.

The newly proposed method combining design-based and model-based approaches, explored here with CM1 and CM2, did not yield satisfactory results. Model-based approaches are superior for this particular simulation study, so there is no reason for combining them with design-based ideas. Also, when finer spatial and temporal scales are desired, the combined methods will be mainly model-based as design-based estimates are largely unattainable. A reason for advocating for the combined methods is that the design-based approach is more accepted within the field of handling complex survey data.

It is of great interest to map the progress towards the SDGs in fine spatial and temporal scale. It is also desired to obtain yearly estimates, resulting in smaller samples. In this thesis, the data were aggregated over a five year period. It is clear that it is not possible to obtain yearly, subnational estimates of neonatal mortality rates with design-based methods. It is difficult to imagine that it is possible to obtain monitor the progress towards the SDGs for all subgroups of the populations with design-based methods. Model-based approaches are a promising direction for achieving the ambitious goal of fine-scale monitoring of mortality.

Bibliography

Alkema, J., New, R. J., 2014. Global estimation of child mortality using a Bayesian B-spline bias-reduction model. The Annals of Applied Statistics 8, 2122–2149.

Banerjee, S., Carlin, B. P., Gelfand, A. E., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC.

Besag, J., Kooperberg, C., 1995. On conditional and intrinsic autoregressions. Biometrika 82 (4), 733–746.

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

Blangiardo, M., Cameletti, M., 2015. Spatial and Spatio-temporal Bayesian Models with R-INLA. Jonh Wiley & Sons.

Bolstad, W. M., 2007. Introduction to Bayesian Statistics. John Wiley & Sond, Inc.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. Journal of statistical software 76 (1).

Casella, B., Berger, R. L., 2001. Statistical Inference. Cengage Learning.

DHS, 2013. Standard recode manual for dhs 6. `https://www.dhsprogram.com/pubs/pdf/DHSG4/Recode6_DHS_22March2013_DHSG4.pdf`, [Accessed:13.02.19].

DHS, 2019. The dhs program. `https://dhsprogram.com/`, [Accessed:13.02.19].

Farhmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression. Springer.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubing, D. B., 2014. Bayesian Data Analysis. Chapman & Hall/CRC.

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

Golding, N., Burstein, R., Longbottom, J., Browne, A., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E., Casey, D., Dwyer-Lindgren, J., Farag, T., Flaxman, A., Fraser, M., Gething, P., Gibson, H., Graetz, N., Krause, L., Kulikoff, X., Lim, S., Mapping, B., Morozoff, C., Reiner, R. J., Sligar, A., Smith, D., Wang, H., Weiss, D., Murray, C., Moyes, C., SI., H., 2017. Mapping under-5 and neonatal mortality in Africa, 2000-15: a baseline analysis for the Sustainable Development Goals. The Lancet 390 (10108), 2171–2182.

IGME, 2019. United Nations inter-agency group for child mortality estimation. http://childmortality.org/, [Accessed:13.02.19].

Kenya National Bureau of Statistics, Ministry of Health and National AIDS Control Council, Kenya Medical Research Institute, National Council for Population and Development, The DHS Program, ICF International, 2015. Kenya demographic and health survey 2014. https://dhsprogram.com/pubs/pdf/fr308/fr308.pdf, [Accessed:11.02.19].

KNBS, 2012. 2009 Population and Housing Census. Nairobi, Kenya.

Lohr, S. L., 2010. Sampling: Design and Analysis. Brooks/Cole.

Lumley, T., 2010. Complex Surveys - A guide to analysis using R. Wiley.

Mercer, L. D., Wakefiled, J., Chen, C., Lumley, T., 2014. A comparison of spatial smoothing methods for small area estimation with sampling weights. Spatial Statistics 8, 69–85.

Mercer, L. D., Wakefiled, J., Pantazis, A., Lutambi, A. M., Masanja, H., Clark, S., 2015. Space-time smoothing of complex survey data: Small area estimation for child mortality. The Annals of Applied Statistics 9 (4), 1889–1905.

Rue, H., 2019. The R-INLA project. http://www.r-inla.org/, [Accessed:25.02.19].

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71 (2), 319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simspson, D. P., Lindgren, F. K., 2017. Bayesian computing with INLA: A review. Annual Review of Statistics and Its Applications 4, 395–421.

UN, 2018. Sustainable Development Goal 3. https://sustainabledevelopment.un.org/sdg3, [Accessed:31.01.19].

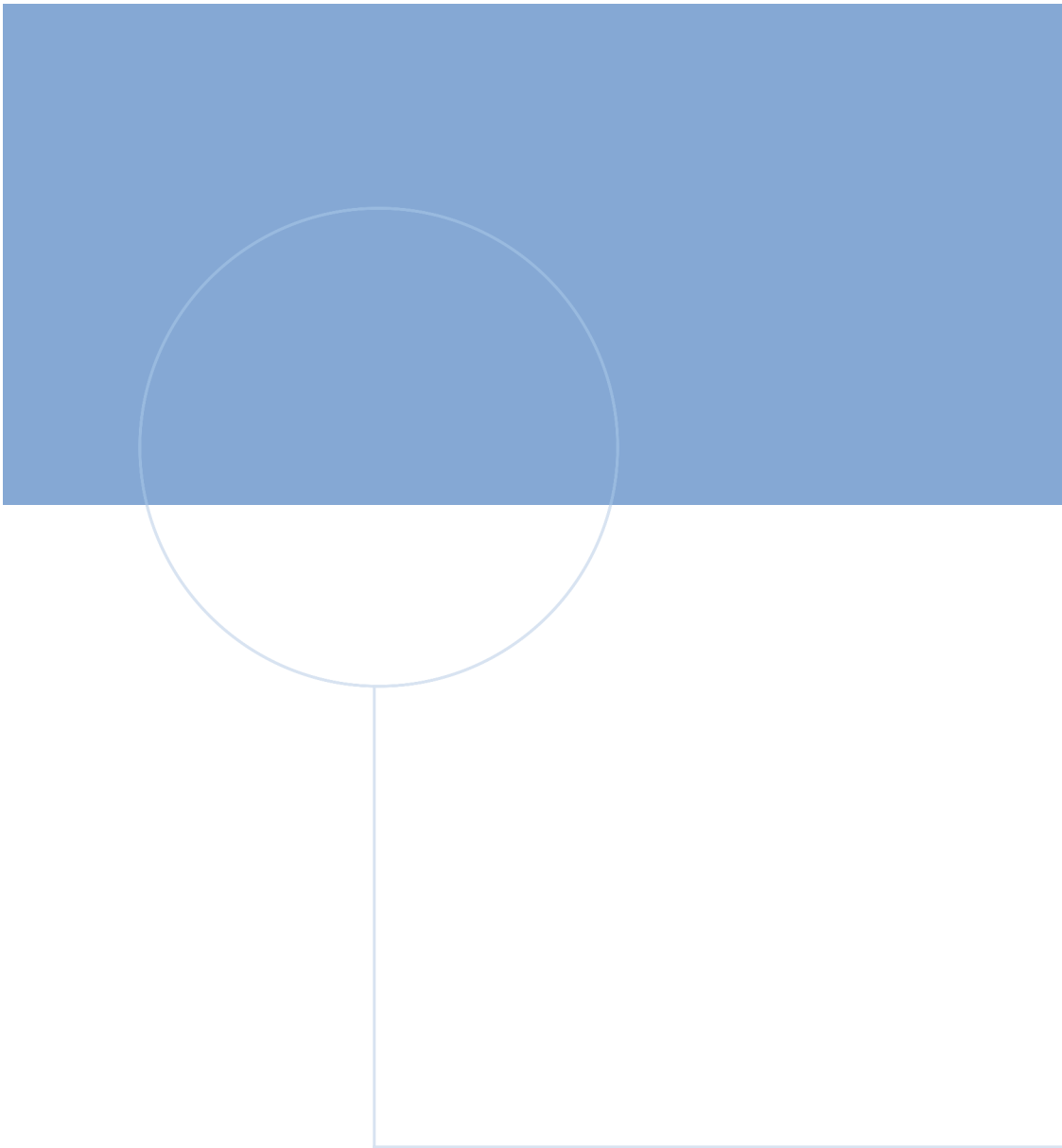
UN, 2019a. Leave no one behind. https://unstats.un.org/sdgs/report/2016/leaving-no-one-behind, [Accessed:12.04.19].

UN, 2019b. Millennium Development Goals and beyond 2015. http://www.un.org/millenniumgoals/, [Accessed:31.01.19].

UNICEF, 2018a. Neonatal mortality. https://data.unicef.org/topic/child-survival/neonatal-mortality/, [Accessed:31.01.19].

UNICEF, 2018b. Under-five mortality. https://data.unicef.org/topic/child-survival/under-five-mortality/, [Accessed:31.01.19].

Zehang, L., Hsiao, Y., Godwin, J., Martin, D. B., Wakefield, J., Clark, J. S., 2019. Changes in the spatial distribution of the under-five mortality rate: Small-are analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. PLoS one 14 (1), e0210645.



NTNU
Norwegian University of
Science and Technology