UNIVERSITY OF
KWAZULU-NATAL

INYUVESI
YAKWAZULU-NATALI

Spatial and spatio-temporal modeling and mapping of self-reported health among individuals between the ages of 15 – 49 years in South Africa.

A thesis submitted in fulfillment of the requirement for the Masters Degree in Statistics

by

Banele P. Mdakane

School of Mathematics, Statistics and Computer Science

University of KwaZulu-Natal

Pietermaritzburg

South Africa

March 20, 2019

Declaration

I, Banele Phumlani Mdakane, declare that this thesis titled, 'Spatial and spatio-temporal modeling and mapping of self-reported health among individuals between the ages of 15 – 49 years in South Africa' and the work presented in it are my own. I confirm that this work was done wholly or mainly while in candidature for a research degree at this University. Where I have consulted the published work for others, this is always clearly attributed. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work. I have acknowledged all the main sources of help. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Mr. B.P. Mdakane (Author) Date

Prof. H. Mwambi (Supervisor) Date

Prof. B. Sartorius (Co-supervisor) Date

Acknowledgments

First and foremost, I thank God the almighty for giving me the strength to complete this research project. I deeply express my profound and genuine gratitude to my supervisor Prof. Henry Mwambi and co-supervisor Prof. Benn Sartorius for giving me the opportunity to do research and providing me with undivided assistance throughout this research. Both their guidance has taught me the approaches to present research work as plainly as possible and the methodology to carry out the research. Special thanks to Dr. Elphas Luchemo and Mr. Melkamu Dedefo for their assistance which made this research possible. Thanks to the School of Mathematics, Statistics and Computer Sciences stuff members who contributed to this research work.

I would like to thank the DELTAS Africa Initiative SSACAB and the National Research Foundation (NRF) for funding this research project, I am truly grateful for the financial support. Many thanks to the DELTAS stuff members for their flexible support. I also want to thank the National Income Dynamic Studies (NIDS) for making the datasets available to use on their website.

Finally, I thank my family and friends for encouraging me to complete my research. Special thanks to my mother (Nonhlanhla Mdakane) for her unconditional support and encouraging me not to give up. I thank my dad (Sizwe Mdakane) for his support, my brother (Lizwe Mdakane) for believing in me and my younger sister (Slindile Mdakane) for reminding me of how much I am her role model. Furthermore, I am thankful to my son (Lwande Mdakane) and my beautiful daughter (Awande Mdakane) for being my pillar of strength to accomplish this work. Thank you.

Abstract

Self-reported health has been commonly used as a measure of individuals health in public health studies. Health presents a complete physical, emotional, and social well-being. It also plays an important role in the development of the country, economically and socially. Poor health still remains a serious problem and it is linked to high burden of diseases in the world. As part of the Healthy People 2020 and Sustainable Development Goals (SDGs) in Sub-Saharan African (SSA), the goals of improving health has not been achieved. Hence, further investigation of the influential factors on health is relevant to improving health inequalities in SSA countries. Disease mapping provides a robust tool to assess geographical variation of disease and has been used in epidemiology and public health studies. The aim of this research is to use two distinct response outcome variables to investigate factors and geographical variations that are associated with self-reported health in South Africa. To accomplish the former and the latter, this research uses data from the National Income Dynamics Study (NIDS). The NIDS datasets are longitudinal data collected every two years from 2008. In this research, several structured additive regression (STAR) models were utilized within a Bayesian methodology, particularly the Bayesian hierarchical models. Models reviewed included Bayesian spatial and spatio-temporal cumulative logit models and logistic regression models, the primary interest was on the conditional autoregressive (CAR) models. Furthermore, the nonlinear effects of individuals age and body mass index (BMI) were part of the research interest. Two applications are discussed; one for the cumulative logit models for the ordinal response, the other for the logistic regression models of the binary response. In the case of the ordinal response, inference was based on the empirical Bayes approach, while for the binary case, a fully Bayesian procedure was used. Similar results were obtained between the two approaches.

Findings reveal that age, gender, household income, education, exercising level, alcohol consumption level, smoking, employment, nutrition status, TB, and depression were associated with self-reported health. The BMI was found to have a nonlinear relationship with self-reported health. Also, the findings show that age has a positive linear effect on self-reported health. In addition, the findings reveal significant spatial variation, with higher poor health prevalence in the Siyanda, John Taoli Gaetsewe, Ngaka Modiri Molema, Dr Ruth Segomotsi Mompati, Dr Kenneth Kaunda, Frances Baard, Lejweleputswa, Xhariep, Thabo Mofutsanyane, Fezile Dabi, Mangaung, Chris Hani, Umgungundlovu, Sisonke, Zululand, Umkhanyakude and Gert Sibande districts. Nevertheless, low poor health prevalence was recorded in the West Coast, Cape Winelands, Overberg, Eden, Central Karoo, Uthungulu, iLembe, and eThekwini districts. Interventions to improve individuals health should include addressing of gender inequalities, education, and income inequalities but altogether with employment status and healthy living lifestyle, in particular, targeting districts identified to have highest poor health prevalence.

Keywords: Self-reported health, conditional autoregressive (CAR) models, spatial and spatio-temporal models, structured additive regression (STAR) models and mapping.

Contents

List of Tables

List of Figures

Acronyms

AIC	Akaike information criterion
BIC	Bayesian information criterion
BMI	body mass index
BYM	Besag, York and Mollie
CAR	conditional autoregressive
CIs	credible intervals
CSMs	Continuing Sample Members
DIC	deviance information criterion
EB	empirical Bayes
EDA	exploratory data analysis
FB	fully Bayesian
GCV	generalized cross-validation
GLMs	generalized linear models
GLMM	generalized linear mixed model
GMRF	Gaussian Markov random field
iCAR	*intrinsic* conditional autoregressive
INLA	integrated nested Laplace approximation
IWLS	iteratively weighted least squares
MCMC	Markov chain Monte Carlo
MGLMs	multivariate generalized linear models
MRF	Markov random field
NIC	network information criterion
NIDS	National Income Dynamics Study
PORs	posterior odds ratios
PSUs	primary sampling units

REML	restricted maximum likelihood
SA	South Africa
SALDRU	Southern Africa Labour and Development Research Unit
SDG	Sustainable Development Goal
SSA	Sub-Saharan African
STAR	structured additive regression
TB	Tuberculosis
TIC	Takeuchi information criterion
TSMs	Temporary Sample Members

"We have distilled what we have learned from examining what has worked in Africa.... It is hoped that countries will take up the challenges of implementation or scale-up of proven effective interventions through strong health systems to move towards universal health coverage."

–Dr. Luis Gomes Sambo,

WHO Regional Director for Africa (WHO et al., 2014).

Chapter 1

Introduction

1.1 Background

Self-reported health is commonly used as a measure of health. Self-reported health also known as self-perceived health refers to individuals rating their own health. The use of self-reported health assessment may be biased due to direct contingency on social experience (Sen, 2002). That is, disadvantaged population tend to report worse health than advantaged population. Despite the biases, self-reported health has been used as a global measure of health (Wu et al., 2013). Health is a very important measure for a human well-being, especially at a young age. Worldwide health outcomes are pinpointed because the infrastructure for health across the lifespan is greatly determined by early exposures in life (Komro et al., 2014). Health and health outcomes are not influenced only by health care and health services, but also by other complex factors (Ataguba et al., 2015). Health is a crucial human right and would devote not only to a better quality of life but also to global peace and security as well as economic and social development (Moodley and Ross, 2015). Poor health is one of the major problems in the world, both in developing and developed countries. Africa has the greatest disease burden and poorest health services compared to any other continents (Grut et al., 2012). Thus, improvement of health inequalities is of great relevance in developing and developed countries.

Sub-Saharan African (SSA) countries are rated as poor countries with low incomes, hence they are likely to experience poor health. African development specialists and policymakers have attempted to improve the quality of life in several African countries but the health of SSA still remained lowest in the world (Fayissa and Gutema, 2005). The life expectancy of several African countries is around 50 years as compared to other countries, such as Japan, Sweden or Brazil. The main cause is that health heterogeneity arises globally depending on where an individual resides. About 42.4% of the SSA population were satisfied with the availability of the high-quality health care in the areas they reside in, which was rated lowest in the world recently (Angus and Robert, 2015). There is a strong link between health and primary health care, and it is acknowledged that individuals who visit health care more often are likely to have a better and healthy life. Nonetheless, other important determinants of health status include social, political, economic, and environmental factors (Fayissa and Gutema, 2005).

South Africa is one of the most established and developed countries in SSA with approximately two-thirds of its national population residing in urban areas (McGranahan and Martine, 2012). South Africa is known to be rated as one of the lower middle-income countries. Also, it is one of the countries that have the highest health inequalities (Ataguba et al., 2015). Colonization and apartheid in South Africa resulted in social and economic injustice which thereafter led to unemployment and increased urbanization (Weimann et al., 2016). Hence, the most relevant improvement of health inequality is social determinants. Unemployment in South Africa has been one of the attributes of increased poverty and has not improved for any better. Reducing the unemployment rate will not only reduce universal poverty but also play a huge role in improved and high-quality education. Both employment and education enhance better health, providing a foundation for economic and social development. The rates of unemployment are high for black African youths with lower levels of education and living in rural areas. In addition, about 51.3% and 30.1% of individuals between the ages of 15-24 and 25-34 years respectively were officially unemployed in 2014 (Motala et al., 2015).

The statistics show that males complete basic education at a slightly higher rate than females in South Africa (Statistics South Africa, 2015). Low level of education has a strong relationship with lower levels of socioeconomic factors. Lower well-being is highly affected by lower incomes in household level. The linkage between alcohol consumption and smoking plays a vast role in affecting individuals well-being, which further impacts populations health. Overall alcohol consumption in South Africa reduced in 2014 due to being expensive (WHO et al., 2014), but heavy drinking has not been successfully dealt with. The reason behind heavy drinking is that individuals tend to think alcohol is the answer to their problems, yet it increases the prevalence of poor health when consumed heavily for a long period of time. Furthermore, smoking is known to cause chronic diseases such as lung cancer among other diseases which negatively impact the health status of an individual. In contrast to the determinants of health, physical activities and a good diet are connected with better health.

Ill-health conditions have a strong link with self-reported health. Communicable and non-communicable diseases are a major cause of poor health. The burden of non-communicable diseases is two to three times higher in South Africa as compared to developed countries (WHO, 2008). The individuals residing in rural communities are the ones affected the most. Depression is considered among other non-communicable diseases as the greatest burden of health in the world (Househam, 2010). In small rural areas of South Africa, the depression rate was estimated to be around 27% and 25.2% in urban regions. Females are known to suffer more from depression than males, especially females who have children in regions which are not well established. Furthermore, other studies have linked depression with common influential factors which impact individuals health, and communicable diseases such as Tuberculosis (TB) is a major one. TB is one of the major causes of poor health and death in South Africa for the past decade, with about 380,000 estimated adult's cases of TB incidence in 2016 (WHO, 2015). Deficiency of safe water and sanitation facilities have a huge impact on the increasing number of infectious disease cases which also affects individuals health.

There is a lack of geographical studies on self-reported health due to the availability of geo-referenced health data. Most studies done on self-reported and self-rated health have applied classical models under frequentist settings to investigate the determinants of health. However, studies in epidemiology focus on analyzing the geographic variation for the severity of the disease. Bayesian hierarchical models are widely used in the context of disease mapping. This research aims at identifying influential factors on self-reported health status in South Africa by developing spatial and spatio-temporal models used in disease mapping.

1.2 Literature review

This section provides a review done by other authors in relation to health. The review discusses different statistical approaches previously used in assessing the relationship between self-reported health and influential factors. This section particularly helps in highlighting the research gap that can be filled based on previous methods and findings. Furthermore, it relates to and informs the current study to other previous studies based on similarities or new findings.

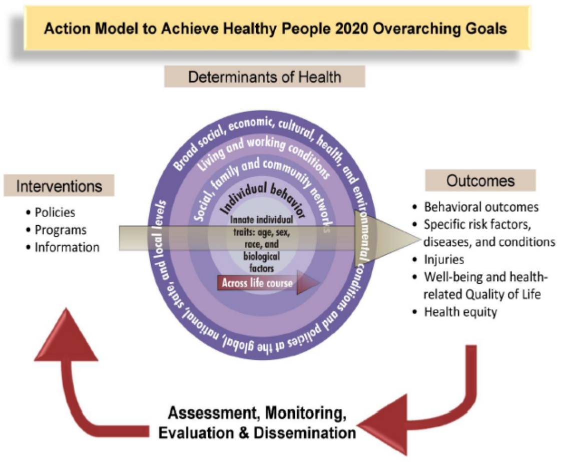

Action Model to Achieve Healthy People 2020 Overarching Goals
Determinants of Health
Interventions
• Policies
• Programs
• Information
Broad social, economic, cultural, health, and environmental conditions and policies at the global, national, state, and local levels
Living and working conditions
Social, family and community networks
Individual behavior
Innate individual traits: age, sex, race, and biological factors
Across life course
Outcomes
• Behavioral outcomes
• Specific risk factors, diseases, and conditions
• Injuries
• Well-being and health-related Quality of Life
• Health equity
Assessment, Monitoring, Evaluation & Dissemination

Figure 1.1: Conceptual framework action model to achieve Health People 2020 overarching goals (Source: https://www.health.ny.gov/statistics/chac/improvement/hp2020_action_model.htm).

Figure 1.1 shows the conceptual framework diagram of the action model adopted by the Healthy People 2020 to achieve its overarching goal by 2020. The figure shows assessment, monitoring, evaluation and dissemination of the outcomes after interventions of several determinants of health. In relation to this research, some of the determinants of health will be investigated accounting for specific regions or locations in general. Similar to other previous literature on the determinants of health some of this factors have been investigated. Hence, this framework provides an insight into which are the influential factors to pose attention on. Next, we review the literature on different methods used to investigate the determinants of health.

Health is globally known as the driving force of development both in developing and developed countries. An individual with good health is less susceptible to the burden of diseases especially when aging. In many studies the concern about health status is related to children, however, adults health is of equal importance as of children, especially at reproductive ages. Health improvement has been one of the world's objective from decades ago, but it requires a lot of resources and practice altogether with more strategic planning and development. This includes the Healthy People 2020 adopted in 2010 and the global development that is linked to the sustainable development goals (SDGs) to improve the lives of the poor by 2030. The health impact assessment (HIA) strategy is also the key to improved health as it aligns the potential factors affecting health. The important influential factors of health include social, environmental, lifestyle and individual factors (Lock, 2000).

There was an increase in obesity in 2012 in the low and middle-income countries which increased several types of cancers and other cardiovascular diseases (WHO, 2013). About nine out of ten individuals rate their health as good in developed countries. However, in Japan, Korea and Portugal about half of the population perceived their health as good or very good (OECD, 2013). On the other hand, the progress of the health-related millennium development goals (MDGs) in the world has been seen, dating back from

1990 to 2012. For example, target on the source of drinking water was met in 2012 with an 89% increase from 76% in 1990, while an improvement on basic sanitation had slow progress (WHO, 2013). In South Africa, the rate of smoking daily in 2011 was below 15%, with males having a steeper decline. However, with all the current improvements the SDGs are still ongoing processes to ensure health inequalities are reduced across countries population. The entire world still suffers from health and socioeconomic inequalities among individuals, more especially in Sub-Saharan Africa (WHO, 2016; Anita, 2013).

Many studies which have been done on adults self-reported health are based on frequencies approach, with the comparison between dichotomous and categorical response outcomes. Within the Bayesian spatial modeling under disease mapping, not much attention has been seen for modeling self-reported health data. In the study by Manor et al. (2000), a comparison of dichotomous and alternative categorical ordered response was conducted using the 1958 British cohort birth dataset. The study aim was to investigate the relationship between age, social class, socioeconomic status, and education on good rated health for both men and women at four age classes (birth, and ages 16, 23 and 33 years). The commonly used statistical models known as the logistic regression and cumulative odds were adopted for binary and ordinal self-rated health responses respectively in their study. Other alternate models for ordinal response were also investigated in their study, including the polytomous regression, continuation ratio, and adjacent categories model. The study noted that both the logistic regression and cumulative odds models which were considered yielded similar results for both men and women. The men and women social class age, socioeconomic status, and education were all significantly associated with increased odds of good self-rated health except for men social class aged 16. These reveal no gender differences in their study. The ordinal model assumption for all the considered ordinal models of parallelism was tested and none were violated. Logistic regression resulted in a more robust method than other models (Manor et al., 2000). This illustrates that different models may have similar results, but a certain model may fit the data better than other alternative models.

A study by Subramanian et al. (2010) was conducted in 69 countries, including Sub-Saharan Africa countries. The study examined the association between education and self-reported poor health while adjusting for age and gender variables at the global level. Self-reported health was assessed using a five-point Likert scale but was analyzed as a collapsed dichotomous response. The assumption that was made in this study was that education with missing values was replaced with no formal education. Data description analysis was conducted followed by a logistic regression application. The study used a simple random sampling and a multistage stratified sampling for 10 and 59 countries respectively from the 2002 World Health Survey. Results showed that respondents who had the highest percentage (48.9%) of reporting poor health were Swaziland. Swaziland is the smallest country which forms a border with South Africa and a northern border with Mozambique. The years of schooling were found to be higher in Belgium, France, and Israel, but lowest in many African countries. The results for the logistic regression model revealed that there was an inverse association between years of schooling and self-rated poor health in all countries for men and women. This implies that years of schooling did not result in similar findings from other studies. One would expect that the higher the years of education the higher the chances of better health. Furthermore, those individuals in the lowest quintile were twice as likely to rate poor health (Subramanian et al., 2010). Other studies have shown a strong link between education and population health (Johanson, 2001; Mirowsky and Ross, 2013; Brunello et al., 2016).

A similar study was done by Hosseinpoor et al. (2012), where they investigated social determinants of self-reported health in men and women from 57 countries using the 2002-2004 World Health Survey datasets. In their study, they examine how gender affects health in African and European countries. However, in their study, self-reported health was measured using the item response theory partial credit method ranging from 0 as worse to 100 as the best health status. The Multivariate linear regression was used in their analysis to assess the relationship between social determinants and health status among men and women. In all considered countries, individuals living in urban areas had better health mean score than those residing in rural areas. In Europe, individuals level of education and employment were seen to contribute less in explaining health inequalities than in African regions. Females were found to have significantly lower health status than males. Other factors like marital status and household income were seen to be significantly associated with health in all considered countries. The SDGs in all countries, which deals with gender inequality and women empowerment is likely to be achieved. These are possible by improving social policies pertaining to females empowerment within regions, women's perceived social status, well-being and aging, and other biological risk factors (Hosseinpoor et al., 2012).

A study by Phillips et al. (2005) was aimed at assessing the determinants of self-rated health for adults in Texas with chronic illness using data from the 2003 Behavioral Risk Factor Surveillance System survey. The data sample in their study was collected using a random digit dialing method and a five-point scale to measure self-rated health. In Texas, it was found that older individuals, women, low-income households, obese and none exercising individuals rated their health as poor. The study noted that higher education, non-Hispanic ethnicity, doing physical activities and a lower Body Mass Index (BMI) were systematically associated with better health status. In addition, the study indicated that potentially modifiable factors such as BMI and physical activities were most powerful to predict self-rated health. The study noted a link between socio-culture and self-rated health, such that individuals who were interviewed in Spanish were much likely to rate their health as poor than those interviewed in English. Phillips et al. (2005) used the multiple logistic regression for a dichotomous self-rated health to assess the predictor variables. The conclusion in their study stated that health care services and delivery based on cultural sensitivity must be considered.

Cau et al. (2016a) examined determinants associated with poor self-rated health among adults in Maputo metropolitan area in Mozambique. The study used data from Health Barometer: Individual and Community Health Promoting Practices in Maputo City.

However, the sample size was not large enough (n = 677) in their study. Their response variable was a six-point scale to rate individuals health. Multiple logistic regression model was adopted in their study. The variables age, gender, and marital status were found to be the key determinants of poor self-rated health. Being female and single respectively had higher odds of reporting poor health. Also, being a widow and separated or divorced had higher odds of reporting poor health. Other variables such as education, type of occupation, drinking water treatment, and physical activities were found to be significantly associated with self-rated health.

Many studies have not considered social capital as a determinant of health, however other determinants have been reviewed concerning self-rated health. A study conducted in South Africa was done by Cramm and Nieboer (2011) to identify the role of social capital along with socioeconomic conditions on self-rated health for economically and health deprived communities. In this study, it was revealed that social capital was significantly associated with self-rated health. Social capital among other predictor variables like employment was found to be aligned with increased odds of rating good health. The study was based on a survey administered in Rhiri, in the Eastern Cape, South Africa. The ordinal logistic regression model was used to assess the covariates. The importance of social capital, education, and employment in low-income regions in South Africa is shown in their study.

The Logistic regression model and the ordinal logistic regression model are the most popular employed models when the nature of the response variable is binary and categorical respectively. Both these models can be extended to multilevel modeling, where the incorporation of random effects is considered. The random effects account for the unobserved heterogeneity of the covariates under study. A recent study by Lau and Ataguba (2015) examined the association between social capital and self-rated health. The study used the data from the NIDS wave 1 and wave 2 in South Africa. The hierarchical linear regression model was adopted to identify the covariates which correlate with self-rated health. The findings suggested that both individual and contextual-level social capital were significantly associated with self-rated health (Lau and Ataguba, 2015).

1.2.1 Spatial and Spatio-temporal modeling

In recent literature, mapping of diseases has been a fundamental objective to identify geographical variation of disease risk. The availability of geo-referenced and time framed data has made disease mapping methodologies broader and important in recent studies. Bayesian hierarchical modeling is commonly used in the context of disease mapping. The use of spatial and spatio-temporal models provide a unified framework to handle complex data with multilevel characteristics. The two models are statistical tools used to estimate disease risk parameters. Application and development of disease mapping techniques date way back from decades ago, these include literature on Clayton and Kaldor (1987), Besag et al. (1991), Bernardinelli and Montomoli (1992) and Best et al. (2005) among others.

In disease mapping, the Bayesian methods are very popular. Clayton and Kaldor (1987) used the empirical Bayes (EB) approach to estimate disease risk based on two simple models, namely the log-Normal and Poisson-gamma models. Their study assumed a conditional autoregressive (CAR) model for the spatial correlation. In contrast to the random effects, they also allowed for a non-parametric form of random effects contrary to the spatial correlation. However, the EB estimation approach provides estimates that are contracted towards the local or global mean and have been criticized. Other studies have considered the use of EB approach for disease risk estimation (Marshall, 1991; Devine and Louis, 1994; Kneib and Fahrmeir, 2006).

The fully Bayesian (FB) approach in disease mapping was proposed by Besag et al. (1991). Their model is generally referred to as the Besag, York and Mollié (BYM) model. In their study, they proposed the incorporation of spatial random components that can split into two spatial random effects. These two spatial components were assumed to be independent and were assigned different Gaussian priors. A study by Best et al. (2005)

also used a similar approach, adopting a FB approach. In both these studies, a CAR model was used for the spatial correlation. A FB approach in the estimation of disease risk for interaction between space and time under spatio-temporal modeling was introduced by Bernardinelli et al. (1995). Their study proposed spatio-temporal modeling which accounts for space and time components. In the mentioned studies, the estimation of parameters was obtained using Markov chain Monte Carlo (MCMC) techniques. Other studies have also considered the FB approach under spatio-temporal modeling in disease mapping (Waller et al., 1997; Knorr-Held, 2000).

A comparison of EB and FB estimation approach was done by Bernardinelli and Montomoli (1992). In their study, they considered models that are used in the geographic variation of diseases. To be more specific, they considered spatial Bayesian hierarchical modeling with an assumption of a Poisson and multivariate distributed models. The comparison of the two estimation procedure was based on the study of cancer mortality at district levels of Sardinia. Bernardinelli and Montomoli (1992) considered two prior models for the spatial random effects, evaluated separately in turn. The prior models that they adopted in their study were the independent identical distributed and the CAR model. In their study, they pinpointed that FB is more powerful and preferable estimation approach than the EB approach. The FB estimation considers the uncertainty of the model parameters while the EB procedure conditions estimation on point estimates by approximation (Bernardinelli and Montomoli, 1992). However, the FB estimation approach is known to be computationally expensive in terms of time as compared to the EB estimation procedures.

1.2.2 Research on self-reported health using spatial and spatio-temporal models

According to our knowledge, few studies have been done on spatial and spatio-temporal modeling of self-reported health, especially in Sub-Saharan African (SSA). However, few recent studies have examined the association between health and geographic factors. The following are some of the studies that have considered the geographical variation of self-rated health.

To examine the spatial distribution of perceived environmental hazards and self-rated health while adjusting for other influential factors in the districts of Beijing, Ma et al. (2017) used Bayesian spatial multilevel logistic regression models. Their model incorporated a CAR model developed by Leroux et al. (2000) for the spatial random effects. Their study found that lower odds of rating poor health were in northern-western and central regions of Beijing. Furthermore, they found that environmental hazards, such as air, noise, and landfill pollution were significantly associated with higher odds of poor self-rated health.

On the other hand, a study by Cabrera-Barona (2017) examined the influence of urban multi-criteria deprivation and spatial accessibility to health on self-reported health in the city of Quito, Ecuador. The study used a multilevel logistic regression model to assess the relationship between the random effects on self-reported health. Mapping was done on urban multi-criteria deprivation and spatial accessibility to investigate area level influence on self-reported health. This study found lower levels of deprivation on self-reported health in Quito. The study also found a spatial variation in health care accessibility, with highest accessibility on the north, central and southern regions of Quito.

The study by Browning et al. (2003) investigated how health status varied across space and time. Their study used data from the city of Chicago in the USA between the period of 1990 to 1999. Health status in this study was first reported on a four-point scale and was then collapsed into a binary response. The study adopted a three-level (individual, temporal, and spatial components) hierarchical logit models to assess the variation and factors influencing the individual's health status. However, the assumption made on this study was that space and time are of parametric form. They found that neighborhoods in Chicago were not spatially dependent. They also found that health status improved across the time periods.

1.3 Significance of the Study

In order to improve and promote good health, more health-care systems should be established. Despite the lack of healthcare institutes and resources, understanding the determinants of health is important to further improving population health. Many major problems in different regions are related to interaction and behavior. Some of these problems are connected to the social, socioeconomic, living environment and lifestyle of individuals as an entire population. The field of biostatistics and epidemiology deals with issues of public health both in developing and developed countries. Poor health is one of the major public health problems in many countries. This issue is one of the most crucial topics, not only in developing countries but also in developed countries. Understanding the key factors which are associated with poor health can lead to problem-solving implementations and programme developments. There is this aspect of elucidating spatial and spatio-temporal patterns or distribution or heterogeneity of poor health. This study will provide concrete statistical approaches including modeling techniques for those key factors linked to reporting poor health among adults. Hence, the contribution of this study will be of interest to biostatisticians and epidemiologists, particularly in health-care practitioners. Furthermore, this study will be beneficial to society, but mostly to the government for intervention and policymakers. Nonetheless, the contributions of this study are not expected to be restricted to health care context, and should not be exclusive in any institution aiming to achieve model development for solving public health problems.

1.4 Problem Statement

Health status is a crucial measure of an individuals well-being both physical and emotional. It plays a huge role in the social and economic development of a country. Poor health has been identified as one of the causes of mortality and morbidity. In South Africa improving health and well-being has been one of the Healthy People 2020 and Sustainable Development Goals (SDGs) implementations, more especially for children. However, these goals have not been achieved for all the provinces and population in South Africa. Hence improving health increases the reduction of mortality and morbidity. In this research, we address the issue of poor self-reported health using the National Income Dynamics Study (NIDS) adults datasets, by developing suitable statistical methods that can help us understand the underlying factors on self-reported health in South Africa. Further, this research explores the geographic variations of poor self-reported health and its changes over time.

1.5 Objectives

The aim of this research project is to investigate the determinants and geographic variation of self-reported health status using adults national income dynamic studies (NIDS) datasets in South Africa.

1.6 Specific objectives

The specific objectives of this research are:

- To review statistical methods for discrete choice outcomes used in disease mapping.
- To identify appropriate spatial models for modeling poor health at the district level

using wave 4 of the NIDS data in South Africa.

- To identify suitable spatio-temporal models for modeling and mapping poor health in South Africa at the district level using waves 1 to 4 of the NIDS data in South Africa.
- To develop and extend suitable models desirable for investigating influential factors associated with self-reported health status in the district level of South Africa.

1.7 Structure of the Thesis

The second chapter of this research focuses on exploring the data and checking of relevant assumptions. In the third chapter, we review Bayesian spatial models used in structured additive regression (STAR) models. In the fourth chapter we review Bayesian spatial models with extension to STAR models for ordinal outcomes, this includes application to NIDS data and model comparison. In the fifth chapter we review Bayesian spatial models with extension to STAR models for binary outcomes, this includes application to NIDS data and model comparison. In the sixth chapter, we discuss spatio-temporal models for ordinal and binary outcomes, this includes mapping self-reported poor health in South Africa using the four available waves of NIDS data and models comparison of several models. In the last chapter, we discuss the findings then give a conclusion and future research recommendations.

Chapter 2

Exploratory Data Analysis

2.1 Introduction

The most essential attribute to consider before making any inference from the data is to examine all the variables in that data. This is known as the exploratory data analysis (EDA). The primary aim of the EDA is to examine the following:

- to capture mistakes,
- to find violations of statistical assumptions,
- to generate hypotheses,
- to observe patterns in the data,
- to figure out relationships between the response and risk factors.

This chapter presents a detailed data description. We describe how the variables were classified. The cross-tabulations for the variable of interest against the selected independent variables were assessed using the Pearson chi-square test. Also, the proportional odds assumptions were checked for the case of the ordinal response variable. Furthermore, we tested for multicollinearity among the independent variables. All the analysis in this chapter were carried out in Stata version 14.1 (StataCorp, 2015) and R statistical software version 3.4.2 (R Core Team, 2017).

2.2 Data Source

The datasets used for this research project were extracted from the representation of the National Income Dynamics Study (NIDS), which is conducted by the Southern Africa Labour and Development Research Unit (SALDRU) situated at the University of Cape Town. The NIDS is the first longitudinal study done in South Africa (SA) that aims at describing and explaining the changes in socioeconomic indexes, such as education, income, expenditures, assets, access to services, health, and well-being (Leibbrandt et al., 2009). The data is an individual level panel survey that has a sample of over 28,000 individuals in 7,300 households across the country which is collected biannually, with four waves available. The first wave of the NIDS was collected in 2008, the second between 2010-2011, the third in 2012, and the fourth between 2014-2015. The survey employed a stratified, two-stage cluster sampling design to collect a representative sample from the private households in all nine provinces in SA and residents in worker's hostels, convents, and monasteries (Leibbrandt et al., 2009). In the first stage of sampling, 400 primary sampling units (PSUs) were chosen from the master sample of 3000 PSUs classified by Stats SA in 2003. In the second stage, the sample was proportionally allocated to the strata based on the master sample of 53 district councils and 400 PSUs were randomly chosen within each stratum (Leibbrandt et al., 2009).

An important component of the longitudinal study is its capability to follow individuals over a period of time. Also, it allows researchers and policy analysts to see how households and individuals are impacted over time (Leibbrandt et al., 2009). However, there are disadvantages to the occurrence of missing data. The NIDS uses trained enumerators to collect the data using three sets of questionnaires at both the individual and household levels. The individual's questionnaire was designed to collect information on adults aged 15 years and older. The second one was designed to collect information for children younger than 15 years. The third one, known as the household questionnaire was designed to collect information about the household head and the characteristics of the dwelling place. Individuals are classified as either the Continuing Sample Members (CSMs) or Temporary Sample Members (TSMs). CSMs are interviewed in every wave of NIDS whereas TSMs are interviewed only in the wave(s) that they are co-resident with a CSMs. In this research project, we use the NIDS adults datasets for individuals between the ages of 15-49 years.

2.3 Study variables

2.3.1 Dependent variable: Self-reported health status

The main variable of interest in this research is the self-reported health status of adults between 15 and 49 years of age. Self-reported health status in the NIDS survey was assessed using a five-point Likert scale through the question: ”How would you describe your health at present? Would you say it is poor, fair, good, very good, or excellent?”. Self-reported health in this research was analyzed using two different techniques. First, we considered self-reported health as an ordinal variable. Secondly, we dichotomized self-reported health status as good health (excellent, very good or good) and poor health (fair or poor) following previous studies by Manor et al. (2000), Kawachi et al. (1999) and Lamarca et al. (2013). The ordinal response variable is classified as 1 = ”excellent”, 2 = ”very good”, 3 = ”good”, 4 = ”fair” and 5 = ”poor” and the binary response is classified as 1 = ”poor” and 0 = ”good” in all the analysis to be performed in this research project.

2.3.2 Explanatory variables

Several explanatory variables were collected in all the NIDS wave surveys. However, the variables that were considered in this research were based on some of the variables used in the previous studies by Reichmann et al. (2009) and Hosseinpoor et al. (2012)

on adults self-reported health. The explanatory variables which were considered are described further. The data includes the demographic, socio-economic, and lifestyle characteristics. The demographic characteristics included the individual's age, gender, race, marital status, and life satisfaction level. Socio-economic characteristics included the individual's type of residence, education level, household income, and employment status. The lifestyle characteristics included alcohol consumption level, exercising level, smoking cigarette, and nutrition status. Moreover, we included comorbidity factors which included depression and Tuberculosis (TB). Lastly, we included environmental factors such as household water source and type of toilet facilities. We derived some of the original variables to newly defined categorical variables. These newly defined variables included education level (no schooling, primary, secondary, high, college and tertiary), marital status (never married, widow/divorced/separated and married/living with partner), household water source (adequate, inadequate and other) and life satisfaction level (very dissatisfied, dissatisfied, neutral, satisfied and very satisfied). Also, the nutrition status was derived from the body mass index (BMI) using the standard formula (weight in kg)/(height in m)^2.

2.4 Preliminary analysis

It is very important to understand the baseline characteristics of the individuals selected in the study. This section presents the description of the NIDS wave 4 data. The sample distribution along with the Pearson's chi-square test p-values of the covariates will be discussed. The data consisted of 22,752 individuals of age 15 years and above in total. Out of 22,752, there were 21,400 individuals between the ages of 15 - 49 years which were considered in this study. The health status of individuals was recorded, where missing values were dropped so that the final sample size was 15,795, representing 73.8% of the original sample.

The distribution of self-reported health is presented in Table 2.1. It can be observed that most of the sample was from the respondents who reported excellent health 5,691 (36.03%), while the smallest sample was from respondents who reported poor health 199 (1.26%). The sample distribution for respondents who reported Very good, Good and Fair health status were 5,169 (32.73%), 4,056 (25.68%) and 680 (4.31%) respectively.

Table 2.1: Frequency distribution of self-reported health status in the National Income Dynamics Study (NIDS) wave 4.

Health status	**Response**	**Frequency**	**Percent**
Poor	1	199	1.26
Fair	2	680	4.31
Good	3	4,056	25.68
Very good	4	5,169	32.73
Excellent	5	5,691	36.03
Total	-	15,795	100.00

This research consists of covariates which are categorical variables, thus we further explore the proportion for each variable. Since the purpose of this research is to investigate factors associated with self-reported health, we discuss the distribution of the sample based on health status for each covariate.

Table 2.2: Multiple univariate two-way contingency table analysis of ordinal outcome health status by covariate categories classification.

	Health status						
Covariates	**Excellent**	**Very good**	**Good**	**Fair**	**Poor**	**Total**	**p-value**
	N (%)	**N (%)**	**N (%)**	**N (%)**	**N (%)**	**N**	
Age group							<0.001
15-19	1424 (40.51)	1204 (34.25)	790 (22.48)	71 (2.02)	26 (0.74)	3515	
20-24	1288 (40.36)	1070 (33.53)	747 (23.41)	68 (2.13)	18 (0.56)	3191	
25-29	1022 (37.98)	895 (33.26)	682 (25.34)	77 (2.86)	15 (0.56)	2691	
30-34	767 (37.20)	663 (32.15)	524 (25.41)	83 (4.03)	25 (1.21)	2062	
35-39	524 (31.97)	525 (32.03)	449 (27.39)	114 (6.96)	27 (1.65)	1639	
40-44	387 (27.60)	449 (32.03)	414 (29.53)	112 (7.99)	40 (2.85)	1402	
45-49	279 (21.54)	363 (28.03)	450 (34.75)	155 (11.97)	48 (3.71)	1295	
Gender							<0.001
Female	2977 (33.55)	2949 (33.23)	2393 (26.97)	435 (4.90)	120 (1.35)	8874	
Male	2714 (39.21)	2220 (32.08)	1663 (24.03)	245 (3.54)	79 (1.14)	6921	
Race							0.040

Table 2.2 *Continues*

	Health status						
Covariates	**Excellent** **N (%)**	**Very good** **N (%)**	**Good** **N (%)**	**Fair** **N (%)**	**Poor** **N (%)**	**Total** **N**	**p-value**
African	4827 (36.05)	4397 (32.84)	3406 (25.44)	587 (4.38)	171 (1.28)	13388	
Asian/Indian	47 (35.61)	49 (37.12)	26 (19.70)	7 (5.30)	3 (2.27)	132	
Coloured	745 (35.90)	661 (31.86)	574 (27.66)	74 (3.57)	21 (1.01)	2075	
White	72 (36.00)	62 (31.00)	50 (25.00)	12 (6.00)	4 (2.00)	200	
Type of residence							<0.001
Urban Informal	490 (38.04)	401 (31.13)	320 (24.84)	57 (4.43)	20 (1.55)	1288	
Rural Formal	536 (34.12)	527 (33.55)	425 (27.05)	60 (3.82)	23 (1.46)	1571	
Urban Formal	2368 (35.66)	2044 (30.78)	1823 (27.45)	319 (4.80)	87 (1.31)	6641	
Tribal Authority Areas	2297 (36.49)	2197 (34.90)	1488 (23.64)	244 (3.88)	69 (1.10)	6295	
Education level							<0.001
No schooling	55 (19.86)	74 (26.71)	91 (32.85)	40 (14.44)	17 (6.14)	277	
Primary	82 (19.52)	106 (25.24)	160 (38.10)	50 (11.90)	22 (5.24)	420	
Secondary	455 (27.74)	541 (32.99)	495 (30.18)	109 (6.65)	40 (2.44)	1640	
High	4042 (37.48)	3619 (33.56)	2654 (24.61)	367 (3.40)	103 (0.96)	10785	
College	251 (36.22)	232 (33.48)	171 (24.68)	32 (4.62)	7 (1.01)	693	
Tertiary	806 (40.71)	597 (30.15)	485 (24.49)	82 (4.14)	10 (0.51)	1980	
Household income							<0.001
Much below average	1100 (40.00)	773 (28.11)	696 (25.31)	133 (4.84)	48 (1.75)	2750	
Below average	1460 (34.93)	1411 (33.76)	1044 (24.98)	205 (4.90)	60 (1.44)	4180	
Average	2453 (34.75)	2368 (33.54)	1872 (26.52)	286 (4.05)	81 (1.15)	7060	
Above average	444 (38.18)	372 (31.99)	294 (25.28)	44 (3.78)	9 (0.77)	1163	
Much above average	234 (36.45)	245 (38.16)	150 (23.36)	12 (1.87)	1 (0.16)	642	
Marital status							<0.001
Never married	4306 (37.74)	3829 (33.56)	2777 (24.34)	387 (3.39)	111 (0.97)	11410	
Widow/Divorced/Seperated	136 (26.77)	121 (23.82)	189 (37.20)	41 (8.07)	21 (4.13)	508	
Married/living with partner	1249 (32.22)	1219 (31.44)	1090 (28.11)	252 (6.50)	67 (1.73)	3877	
Alcohol consumption level							<0.001
Never drunk alcohol	3373 (36.14)	3216 (34.46)	2297 (24.61)	355 (3.80)	92 (0.99)	9333	
No longer drink alcohol	568 (33.89)	512 (30.55)	451 (26.91)	104 (6.21)	41 (2.45)	1676	
Drink very rarely	1019 (37.75)	779 (28.86)	750 (27.79)	118 (4.37)	33 (1.22)	2699	
Less than once a week	193 (33.68)	196 (34.21)	161 (28.10)	21 (3.66)	2 (0.35)	573	
1 or 2 days a week	431 (37.48)	351 (30.52)	289 (25.13)	57 (4.96)	22 (1.91)	1150	
3 or 4 days a week	75 (29.88)	78 (31.08)	75 (29.88)	17 (6.77)	6 (2.39)	251	
5 or 6 days a week	15 (24.59)	24 (39.34)	18 (29.51)	3 (4.92)	1 (1.64)	61	
Every day	17 (32.69)	13 (25.00)	15 (28.85)	5 (9.62)	2 (3.85)	52	
Employment status							0.020
Unemployed_Strict	912 (38.11)	749 (31.30)	608 (25.41)	98 (4.10)	26 (1.09)	2393	

Table 2.2 *Continues*

	Health status						
Covariates	**Excellent** **N (%)**	**Very good** **N (%)**	**Good** **N (%)**	**Fair** **N (%)**	**Poor** **N (%)**	**Total** **N**	**p-value**
Unemployed_Discouraged	89 (38.86)	91 (39.74)	43 (18.78)	4 (1.75)	2 (0.87)	229	
Not Economically Active	2401 (35.99)	2216 (33.22)	1681 (25.20)	280 (4.20)	93 (1.39)	6671	
Employed	2289 (35.20)	2113 (32.50)	1724 (26.51)	298 (4.58)	78 (1.20)	6502	
Exercising level							<0.001
Never	3477 (33.90)	3365 (32.81)	2749 (26.80)	509 (4.96)	156 (1.52)	10256	
Less than once a week	511 (41.34)	409 (33.09)	272 (22.01)	34 (2.75)	10 (0.81)	1236	
Once a week	326 (39.52)	245 (29.70)	211 (25.58)	30 (3.64)	13 (1.58)	825	
Twice a week	385 (37.63)	322 (31.48)	272 (26.59)	36 (3.52)	8 (0.78)	1023	
Three or more times a week	992 (40.41)	828 (33.73)	552 (22.48)	71 (2.89)	12 (0.49)	2455	
Type of toilet facility							<0.001
None	159 (30.46)	181 (34.67)	146 (27.97)	28 (5.36)	8 (1.53)	522	
Flush toilet with offsite disposal	1359 (37.26)	1068 (29.28)	998 (27.36)	168 (4.61)	54 (1.48)	3647	
Flush toilet with onsite disposa	1514 (34.84)	1392 (32.03)	1194 (27.47)	197 (4.53)	49 (1.13)	4346	
Bucket toilet	178 (36.55)	186 (38.19)	97 (19.92)	21 (4.31)	5 (1.03)	487	
Chemical toilet	109 (29.95)	100 (27.47)	130 (35.71)	17 (4.67)	8 (2.20)	364	
Pit latrine with ventilation pipe	993 (39.06)	861 (33.87)	548 (21.56)	108 (4.25)	32 (1.26)	2542	
Pit latrine without ventilation pipe	1376 (35.65)	1366 (35.39)	937 (24.27)	138 (3.58)	43 (1.11)	3860	
Other	3 (11.11)	15 (55.56)	6 (22.22)	3 (11.11)	0 (0.00)	27	
Smokes cigarette							<0.001
No	4704 (36.47)	4263 (33.05)	3264 (25.31)	515 (3.99)	152 (1.18)	12898	
Yes	987 (34.07)	906 (31.27)	792 (27.34)	165 (5.70)	47 (1.62)	2897	
Felt depressed in past week?							<0.001
Less than 1 day	3519 (39.11)	2946 (32.74)	2175 (24.17)	289 (3.21)	68 (0.76)	8997	
Little of the time (1-2 days)	1540 (31.78)	1666 (34.38)	1327 (27.38)	237 (4.89)	76 (1.57)	4846	
Occasionally (3-4 days)	530 (33.69)	464 (29.50)	424 (26.95)	118 (7.50)	37 (2.35)	1573	
All of the time (5-7 days)	102 (26.91)	93 (24.54)	130 (34.30)	36 (9.50)	18 (4.75)	379	
Diagnosed with tuberculosis (TB)?							<0.001
No	5575 (36.61)	4994 (32.80)	3898 (25.60)	604 (3.97)	156 (1.02)	15227	
Yes	116 (20.42)	175 (30.81)	158 (27.82)	76 (13.38)	43 (7.57)	568	
Household water source							0.280
Adequate	4028 (36.01)	3621 (32.37)	2909 (26.01)	490 (4.38)	137 (1.22)	11185	
Inadequate	1585 (35.79)	1493 (33.72)	1110 (25.07)	182 (4.11)	58 (1.31)	4428	
Other	78 (42.86)	55 (30.22)	37 (20.33)	8 (4.40)	4 (2.20)	182	
Life satisfaction level							<0.001
Very dissatisfied	514 (35.16)	510 (34.88)	328 (22.44)	77 (5.27)	33 (2.26)	1462	
Dissatisfied	1354 (34.41)	1378 (35.02)	993 (25.24)	162 (4.12)	48 (1.22)	3935	
Neutral	1794 (35.14)	1647 (32.26)	1350 (26.44)	248 (4.86)	66 (1.29)	5105	

Table 2.2 *Continues*

	Health status						
Covariates	**Excellent**	**Very good**	**Good**	**Fair**	**Poor**	**Total**	**p-value**
	N (%)	**N (%)**	**N (%)**	**N (%)**	**N (%)**	**N**	
Satisfied	1294 (36.97)	1152 (32.91)	891 (25.46)	125 (3.57)	38 (1.09)	3500	
Very satisfied	735 (40.99)	482 (26.88)	494 (27.55)	68 (3.79)	14 (0.78)	1793	
Nutrition status							<0.001
Normal	2870 (38.02)	2435 (32.26)	1865 (24.71)	283 (3.75)	96 (1.27)	7549	
Underweight	332 (34.16)	315 (32.41)	263 (27.06)	47 (4.84)	15 (1.54)	972	
Overweight/obese	2377 (34.24)	2311 (33.29)	1854 (26.70)	328 (4.72)	73 (1.05)	6943	
Severe	112 (33.84)	108 (32.63)	74 (22.36)	22 (6.65)	15 (4.53)	331	

Table 2.2 presents the results of the multiple univariate two-way classifications, along with the chi-square (χ^2) test of association of the selected covariates. The covariates are derived from the respondents demographic, socioeconomic status, environmental and lifestyle factors. The results reveal that among individuals aged 40-44 and 45-49 years 2.85% and 3.71% of them respectively reported poor health. For gender, 1.35% of females and 1.14% of males reported poor health. Among individuals who are African, 1.28% of them reported poor health and 36.05% of them reported excellent health. On the other hand, 2.27% of respondents who reported poor health and 35.61% who reported excellent health were Asian/Indian. Among individuals residing in Urban informal areas, 38.04% of them reported excellent health and 1.55% of them reported poor health. Those residing in Rural formal areas 34.12% of them reported excellent health and 1.46% of them reported poor health. It can be observed that there is an increasing trend of reporting excellent health and a decreasing trend of reporting poor health as education level increases. Among individuals with tertiary education, 40.71% of them reported excellent health and 0.51% of them reported poor health. Individuals from the household with much below average income 25.31% of them reported good health and 4.84% of them reported fair health. Those from the household with average income 34.75% of them reported excellent health and 1.15% of them reported poor health. Furthermore, among individuals from the household with much above average income 36.45% of them reported excellent health

and 0.16% of them reported poor health. Individuals who were never married reported less poor health (0.97%) than those who we married or living with a partner (1.73%). On the other hand, among those who were either widow/divorced/separated, 4.13% of them reported poor health and 26.77% of them reported excellent health. We observed that among individuals who drank every day, 3.85% of them reported poor health while those who never drunk alcohol 36.14% of them reported excellent health. Those who drank less than once a week 34.21% of them reported very good health and 3.66% of them reported fair health. We also observed that respondents who were not economically active 1.39% of them reported poor health. Those individuals who never exercise 1.52% of them reported poor health, while those who exercise twice a week 0.78% of them reported poor health. Respondents with flushing toilets that have offsite and onsite disposal had an almost equal proportion of poorly reported health (1.48% and 1.13% respectively). Whereas households with other type of toilet 11.11% of them reported excellent health and none reported poor health. Among those respondents who smoke a cigarette, 34.07% of them reported excellent health and 1.62% of them reported poor health. However, those who do not smoke a cigarette, 1.18% of them reported poor health. This reveals that people who smoke a cigarette are more likely to have poor health than those who do not smoke. It can be observed that the highest proportion of reporting poor health is among those who felt depressed all the time (4.75%) while those who felt depressed less than a day had the highest proportion of reporting excellent health (39.11%). Among individuals who were diagnosed with TB, 20.42% of them reported excellent health and 7.57% of them reported poor health. Respondents living in households with adequate water, about 1.22% of them reported poor health, while those living in households with inadequate water 1.31% of them reported poor health. The results reveal that as life satisfaction level increases, the proportion of individuals reporting excellent health also increases while reporting poor decreases. Among individuals with overweight/obese nutrition status, 34.24% of them had excellent health and 1.05% of them had poor health. Furthermore, those individuals with severe malnutrition 4.53% of them had poor health and 33.84% of them had excellent health.

Table 2.3: Multiple univariate two-way contingency table analysis of binary outcome health status by covariate categories classification.

	Health status			
Covariates	**Good** **N (%)**	**Poor** **N (%)**	**Total** **N**	**p-value**
Age group (years)				<0.001
15-19	3418 (97.24)	97 (2.76)	3515	
20-24	3105 (97.30)	86 (2.70)	3191	
25-29	2599 (96.58)	92 (3.42)	2691	
30-34	1954 (94.76)	108 (5.24)	2062	
35-39	1498 (91.40)	141 (8.60)	1639	
40-44	1250 (89.16)	152 (10.84)	1402	
45-49	1092 (84.32)	203 (15.68)	1295	
Gender				<0.001
Female	8319 (93.75)	555 (6.25)	8874	
Male	6597 (95.32)	324 (4.68)	6921	
Race				0.050
African	12630 (94.34)	758 (5.66)	13388	
Asian/Indian	122 (92.42)	10 (7.58)	132	
Coloured	1980 (95.42)	95 (4.58)	2075	
White	184 (92.00)	16 (8.00)	200	
Type of residence				0.030
Urban Informal	1211 (94.02)	77 (5.98)	1288	
Rural Formal	1488 (94.72)	83 (5.28)	1571	
Urban Formal	6235 (93.89)	406 (6.11)	6641	
Tribal Authority Areas	5982 (95.03)	313 (4.97)	6295	
Education level				<0.001
No schooling	220 (79.42)	57 (20.58)	277	
Primary	348 (82.86)	72 (17.14)	420	
Secondary	1491 (90.91)	149 (9.09)	1640	
High	10315 (95.64)	470 (4.36)	10785	
College	654 (94.37)	39 (5.63)	693	
Tertiary	1888 (95.35)	92 (4.65)	1980	
Household income				<0.001
Much below average	2569 (93.42)	181 (6.58)	2750	
Below average	3915 (93.66)	265 (6.34)	4180	
Average	6693 (94.80)	367 (5.20)	7060	
Above average	1110 (95.44)	53 (4.56)	1163	
Much above average	629 (97.98)	13 (2.02)	642	

Table 2.3 *Continues*

	Health status			
Covariates	**Good** **N (%)**	**Poor** **N (%)**	**Total** **N**	**p-value**
Marital status				<0.001
Never married	10912 (95.64)	498 (4.36)	11410	
Widow/Divorced/Seperated	446 (87.80)	62 (12.20)	508	
Married/living with partner	3558 (91.77)	319 (8.23)	3877	
Alcohol consumption level				<0.001
Never drunk alcohol	8886 (95.21)	447 (4.79)	9333	
No longer drink alcohol	1531 (91.35)	145 (8.65)	1676	
Drink very rarely	2548 (94.41)	151 (5.59)	2699	
Less than once a week	550 (95.99)	23 (4.01)	573	
1 or 2 days a week	1071 (93.13)	79 (6.87)	1150	
3 or 4 days a week	228 (90.84)	23 (9.16)	251	
5 or 6 days a week	57 (93.44)	4 (6.56)	61	
Every day	45 (86.54)	7 (13.46)	52	
Employment status				0.160
Unemployed_Strict	2269 (94.82)	124 (5.18)	2393	
Unemployed_Discouraged	223 (97.38)	6 (2.62)	229	
Not Economically Active	6298 (94.41)	373 (5.59)	6671	
Employed	6126 (94.22)	376 (5.78)	6502	
Exercising level				<0.001
Never	9591 (93.52)	665 (6.48)	10256	
Less than once a week	1192 (96.44)	44 (3.56)	1236	
Once a week	782 (94.79)	43 (5.21)	825	
Twice a week	979 (95.70)	44 (4.30)	1023	
Three or more times a week	2372 (96.62)	83 (3.38)	2455	
Type of toilet facilities				0.090
None	486 (93.10)	36 (6.90)	522	
Flush toilet with offsite disposal	3425 (93.91)	222 (6.09)	3647	
Flush toilet with onsite disposal	4100 (94.34)	246 (5.66)	4346	
Bucket toilet	461 (94.66)	26 (5.34)	487	
Chemical toilet	339 (93.13)	25 (6.87)	364	
Pit latrine with ventilation pipe	2402 (94.49)	140 (5.51)	2542	
Pit latrine without ventilation pipe	3679 (95.31)	181 (4.69)	3860	
Other	24 (88.89)	3 (11.11)	27	
Smokes cigarette				<.0.001
No	12231 (94.83)	667 (5.17)	12898	
Yes	2685 (92.68)	212 (7.32)	2897	
Felt depressed in the past week?				<.0.001

Table 2.3 *Continues*

	Health status			
Covariates	**Good** **N (%)**	**Poor** **N (%)**	**Total** **N**	**p-value**
Less than 1 day	8640 (96.03)	357 (3.97)	8997	
Little of the time (1-2 days)	4533 (93.54)	313 (6.46)	4846	
Occasionally (3-4 days)	1418 (90.15)	155 (9.85)	1573	
All of the time (5-7 days)	325 (85.75)	54 (14.25)	379	
Was diagnosed with tuberculosis (TB)?				<.0.001
No	14467 (95.01)	760 (4.99)	15227	
Yes	449 (79.05)	119 (20.95)	568	
Household water source				0.740
Adequate	10558 (94.39)	627 (5.61)	11185	
Inadequate	4188 (94.58)	240 (5.42)	4428	
Other	170 (93.41)	12 (6.59)	182	
Life satisfaction level				<.0.001
Very dissatisfied	1352 (92.48)	110 (7.52)	1462	
Dissatisfied	3725 (94.66)	210 (5.34)	3935	
Neutral	4791 (93.85)	314 (6.15)	5105	
Satisfied	3337 (95.34)	163 (4.66)	3500	
Very satisfied	1711 (95.43)	82 (4.57)	1793	
Nutrition status				<.0.001
Normal	7170 (94.98)	379 (5.02)	7549	
Underweight	910 (93.62)	62 (6.38)	972	
Overweight/obese	6542 (94.22)	401 (5.78)	6943	
Severe	294 (88.82)	37 (11.18)	331	

Table 2.3 presents the results of the collapsed self-reported health status, with good and poor health status. The categorical covariates presented are still the same as in Table 2.2. However, it is the health status that has been collapsed into two categories. It is observed that among individuals aged 15-19 years 2.76% of them reported poor health and 97.24% of them reported good health. Those aged 30-34 years, 94.76% of them reported good health and 5.24% of them reported poor health. Moreover, those individuals aged 45-49 years, 15.68% of them reported poor health and 84.32% of them reported good health. We can see an increasing trend of reporting poor health as age increases and a decreasing trend of reporting good health as age increases. Among individuals who are female, 6.25% of them reported poor health. On the other hand, those individuals who

are male, 4.68% of them reported poor health and 95.32% of them reported good health. Among individuals who are African, 94.34% of them reported good health and 5.66% of them reported poor health. Those individuals who are coloured, 95.42% of them reported good health and 4.58% of them reported poor health. Individuals residing in rural areas, about 94.72% of them reported good health and 5.98% of them reported poor health. Those residing in tribal authority areas, 95.03% of them reported good health and 4.97% of them reported poor health. There is a decreasing trend of reporting poor health as the education level increases. Among individuals with high education, 95.64% of them reported good health. Moreover, those individuals with tertiary education, 95.35% of them reported good health. An increasing trend of reporting good health as household income level increase can be observed. We can also observe a decreasing trend of reporting poor health as household income level increases. Those individuals who were either widow/divorced/separated, 12.20% of them reported poor health and 87.80% of them reported good health. About 4.79% of respondents who never drunk alcohol reported poor health while 13.46% of respondents who reported poor health were those who drank everyday. About six percent of the respondents who reported poor health were not economical active (5.59%) and employed (5.78%) respectively. Among individuals who never exercised, 93.52% of them reported good health and 6.48% of them reported poor health. Those who exercised twice a week, 95.70% of them reported good health and 4.30% of them reported poor health. Furthermore, those who exercised more than twice a week, 3.38% of them reported poor health and 96.62% of them reported good health. Those individuals with no toilets, 6.90% of them reported poor health and 93.10% of them reported good health. Among respondents with flushing toilets with onsite disposal, 94.34% of them reported good health while 5.66% of them reported poor health. Conversely, those with other type of toilet, 11.11% of them reported poor health and 88.89% of them reported good health. Individuals who do not smoke a cigarette, 94.83% of them reported good health and 5.17% of them reported poor health. On the other, those who smoke a cigarette, 7.32% of them reported poor health and 92.68% of them reported good health. Among individuals who felt depressed less than a day, 96.03% of them reported good health and 3.97% of them reported poor health. In addition, those individuals who were depressed all the time, 14.25% of them reported poor health. Among individuals who were diagnosed with TB, 20.95% of them reported poor health while 79.05% of them reported good health. We can also observe that those who were not diagnosed with TB had the highest proportion of reporting good health. About 5.61% of respondents with adequate household water source reported poor health while about 6.59% with other source of water reported poor health. It can be observed that among individuals with very dissatisfied life, 7.25% of them reported poor health. On the other hand, those individuals with a very satisfied life, 4.57% of them reported poor health. Among individuals with a normal nutrition status, 94.98% of them had good health and 5.02% of them had poor health. Those individuals who were underweight, 6.38% of them had poor health. Moreover, those individuals with severe malnutrition, 11.8% of them had poor health.

2.5 Chi-Square test of independence

The Pearson's Chi-squared test of independence can be used to examine if there is a significant relationship between a categorical response variable and the categorical explanatory variables. The results from Table 2.2 and Table 2.3 provides the p-values for the Pearson's chi-squared test of association between self-reported health and several covariates. All the covariates with the p-value less than or equal to 5% were considered to be statistically significantly associated with self-reported health. From Table 2.2 and Table 2.3, we can deduce that there is a statistically significant association between respondent's age, gender, education level, household income, marital status, alcohol consumption, exercising level, smoking status, depression, life satisfaction level, nutrition status and self-reported health status ($p < 0.001$). The respondent's race group, type of resident, employment status, type of toilet facilities were also significantly associated with self-reported status.

However, household water source of the respondents did not show statistically significant association with self-reported health. Furthermore, multicollinearity was tested and Table D.1 reveal that none was found between any of the considered variables.

Chapter 3

Bayesian Hierarchical Models for Disease Mapping

3.1 Introduction

Bayesian disease mapping of incidence or prevalence is one of the robust and interesting areas in biostatistics and epidemiology. It covers a series of topics where the spatial or geographical variation of a disease is of significance. The variations can be mapped to assist in the detection of areas where the disease is prevalent particularly. The disease mapping techniques have received much attention in discrete data cases, with the aim of describing the variation in health outcomes across geographic regions. Moreover, these are very robust methods for the initial identification of potential health problems (Kistemann et al., 2002).

This chapter introduces the fundamental idea behind Bayesian hierarchical modeling, as a commonly effective tool for modeling complex spatial correlated data in various epidemiological and public health research. In the Bayesian framework, the estimation is based on the posterior distribution of unknown parameters given the observations, implemented by combining the likelihood function and the prior distribution of the parameters considered as random variables. The prior distribution basically prescribes ones belief about the unknown parameters of interest. While the information content of the data is exclusively carried by the likelihood function (Lawson, 2008). According to (Banerjee et al., 2014), the practical issue in applying Bayesian methods is the computational challenges. However, the use of Markov chain Monte Carlo and other variety computing methods have resolved the issues.

3.2 The Likelihood Model

Suppose that $\mathbf{y} = (y_1, \ldots, y_n)$ are observed random variables with the probability density $p(\mathbf{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a p length vector of unknown parameters. The likelihood function for the observations y_i, $i = 1, \ldots, n$ is defined as

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}). \tag{3.1}$$

Here the underlying assumption is that the observations y_i's are conditionally independent given all the θ parameters. The assumption made in the formulation of the likelihood is that the individual's contribution to the likelihood is independent and allows the derivation of the likelihood to be expressed as Equation (3.1). Hence the observations are assumed to be conditionally independent and this assumption is fundamental to numerous disease mapping applications (Lawson, 2013). Nevertheless, it is possible to have correlated observations, thus different approaches may be considered.

3.3 Prior Distributions

The Bayesian approach provides a cohesive framework for mixing complex data and external knowledge (Sudipto et al., 2004). Within this framework, the models are assigned appropriate mixing probability distribution. This probability is determined by the prior distribution, assigned before the data are observed. When the data is too large, the likelihood of the observations dominates any prior expectation that is assigned. Contrarily, when the data is less informative, the prior expectation will be more dominant.

In the Bayesian paradigm, all parameters are considered random, hence are assigned prior distributions. These prior distributions provide additional information and, they are used to improve the identification of parameters. Consider a stochastic single parameter, θ, the prior distribution is denoted by $p(\theta)$. However, consider a vector of parameters $\boldsymbol{\theta}$, the joint prior distribution is donated by $p(\boldsymbol{\theta}|\boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta}$ is a vector of hyperparameters, and if it is unknown, hyperpriors are assigned to it. Next, we discuss the properties of different prior distributions.

3.3.1 Propriety

The condition where the integration of the prior distribution of a random variable θ over its range (Ω) is infinity is known as the impropriety (Lawson, 2008). This is expressed as

$$\int_{\Omega} p(\theta) d\theta = \infty, \tag{3.2}$$

and if its normalizing constant (see Section 3.4) is infinite, then a prior distribution is considered as proper (Lawson, 2008). Though impropriety is a restriction of any prior distribution, hence it does not necessarily lead to improper posterior distribution (Lawson, 2008). This simply means that the posterior distribution can be much proper even with an improper prior specification. However, according to Morris and Normand (1992) with regards to improper prior, stated that it is important to avoid the informal use of standard improper priors, since they may result in improper posterior distributions.

3.3.2 Conjugate Priors

A conjugate prior is a prior distribution that leads to the same posterior distribution for θ. It is accessible in closed form and is a member of the same distributional family as the prior. The conjugate families are convenient and allow a variety of shapes that are wide enough to capture analyst's prior beliefs (Sudipto et al., 2004). For example, the Poisson likelihood with the mean parameter θ that has a Gamma prior distribution for θ leads to a posterior distribution of θ that is also Gamma. Similarly, for a Binomial likelihood with Beta prior distribution, results in a beta posterior distribution, and for a Normal data likelihood with Normal prior distribution, the posterior distribution is normal (Lawson, 2013). The conjugacy can be identified by probing the kernel of the prior-likelihood product. The kernel which is not normalized should have an identifiable form related to the conjugate distribution (Lawson, 2008). In addition, Gutirrez-Pea (1997) stated that these conjugate families often provide prior distributions which are tractable in at least two other respects. The first is that the normalizing constant of the conjugate density is readily found for many exponential family likelihoods. Secondly, it is that for some important functions of the parameters it is often possible to express the expectations in a convenient form. Moreover, Lawson (2008), stated that conjugacy always assures a proper posterior distribution.

3.3.3 Noninformative priors

In the Bayesian framework, the prior is often specified such that, even for average sample sizes, the information provided by the data dominates the prior because of the nature of the prior knowledge about the parameter of interest. The noninformative prior is the type of prior distributions that are assumed not to make strong preference over the data (Lawson, 2013). These type of prior distributions are generally referred to as vague or flat priors. The often used noninformative prior is the flat prior, which is denoted by the probability density function

$$p(\theta) \propto 1, \text{ with } \theta \in [0, 1].$$

Another widely used noninformative prior is the Jefferys' prior, this prior is based on the Fisher information, hence the probability density function is denoted by

$$p(\theta) \propto \sqrt{|I(\theta)|},$$

where the Fisher information $I(\theta)$ is denoted by

$$I(\theta) = -E_\theta \left[\frac{\partial^2 \ln p(\mathbf{y}|\theta)}{\partial \theta^2} \right],$$

and this prior is locally flat but can be improper. According to Lawson (2013), the prior choice can be usually made based on some general understanding of the range and behavior of the variable. In addition, the variance parameters must have distributions on the positive real line. The gamma, inverse gamma, or uniform families are often the noninformative distributions in this range.

3.4 Posterior Distribution

The conditional probability density of random unknown parameters given the data is known as the posterior distribution. This distribution is proportional to the product of the prior distributions and the likelihood function. The posterior distribution identifies the behavior of the parameters after the data are observed and prior assumption are made about the parameters. The posterior distribution is defined as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{3.3}$$

where $\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is considered as a normalizing constant, and it is equal to the marginal distribution $p(\mathbf{y})$, which does not depend on $\boldsymbol{\theta}$ and, thus can be considered as a constant. Alternatively, the posterior distribution in Equation (3.3) can be specified as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

where the posterior distribution is proportional to the likelihood and the prior distribution.

3.5 Model Criterion

When dealing with the Bayesian hierarchical models, there are varying goodness of fit criteria, depending on the nature of the model and properties of the criteria. Criteria such as the deviance information criterion (DIC), Bayesian information criterion (BIC), Akaike Information Criterion(AIC), Takeuchi information criterion (TIC) (Takeuchi, 1976) and network information criterion (NIC) (Murata et al., 1994) are widely used in the Bayesian application. In this research, we discuss only a few of these criteria, that will be used later in the subsequent chapters.

3.5.1 Akaike Information Criterion (AIC)

A commonly used measure of goodness-of-fit external of the Bayesian framework is the Akaike information criterion (AIC) proposed by Akaike (1974). The AIC avoids overfitting of the data by penalizing high model complexity. For an estimated vector of parameters $\boldsymbol{\theta}$, the AIC is given as

$$\text{AIC} = -2[\ell(\hat{\boldsymbol{\theta}}) + p], \tag{3.4}$$

where $\ell(\hat{\boldsymbol{\theta}}) = \log(f(\mathbf{y}|\boldsymbol{\theta}))$ is the model maximal log-likelihood value, and p is the number of covariates in the model with the intercept included which penalizes excessively complex models. In the choice of model selection, the models with small AIC are favored. With regards to the AIC, Gill (2014), stated that the AIC is preferred in comparison and selecting of non-nested model specifications. However, the AIC has a robust bias towards models that overfit with additional parameters, due to the penalty component, which linearly increases with the number of covariates, and hence the log likelihood increases more.

3.5.2 The Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) is commonly used as a model choice criterion within Bayesian and hierarchical models, also known as the Schwarz criterion proposed by Schwarz (1978). It is closely related to the AIC discussed above, but it is strongly linked to the Bayesian theory. The BIC introduces the penalty term for the number of parameters in the models whenever overfitting occurs. This is when the likelihood is increased by adding the parameters. The BIC is given by

$$\mathrm{BIC} = -2\ell(\hat{\boldsymbol{\theta}}) + p\log(n), \tag{3.5}$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the maximized value of the log-likelihood function, with p and n presenting the number of parameters and sample size respectively. The BIC is more appropriate in model comparison where sample size differs because it explicitly includes n. The penalty term depends on the sample size, the larger the n the larger the penalty and the less n the less penalty. The assumption under Equation (3.5) is that the model errors are Normally independent and identically distributed. According to Gill (2000), regarding the AIC and BIC, it is stated that though the two measures are similar, they indicate different model specifications. The BIC favors models with few covariates and poorer fit and the AIC favors models with more covariates and better fit.

3.5.3 Generalized Cross Validation (GCV)

The generalized cross-validation (GCV) focuses on the model optimality instead of complexity. This entails that the GCV is not a likelihood-based criterion like the AIC and BIC. The GCV can adapt the use of the residual sum of squares based on the last step of the scoring algorithm suitable for non-normal outcomes (Wood, 2006). Additionally, it can possibly use the squared Pearson residuals (Fahrmeir and Tutz, 2001) or *deviance residuals* (Hastie and Tibshirani, 1990). In this section, we only discuss the use of GCV based on the deviance residuals.

In the generalized linear models (GLMs), the goodness of fit measures can be defined in terms of the deviance residuals, denoted by

$$D_i = D(y_i, \mu_i) = 2[\ell_i(y_i) - \ell_i(\mu_i)] \tag{3.6}$$

where $\ell_i(y_i)$ is the log-likelihood of observation i assessed for the observation itself, and $\ell_i(\mu_i)$ is the log-likelihood of observation i assessed for the predicted mean (μ_i) from the actual model. Now take the sum of Equation (3.6), the expression results to

$$D = \sum_{i=1}^{n} D_i = 2\left[\sum_{i=1}^{n} \ell_i(y_i) - \sum_{i=1}^{n} \ell_i(\mu_i)\right], \tag{3.7}$$

which is known as the *deviance* (Fahrmeir and Tutz, 2001), and based on the deviance, the GCV is defined as

$$\text{GCV} = \frac{n}{(n-(p+1))^2} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) \tag{3.8}$$

where $p+1$ is the number of covariates associated with the actual model.

3.5.4 The Deviance Information Criterion (DIC)

The deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) is a commonly used tool for model comparison and assessment. It is basically a generalization of the AIC. The DIC has become a popular model comparison criterion in a fully Bayesian (FB) context. It comprises two components, the measure of fit and complexity of the model. The model fit is measured through the posterior expectation of the deviance for the data $\mathbf{y}$ and parameter vector $\boldsymbol{\theta}$. The deviance of the model is specified as

$$D(\boldsymbol{\theta}) = -2\log[p(\mathbf{y}|\boldsymbol{\theta})], \tag{3.9}$$

and now taking the expectation of Equation (3.9) over $\boldsymbol{\theta}$, the posterior expected deviance is given by

$$\overline{D(\boldsymbol{\theta})} = \mathrm{E}_{\boldsymbol{\theta}|\mathbf{y}}\{D(\boldsymbol{\theta})\}.$$

The model complexity is measured by the effective number of parameters p_D. The effective number of parameters defined by the posterior expected deviance (mean deviance) minus the deviance of the posterior mean. The posterior mean of parameters are given by $\bar{\boldsymbol{\theta}}$, hence the deviance evaluated at the posterior mean of parameters is $D(\bar{\boldsymbol{\theta}})$. Thus, Spiegelhalter et al. (2002) defined the effective dimension of the model as

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}), \tag{3.10}$$

hence, with the information described above, the deviance information criterion (DIC) can be defined as

$$\mathrm{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D \quad \text{or} \quad \mathrm{DIC} = \overline{D(\boldsymbol{\theta})} + p_D \tag{3.11}$$

In the case of weak prior information, the DIC is relatively equivalent to the AIC. Analogously to the AIC and BIC, smaller values of the DIC indicate better model fit supported by the data.

3.6 Spatial Bayesian Hierarchical Models

The mapping of health outcomes has long been part of the application in public health, epidemiology and other studies of disease. This technique uses the Bayesian hierarchical methods which allow overdispersion and spatial correlation. Spatial units tend to exhibit an inherent correlation where units close to each other have similarities than those further apart. These methods offer a mechanism to borrow strength or information across neighboring areas, both locally and globally. To capture heterogeneity among the areas or regions, including spatial random effects in the model is required. There are various structures that can be specified for random effects, but we focus on only the ones employed in this research project. In the next section, we review in details the basic model structure used typically in Bayesian disease mapping, in particular, the Besag, York and Mollie (BYM) (Besag et al., 1991) model.

3.6.1 The Besag, York, and Mollie (BYM) Model

The most commonly used tool under spatial Bayesian hierarchical models for disease mapping of a single disease is the Besag, York, and Mollie (BYM) model proposed by Besag et al. (1991). The BYM has two random effects that can be split into two random components, namely the spatially structured $\boldsymbol{u}$ and the spatially unstructured $\boldsymbol{v}$ components, which are interpreted as surrogates for unknown or unobserved covariates. The spatially structured component $\boldsymbol{u}$ is due to the fact that spatial units are correlated with neighboring spatial units, thus observed would regulate the spatial structure, whereas the spatially unstructured component $\boldsymbol{v}$ represent the uncorrelated extra variation. Furthermore, it is worth noting that the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ contain individual unit random effects u_i $(i = 1, \ldots, n)$ and v_i $(i = 1, \ldots, n)$ respectively.

3.6.1.1 Conditional Autoregressive Model

The conditional autoregressive (CAR) model has been widely used in the field of epidemiology and other studies of diseases and was developed by Besag (1974), and later introduced by Clayton and Kaldor (1987). The CAR models are also known as the Markov random field (MRF) model and are in the class of the Gaussian Markov random field (GMRF) models. In the spatial modeling for administrative districts areal data, such as disease mapping, the MRF models are commonly employed. The MRF is based on the conditional specifications known as the CAR, with both referring to the same model structure. The virtual common form of CAR incorporates the structure of spatial dependence, based on the idea that areas that share a border or boundary are regarded as neighbours. The neighboring areas are bound to have too many similarities than those far apart. Thus, smoothing of the health outcome risk for an areal unit depends on its neighbour's risk. In this research, we focus on the CAR model also known as the MRF model for the incorporation of the spatially structured random term. The CAR models are generally used to model the spatial effects of an areal unit. They are employed in the models as a specification in certain classes of hierarchical spatial models, in particular, the second stage of such models. As proposed in this dissertation, the CAR model can be defined as follows.

Consider $\boldsymbol{u} = (u_1, \ldots, u_n)$ to be the vector of univariate random variables associated with the observed districts or spatial location under study and denote $\{\partial(i)\colon\ i = 1, \ldots, n\}$ as the districts sharing a common boundary with district i. That is, for any $i,\ j = 1, \ldots, n,\ j \in \partial(i)$ only if $i \in \partial(j)$ and $i \notin \partial(i)$ must be satisfied.

Now assuming that the conditional density of u_i, for $i = 1, \ldots, n$, follows the conditional normal variable denoted as

$$u_i | u_j, (i \neq j) \sim N\left(\mu_i + \sum_{j\in\partial(i)} c_{ij}(u_j - \mu_j), d_i^2\right), \quad i, j = 1, \ldots, n, \tag{3.12}$$

where μ_j is the mean for district j, μ_i represent spatial trend at location i and $d_i^2 = \sigma_u^2/\partial(i)$ is the conditional variance of the ith district, which depends on the number of neighbors. Thus, the more the number of district neighbors the smaller the variance for the current district. The c_{ij} denote the spatial dependence parameters for $i = 1, \ldots, n$, such that $c_{ii} = 0$ for all i's, and σ_u^2 is the variance parameter that controls the amount of variation between spatial similarity. In particular, the quantity c_{ij} indicates the spatial dependency. The matrix form of Equation (3.12) is given by (Cressie, 1993);

$$\boldsymbol{u} \sim N(\boldsymbol{\mu}, \boldsymbol{B}^{-1}\boldsymbol{M}), \tag{3.13}$$

which represent the joint distribution of $\boldsymbol{u}$, where $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{C})$, with $\boldsymbol{C} = [c_{ij}]_{n\times n}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and $\boldsymbol{M} = diag(d_1^2, \ldots, d_n^2)$ is an $n \times n$ diagonal matrix. The Equation (3.13) is proper if $\boldsymbol{B}$ is invertible and $\boldsymbol{B}^{-1}\boldsymbol{M}$ is symmetric and the symmetry is guaranteed by the conditional constraints $c_{ij}d_j^2 = c_{ji}d_i^2$ for all $i \neq j$, and must be positive definite. The elements of invertible matrix $\boldsymbol{B}$ are defined as

$$b_{(ij)} = \begin{cases} 1 \text{ for } i = j \\ -c_{ij} \text{ for } j \in \partial(i), \\ 0 \text{ otherwise,} \end{cases} \tag{3.14}$$

since the covariance matrix in Equation (3.13) must not only be symmetric but also be positive definite as stated above for it to be a valid joint distribution. Thus, a common way to construct this is to define a symmetric weighted adjacency matrix $\boldsymbol{W} = (W_{ij})$, and set $c_{ij} = \phi W_{ij}$ where

$$W_{ij} = \begin{cases} 1 \;\; \text{if } i \text{ and } j \text{ share a common boundary} \\ 0 \;\; \text{otherwise,} \end{cases} \tag{3.15}$$

and the parameter ϕ controls the properness of the distribution. Now the covariance matrix $\boldsymbol{\Sigma_u} = (\boldsymbol{I} - \phi\boldsymbol{W})^{-1}\boldsymbol{M}$ must be positive definite such that $d_i^2 > 0$ and $\phi \in (\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}})$, with $\lambda_{(1)} \geq \lambda_{(2)} \ldots \geq \lambda_{(n)}$ are the ordered eigenvalues of the weight matrix $\boldsymbol{W}$. The proper joint density auto-regressive specification $\boldsymbol{u}$ is said to be a multivariate Gaussian distribution defined as

$$\boldsymbol{u} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma_u}). \tag{3.16}$$

Now let $\boldsymbol{W_D}$ denote the diagonal adjacent matrix of standardization or normalization given by

$$\boldsymbol{W_D} = diag(W_{1+}, W_{2+}, \ldots, W_{n+}), \tag{3.17}$$

where

$$W_{i+} = \sum_{j \in \partial(i)} W_{ij}, \; i, j = 1, 2, \ldots, n. \tag{3.18}$$

We then assign the matrix of interaction $\boldsymbol{C}$ as a normalized adjacent matrix given by

$$\boldsymbol{C} = \boldsymbol{W_D}^{-1}\boldsymbol{W} \text{ and } \boldsymbol{W_D} = \sigma^2 \boldsymbol{W}_D^{-1},$$

with $c_{ij} = W_{ij}/W_{i+}$ and $d_i^2 = \sigma^2/W_{i+}$ respectively. Hence, Equation (3.16) can be expressed as

$$\boldsymbol{u} \sim N\left(\boldsymbol{\mu}, \left[\frac{1}{\sigma^2}(\boldsymbol{W_D} - \phi \boldsymbol{W})\right]^{-1}\right), \tag{3.19}$$

and the conditional distribution of $u_i|u_j$ is now given by

$$u_i|u_j \sim N\left(\mu_i + \phi \sum_{j \in \partial(i)} W_{ij}(u_j - \mu_j), d_i^2\right), \tag{3.20}$$

and the correlation between areas i and j depends only on ϕ and $\boldsymbol{W}$, where $W_{ij} = c_{ij}/\partial(i)$. When ϕ is fixed to one, the covariance matrix $\boldsymbol{\Sigma_u}$ is not positive definite, this leads to a conditional distribution expressed as

$$u_i|u_j \sim N\left(\mu_i + \frac{1}{\partial(i)} \sum_{j \in \partial(i)} c_{ij}(u_j - \mu_j), d_i^2\right). \tag{3.21}$$

This CAR specification is known as the *intrinsic* conditional autoregressive (iCAR) and has no proper joint density due to the non-positive resolution. Now if we assume $\mu_i = 0$ for each i the conditional distribution is similar to the form given by Besag et al. (1991)

$$u_i | u_j, i \neq j \sim N\left(\frac{1}{\partial(i)} \sum_{j \in \partial(i)} c_{ij} u_j, \frac{\sigma_u^2}{\partial(i)}\right). \tag{3.22}$$

We employ this specification under structured spatial random effects. The hyperparameter σ_u^2 is assigned an improper inverse exponential hyperprior which was proposed by Besag et al. (1991). Again the hyperparameter σ_u^2 is the variance that controls the degree of smoothness. The $\tau_u = 1/\sigma_u^2$ is the precision parameter, where a large τ_u means a high precision.

3.6.1.2 Convolution Model

The BYM model is the extension of the CAR model, which was proposed by Clayton and Kaldor (1987) and later was developed by Besag et al. (1991). It is also known as the convolution model, consisting of the two random components, the spatially structured and unstructured random components. In general, the spatial random components can be decomposed into two components using a convoluted structure. Besag et al. (1991) formulated this model by assuming that observations y_i are Poisson distributed, denoted as

$$y_i \sim Poisson(e_i \exp(\eta_i)),$$

where $\eta_i = \log(\lambda_i)$ is the log relative risk and is given by

$$\eta_i = u_i + v_i, \text{ thus, } \lambda_i = \exp(u_i + v_i), \tag{3.23}$$

and v_i is assumed to be a Gaussian prior with zero mean, hence

$$v_i \sim N(0, \sigma_v^2), \tag{3.24}$$

where σ_v^2 is also assigned a gamma hyperprior. Again $\tau_v = 1/\sigma_v^2$ is the precision parameter. It is worth noting that either u_i and v_i will dominate the other such that for strong u_i, the estimated mean results to a spatial structure, and vice-versa. These models can be extensions of the classical models to allow for a more flexible approach.

3.7 The Random Walk Models

In models where parametric modeling is not sufficient, a more flexible approach is adopted. Ideally, this approach is used to handle covariates differently, such as allowing for non-linear effects for continuous covariates which the data may contain. These continuous covariates are modeled with semiparametric and generalized additive approach (Fahrmeir and Lang, 2001). Such models are used to describe smooth curves in time or surface in space (Fahrmeir and Lang, 2001). Similar to spatial area effects in Section 3.6.1, metrical covariates are assigned specific priors to allow smoothing. Several alternatives specifications are available for smoothness prior functions of metrical covariates, but we will distinguish few main approaches under Bayesian modeling. The commonly used priors for smooth functions are first or second order random walk models, but we focus on the second-order random walk model.

Consider a case of a continuous covariate x with equidistant design points or observations x_i, $i = 1, \ldots, m$, $(m \leq n)$. Then define an equidistant grid on the x-axis as an ordered sequence of distinct covariate values $x_{(1)} < \ldots < x_{(t)} < \ldots < x_{(m)}$. Let $f = (f(1), \ldots, f(t), \ldots, f(m))'$ denote the vector of function evaluations at these points and define $f(t) := f(x_{(t)})$, then a first-order (RW1) and second-order (RW2) random walk models or functions are defined by

$$f(t) = f(t-1) + u(t) \text{ and } f(t) = 2f(t-1) - f(t-2) + u(t) \tag{3.25}$$

respectively, where $u(t)$ has the Gaussian error with mean zero and known variance τ^2 and diffuse priors for initial values are assigned, such that $f(1) \propto const$, and $f(1)$ and $f(2) \propto const$ respectively. The variance τ^2 controls the smoothness of the function. In addition, the RW2 penalizes large deviations from the $2f(t-1)-f(t-2)$ linear trend. Rue and Held (2005) defined the joint density of the RW2 model with the forward difference approach, assuming an independent second order increment as

$$\Delta^2 x_i = x_i - 2x_{i+1} + x_{i+2},\ \Delta^2 x_i \sim N(0, \tau^{-1}),\ i = 1, \dots, n-2, \tag{3.26}$$

and the joint density is defined as

$$\begin{aligned} p(\boldsymbol{x}) &\propto \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n-2}(\Delta^2 x_i)\right) \\ &= \tau^{(n-2)/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}\right), \end{aligned} \tag{3.27}$$

where $\boldsymbol{Q} = \tau\boldsymbol{R}$ is the precision matrix with rank $n-2$ and $\boldsymbol{R}$ is the structure matrix.

Now consider the case where the covariates x is with non-equally spaced observations, the modification must be done on the random walk priors to account for non-equal distances. In such cases, the RW2 models are modified to weight each point differently such that $\delta_t = x_{(t)} - x_{(t-1)}$ between observations. Let $x_{(1)} < \ldots < x_{(t)} < \ldots < x_{(m)}$ be strictly ordered different observations of x, and define the vector of unknown function evaluations as $f = (f(1), \dots, f(t), \dots, f(m))'$. Then the RW1 and the RW2 models respectively generalize to

$$f(t) = f(t-1) + u(t) \text{ and } f(t) = \left(1 + \frac{\delta_t}{\delta_{t-1}}\right) f(t-1) - \left(\frac{\delta_t}{\delta_{t-1}}\right) f(t-2) + u(t), \tag{3.28}$$

$u(t) \sim N(0, \omega_t \tau^2)$, where ω_t is the appropriate weight. The case of $\omega_t = \delta_t$ denotes the simplest appropriate weight. However, other more complex weight forms may be used.

3.8 Bayesian Penalized Splines

The penalized (P) splines are another commonly used smoothness priors and are employed in this research. The P-splines were introduced by Eilers and Marx (1996) in the frequentist framework and extended by Fahrmeir and Tutz (2001) and Lang and Brezger (2000) in the Bayesian setting.

For an unknown function f of a particular covariate x, it is assumed that it can be approximated by the polynomial spline of degree l defined on a set of equally spaced knots $\zeta_0 = x_{\min} < \ldots < \zeta_{s-1} < \zeta_s = x_{\max}$ within the surface of x. Such Bayesian splines can be written as a linear combination of $d = s + l$ B-spline basis functions B_t, which is denoted as

$$f(x) = \sum_{t=1}^{d} \varsigma_t B_t(x), \tag{3.29}$$

where on a domain spanned by $2 + l$ the basis functions are defined locally in the sense that they are non zero (Kandala et al., 2001). The properties of the B-spline basis function is not discussed in this research as it is beyond the scope of this research. The vector $\boldsymbol{\varsigma} = (\varsigma_1, \varsigma_2, \ldots, \varsigma_d)'$ correspond to the vector of unknown regression coefficients to be estimated from the data under a frequentist point of view. An essential choice is the number of knots in the simple regression setting in order to ensure flexibility in capturing the variability of the data. According to Eilers and Marx (1996), the number of knots d should be moderately large enough, between 20 and 40 to ensure flexibility, while Ruppert (2002) suggested knots from 5 to 20. In the Bayesian framework, penalized splines are equivalent in the estimation model parameters $\boldsymbol{\varsigma}$ by assigning them a first or second random walk prior for equidistant knots, given in Equation (3.25). It is also worth noting that the random walk models form a special case of B-spline with degree zero.

Moreover, a large number of non-linear effects can be estimated simultaneously based on P-splines.

3.9 Empirical Bayes (EB) estimation based on GLMM

In this section, we briefly discuss inference for structured additive regression (STAR) models approach based on the generalized linear mixed model (GLMM) representation. This method involves treatment of all functions and effects within a unified framework by assigning different forms of appropriate priors and degrees of smoothness. In general, we discuss the approach proposed by Fahrmeir et al. (2004) which is an extension from Lin and Zhang (1999), derived to handle large datasets. In the Bayesian STAR models the linear predictor $\eta_i = \mathbf{x}_i'\beta$ is replaced with a structured additive predictor under the assumption that y_i belong to the exponential family, with mean θ_i linked to the function η_i by $h(\theta_i)$. Hence, the structured additive predictor can be expressed as

$$\eta_i = f_1(\nu_{i1}) + f_2(\nu_{i2}) + \cdots + f_{l+1}(\nu_{i(l+1)}) + \cdots + f_p(\nu_{ip}) + \mathbf{x}_i'\beta, \tag{3.30}$$

where i is the generic observation index, the ν_j are different types and dimension of generic covariates, and $f_1, \ldots, f_{l+1}, \ldots, f_p$ are functions of covariates which usually accounts for different types of effects such as spatial effects, temporal effects, non-linear effects of continuous covariates, random effects or interaction effects. This approach can be viewed as a special case of several models, including the semiparametric ordinal models introduced by Tutz (2003). Since this research adopts a Bayesian setting, the unknown functions $f_1, \ldots, f_p$ and covariates effects β are considered as random variables. Thus, they are assigned different appropriate priors as mentioned earlier.

For the empirical Bayes (EB) approach the predictor model in Equation (3.30) needs to be changed into a particular form of the GLMM with appropriate reparametrization. Now express $f_j = (f_j(\nu_{1j}), \ldots, f_j(\nu_{nj}))'$ the vector of function evaluations of unknown f_j as the matrix product defined by

$$f_j = \Psi_j \gamma_j, \; j = 1, \ldots, p \tag{3.31}$$

where Ψ_j is the design matrix and γ_j is a vector of unknown parameters. Hence, Equation (3.30) can be expressed in matrix notation as

$$\eta = \Psi_1 \gamma_1 + \cdots + \Psi_p \gamma_p + X\beta, \tag{3.32}$$

where X is the common design matrix for fixed effects. The vector γ_j is assigned a multivariate Gaussian prior of the form

$$p(\gamma_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} \gamma_j' K_j \gamma_j\right), \; j = 1, \ldots, p$$

where τ_j^2 is considered as a fixed variance which controls the trade off between smoothness and flexibility, and K_j is a precision matrix that penalizes too abrupt jumps among neighboring parameter or shrinks them towards zero (Fahrmeir et al., 2004). Furthermore, in the EB approach, the predictor in Equation (3.32) can be reconstructed as a GLMM that provides the fundamentals of estimating functions f_j and the variance parameters τ_j^2 simultaneously. Assume that the vector parameter γ_j has dimension d_j and the corresponding penalty matrix K_j has rank $r_j < d_j$. Hence, the vector parameters γ_j are partitioned into penalized and unpenalized part as

$$\gamma_j = \Psi_j^{unp} \gamma_j^{unp} + \Psi_j^{pen} \gamma_j^{pen}, \tag{3.33}$$

for some $d_j \times (d_j - r_j)$ well defined matrix Ψ_j^{unp} and a $d_j \times r_j$ matrix Ψ_j^{pen}. The vector of parameters γ_j^{pen} denotes the deviations of parameters γ_j from K_j while the vector γ_j^{unp}

denotes the unpenalized part γ_j by K_j. In Equation (3.33), the priors assumed for the penalized part and for the unpenalized part are as follow;

$$p(\gamma_j^{pen}) \sim N(0, \tau_j^2 I_{hj}) \quad \text{and} \quad p(\gamma_j^{unp}) \propto const \tag{3.34}$$

corresponding to the i.i.d Gaussian prior and a flat prior respectively. Hence, the penalized part γ_j^{pen} are i.i.d random effects and the γ_j^{unp} are considered as a fixed effect. Now decomposing the model in Equation (3.33) to form components of Equation (3.32) yields a variance component model defined as

$$\eta = U\gamma^{unp} + Z\gamma^{pen} \tag{3.35}$$

where $U = (X, \Psi_1^{unp}, \ldots, \Psi_p^{unp})$, $Z = (\Psi_1^{pen}, \ldots, \Psi_p^{pen})$, $\gamma^{unp} = (\beta', \gamma_1^{unp\prime}, \ldots, \gamma_p^{unp\prime})'$ and $\gamma^{unp} = (\gamma_1^{pen\prime}, \ldots, \gamma_p^{pen\prime})'$, with

$$p(\gamma^{pen}) \sim N(0, Q) \quad \text{and} \quad p(\gamma^{unp}) \propto const, \tag{3.36}$$

where $Q = \text{blockdiag}(\tau_1^2 I, \ldots, \tau_p^2 I)$ is the covariance matrix of block diagonal of p-dimension τ_j^2. Thus, this results in a GLMM with a stacked vector of random effects γ^{pen} and stacked vector of fixed effects γ^{unp}. Finally, the function evaluations f_j and the variance parameters can be further estimated simultaneously. Next, we discuss inference for parameters to be estimated given the previous mixed model setting.

3.9.1 Inference

Parameter estimation under Bayesian inference is based on the posterior distribution of the model. The posterior distribution for the EB inference in terms of the GLMM representation is denoted as

$$p(\gamma^{unp}, \gamma^{pen}|y) \propto L(y, \gamma^{unp}, \gamma^{pen}) \prod_{j=1}^{p} \left(p(\gamma_j^{pen}|\tau_j^2)\right) \tag{3.37}$$

where the variances τ_j^2 are unknown constants and $p(\gamma_j^{pen})$ is defined in Equation (3.34). Further taking the log of the posterior results to

$$l_{pen}(\gamma^{unp}, \gamma^{pen}|y) = l(y, \gamma^{unp}, \gamma^{pen}) - \frac{1}{2}\gamma^{pen\prime}Q^{-1}\gamma^{pen} \tag{3.38}$$

the form of a penalized likelihood that needs to be maximized to obtain posterior mode estimates. Estimation of regression coefficients and variance parameter can be done via iteration and approximation. Among several alternatives estimation procedures, here we make use of iteratively weighted least squares (IWLS) and approximate restricted maximum likelihood (REML) constructed under GLMM. The former and the latter estimation is achieved in two steps, the first step is to maximize Equation (3.38) by adopting a Fisher scoring algorithm and rewrite it as an IWLS scheme that is given by

$$\begin{pmatrix} U'WU & U'WZ \\ Z'WU & Z'WZ + Q^{-1} \end{pmatrix} \begin{pmatrix} \gamma^{unp} \\ \gamma^{pen} \end{pmatrix} = \begin{pmatrix} UW\tilde{y} \\ ZW\tilde{y} \end{pmatrix} \tag{3.39}$$

which yields a system of equations to be solved iteratively to obtain estimates (Kneib, 2006). The $W = DS^{-1}D$ is a block diagonal structure weight matrix built by the block-diagonal matrices $D = blockdiag(D_1, \ldots, D_n)$ and $S = blockdiag(S_1, \ldots, S_n)$, while the $\tilde{y}$ is a $n \times 1$ vector of the working observations defined by $\tilde{y} = \hat{\eta} + (D^{-1})(y - \pi)$. The $q \times q$ matrices S_i and D_i are denoted as

$$S_i = \begin{pmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i1} & \cdots & -\pi_{i1}\pi_{iq} \\ -\pi_{i1}\pi_{i2} & \ddots & & \vdots \\ \vdots & & \ddots & -\pi_{i(q-1)}\pi_{iq} \\ -\pi_{i1}\pi_{iq} & \cdots & -\pi_{i(q-1)}\pi_{iq} & \pi_{iq}(1-\pi_{iq}) \end{pmatrix}, \; D_i = \frac{\partial h(\eta_i)}{\partial \eta} \begin{pmatrix} \frac{\partial h_i(\eta_i)}{\partial \eta_1} & \cdots & \frac{\partial h_q(\eta_i)}{\partial \eta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_i(\eta_i)}{\partial \eta_q} & \cdots & \frac{\partial h_q(\eta_i)}{\partial \eta_q}, \end{pmatrix}$$

where the derivatives of h depend on the density of f of the latent variable such that

$$\frac{\partial h(\eta)}{\partial \eta_j} = \begin{cases} f(\eta_j) \quad \text{for} \quad j = r, \\ -f(\eta_j) \quad \text{for} \quad j = r-1, \\ 0 \text{ elsewhere.} \end{cases}$$

Thus the system of equations in Equation (3.39) is solved to obtain updated estimates of $\hat{\gamma}^{unp}$ and $\hat{\gamma}^{pen}$ given the current variance parameters using IWLS. Furthermore, it is worth noting that the credible intervals are also constructed under the estimates of Equation (3.39).

The second step is estimating the variance parameters by maximizing the marginal likelihood defined by

$$L^{\text{marg}}(Q) = \int L(\gamma^{unp}, \gamma^{pen}, Q) d\gamma^{unp} d\gamma^{pen}, \tag{3.40}$$

which is of the REML form. Kneib and Fahrmeir (2006) suggested Laplace approximation to Equation (3.40) since direct integration is generally not possible. Hence, the approximation is applied to $L(\gamma^{unp}, \gamma^{pen}, Q)$ which results in a restricted log-likelihood defined as

$$l^{\text{marg}}(Q) \approx -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\log(|U'\Sigma^{-1}U|) - \frac{1}{2}(\tilde{y} - U\hat{\gamma}^{unp})'\Sigma^{-1}(\tilde{y} - U\hat{\gamma}^{unp}), \tag{3.41}$$

where an approximation to the marginal covariance of $\tilde{y}|\gamma^{pen}$ is given by $\Sigma = W^{-1} + Z'QZ$. The maximization of Equation (3.41) can be done via Fisher's scoring derived by Fahrmeir et al. (2004) that allows computation of REML estimation for large datasets. Now based on a conditional point of view of the GLMM by Lin and Zhang (1999), the elements of the score vector and expected Fishers information can be defined as

$$s_j^*(\tau^2) = -\frac{1}{2}\text{tr}(PZU_jU_j') + \frac{1}{2}||U_j'W(\tilde{y} - U\hat{\gamma}^{unp} - Z\hat{\gamma}^{pen})||^2, \quad j = 1, \ldots, p \tag{3.42}$$

with

$$P = W - W(UZ)H^{-1}(UZ)'W \tag{3.43}$$

and

$$F^*_{jk}(\tau^2) = \frac{1}{2}\mathrm{tr}(PU_jU_j'PU_kU_k'), \quad j, k = 1, \ldots, p \tag{3.44}$$

hence to obtain updated variances in Q, the restricted log-likelihood in Equation (3.41) is maximized by

$$\hat{\tau}^2 = \tilde{\tau}^2 + F^*(\tilde{\tau}^2)^{-1}s^*(\tilde{\tau}^2) \tag{3.45}$$

where $\tilde{\tau}^2$ are the last iteration variance parameters. Therefore, the iterations of the two steps are performed until convergence. For further details on the EB inference algebraic part, the reader can refer to Fahrmeir et al. (2004); Lin and Zhang (1999) among others. The implementation of the EB approach in this research is used for all the analyses of the ordinal response.

3.10 Integrated Nested Laplace Approximation (INLA)

The Markov chain Monte Carlo (MCMC) methods have been widely used in the past and recent years. Part of this is due to Bayesian inference being very popular in spatial and spatio-temporal statistics. However, MCMC uses a sampling technique which can be slow, and reaching the required number of samples can take a long time for complex models or large datasets. In this section, we introduce a more flexible widely used technique for inference within the Bayesian framework. The integrated nested Laplace approximation (INLA) is a recent alternative to MCMC, proposed by Rue et al. (2009). What makes INLA a more beneficial method for Bayesian inference is because it returns similarly accurate results to MCMC methods in significantly less time. However, a few relevant models have been implemented in the INLA R software package.

The main principle behind INLA lies in the approximating the posterior marginals of the wide range of Bayesian hierarchical models. In general, the focus is on the approximation of posterior marginals for latent Gaussian models. These class of models is a subset of all the flexible and extensively used Bayesian additive regression models. Let $\boldsymbol{y}$ denote the observed data points, $\boldsymbol{x}$ be the vector of all the latent Gaussian variables, and $\boldsymbol{\theta}$ denotes the vector of hyperparameters. Assuming conditional independence, the likelihood of the n observations $\boldsymbol{y}$ is denoted by

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|x_i,\boldsymbol{\theta}), \quad i = 1,\dots,n, \tag{3.46}$$

and assuming a multivariate Gaussian prior on $\boldsymbol{x}$ with $\mathbf{0}$ mean and precision matrix $\boldsymbol{Q(\theta)}$, the density function of the latent effects are given by

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}}|\boldsymbol{Q(\theta)}|^{0.5} \exp\left(-\frac{1}{2}\boldsymbol{x'Q(\theta)x}\right) \tag{3.47}$$

where $|\cdot|$ is the matrix determinant (Blangiardo and Cameletti, 2015). The properties of $\boldsymbol{x}$ are that they are conditionally independent such that $\boldsymbol{Q(\theta)}$ is a sparse matrix which allows inference with Gaussian Markov random fields (GMRFs). The joint posterior distribution for the latent Gaussian models due to Rue et al. (2009) focused on estimating equation given by the product of Equations (3.46), (3.47) and of the hyperparameter prior $p(\boldsymbol{\theta})$ as

$$\begin{aligned} p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{\theta})|\boldsymbol{Q(\theta)}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x'Q(\theta)x} + \sum_{i=1}^{n} \log(p(y_i|x_i,\boldsymbol{\theta})\right). \end{aligned} \tag{3.48}$$

In INLA, the main objectives are to approximate each element of the parameter vector of the posterior marginals $p(x_i|\boldsymbol{y})$ and $p(\theta_j|\boldsymbol{y})$. Note that, the posterior marginals of the latent Gaussian components can be written as

$$p(x_i|\boldsymbol{y}) = \int p(x_i|\boldsymbol{\theta}, \boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \tag{3.49}$$

and of the hyperparameter vector for each element of $\boldsymbol{\theta}$ can be written as

$$p(\theta_j|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}, \quad j = 1, \ldots, m, \tag{3.50}$$

where subscript θ_{-j} denotes all the parameter elements $\boldsymbol{\theta}$ except current θ_j. The key approach is to construct and compute nested approximations of

$$\tilde{p}(x_i|\boldsymbol{y}) = \int \tilde{p}(x_i|\boldsymbol{\theta}, \boldsymbol{y})\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \tag{3.51}$$

and

$$\tilde{p}(\theta_j|\boldsymbol{y}) = \int \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j} \tag{3.52}$$

using the form of Equations (3.49) and (3.50). Therefore, the approximations to $p(x_i|\boldsymbol{y})$ are based on a computational approximation of $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $p(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ with the aid of numerical integration methods to integrate out the hyperparameter $\boldsymbol{\theta}$. Based on the Laplace approximation, INLA approach exploits the assumptions of the model to produce the posteriors of interest using numerical approximation (Tierney and Kadane, 1986).

The approximation is divided into threefold. Firstly, is the computation of an approximation to the joint posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ of the hyperparameters as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} \approx \left.\frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} =: \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}), \tag{3.53}$$

where $\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation to the full conditional of $\boldsymbol{x}$ based on the Gaussian distribution, and $\boldsymbol{x}^*$ denotes the mode of the full conditional of $\boldsymbol{x}$, for a given $\boldsymbol{\theta}$.

Secondly, a good approximation to the conditional distribution $p(x_i, \boldsymbol{\theta}|\boldsymbol{y})$ is required. The use of the Gaussian approximation can be adapted, but this approach led to many issues of not resulting in a very good approximation. Hence, Rue et al. (2009) developed a better approximation by rewriting $\boldsymbol{\theta} = (\theta_j, \boldsymbol{\theta_{-j}})$ and using the Laplace approximation again to obtain

$$\tilde{p}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{p}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^*_{-i}(x_i,\boldsymbol{\theta})}, \tag{3.54}$$

where $\tilde{p}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation to the conditional distribution of $\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y}$ and the entire expression is centered around the mode $\boldsymbol{x}^*_{-i}(x_i, \boldsymbol{\theta})$, for the given $\boldsymbol{\theta}$.

Lastly, once $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$ and $\tilde{p}_{LA}(x_i|\boldsymbol{\theta, y})$ are obtained, the marginal posterior distributions for Equation (3.51) can be computed via numerical integration

$$\tilde{p}(x_i|\boldsymbol{y}) \approx \sum_{k=1}^{n} \tilde{p}(x_i|\boldsymbol{\theta}_k\boldsymbol{y})\tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{y})\Delta_k, \tag{3.55}$$

here the $\boldsymbol{\theta}_k$ is a set of grid points corresponding to the set of weights Δ_k. In addition, the computation of $\tilde{p}(\theta_j|\boldsymbol{y})$ in Equation (3.52) can be achieved by integrating out $\boldsymbol{\theta}_{-j}$ from the approximation $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$.

The INLA approach has been implemented in several disease mapping analyses within a fully Bayesian inference, as it makes it possible to compare and assess models (Gomez-Rubio et al., 2014). However, it has limitations when it comes to implementation of ordinal response outcomes due to the availability of the likelihood specifications. The implementation of the INLA approach in this research is adopted for a dichotomous self-reported health response.

This chapter introduced the fundamental idea behind Bayesian hierarchical modeling, as a commonly effective tool for modeling complex spatial correlated data. The Bayesian approach provides a robust alternative approach to a frequentist approach under disease mapping. Different models for prior assumptions of generic covariates were discussed. These models are a class of Gaussian Markov random fields. Model selection criterion was also discussed, namely the DIC, AIC, BIC and the GCV. Furthermore, the competing estimation procedure was also given.

Chapter 4

Spatial Modeling of self-reported health Ordinal response outcome

4.1 Introduction

This research project comprises two distinct types of response outcomes, namely the ordinal and the collapsed (binary). In this chapter, we review models for the ordinal response, and for the dichotomous response will be reviewed in the next chapter. In particular, the cumulative logit models will be extended by accounting for the spatial Bayesian hierarchy structure. Furthermore, a flexible approach will be used to allow for continuous covariates. All the models will be fitted with applications to the wave 4 NIDS dataset. Model comparison and selection will be done then the results will be interpreted and displayed in form of maps and graphs.

4.2 Multivariate Generalized Linear Models

The multivariate generalized linear models (MGLMs) are a class of models extended from univariate generalized linear models (GLMs), which are well-known distribution dependent models belonging to the exponential family. The basic assumption is that the observations y_i, $i = 1, \ldots, n$ are independent and have a distribution that belongs to the exponential family, and their probability density can be written as

$$f(y_i|\mu_i, \phi, \omega_i) = \exp\left(\frac{y_i'\mu_i - b(\mu_i)}{\phi}\omega_i + c(y_i, \phi, \omega_i)\right),\ i = 1, \dots, n, \tag{4.1}$$

where ω_i is a weight, ϕ is a scale parameter common to all the y_i's, μ_i is the natural parameter of the exponential family, and $b(\mu_i)$ is the normalizing function (Fahrmeir and Tutz, 2001). The mean $\theta_i = E(y_i|\mathbf{x}_i)$ for a given covariate $\mathbf{x}_i$, can be determined by a linear predictor

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta},\ i = 1, \dots, n$$

where $\mathbf{x}_i'$ is the corresponding ith row of the design matrix $X_i^{(n\times p)}$, hence the mean θ_i and η_i are linked via a link function $g(\cdot)$ as

$$g(\theta_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta},\ i = 1, \dots, n$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression coefficients, then it follows that $\theta_i = \eta_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$. Models used in this chapter are special cases of MGLMs such that y_i follows a multinomial distribution given by

$$y_i \sim Multinomial(m_i, \pi_i) \text{ and } \pi_i = (\pi_{i1}, \dots, \pi_{ik}).$$

Next, we discuss a special case of MGLMs, which are used to examine the impact of different type of covariates on self-reported health.

4.3 Spatial Cumulative Logit Models

Multicategorical variables can be either nominal or ordinal. The models to accommodate ordered response data have been widely used in both non-spatial and spatial settings. In health and social surveys, econometric and psychometric applications, data with ordinal outcomes occur frequently (Congdon, 2005). Particularly a wide attention has been brought to the use of these models under disease mapping. There are several models for fitting ordinal data, such as the cumulative threshold logit, sequential, continuation ratio, partial proportional odds, unconstrained and constrained partial proportional odds and stereotype models. The most widely used model to analyze data with the ordered categorical response is the cumulative threshold model (Fahrmeir and Tutz, 2001). Next, we focus on the cumulative logit model with its extensions, where the ordering is taken into consideration.

4.3.1 Cumulative Logit Models

Let y_{ij} be an ordinal response variable for individuals $j = 1, \ldots, n$ in districts $i = 1, \ldots, 52$ taking values in the range $1, \ldots, k$ with $(k > 2)$. In this case, y_{ij} is the self-reported health with five levels (poor, fair, good, very good and excellent). This model assumes that the observable self-reported health response y_{ij} of individual j in district i is a categorized version of the latent continuous variable U_{ij} determined by the cutpoints. The assumed linear form of the latent variable is given by

$$U_{ij} = \eta_{ij} + \epsilon_{ij} \tag{4.2}$$

where $\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}$ is the linear predictor with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ vector of coefficients and ϵ_{ij} are the random error terms with some continuous cumulative distribution function F. Now for a giving vector $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ of predictor variables, U_{ij} and the observable y_{ij} are linked by

$$y_{ij} = r \quad \text{if and only if} \quad \theta_{r-1} < U_{ij} \leq \theta_r, \qquad r = 1, 2, \ldots, k$$

where r are categories of y_{ij} and θ_r are unknown cutpoints satisfying $-\infty = \theta_0 < \theta_1 < \cdots < \theta_k = \infty$. From the assumption of the linear form for the latent variable U_{ij} it follows that the observed y_{ij} is given by the model

$$\begin{aligned} P(y_{ij} \leq r|\eta_{ij}) &= F(\theta_r - \eta_{ij}) \\ F^{-1}\left[P(y_{ij} \leq r|\eta_{ij})\right] &= \theta_r - \eta_{ij}, \quad r = 1, \ldots, k-1 \end{aligned} \tag{4.3}$$

where F^{-1} is a link function and $P(y_{ij} \leq r)$ are cumulative probabilities. Specific choices of distribution of the errors ϵ_{ij} in Equation (4.2) lead to specific cumulative models based on the link function, with the common choice being the logistic distribution. In this research, we assumed that errors have a logistic distribution. Consequently, the link function is the logit link and hence the model is the cumulative logit model also known as the proportional odds model (McCullagh, 1980). The proportionality of the odds was derived from the basic assumption that the regression coefficients $\boldsymbol{\beta}$ are the same across the categories and this assumption needs to be verified. Furthermore, the linear predictor η_{ij} does not contain an intercept for identifiability, otherwise, one of the cutpoints must be set to zero.

In general, for self-reported health of individual j in district i, we assume that

$$y_{ij}|\eta_{ij} \sim \text{Categorical}(\pi_{ij}), \quad i = 1, \ldots, 52, \ j = 1, \ldots, n$$

with a latent continuous variable of $k-1$ cutpoints and $\pi_{ij} = (\pi_{ij1}, \pi_{ij2}, \ldots, \pi_{ijk})$ is the vector of probabilities of the model. In this research, we classified the self-reported health categories as

$$y_{ij} = \begin{cases} 1: & \text{excellent} \\ 2: & \text{very good} \\ 3: & \text{good} \\ 4: & \text{fair} \\ 5: & \text{poor (used as reference category),} \end{cases}$$

where y_{ij} is a five-category outcome. In the model formulation in Equation (4.3), the linear predictor η_{ij} only accounts for the parametric form. However, it can be extended to account for a non-parametric part of the model. Since the data are associated with the location, it is worthy to account for spatial heterogeneity and correlation. These may be obtained by introducing random effects in the models. The models that incorporate spatial random effects are known as the geoadditive models (Kammann and Wand, 2003). We now consider such models in the order of complexity by extending η_{ij} in Equation (4.3). In general, we consider models with spatial random effects by adopting the models discussed in Section 3.6. Considering the cumulative logit model, the predictor η_{ij} can be extended by adopting a geoadditive predictor form and introducing random effects, then it follows that

$$\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_{str}(s_i) \tag{4.4}$$

where $u_i = f_{str}(s_i)$ with spatial index $s_i \in \{1, \ldots, 52\}$ are structured spatial random effects of districts modeled as Markov random fields (MRFs) (Equation (3.22)). Another alternative model introducing area-specific random effects can be specified as

$$\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_{unstr}(s_i) \tag{4.5}$$

where $v_i = f_{unstr}(s_i)$ are unstructured spatial random effects of districts modeled as an i.i.d Gaussian distribution (Equation (3.24)). The last model accounts for both unstructured and structured random effects and is known as the convolution model specified in Section (3.23). The model is then extended as following

$$\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_{spat}(s_i) \tag{4.6}$$

where $f_{spat}(s_i) = f_{str}(s_i)+f_{unstr}(s_i)$ following Besag et al. (1991). The two spatial random effects are assumed to be independent and they are assigned independent priors. The $f_{str}(s_i)$ are assumed to follow the Markov random field (MRF) structure and the $f_{unstr}(s_i)$

are assigned the i.i.d Gaussian distribution.

4.3.2 Cumulative Structured Additive Regression Models

The previous section discussed models where the linear predictor caters for linear and area-specific effects respectively. However, it is of interest to investigate how other covariates, such as continuous covariates influence the dependent variable, which could not be displayed if modeled linearly. This section discusses models for a more flexible approach, known as semiparametric models. The semiparametric models were proposed by Kammann and Wand (2003) and Fahrmeir and Lang (2001) under the empirical and fully Bayesian framework respectively. These models are commonly known as structured additive regression (STAR) models. The linear predictor in Equation (4.3) is extended by replacing it with a structured additive predictor to allow for additional nonlinear covariates, this yields a semiparametric predictor

$$\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \sum_{l=1}^{q} f_l(\boldsymbol{z}_{ijl}) + f_{spat}(s_i), \quad i = 1, \ldots, 52,\ j = 1, \ldots, n \tag{4.7}$$

where $\mathbf{x}_{ij}$ is a vector of categorical covariates accounting for linear effects, f_l are unknown nonlinear functions of continuous covariates $\boldsymbol{z}_{ijl}$ and f_{spat} are spatial functions that captures area specific effects and can be splits into structured and unstructured spatial components as mentioned previously. In general, the predictor in Equation (4.3) is expanded to include all possible covariates such as fixed, nonlinear and spatial variables.

4.3.3 Parameter Estimation

The initial step in the estimation of unknown parameters under empirical Bayes (EB) approach via generalized linear mixed model (GLMM) is to formulate the overall vector of parameters for all components of the model. Let $\boldsymbol{\gamma} = (\theta_1, \ldots, \theta_k, \boldsymbol{\beta}')'$ define the overall vector of fixed regression coefficients, and let

$$\mathbf{U} = \begin{pmatrix} u_1' \\ \vdots \\ u_p' \end{pmatrix} = \begin{pmatrix} 1 & & & -x' \\ & \ddots & & \vdots \\ & & 1 & -x' \end{pmatrix}$$

define the corresponding design matrix constructed from covariates $\mathbf{x}_{ij}'$. Hence, we can rewrite the predictor in Equation (4.7) in generic matrix notation as

$$\boldsymbol{\eta}_j = \mathbf{U}\boldsymbol{\gamma} + \mathbf{X}_1\boldsymbol{\beta}_{j1} + \mathbf{X}_2\boldsymbol{\beta}_{j2} + \cdots + \mathbf{X}_l\boldsymbol{\beta}_{jl} + \mathbf{X}_{unstr}\boldsymbol{\beta}_{j,unstr} + \mathbf{X}_{str}\boldsymbol{\beta}_{j,str} \tag{4.8}$$

where $(\mathbf{U}, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_l, \mathbf{X}_{unstr}, \mathbf{X}_{str})$ are appropriate design matrices for each fixed, continuous and spatial effect respectively, and $(\boldsymbol{\gamma}, \boldsymbol{\beta}_{j1}, \boldsymbol{\beta}_{j2}, \ldots, \boldsymbol{\beta}_{j,unstr}, \boldsymbol{\beta}_{j,str})$ is a high dimensional parameter vector, such that $\boldsymbol{f}_j = \mathbf{X}_j\boldsymbol{\beta}_{jl}$. In addition, the predictor from the previous equation can be defined in a more compact form as

$$\boldsymbol{\eta} = \mathbf{U}\boldsymbol{\gamma} + \mathbf{X}_{unstr}\boldsymbol{\beta}_{unstr} + \mathbf{X}_{str}\boldsymbol{\beta}_{str} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \cdots + \mathbf{X}_l\boldsymbol{\beta}_l. \tag{4.9}$$

Therefore this yields a form of a variance component in Equation (3.35). Parameter estimation in all the models of this chapter was done via an empirical Bayesian approach based on mixed models methodology. All the analyses were performed using BayesX (Belitz et al., 2009) with combination to a Bayesian inference R package known as *R2BayesX* (R Core Team, 2017). All the R codes are presented in Appendix A. In the next consecutive sections we consider two applications. The first application is on the spatial models and the second application is on the spatial models with an extension of nonlinear effects.

4.4 Application of the spatial models to NIDS wave 4 data

This section considers the application of various models. These models are used to model the association between self-reported health and covariates, and further examines geographic variation. The first model (Model 1) is the standard cumulative logit model represented by Equation (4.3). Model 2 is represented by Equation (4.4) which is a parametric cumulative logit regression model with spatially structured random effects that account for unobserved covariates which vary spatially across the districts and it was modeled with a Markov random field (MRF) model. Model 3 denoted by Equation (4.6) is similar to Model 2 but caters for unstructured random effects which accounts for unobserved influential covariates that are inherent among the districts and was assigned an i.i.d Gaussian distribution. Model 4 assumes a linear effect of categorical covariates, and it incorporates both spatially structured and unstructured random effects. Each of these models was fitted to the NIDS wave 4 datasets. Model selection was done based on the Akaike information criterion (AIC), Bayesian information criterion (BIC), and by means of generalized cross-validation (GCV). The better fitting model is considered to be that with the smallest AIC, BIC, and GCV (Kneib et al., 2008). However, the BIC favors simple models because it gives a large penalty to overfitting than the AIC (Ngwira and Kazembe, 2016). The GCV is known to select the best fitting model, without giving any penalty to overfitting.

Table 4.1 provides the analyses results together with values of the AIC, BIC, and GCV of the fitted models. The results suggest that the AIC and the GCV values favor Model 4, while the BIC values favor Model 1. Thus, interpretation of results will be based on Model 4 which is suggested by the AIC and GCV based on majority vote. The cumulative posterior odds ratios (PORs) estimates and their corresponding 95% credible intervals (CIs) for the covariates of the fitted models are also provided in the table.

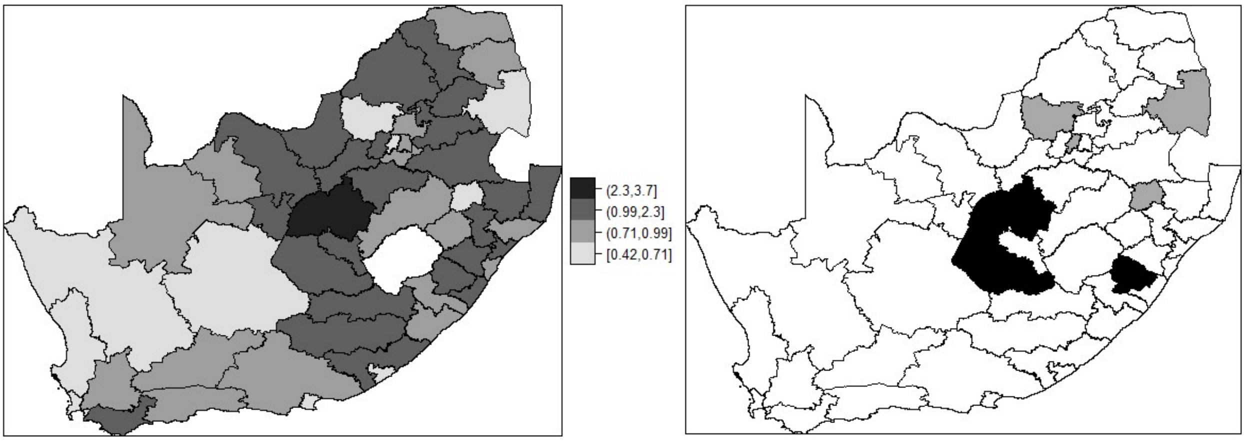
Table 4.1: Parameter estimates of the multivariable Bayesian spatial cumulative logit models.

Covariates	Model 1 POR (95% CI)	Model 2 POR (95% CI)	Model 3 POR (95% CI)	Model 4 POR (95% CI)
Age group (ref = 15-19)				
20-24	1.11 (1.00, 1.22)	1.12 (1.02, 1.24)	1.12 (1.02, 1.24)	1.12 (1.02, 1.24)
25-29	1.28 (1.14, 1.43)	1.30 (1.16, 1.45)	1.30 (1.16, 1.45)	1.30 (1.16, 1.45)
30-34	1.34 (1.18, 1.51)	1.37 (1.21, 1.55)	1.37 (1.21, 1.55)	1.37 (1.21, 1.55)
35-39	1.70 (1.49, 1.95)	1.70 (1.49, 1.95)	1.71 (1.49, 1.95)	1.71 (1.49, 1.95)
40-44	1.94 (1.68, 2.24)	1.98 (1.72, 2.29)	1.99 (1.72, 2.29)	1.99 (1.72, 2.29)
45-49	2.71 (2.33, 3.15)	2.75 (2.37, 3.20)	2.75 (2.37, 3.20)	2.75 (2.37, 3.20)
Gender (ref = Female)				
Male	0.79 (0.74, 0.85)	0.79 (0.73, 0.85)	0.79 (0.73, 0.85)	0.79 (0.73, 0.85)
Race (ref = African)				
Asian/Indian	0.87 (0.63, 1.22)	0.86 (0.61, 1.21)	0.87 (0.62, 1.22)	0.86 (0.62, 1.21)
Coloured	0.83 (0.75, 0.92)	1.04 (0.90, 1.20)	1.01 (0.88, 1.15)	1.02 (0.89, 1.17)
White	0.96 (0.73, 1.26)	1.08 (0.82, 1.42)	1.06 (0.81, 1.40)	1.07 (0.81, 1.41)
Type of residence (ref = Urban informal)				
Rural Formal	1.03 (0.89, 1.18)	0.99 (0.85, 1.15)	0.99 (0.85, 1.15)	0.99 (0.85, 1.15)
Urban Formal	1.14 (1.01, 1.28)	1.10 (0.97, 1.25)	1.11 (0.98, 1.25)	1.11 (0.98, 1.25)
Tribal Authority Areas	1.03 (0.91, 1.17)	0.96 (0.84, 1.11)	0.97 (0.85, 1.12)	0.97 (0.84, 1.12)
Education level (ref = No formal education)				
Primary	1.14 (0.85, 1.51)	1.15 (0.86, 1.53)	1.14 (0.86, 1.52)	1.14 (0.86, 1.52)
Secondary	0.74 (0.58, 0.94)	0.74 (0.58, 0.95)	0.74 (0.58, 0.94)	0.74 (0.58, 0.94)
High	0.53 (0.42, 0.66)	0.53 (0.42, 0.67)	0.53 (0.42, 0.66)	0.53 (0.42, 0.66)
College	0.50 (0.38, 0.66)	0.52 (0.40, 0.68)	0.51 (0.39, 0.67)	0.52 (0.39, 0.68)
Tertiary	0.44 (0.34, 0.57)	0.44 (0.34, 0.56)	0.44 (0.34, 0.56)	0.44 (0.34, 0.56)
Household income (ref = Much below average)				
Below average	1.18 (1.08, 1.30)	1.16 (1.05, 1.27)	1.16 (1.05, 1.27)	1.16 (1.05, 1.27)
Average	1.31 (1.20, 1.43)	1.29 (1.18, 1.41)	1.29 (1.18, 1.41)	1.29 (1.18, 1.41)
Above average	1.12 (0.98, 1.27)	1.14 (1.00, 1.31)	1.14 (1.00, 1.31)	1.14 (1.01, 1.31)
Much above average	1.05 (0.84, 1.18)	1.07 (0.90, 1.27)	1.07 (0.90, 1.27)	1.07 (0.90, 1.27)
Marital status (ref = Not married)				
Widow/Divorced/Seperated	1.19 (1.00, 1.42)	1.17 (0.98, 1.40)	1.17 (0.98, 1.40)	1.17 (0.98, 1.40)
Married/living with partner	1.01 (0.93, 1.09)	1.01 (0.92, 1.08)	0.99 (0.92, 1.08)	0.97 (0.92, 1.08)
Life satisfaction level (ref = Very dissatisfied)				
Dissatisfied	0.98 (0.87, 1.09)	0.94 (0.84, 1.06)	0.95 (0.84, 1.06)	0.95 (0.84, 1.06)
Normal	1.03 (0.92, 1.16)	1.02 (0.91, 1.15)	1.03 (0.91, 1.15)	1.03 (0.91, 1.15)
Satisfied	0.95 (0.85, 1.08)	0.97 (0.86, 1.10)	0.98 (0.86, 1.11)	0.98 (0.86, 1.11)
Very satisfied	0.92 (0.81, 1.06)	0.93 (0.81, 1.07)	0.93 (0.81, 1.07)	0.93 (0.81, 1.07)
Exercise (ref = Never)				
Less than once a week	0.74 (0.66, 0.83)	0.78 (0.70, 0.88)	0.78 (0.70, 0.88)	0.78 (0.70, 0.88)
Once a week	0.93 (0.81, 1.06)	0.96 (0.84, 1.10)	0.96 (0.84, 1.10)	0.96 (0.84, 1.10)
Twice a week	1.01 (0.89, 1.14)	1.03 (0.91, 1.16)	1.03 (0.91, 1.16)	1.03 (0.91, 1.16)
Three or more times a week	0.86 (0.79, 0.94)	0.88 (0.80, 0.96)	0.88 (0.80, 0.96)	0.88 (0.80, 0.96)
Alcohol consumption level (ref = Never drunk alcohol)				

Table 4.1 *Continues*

	Model 1	Model 2	Model 3	Model 4
Covariates	**POR (95% CI)**	**POR (95% CI)**	**POR (95% CI)**	**POR (95% CI)**
No longer drink	1.19 (1.07, 1.32)	1.18 (1.07, 1.31)	1.18 (1.07, 1.31)	1.18 (1.07, 1.31)
Drink very rarely	1.10 (1.01, 1.20)	1.07 (0.98, 1.17)	1.07 (0.98, 1.17)	1.07 (0.98, 1.17)
Less than once a week	1.13 (0.96, 1.34)	1.14 (0.96, 1.34)	1.14 (0.96, 1.35)	1.14 (0.96, 1.35)
1 or 2 days a week	1.04 (0.91, 1.18)	1.05 (0.92, 1.19)	1.04 (0.92, 1.19)	1.04 (0.92, 1.19)
3 or 4 days a week	1.26 (0.98, 1.61)	1.30 (1.02, 1.66)	1.30 (1.02, 1.66)	1.30 (1.02, 1.66)
5 or 6 days a week	1.09 (0.67, 1.75)	1.08 (0.67, 1.74)	1.08 (0.67, 1.74)	1.08 (0.67, 1.74)
Every day	1.04 (0.62, 1.74)	1.05 (0.59, 1.67)	1.05 (0.59, 1.67)	1.06 (0.59, 1.67)
Smokes (ref = No)				
Yes	1.10 (1.00, 1.21)	1.09 (1.01, 1.19)	1.08 (0.99, 1.19)	1.09 (1.00, 1.19)
Type of toilet (ref = None)				
Flush toilet with offsite disposal	0.88 (0.74, 1.06)	0.88 (0.73, 1.06)	0.87 (0.73, 1.05)	0.88 (0.73, 1.05)
Flush toilet with onsite disposal	0.93 (0.77, 1.11)	0.98 (0.81, 1.18)	0.97 (0.81, 1.17)	0.97 (0.81, 1.17)
Bucket toilet	0.80 (0.64, 1.02)	0.79 (0.62, 1.00)	0.78 (0.62, 0.99)	0.78 (0.62, 0.99)
Chemical toilet	1.29 (1.00, 1.67)	1.33 (1.03, 1.72)	1.33 (1.02, 1.72)	1.33 (1.03, 1.72)
Pit latrine with ventilation pipe	0.78 (0.65, 0.94)	0.82 (0.68, 0.98)	0.82 (0.68, 0.98)	0.82 (0.68, 0.98)
Pit latrine without ventilation pipe	0.87 (0.73, 1.04)	0.91 (0.76, 1.08)	0.90 (0.76, 1.08)	0.90 (0.76, 1.08)
Other	1.55 (0.75, 3.19)	1.70 (0.82, 3.52)	1.71 (0.83, 3.52)	1.71 (0.83, 3.53)
Employment status (ref = Unemployed strict)				
Unemployed Discouraged	0.81 (0.62, 1.05)	0.84 (0.65, 1.09)	0.84 (0.65, 1.09)	0.84 (0.65, 1.09)
Not Economically Active	1.23 (1.12, 1.35)	1.16 (1.06, 1.28)	1.17 (1.06, 1.29)	1.17 (1.06, 1.29)
Employed	0.99 (0.90, 1.09)	0.98 (0.90, 1.08)	0.98 (0.90, 1.08)	0.98 (0.90, 1.08)
Nutrition status (ref = Normal)				
Underweight	1.24 (1.09, 1.40)	1.23 (1.08, 1.40)	1.23 (1.08, 1.40)	1.23 (1.08, 1.40)
Overweight/obese	0.94 (0.88, 1.01)	0.93 (0.87, 1.00)	0.93 (0.87, 1.00)	0.93 (0.87, 1.01)
Severe	1.32 (1.07, 1.63)	1.29 (1.05, 1.59)	1.29 (1.05, 1.59)	1.29 (1.05, 1.59)
Was diagnosed with TB? (ref = No)				
Yes	2.08 (1.78, 2.45)	2.10 (1.79, 2.47)	2.10 (1.79, 2.46)	2.10 (1.79, 2.46)
Felt depressed in past week? (ref = Less than 1 day)				
Little of the time (1-2 days)	1.32 (1.23, 1.41)	1.31 (1.22, 1.40)	1.31 (1.22, 1.40)	1.31 (1.22, 1.40)
Moderate amount of time (3-4 days)	1.31 (1.19, 1.46)	1.34 (1.21, 1.49)	1.34 (1.21, 1.49)	1.34 (1.21, 1.49)
All the time (5-7 days)	2.05 (1.68, 2.48)	2.05 (1.69, 2.49)	2.04 (1.68, 2.48)	2.05 (1.68, 2.49)
	Est. (95 % CI)	**Est. (95 % CI)**	**Est. (95 % CI)**	**Est. (95 % CI)**
θ_1	-0.65 (-0.99, -0.32)	-0.71 (-1.05, -0.37)	-0.72 (-1.07, -0.36)	-0.71 (-1.07, -0.36)
θ_2	0.77 (0.44, 1.10)	0.76 (0.42, 1.10)	0.75 (0.40, 1.11)	0.76 (0.40, 1.11)
θ_3	2.90 (2.57, 3.24)	2.94 (2.60, 3.28)	2.93 (2.57, 3.29)	2.94 (2.58, 3.29)
θ_4	4.47 (4.11, 4.83)	4.51 (4.15, 4.87)	4.5 (4.12, 4.88)	4.51 (4.13, 4.88)
Additional model parameters				
Spatially structured variation (σ^2_{str})	-	0.6124	-	0.0252
Spatially unstructured variation (σ^2_{unstr})	-	-	0.1525	0.1404
Model fit				
AIC	39167.9	38597.2	38596.5	38596.2
BIC	39635.6	39418.9	39420.1	39420.1
GCV	2.453	2.397	2.397	2.397

The association was considered significant at 5% level of significance. Next, we only discuss the results for the significant covariates. Based on Model 4 in Table 4.1, the results show that all the considered covariates were significantly associated with self-reported health except for race, type of residence, marital status, and life satisfaction level. The results in Table E.1 revealed that the proportional odds assumption was found to be insignificant at 5% level of significance (p-value = 0.1151), which means that the proportional odds assumption was not violated. Thus, the results are based on the proportional odds assumption. The odds of reporting poor health for individuals between the age groups 20-24, 25-29, 30-34, 35-39, 40-44 and 45-49 years were respectively 1.12 (with 95% CI: 1.02 to 1.24), 1.30 (with 95% CI: 1.16 to 1.45), 1.37 (with 95% CI: 1.21 to 1.55), 1.71 (with 95% CI: 1.49 to 1.95), 1.99 (with 95% CI: 1.72 to 2.29) and 2.75 (with 95% CI: 2.37 to 3.20) times the odds of reporting poor health for individuals between 15-19 years of age. This demonstrates a linear odds increase trend of reporting poor health as age increases. The odds of reporting poor health among individuals who are male was 0.79 (with 95% CI: 0.73 to 0.85) times the odds of reporting poor health for individuals who are female. This means the prevalence of poor health is significantly high in females than males. The odds of reporting poor health among individuals with secondary education were 0.74 times the odds of reporting poor health for individuals with no formal education (POR: 0.74, 95% CI: 0.58 to 0.94). The odds of reporting poor health for individuals with high education were 0.53 times the odds of reporting poor health for individuals with no formal education (POR: 0.53, 95% CI: 0.42 to 0.66). The odds of reporting poor health among individuals with college and tertiary education were respectively 0.52 (with 95% CI: 0.39 to 0.68) and 0.44 (with 95% CI: 0.34 to 0.56) times the odds of reporting poor health for individuals with no formal education. This means the prevalence of poor health is less the higher the education level. The odds of reporting poor health for individuals with below average household income were 1.16 times the odds of reporting poor health for individuals with much below average household income, (POR: 1.16, 95% CI: 1.05 to 1.27). The odds of reporting poor health for individuals with average household income were 1.29 times the odds of reporting poor health for individuals with much below average household income, (POR: 1.29: 95% CI: 1.18 to 1.41). The odds of reporting poor health among individuals with above average household income were 1.14 times the odds of reporting poor health for individuals with much below average household income, (POR: 1.14, 95% CI: 1.01 to 1.31). Moreover, the odds of reporting poor health among individuals with much above average were 1.07 times the odds of reporting poor health for individuals with much below average household income, but this was insignificant (POR: 1.07, 95% CI: 0.90 to 1.27). The odds of reporting poor health for individuals who exercise less than once a week were 0.78 (with 95% CI: 0.70 to 0.88) times the odds of reporting poor health for individuals who never exercise. The odds of reporting poor health among individuals who exercise three or more times a week were 0.88 (with 95% CI: 0.80 to 0.96) times the odds of reporting poor health for individuals who never exercise. However, exercising once a week and twice a week were found not to be insignificantly associated with self-reported health. The corresponding odds ratios were (POR: 0.96, 95% CI: 0.84 to 1.10) and (POR: 1.03, 95% CI: 0.91 to 1.16) respectively. The odds of reporting poor health for individuals who no longer drink alcohol were 1.18 times the odds of reporting poor health for individuals who never drunk alcohol, (POR: 1.18, 95% CI: 1.07 to 1.31). The odds of reporting poor health among individuals who drink on 3 or 4 days a week were 1.30 times the odds of reporting poor health for individuals who have never drunk alcohol, (POR: 1.30, 95% CI: 1.02 to 1.66). However, individuals who drink alcohol very rarely, less than once a week, 1 or 2 days a week, 5 or 6 days a week and every day were found to be insignificantly associated with self-reported health. The corresponding odds ratio given by (POR: 1.07, 95% CI: 0.98 to 1.17), (POR: 0.96, 95% CI: 0.96 to 1.35), (POR: 1.04, 95% CI: 0.92 to 1.19), (POR: 1.08, 95% CI: 0.67 to 1.74) and (POR: 1.06, 95% CI: 0.59 to 1.67) respectively. The odds of reporting poor health for individuals who smoke a cigarette was 1.09 times the odds of reporting poor health for individuals who do not smoke a cigarette (POR: 1.09, 95%CI: 1.00 to 1.19). Individuals with the flush type of toilets with offsite and onsite disposal were found to be insignificantly associated with self-reported health. The corresponding odds ratios were (POR: 0.88, 95% CI: 0.73 to 1.05) and (POR: 0.97, 95% CI: 0.81 to 1.17) respectively. The odds of reporting poor health for individuals with a bucket and chemical type of toilets were respectively 0.78 (with 95% CI: 0.62 to 0.99) and 1.33 (with 95% CI: 1.03 to 1.72) times the odds of reporting poor health for individuals with no toilet. Employment status was found to be significantly associated with self-reported health. The odds of reporting poor health among individuals who are not economically active were 1.17 times the odds of reporting poor health for individuals who are strictly unemployed (POR: 1.71, 95% CI: 1.06 to 1.29). Nutrition status was found to be significantly associated with self-reported health. The odds of reporting poor health for individuals with underweight nutrition status were 1.23 times the odds of reporting poor health for individuals with normal nutrition status (POR: 1.23, 95% CI: 1.08 to 1.40). The odds of reporting poor health among individuals with severe nutrition status were 1.29 times the odds of reporting poor health for individuals with normal nutrition status (POR: 1.29, 95% CI: 1.05 to 1.59). The odds of reporting poor health among individuals with overweight/obese nutrition status were 0.93 times the odds of reporting poor health for individuals with normal nutrition status, but this was insignificant (POR: 0.93, 95% CI: 0.87 to 1.01). Being diagnosed with TB was found to be significantly associated with self-reported health. The odds of reporting poor health among individuals who were diagnosed with TB was 2.10 times the odds of reporting poor health for individuals who were not diagnosed with TB (POR: 2.10, 96% CI: 1.79 to 2.46). Depression was also found to be significantly associated with self-reported health. The odds of reporting poor health among individuals who felt depressed in past week for little of the time, a moderate amount of the time and all the time were respectively 1.13 (with 95% CI: 1.22, 1.40), 1.34 (with 95% CI: 1.21 to 1.49) and 2.05 (with 95% CI: 1.68 to 2.49) times the odds of reporting poor health for individuals who felt depressed in past week for less than one day. This shows a linear odds increase trend of reporting poor health as depression level increases. However, this is expected because depression is one of the factors that lead to poor health. The cutpoint estimates are also provided in the results. The cutpoint between reporting excellent and very good health is given by θ_1, the cutpoint between reporting very good and good health is θ_2, the cutpoint between reporting good and fair health is θ_3 and the cutpoint between reporting fair and poor health is θ_4. The sign of the cutpoint parameters signifies a shift towards a latent scale, with a positive sign signifying a shift to a higher probability, and the negative signifying a shift towards lower probability. The results reveal that θ_1 is negative, which means reporting excellent health corresponds to reduced odds of reporting poor health. The other three cutpoints (θ_2, θ_3 and θ_4) are all positive, which suggests that reporting very good, good and fair health categories is respectively associated with higher odds of reporting poor health.



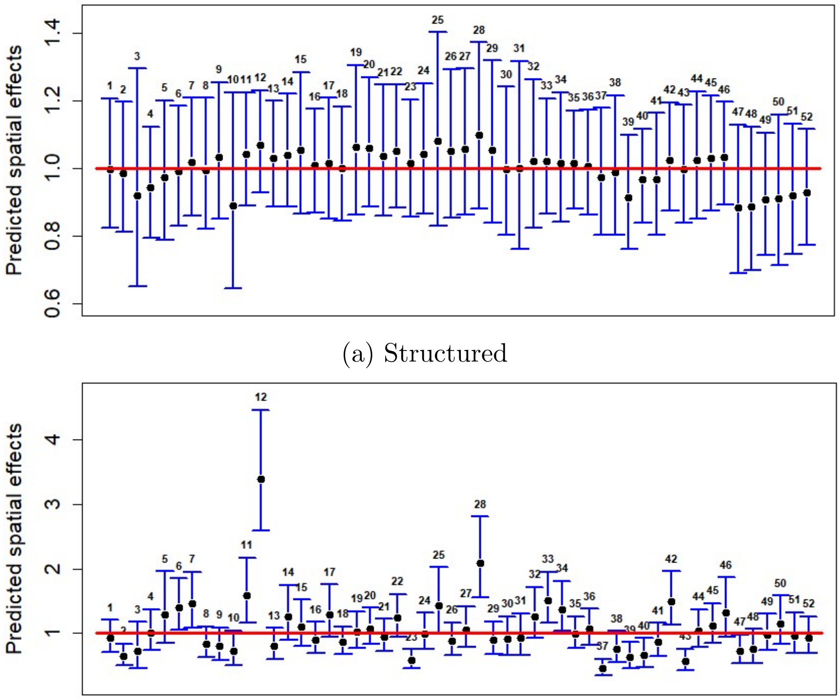
(a)

(b)

Figure 4.1: Map of South Africa showing total spatial district residual effects estimates (a) and their corresponding 95% map of significance (b) of the spatial effects based on Model 4.

The residual total spatial effects are presented based on Model 4. Figure 4.1 shows the map of residuals spatial effects and their corresponding map of significance. All the district names and their corresponding codes are presented in Figure C.1 and Table C.1 respectively. In Figure 4.1(a) black and dark grey coloured districts indicate a higher association of reporting poor health and the grey and light grey coloured districts indicate a lower association of reporting poor health. In addition, looking at the map on the right the white colored districts indicate not significant, the black indicate significantly high odds of reporting poor health and grey indicate significantly low odds of reporting poor health. There is clear evidence of spatial variations at the district level of self-reported health in South Africa. Figure 4.1(a) reveal that there is a significantly higher concentration in the central regions. Figure 4.1(b) shows that the Lejweleputswa, Xhariep, Sisonke, Amajuba, City of Johannesburg, Bojanala and Ehlanzeni districts were significantly associated with self-reported health. The Lejweleputswa, Xhariep and Sisonke districts had high poor health prevalence. Moreover, the Amajuba, City of Johannesburg, Bojanala and Ehlanzeni districts showed low prevalence poor health.

Figure 4.2 display the predicted structured (a) and unstructured (b) residual spatial effects and their corresponding 95% credible intervals (CIs) based on Model 4. The numbers



(a) Structured

(b) Unstructured

Figure 4.2: Predicted residuals of the spatially structured (a) and unstructured (b) effects with their 95% credible intervals based on Model 4.

on top of each plot correspond to the district codes presented in Table C.1. Each credible interval has a length inversely related to the number of the collected odds ratios of reporting poor health and it can be used to test whether the spatial effects are significantly different from one. It can be seen that all the spatially structured random effects in Figure 4.2(a) were found to be insignificantly associated with self-reported health. On the other hand, the spatially unstructured random effects in Figure 4.2(b) were found to be significantly associated with self-reported health, with Chris Hani, Joe Gqabi, Xhariep, Lejweleputswa, Sisonke, Waterberg, Great Sekhukhune and Frances Baard districts significantly having the effects of increasing the individuals odds of reporting poor health, while City of Johannesburg, Amajuba, Ehlanzeni, Namakwa, Pixley ka Seme, Bojanala and City of Cape Town districts significantly having the effects of decreasing the individuals odds of reporting poor health.

4.5 Application of Structured Additive Regression (STAR) models to NIDS wave 4 data

In this section, we propose various models that are extensions to the spatial models from the previous section. These models provide a more flexible approach that accounts for generic covariates. To be briefer we account for nonlinear effects of continuous covariates. Hence in this section, we proposed the following models to further capture the effects of categorical and continuous covariates on self-reported health, also we assess the spatial variations. Furthermore, it is worth noting that all the models in this section are an extension of Equation (4.3), the linear predictor is replaced with a structured linear predictor defined in Equation (4.7).

The first model is the standard cumulative logit regression model which incorporates fixed and nonlinear effects for age and body mass index (BMI), but it does not account for spatial random effects. The model is given by

$$\text{Model A1} : \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(age_{ij1}) + f_2(bmi_{ij2}),$$

in this model, age and BMI are continuous covariates of an individual. The second model is similar to Model A1 with an additional of spatially structured random effects, given by

$$\text{Model A2} : \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(age_{ij1}) + f_2(bmi_{ij2}) + f_{str}(s_i)$$

to capture unobserved influential factors that may vary across districts or spatial location in general. In the third model, we propose a similar model to Model A2 but instead incorporates spatially unstructured random effects. The model is given by

$$\text{Model A3} : \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(age_{ij1}) + f_2(bmi_{ij2}) + f_{unstr}(s_i)$$

which captures unstructured heterogeneity. The final model examines the effect of fixed and nonlinear effects and accounts for both spatially structured and unstructured random effects. The model is given by

$$\text{Model A4} : \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(age_{ij1}) + f_2(bmi_{ij2}) + f_{str}(s_i) + f_{unstr}(s_i),$$

which captures spatial heterogeneity. In all the four models defined above the vector $\boldsymbol{\beta}$ are regression coefficients which were assigned independent diffuse priors ($\boldsymbol{\beta} \propto const$), the smooth functions (f_1 and f_2) of continuous covariates age and BMI were assumed to have a nonlinear effect on self-reported health and were both assigned second-order random walk priors discussed in Section 3.25. Furthermore, the spatially structured effects were assigned Markov random fields (MRFs) prior and the spatially unstructured were assigned the i.i.d Gaussian prior.

Table 4.2 presents the cumulative posterior odds ratios (PORs) and their corresponding 95% credible intervals (CIs) for all the fitted models. The table also shows the results of the model fit statistics. The model fit values suggest that the AIC value favors Model A4 while the GCV values are the same for Model A3 and Model A4.

Table 4.2: Parameter estimates of the multivariable Bayesian spatial cumulative logit models with nonlinear effects.

	Model A1	**Model A2**	**Model A3**	**Model A4**
Covariates	**POR (95% CI)**	**POR (95% CI)**	**POR (95% CI)**	**POR (95% CI)**
Gender (ref = Female)				
Male	0.79 (0.88, 0.85)	0.78 (0.73, 0.84)	0.78 (0.73, 0.84)	0.78 (0.73, 0.84)
Race (ref = African)				
Asian/Indian	0.88 (0.84, 1.23)	0.87 (0.62, 1.22)	0.88 (0.62, 1.23)	0.87 (0.62, 1.22)
Coloured	0.84 (0.97, 0.92)	1.05 (0.91, 1.21)	1.02 (0.89, 1.16)	1.03 (0.90, 1.18)
White	0.97 (1.02, 1.28)	1.10 (0.83, 1.44)	1.08 (0.82, 1.42)	1.09 (0.83, 1.43)
Type of residence (ref = Urban informal)				
Rural Formal	1.02 (1.13, 1.18)	0.98 (0.85, 1.15)	0.99 (0.85, 1.15)	0.99 (0.85, 1.15)
Urban Formal	1.13 (1.03, 1.28)	1.10 (0.97, 1.25)	1.10 (0.97, 1.25)	1.10 (0.97, 1.25)
Tribal Authority Areas	1.03 (1.14, 1.16)	0.96 (0.83, 1.11)	0.97 (0.84, 1.12)	0.97 (0.84, 1.11)
Education level (ref = No formal education)				
Primary	1.14 (0.75, 1.52)	1.15 (0.87, 1.53)	1.15 (0.86, 1.53)	1.15 (0.86, 1.53)
Secondary	0.75 (0.53, 0.96)	0.75 (0.59, 0.96)	0.75 (0.59, 0.95)	0.75 (0.59, 0.95)
High	0.53 (0.51, 0.67)	0.54 (0.42, 0.68)	0.53 (0.42, 0.67)	0.53 (0.42, 0.67)
College	0.51 (0.45, 0.67)	0.53 (0.40, 0.69)	0.52 (0.40, 0.69)	0.52 (0.40, 0.69)
Tertiary	0.45 (1.18, 0.57)	0.44 (0.35, 0.57)	0.44 (0.34, 0.56)	0.44 (0.34, 0.56)
Household income (ref = Much below average)				
Below average	1.18 (1.31, 1.30)	1.15 (1.05, 1.27)	1.15 (1.05, 1.27)	1.15 (1.05, 1.27)
Average	1.31 (1.11, 1.43)	1.28 (1.17, 1.40)	1.28 (1.17, 1.40)	1.28 (1.17, 1.40)
Above average	1.11 (1.00, 1.27)	1.14 (0.99, 1.30)	1.14 (1.00, 1.30)	1.14 (1.00, 1.30)
Much above average	1.07 (1.17, 1.18)	1.07 (0.90, 1.27)	1.06 (0.90, 1.26)	1.06 (0.89, 1.26)
Marital status (ref = Not married)				
Widow/Divorced/Seperated	1.17 (0.99, 1.39)	1.14 (0.96, 1.37)	1.15 (0.96, 1.37)	1.15 (0.96, 1.37)
Married/living with partner	0.99 (0.98, 1.08)	0.98 (0.90, 1.07)	0.98 (0.90, 1.06)	0.98 (0.90, 1.06)
Life satisfaction level (ref = Very dissatisfied)				
Dissatisfied	0.98 (1.03, 1.10)	0.95 (0.84, 1.06)	0.95 (0.84, 1.06)	0.95 (0.84, 1.06)
Neutral	1.03 (0.96, 1.16)	1.03 (0.91, 1.15)	1.03 (0.92, 1.15)	1.03 (0.92, 1.15)
Satisfied	0.96 (0.93, 1.08)	0.98 (0.86, 1.11)	0.98 (0.86, 1.11)	0.98 (0.87, 1.11)
Very Satisfied	0.93 (0.75, 1.06)	0.93 (0.81, 1.08)	0.94 (0.81, 1.08)	0.94 (0.81, 1.08)
Exercise (ref = Never)				
Less than once a week	0.75 (0.93, 0.84)	0.78 (0.70, 0.88)	0.78 (0.70, 0.88)	0.78 (0.70, 0.88)
Once a week	0.93 (1.01, 1.06)	0.97 (0.84, 1.11)	0.96 (0.84, 1.10)	0.96 (0.84, 1.10)
Twice a week	1.01 (0.87, 1.14)	1.03 (0.91, 1.17)	1.03 (0.91, 1.17)	1.03 (0.91, 1.17)
Three or more times a week	0.87 (1.19, 0.95)	0.88 (0.81, 0.96)	0.88 (0.8, 0.96)	0.88 (0.80, 0.96)
Alcohol consumption level (ref = Never drunk alcohol)				
No longer drink alcohol	1.19 (1.10, 1.32)	1.18 (1.07, 1.31)	1.18 (1.07, 1.31)	1.18 (1.07, 1.31)
Drink very rarely	1.10 (1.14, 1.02)	1.07 (0.98, 1.17)	1.07 (0.98, 1.17)	1.08 (0.98, 1.18)
Less than once a week	1.14 (1.04, 1.35)	1.14 (0.97, 1.35)	1.14 (0.97, 1.35)	1.15 (0.97, 1.35)
1 or 2 days a week	1.04 (1.25, 1.18)	1.05 (0.93, 1.20)	1.05 (0.92, 1.19)	1.05 (0.92, 1.19)
3 or 4 days a week	1.25 (1.09, 1.60)	1.30 (1.02, 1.66)	1.30 (1.02, 1.66)	1.30 (1.02, 1.66)
5 or 6 days a week	1.09 (1.04, 1.76)	1.09 (0.68, 1.76)	1.09 (0.68, 1.76)	1.09 (0.68, 1.76)
Every day	1.04 (1.08, 1.74)	0.99 (0.59, 1.67)	1.04 (0.59, 1.67)	1.05 (0.59, 1.68)

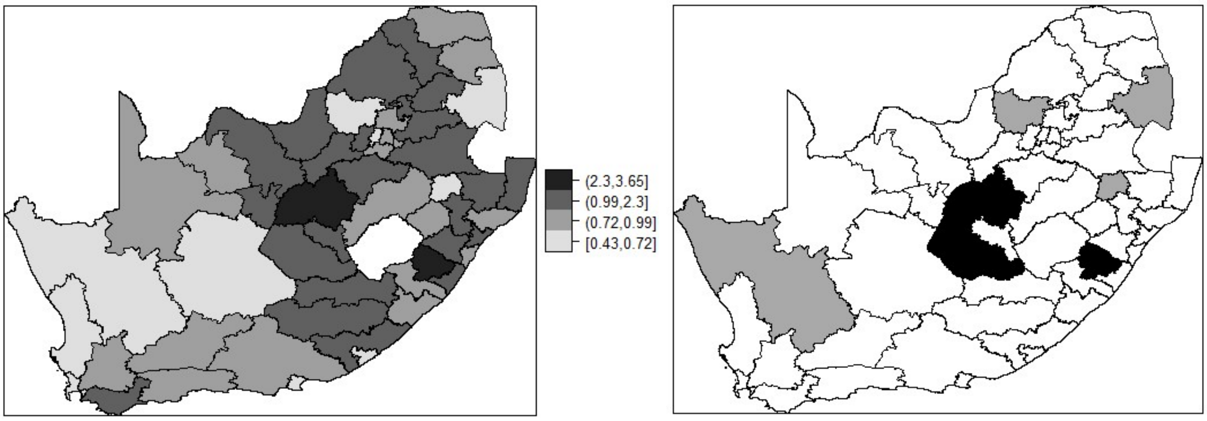
Table 4.2 *Continues*

Covariates	**Model A1** **POR (95% CI)**	**Model A2** **POR (95% CI)**	**Model A3** **POR (95% CI)**	**Model A4** **POR (95% CI)**
Smokes (ref = No)				
Yes	1.08 (0.88, 1.19)	1.07 (0.98, 1.18)	1.07 (0.98, 1.18)	1.07 (1.01, 1.18)
Type of toilet (ref = None)				
Flush toilet with offsite disposal	0.88 (0.92, 1.05)	0.87 (0.72, 1.05)	0.87 (0.72, 1.04)	0.87 (0.72, 1.05)
Flush toilet with onsite disposal	0.92 (0.80, 1.11)	0.97 (0.81, 1.17)	0.96 (0.80, 1.16)	0.97 (0.80, 1.16)
Bucket toilet	0.80 (1.29, 1.01)	0.78 (0.62, 0.99)	0.78 (0.61, 0.99)	0.78 (0.61, 0.99)
Chemical toilet	1.29 (0.78, 1.66)	1.32 (1.02, 1.71)	1.32 (1.02, 1.71)	1.33 (1.02, 1.71)
Pit latrine with ventilation pipe	0.78 (0.86, 0.93)	0.81 (0.68, 0.97)	0.81 (0.68, 0.97)	0.81 (0.68, 0.97)
Pit latrine without ventilation pipe	0.86 (1.51, 1.03)	0.90 (0.75, 1.07)	0.90 (0.75, 1.07)	0.90 (0.75, 1.07)
Other	1.51 (0.81, 3.11)	1.67 (0.81, 3.44)	1.67 (0.81, 3.44)	1.67 (0.81, 3.45)
Employment status (ref = Unemployed strict)				
Unemployed Discouraged	0.81 (1.24, 1.05)	0.84 (0.65, 1.09)	0.84 (0.65, 1.09)	0.84 (0.64, 1.09)
Not Economically Active	1.24 (0.99, 1.37)	1.17 (1.06, 1.29)	1.17 (1.07, 1.29)	1.17 (1.07, 1.29)
Employed	0.99 (2.08, 1.09)	0.98 (0.90, 1.08)	0.98 (0.90, 1.08)	0.98 (0.90, 1.08)
Was diagnosed with TB? (ref = No)				
Yes	2.08 (1.32, 2.44)	2.10 (1.79, 2.46)	2.09 (1.78, 2.46)	2.10 (1.79, 2.46)
Felt depressed in past week? (ref = Less than 1 day)				
Little of the time (1-2 days)	1.32 (1.31, 1.41)	1.30 (1.22, 1.40)	1.31 (1.22, 1.40)	1.30 (1.22, 1.40)
Moderate amount of time (3-4 days)	1.31 (2.04, 1.45)	1.34 (1.21, 1.49)	1.34 (1.21, 1.49)	1.34 (1.21, 1.49)
All of the time (5-7 days)	2.04 (1.68, 2.47)	2.04 (1.68, 2.48)	2.03 (1.68, 2.47)	2.04 (1.68, 2.47)

	Est. (95 % CI)	**Est. (95 % CI)**	**Est. (95 % CI)**	**Est. (95 % CI)**
θ_1	-1.18 (-1.50, -0.86)	-1.25 (-1.60, -0.91)	-1.26 (-1.62, -0.90)	-1.26 (-1.61, -0.90)
θ_2	0.26 (-0.06, 0.58)	0.23 (-0.12, 0.57)	0.22 (-0.14, 0.58)	0.22 (-0.13, 0.58)
θ_3	2.39 (2.07, 2.72)	2.41 (2.06, 2.75)	2.40 (2.04, 2.76)	2.4 (2.04, 2.76)
θ_4	3.96 (3.62, 4.31)	3.98 (3.61, 4.34)	3.97 (3.59, 4.35)	3.97 (3.59, 4.35)
Additional model parameters				
Spatially structured variation (σ^2_{str})	-	0.613	-	0.03
Spatially unstructured variation (σ^2_{unstr})	-	-	0.153	0.139
Age effect (σ^2_{age})	0.0001	0.0002	0.0002	0.0002
BMI effect (σ^2_{BMI})	0.0001	0.0001	0.0001	0.0001
Model fit				
AIC	39127.9	38583.3	38582.8	38582.4
BIC	39582.4	39608.0	39609.7	39609.5
GCV	2.4512	2.3850	2.3848	2.3848

However, the BIC value favors Model A1. Therefore, the results are interpreted based on Model A4 which incorporates both the spatially structured and unstructured random effects. The categorical covariates were assumed to have a linear relationship with self-reported health. Based on Model A4 in Table 4.2, the results show that all the considered categorical covariates were statistically significant except for race, place of residence, marital status and, life satisfaction level. Gender was found to be significantly associated with self-reported health. The odds of reporting poor health for male individuals were 0.78 times the odds of reporting poor health for female individuals. Education was found to be significantly associated with self-reported health. The odds of reporting poor health for individuals with secondary education were 0.75 (with 95% CI: 0.59 to 0.95) times the odds of reporting poor health for individuals with no education. The odds of reporting poor health for individuals with high education were 0.53 (with 95% CI: 0.42 to 0.67) times the odds of reporting poor health for individuals with no education. The odds of reporting poor health for individuals with a college education were 0.52 (with 95% CI: 0.40 to 0.69) times the odds of reporting poor health for individuals with no education. Furthermore, the odds of reporting poor health for individuals with tertiary education were 0.44 (with 95% CI: 0.34 to 0.56) times the odds of reporting poor health for individuals with no education. These means the prevalence of poor health still remains less the higher the education. Household income was also found to be significantly associated with self-reported health. The odds of reporting poor health for individuals with below average household income were 1.15 (with 95% CI: 1.05 to 1.27) times the odds of reporting poor health for individuals with much below average household income. The odds of reporting poor health among individuals with average household income were 1.28 times the odds of reporting poor health for individuals with much below average household income (POR: 1.28, 95% CI: 1.17 to 1.40). Moreover, the odds of reporting poor health among individuals with above average household income were 1.14 times the odds of reporting poor health for individuals with much below average household income (POR: 1.14, 95% CI: 1.00 to 1.30). However, one would expect that the higher the household income the less poor health prevalence. The odds of reporting poor health among individuals exercising less than once a week were 0.78 times the odds of reporting poor health for individuals who never exercise (POR: 0.78, 95% CI: 0.70 to 0.88). The odds of reporting poor health among individuals who exercise three or more times a week were 0.88 times the odds of reporting poor health among individuals who never exercise (POR: 0.88, 95% CI: 0.80 to 0.96). Alcohol was found to be significantly associated with self-reported health. The odds of reporting poor health among individuals who no longer drink alcohol were 1.18 times the odds of reporting poor health for individuals who never drunk alcohol (POR: 1.18, 95% CI: 1.07 to 1.31). The odds of reporting poor health among individuals who drink on 3 or 4 days a week were 1.30 times the odds of reporting poor health for individuals who never drunk alcohol (POR: 1.30, 95% CI: 1.02 to 1.66). Smoking was found to be positively associated with self-reported health, with the odds ratio given by (POR: 1.07, 95% CI: 1.01 to 1.18). The odds of reporting poor health for individuals who smoke a cigarette was 1.07 times the odds of reporting poor health among individuals who do not smoke a cigarette. Type of toilet was found to be significantly associated with self-reported health. The odds of reporting poor health among individuals with bucket, chemical and pit latrine with ventilation pipe toilets were respectively 0.78 (with 95% CI: 0.61 to 0.99), 1.33 (with 95% CI: 1.02 to 1.71) and 0.81 (with 95% CI: 0.68 to 0.97) times the odds of reporting poor health for individuals who have no toilet. The odds of reporting poor health among individuals who are not economically active were 1.17 times the odds of reporting poor health for individuals who are unemployed strictly (POR: 1.17, 95% CI: 1.07 to 1.29). The odds of reporting poor health for individuals who were diagnosed with TB was 2.10 times the odds of reporting poor health among individuals who were not diagnosed with TB (POR: 2.10, 95% CI: 1.79 to 2.46). This means the prevalence of poor health is high for individuals who were previously diagnosed with TB. Depression was found to be positively associated with self-reported health. The odds of reporting poor health among individuals who felt depressed in past week little of the time were 1.30 (with 95% CI: 1.22 to 1.40) times the odds of reporting poor health for individuals who felt depressed in past week for less than one day. The odds of reporting poor health among individuals who felt depressed in past week a moderate amount of time were 1.34 (with 95% CI: 1.21 to 1.49) times the odds of reporting poor health for individuals who felt depressed in past week for less than one day. The odds of reporting poor health among individuals who felt depressed in past week all of the time were 2.04 (with 95% CI: 1.68 to 2.47) times the odds of reporting poor health for individuals who felt depressed in past week for less than one day. A linear odds increase of reporting poor health as individuals depression level increases is observed. These means the prevalence of poor health is high the higher the depression level. Furthermore, the results show that the cutpoint θ_1 is negative, which means reporting excellent health status corresponds to reduced odds of poor reported health. The other three cutpoints (θ_2, θ_3 and θ_4) are all positive, which means that reporting very good, good and fair health categories respectively are associated with higher odds of poor reported health.

Figure 4.3 display the maps of the total residual spatial effects and the corresponding 95% posterior probability map of significance. Similar results as in Figure 4.1 can be observed for both maps. Figure 4.3(a) shows the spatial variation of self-reported health. Spatial variation can be observed in the districts of South Africa. The districts within the northern, central and southern regions had higher poor health prevalence. The Lejweleputswa and Sisonke remaining the significantly highest districts with poor health prevalence. The map in Figure 4.3(b) shows that the Namakwa district was significantly associated with self-reported health. Figure 4.3(a) show that districts in the western regions had



(a)

(b)

Figure 4.3: Map of South Africa showing total spatial district residual effects estimates (a) and the corresponding 95% map of significance (b) of spatial effect estimates of Model $A4$.

lower poor health prevalence. Namakwa district recorded significantly lower poor health prevalence.

Figure 4.4 shows the nonlinear association between age (left), and BMI (right) and self-reported health. Shown is the posterior mean of the smooth functions together with their 95% pointwise credible intervals. We assumed a nonlinear relationship between the respondents' age, BMI and self-reported health. However, Figure 4.4 reveals that age had a linear relationship with self-reported health while the BMI had a clear nonlinear relationship with self-reported health. An increasing age effect can be observed, which is in line with the categorized age. It can also be observed that decreasing to a minimum of BMI between 20 - 25 then starts increasing again. This reveals the same effect as the categorical nutrition status. The age plot reveals that individuals age at 15 to 34 years reduced the odds of reporting poor health, while 35 years and above increases the odds of reporting poor health. These linear trend increase of respondents age (35 years and above) confirms the observed Model 4 findings shown in Table 4.1. The BMI plot reveals that the effect of BMI on individuals self-reported health is approximate to the U shape form. This appears absolutely reasonable as the normal nutrition status is likely to reduce the odds of reporting poor health. The individuals BMI at severity (BMI < 17) increased the odds of reporting poor health while between approximately 18 - 43 the individuals BMI reduced the odds of reporting poor health. An individual with BMI around (43 and higher) was associated with increased odds of reporting poor health.

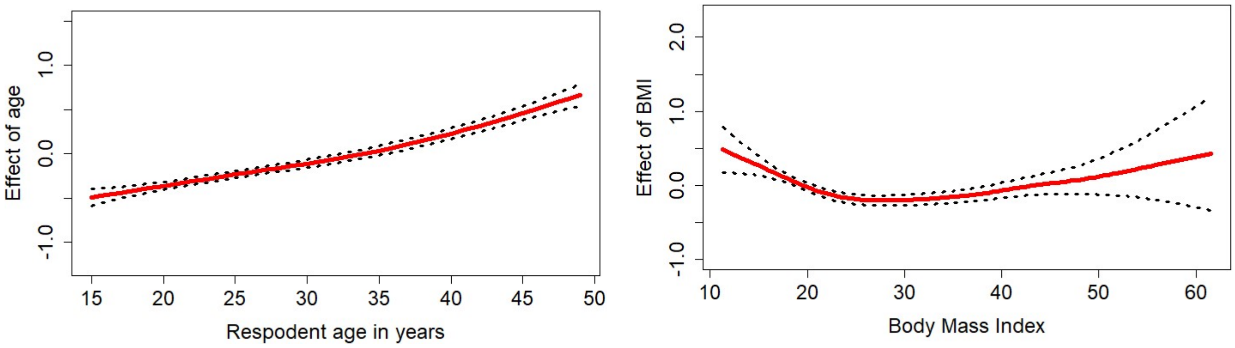


Figure 4.4: Estimated mean (red) of the non-linear effects of age (left) and body mass index (BMI) (right) with 95% credible interval (dotted black lines) of Model $A4$.

The Cumulative regression model is often used when the response is categorical of ordered nature. The assumption under this traditional model is that the predictor is strictly linear. However, the STAR model allows for generic covariates to be added in the predictor in an additive manner. The spatial effects account for unobserved influential factors. The convolution models showed better fitting models. There was a slight difference between the strictly spatial model and the STAR models. The inclusion of continuous covariates further improved the results. There is evidence of spatial variation of self-reported health in South Africa.

Chapter 5

Spatial Modeling of self-reported health Binary response outcome

5.1 Introduction

In Chapter 4 we reviewed and discussed the models where the response variable y_i is an ordinal outcome. In this chapter, we will review models that are commonly used when the outcome is binary (dichotomous) as part of the objective for this research. In particular, a flexible approach is adopted for such models that allows capturing of different types of covariates. One of interest is the incorporation of spatial random effects which allow for correlated and uncorrelated heterogeneity. The generalized linear models (GLMs) are the class of models used for binary response outcome.

5.2 Generalized Linear Models

Similar to the multivariate generalized linear models (MGLMs) discussed in Section 4.2, the generalized linear models (GLMs) are classes of models which were introduced by McCullagh and Nelder (1989). They are used for modeling non-Gaussian response variables. In the GLMs it is assumed that for a given vector of covariates $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ and unknown regression parameters $\boldsymbol{\beta}$ given by $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$. The responses

$y_i = (y_1, \ldots, y_n)'$ are assumed to be independent observations and have a distribution that belongs to the exponential family. Hence the probability density of the y_i's is similar to the one in Equation (4.1). It is given by

$$f(y_i|\mu_i, \phi, \omega_i) = \exp\left(\frac{y_i'\mu_i - b(\mu_i)}{\phi}\omega_i + c(y_i, \phi, \omega_i)\right),\ i = 1, \ldots, n, \tag{5.1}$$

under the univariate response properties. Analogous to the MGLMs, where $\phi = 1$ is the dispersion parameter, the ω_i represent a weight for the observations and μ_i is the natural parameter of the exponential family. Furthermore, the $b(\mu_i)$ and $c(y_i, \phi, \omega_i)$ are exponential family specific dependent functions. Now it follows that given covariates $\mathbf{x}_i'$ the linear predictor is given by

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta},\ i = 1, \ldots, n,$$

and it is linked to the conditional mean $\theta_i = E(y_i|\mathbf{x}_i, \boldsymbol{\beta})$ via a link function as

$$g(\theta_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta},\ i = 1, \ldots, n,$$

where $g(\cdot)$ is the natural link function. There are other possible choices for the link function such that when $y_i \in \{0, 1\}$, a Bernoulli distribution is assumed and the link function can be chosen to be a logit link function. This lead to a logit model which represents the systematic logistic distribution function. Hence, this fulfills the GLMs components. The link between the structure and the distribution assumptions above are determined by that the mean of y_i is also assumed to be of the distributional assumption given by

$$E(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \theta_i = b'(\mu_i) \quad \text{and} \quad Var(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = b''(\mu_i)/\omega_i.$$

For more comprehensive details on the theory of the GLMs, the reader may refer to McCullagh and Nelder (1989), Agresti (2007) and Fahrmeir and Tutz (2001). We now look at a special case of the univariate GLMs used to model binary response data.

5.3 Spatial Logistic Regression Models

Logistic regression models are widely used under frequentist and Bayesian framework to examine the association between covariates and the binary response outcome. Further, it has received too much attention in disease mapping to model dichotomous response data to explain geographic variation that arises in the data. In this research, we propose an extension to a logistic regression model. Next, we shall discuss such models.

5.3.1 Logistic Regression Models

Let y_{ij} be a binary self-reported health for individual j located in district i: $i = 1, \ldots, 52$, whose response is either 0 or 1 such that

$$y_{ij} = \begin{cases} 1: & \text{poor health} \\ 0: & \text{good health.} \end{cases}$$

The response y_{ij} was assumed to be independent Bernoulli distributed with the likelihood given by

$$y_{ij} \sim Bernoulli(\pi_{ij}), \;\; i = 1, \ldots, 52, \; j = 1, \ldots, n$$

where $\pi_{ij} = P(y_{ij} = 1)$ are unknown probabilities and $E(y_{ij}) = \pi_{ij}$ relates to predictor via a logit link function as

$$\text{logit}(\pi_{ij}) = \log\left(\frac{P(y_{ij} = 1)}{1 - P(y_{ij} = 1)}\right) = \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}. \tag{5.2}$$

The vector $\mathbf{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijp})'$ are categorical covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is a vector of regression coefficients. This model only allows for a parametric form of categorical covariates. The main aim of this section is to extend the linear predictor η_{ij} of the logistic regression model in Equation (5.2) to account for a more flexible approach.

Hence, to increase the model complexity by including different forms of covariates. We first extend the logistic regression model to allow for area-specific random effects by replacing the linear predictor in Equation (5.2) with a geoadditive predictor. These random effects are incorporated in the model to capture extra variation. Thus to capture unobserved influential factors that vary across the districts, the model accounts for the structured random effects. This model is given by

$$\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_{str}(s_i). \tag{5.3}$$

Another alternative model is the one which incorporates unstructured random effects instead. The model is given by

$$\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_{unstr}(s_i), \tag{5.4}$$

where f_{unstr} accounts for unobserved heterogeneity within each district. The last model is the convolution model which accounts for both spatial random components as follows:

$$\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_{spat}(s_i), \tag{5.5}$$

where the spatial random effects are decomposed into two components, i.e. $f_{spat}(s_i) = f_{str}(s_i) + f_{unstr}(s_i)$. In addition, the two components are assumed to have independent prior distributions (Besag et al., 1991). In all the model's formulation above the regression coefficients $\boldsymbol{\beta}$ were assumed to have diffuse prior ($\boldsymbol{\beta} \propto const$), the spatially unstructured random effects were assumed to follow an i.i.d Gaussian distribution and the spatially structured random effects were modeled with an *intrinsic* conditional autoregressive (iCAR) defined in Equation (3.21).

5.3.2 Parameter Estimation

In this section, parameter estimation is obtained using a fully Bayesian (FB) procedure. In a FB approach, all the unknown parameters are assumed to be random variables and are assigned priors and further hyperparameters are assigned hyperpriors. The parameter estimation can be obtained by sampling from the posterior distribution, with Markov chain Monte Carlo (MCMC) simulation being a commonly used technique. However, the estimation of parameters for this research was carried out using integrated nested Laplace approximation (INLA) discussed in Section 3.10. The latent Gaussian variables for all the above formulated models under this section is given by $\boldsymbol{\varrho} = \{\{\boldsymbol{\beta}\}, \{f_{str}(\cdot)\}, \{f_{unstr}(\cdot)\}\}$ and the hyperparameters are denoted by a set of precision parameters $\boldsymbol{\psi} = \{\tau_{str}, \tau_{unstr}\}$. Hence the posterior distribution is given by

$$p(\boldsymbol{\varrho}, \boldsymbol{\psi}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\varrho}, \boldsymbol{\psi})p(\boldsymbol{\varrho}, \boldsymbol{\psi}), \tag{5.6}$$

here the hyperparameters τ_{str} and τ_{unstr} are assigned conjugate gamma priors, with $\tau_{str} \sim Gamma(1, 0.00005)$ and $\tau_{unstr} \sim Gamma(1, 0.00005)$. The variability of structured and unstructured spatial random effects are determined by $\sigma^2_{str} = \frac{1}{\tau_{str}}$ and $\sigma^2_{unstr} = \frac{1}{\tau_{unstr}}$ respectively. Furthermore, a sum to zero constraints was imposed on both the functions of spatial random effects for identification.

5.3.3 Application of the spatial models to NIDS wave 4 data

Here we consider several models with applications to wave 4 NIDS data in South Africa. In the analysis, Model 01 is similar to the classical model given in Equation (5.2), it accounts for categorical covariates which are assumed to have linear effects on self-reported health. Model 02 is given by Equations (5.4), this model is also similar to Model 01 and accounts for spatially structured random effects which cater for unobserved influential factors that vary among districts. Model 03 is given by Equation (5.3) and caters for spatially unstructured random effects which accounts for unobserved influential factors that are inherent within the districts. Furthermore, Model 04 examines the effects of linear effects of categorical covariates and incorporates both the spatially structured and spatially unstructured random effects known as the convolution model. All models in this chapter were implemented in the R-INLA package and the corresponding R codes are presented in Appendix B.

For the fitted models in this section, the selection of the better fitting model was done based on the deviance information criterion (DIC) suggested by (Spiegelhalter et al., 2002). The model with the smallest DIC value is considered as the best fitting model. Table 5.1 presents the results for all the fitted spatial logistic regression models. The DIC, $\overline{D}$ and p_D model fit statistics are also shown in the table. The results suggest that Model 02 and Model 04 are generally the same except for a slight difference of 0.09. Therefore, the results are further interpreted based on Model 04 which incorporate both the spatially structured and unstructured effects.

Table 5.1: Parameter estimates of multivariable Bayesian spatial logistic regression models.

	Model 01	**Model 02**	**Model 03**	**Model 04**
Covariates	**POR (95% CI)**	**POR (95% CI)**	**POR (95% CI)**	**POR (95% CI)**
Age group (ref = 15-19)				
20-24	1.05 (0.77, 1.44)	1.06 (0.78, 1.46)	1.07 (0.78, 1.46)	1.06 (0.78, 1.46)
25-29	1.39 (1.00, 1.93)	1.40 (1.01, 1.94)	1.41 (1.02, 1.96)	1.40 (1.01, 1.94)
30-34	2.09 (1.51, 2.89)	2.07 (1.49, 2.88)	2.11 (1.52, 2.93)	2.07 (1.49, 2.88)
35-39	3.55 (2.57, 4.90)	3.57 (2.59, 4.94)	3.62 (2.62, 5.00)	3.57 (2.59, 4.94)
40-44	3.99 (2.87, 5.55)	4.05 (2.91, 5.64)	4.08 (2.93, 5.68)	4.05 (2.91, 5.64)
45-49	5.72 (4.13, 7.95)	5.77 (4.16, 8.02)	5.84 (4.20, 8.12)	5.77 (4.16, 8.02)
Gender (ref = Female)				
Male	0.67 (0.56, 0.80)	0.70 (0.58, 0.84)	0.69 (0.57, 0.83)	0.70 (0.58, 0.84)
Race (ref = African)				
Asian/Indian	1.22 (0.58, 2.34)	1.90 (0.88, 3.77)	1.78 (0.83, 3.56)	1.90 (0.88, 3.77)
Coloured	0.52 (0.40, 0.67)	0.68 (0.49, 0.95)	0.61 (0.45, 0.83)	0.68 (0.49, 0.95)
White	1.27 (0.71, 2.16)	1.59 (0.87, 2.74)	1.51 (0.83, 2.61)	1.59 (0.87, 2.74)
Place of residence (ref = Urban informal)				
Rural Formal	0.79 (0.56, 1.12)	0.75 (0.52, 1.07)	0.77 (0.54, 1.10)	0.75 (0.52, 1.07)
Urban Formal	1.11 (0.84, 1.47)	0.97 (0.73, 1.30)	0.97 (0.75, 1.34)	0.97 (0.73, 1.30)
Tribal Authority Areas	0.87 (0.65, 1.17)	0.84 (0.61, 1.16)	0.84 (0.61, 1.17)	0.84 (0.61, 1.16)

Table 5.1 *Continues*

Covariates	Model 01 POR (95% CI)	Model 02 POR (95% CI)	Model 03 POR (95% CI)	Model 04 POR (95% CI)
Education level (ref = No formal education)				
Primary	0.93 (0.61, 1.42)	0.94 (0.62, 1.44)	0.93 (0.61, 1.42)	0.94 (0.62, 1.44)
Secondary	0.67 (0.46, 0.97)	0.65 (0.45, 0.96)	0.66 (0.45, 0.96)	0.65 (0.45, 0.96)
High	0.45 (0.32, 0.65)	0.44 (0.31, 0.63)	0.44 (0.31, 0.63)	0.44 (0.31, 0.63)
College	0.46 (0.28, 0.74)	0.43 (0.26, 0.70)	0.44 (0.27, 0.71)	0.43 (0.26, 0.70)
Tertiary	0.44 (0.29, 0.67)	0.43 (0.28, 0.65)	0.43 (0.28, 0.65)	0.43 (0.28, 0.65)
Household income (ref = Much below average)				
Below average	1.10 (0.89, 1.37)	1.09 (0.88, 1.36)	1.09 (0.88, 1.35)	1.09 (0.88, 1.36)
Average	1.11 (0.90, 1.37)	1.11 (0.90, 1.37)	1.11 (0.90, 1.37)	1.11 (0.90, 1.37)
Above average	0.91 (0.64, 1.27)	0.88 (0.62, 1.23)	0.89 (0.63, 1.25)	0.88 (0.62, 1.23)
Much above average	0.37 (0.19, 0.64)	0.36 (0.19, 0.63)	0.36 (0.19, 0.64)	0.36 (0.19, 0.63)
Marital status (ref = Not married)				
Widow/Divorced/Seperated	1.04 (0.75, 1.43)	0.98 (0.70, 1.34)	0.99 (0.71, 1.36)	0.98 (0.70, 1.34)
Married/living with partner	1.10 (0.92, 1.31)	1.04 (0.86, 1.24)	1.04 (0.87, 1.25)	1.04 (0.86, 1.24)
Life satisfaction level (ref = Very dissatisfied)				
Dissatisfied	0.73 (0.56, 0.94)	0.73 (0.56, 0.95)	0.73 (0.56, 0.95)	0.73 (0.56, 0.95)
Normal	0.93 (0.73, 1.20)	0.92 (0.72, 1.19)	0.92 (0.72, 1.19)	0.92 (0.72, 1.19)
Satisfied	0.76 (0.57, 1.00)	0.75 (0.57, 1.00)	0.76 (0.57, 1.00)	0.75 (0.57, 1.00)
Very satisfied	0.82 (0.59, 1.13)	0.80 (0.58, 1.12)	0.81 (0.58, 1.12)	0.80 (0.58, 1.12)
Exercise (ref = Never)				
Less than once a week	0.65 (0.47, 0.90)	0.62 (0.44, 0.85)	0.63 (0.45, 0.87)	0.62 (0.44, 0.85)
Once a week	1.10 (0.78, 1.52)	1.05 (0.75, 1.46)	1.07 (0.76, 1.48)	1.05 (0.75, 1.46)
Twice a week	0.86 (0.61, 1.18)	0.83 (0.59, 1.15)	0.84 (0.60, 1.17)	0.83 (0.59, 1.15)
Three or more times a week	0.77 (0.59, 0.98)	0.74 (0.57, 0.95)	0.74 (0.57, 0.95)	0.74 (0.57, 0.95)
Alcohol consumption level (ref = Never drunk alcohol)				
No longer drink	1.71 (1.37, 2.12)	1.62 (1.30, 2.02)	1.63 (1.30, 2.03)	1.62 (1.30, 2.02)
Drink very rarely	1.28 (1.02, 1.58)	1.17 (0.94, 1.46)	1.19 (0.95, 1.48)	1.17 (0.94, 1.46)
Less than once a week	0.83 (0.51, 1.29)	0.75 (0.46, 1.17)	0.77 (0.47, 1.20)	0.75 (0.46, 1.17)
On 1 or 2 days a week	1.38 (1.02, 1.85)	1.31 (0.97, 1.76)	1.32 (0.98, 1.78)	1.31 (0.97, 1.76)
On 3 or 4 days a week	1.41 (0.84, 2.29)	1.30 (0.77, 2.13)	1.34 (0.79, 2.19)	1.30 (0.77, 2.13)
On 5 or 6 days a week	0.62 (0.19, 1.66)	0.57 (0.17, 1.54)	0.57 (0.17, 1.55)	0.57 (0.17, 1.54)
Every day	1.71 (0.67, 3.93)	1.51 (0.59, 3.49)	1.55 (0.60, 3.57)	1.51 (0.59, 3.49)
Smokes (ref = No)				
Yes	1.23 (0.98, 1.53)	1.25 (1.00, 1.56)	1.24 (0.99, 1.55)	1.25 (1.00, 1.56)
Type of toilet (ref = None)				
Flush toilet with offsite disposal	0.96 (0.64, 1.46)	0.88 (0.58, 1.35)	0.89 (0.59, 1.37)	0.88 (0.58, 1.35)
Flush toilet with onsite disposal	0.88 (0.58, 1.34)	0.82 (0.54, 1.26)	0.82 (0.54, 1.27)	0.82 (0.54, 1.26)
Bucket toilet	0.86 (0.49, 1.49)	0.78 (0.44, 1.37)	0.77 (0.43, 1.36)	0.78 (0.44, 1.37)
Chemical toilet	1.24 (0.70, 2.17)	1.22 (0.68, 2.16)	1.20 (0.67, 2.12)	1.22 (0.68, 2.16)
Pit latrine with ventilation pipe	0.99 (0.66, 1.50)	1.03 (0.69, 1.58)	1.02 (0.68, 1.55)	1.03 (0.69, 1.58)
Pit latrine without ventilation pipe	0.82 (0.56, 1.23)	0.79 (0.53, 1.19)	0.80 (0.54, 1.21)	0.79 (0.53, 1.19)
Other	2.03 (0.47, 6.95)	1.79 (0.41, 6.21)	1.90 (0.44, 6.54)	1.79 (0.41, 6.21)
Employment status (ref = Unemployed strict)				
Unemployed Discouraged	0.51 (0.20, 1.13)	0.54 (0.21, 1.21)	0.55 (0.22, 1.23)	0.54 (0.21, 1.21)

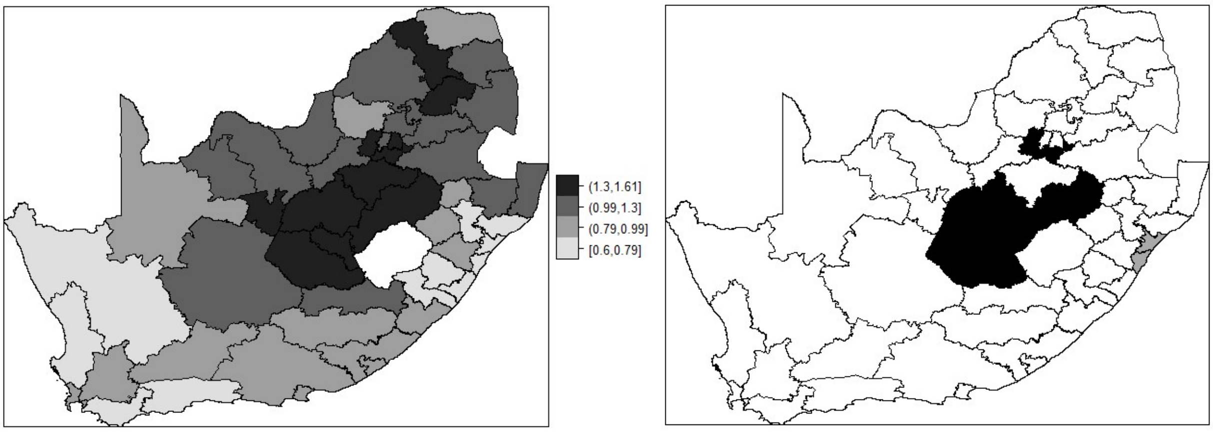
Table 5.1 *Continues*

Covariates	**Model 01** **POR (95% CI)**	**Model 02** **POR (95% CI)**	**Model 03** **POR (95% CI)**	**Model 04** **POR (95% CI)**
Not Economically Active	1.42 (1.13, 1.80)	1.46 (1.15, 1.84)	1.46 (1.16, 1.85)	1.46 (1.15, 1.84)
Employed	0.91 (0.72, 1.14)	0.93 (0.74, 1.17)	0.93 (0.74, 1.17)	0.93 (0.74, 1.17)
Nutrition status (ref = Normal)				
Underweight	1.49 (1.10, 2.00)	1.41 (1.04, 1.90)	1.42 (1.04, 1.91)	1.41 (1.04, 1.90)
Overweight/obese	0.81 (0.68, 0.97)	0.84 (0.70, 1.00)	0.83 (0.70, 0.99)	0.84 (0.70, 1.00)
Severe	2.48 (1.64, 3.68)	2.25 (1.48, 3.35)	2.31 (1.52, 3.44)	2.25 (1.48, 3.35)
Was diagnosed with TB? (ref = No)				
Yes	3.11 (2.44, 3.94)	3.24 (2.53, 4.13)	3.17 (2.47, 4.03)	3.24 (2.53, 4.13)
Felt depressed in past week? (ref = Less than 1 day)				
Little of the time (1-2 days)	1.51 (1.28, 1.78)	1.49 (1.26, 1.76)	1.49 (1.26, 1.76)	1.49 (1.26, 1.76)
Moderate amount of time (3-4 days)	2.21 (1.78, 2.73)	2.20 (1.77, 2.72)	2.19 (1.76, 2.71)	2.20 (1.77, 2.72)
All the time (5-7 days)	3.07 (2.18, 4.27)	3.08 (2.18, 4.30)	3.07 (2.17, 4.29)	3.08 (2.18, 4.30)
	Est. (95 % CI)	**Est. (95 % CI)**	**Est. (95 % CI)**	**Est. (95 % CI)**
Additional model parameters				
Spatially structured variation (σ^2_{str})	-	0.18 (0.06, 0.50)	-	1.83 (0.07, 0.52)
Spatially unstructured variation (σ^2_{unstr})	-	-	0.09 (0.04, 0.22)	0.05 (0.001, 0.077)
Model fit				
DIC	6020.12	5974.29	5976.87	5974.20
$\overline{D}$	5963.07	5895.21	5891.26	5895.08
p_D	57.05	79.08	85.61	79.12

The posterior odds ratio (POR) estimates and their corresponding 95% credible intervals (CI) are presented in Table 5.1. Covariates were considered statistically significant at 5% level of significance. Based on Model 04 it can be observed that all the considered covariates were found to be significantly associated with self-reported health except for marital status, place of residence and type of toilet facility. Age was found to be significantly associated with self-reported health. The odds of reporting poor health among individuals between ages 25-29 years were 1.01 (with 95% CI: 0.78 to 1.46) times the odds of reporting poor health for individuals between 15-19 years. The odds of reporting poor health for individuals between ages 30-34, 35-39, 40-44 and 45-49 years were respectively 2.07 (with 95% CI: 1.49 to 2.88), 3.57 (with 95% CI: 2.59 to 4.94), 4.05 (with 95% CI: 2.91 to 5.64) and 5.77 (with 95% CI: 4.16 to 8.02) times the odds of reporting poor health for individuals between 15-19 years. A linear odds increase trend of reporting poor health can be observed as individuals age increases, similar to the results found in

Section 4.4. The odds of reporting poor health for male individuals were 0.70 times the odds of reporting poor health for individuals who are females (POR: 0.70, 95% CI: 0.58 to 0.84). This means the prevalence of poor health still remains high for females than males. The odds of reporting poor health among individuals who are coloured were 0.68 times the odds of reporting poor health for individuals who are African (POR: 0.68, 95% CI: 0.49 to 0.95). Education was found to be significantly associated with self-reported health. The odds of reporting poor health among individuals with secondary and high education were respectively 0.65 (with 95% CI: 0.45 to 0.96) and 0.44 (with 95% CI: 0.31 to 0.63) times the odds of reporting poor health for individuals with no education. Furthermore, the odds of reporting poor health for individuals with college and tertiary education were respectively 0.43 (with 95% CI: 0.26 to 0.70) and 0.43 (with 95% CI: 0.28 to 0.65) times the odds of reporting poor health for those individuals with no education. This means poor health prevalence is low for individuals with higher education. The odds of reporting poor health for individuals with much above average household income were 0.36 times the odds of reporting poor health among individuals with much below average household income (POR: 0.36, 95% CI: 0.19 to 0.63). Life satisfaction level was also found to be significantly associated with self-reported health. The odds of reporting poor health among individuals with dissatisfied life were 0.73 times the odds of reporting poor health for individuals with very dissatisfied life (POR: 0.73, 95% CI: 0.56 to 0.95). The odds of reporting poor health for individuals who exercise less than once a week were 0.62 times the odds of reporting poor health among individuals who never exercise (POR: 0.62, 95% CI: 0.44 to 0.85). The odds of reporting poor health for individuals who exercise three or more times a week were 0.74 (with 95% CI: 0.57 to 0.95) times the odds of reporting poor health for individuals who never exercise. The odds of reporting poor health among individuals who no longer drink alcohol were 1.62 times the odds of reporting poor health for individuals who never drunk alcohol (POR: 1.62, 9% CI: 1.30 to 2.02). The smoking of cigarette was also found to be significantly associated with self-reported health. The odds of reporting poor health among individuals who smoke a cigarette was 1.25 times the odds of reporting poor health for individuals who do not smoke (POR: 1.25, 95% CI: 1.00 to 1.56). This means smoking of cigarette is associated with high poor health prevalence. The odds of reporting poor health among individuals who are not economically active were 1.46 times the odds of reporting poor health for individuals who are unemployed strictly (POR: 1.46, 95% CI: 1.15 to 1.84). The odds of reporting poor health among individuals who are underweight were 1.41 times the odds of reporting poor health for individuals with normal nutrition status (POR: 1.41, 95% CI: 1.04 to 1.90). Moreover, the odds of reporting poor health for individuals who are severe were 2.25 times the odds of reporting poor health for individuals with normal nutrition status (POR: 2.25, 95% CI: 1.48 to 3.35). The odds of reporting poor health among individuals who were previously diagnosed with TB was 3.24 times the odds of reporting poor health for individuals who were not diagnosed with TB (POR: 3.24, 95% CI: 2.53 to 4.13). Depression was also found to be significantly associated with self-reported health. Similarly to Section 4.4, the results demonstrate that the odds of reporting poor health increased with increasing depression level. The odds of reporting poor health among individuals who were depressed little of the time were 1.49 times the odds of reporting poor health for individuals who were depressed less than one day (POR: 1.49, 95% CI: 1.26 to 1.76). The odds of reporting poor health among individuals who were depressed the moderate amount of the time were 2.20 times the odds of reporting poor health for individuals who were depressed less than one day (POR: 2.20, 95% CI: 1.77 to 2.72). The odds of reporting poor health among individuals who were depressed all the time were 3.08 times the odds of reporting poor health for individuals who were depressed less than one day (POR: 3.08, 95% CI: 2.18 to 4.30).

The residual total spatial effect estimates were mapped based on the better fitting model, Model 04. Figure 5.1 display the map of residual spatial district effects (a) and their 95% posterior probability map of significance (b). All the district names and their corresponding codes are presented in Figure C.1 and Table C.1 respectively. In Figure 5.1(a) black and dark grey color indicate districts with higher odds of reporting poor



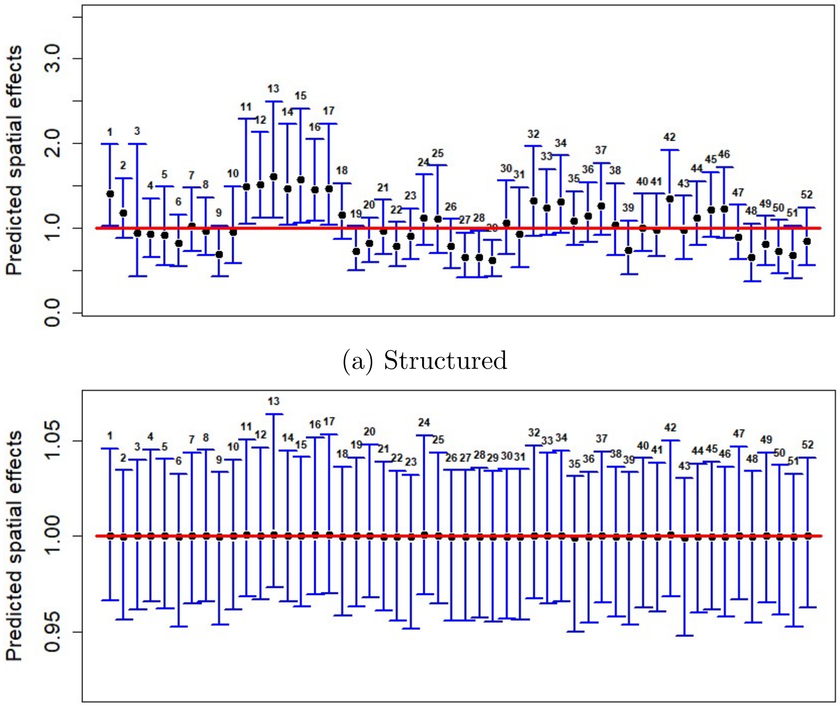
(a)

(b)

Figure 5.1: Map of South Africa showing total spatial district residual effects estimates (a) and the corresponding 95% map of significance (b) of spatial effect estimates based on Model 04.

health and the grey and light grey color indicate districts with lower odds of reporting poor health. In addition, the map on the right indicates the significance of the spatial effects (white = not significant; black = significantly positive (high prevalence); and grey = significantly negative (low prevalence). The maps yielded similar results as the maps for Model 4 in Figure 4.1, except for a slight difference. There is clear evidence of spatial variations at the district level of reporting poor health in South Africa. There is a high prevalence of poor health in the districts situated on the northern and central side of the country. However, in the western and southern regions, there is a low prevalence of poor health. The Xhariep, Lejweleputswa, Thabo Mofutsanyane, Mangaung, Sedibeng, West Rand, iLembe and eThekwini districts were found to be significantly associated with self-reported health. The Xhariep, Lejweleputswa, Thabo Mofutsanyane, Mangaung, Sedibeng and West Rand districts significantly recorded higher odds of reporting poor health. The iLembe and eThekwini districts significantly reduced the odds of reporting poor health. Furthermore, it can be observed that the districts within central and northern regions increased the odds of reporting poor health, while the districts within the Southern Western regions reduced the odds of reporting poor health.

Figure 5.2 display the predicted structured (a) and unstructured (b) residual spatial effects and their corresponding 95% credible intervals (CIs) based on Model 04. The numbers on top of each plot correspond to the districts names in Table C.1. Each credible interval has a length inversely related to the number of the odds ratios of reporting poor health and it can be used to test whether the spatial effects are significantly different from one. The plot yielded different results from the plots in Figure 4.2, the CIs are much wider for the unstructured spatial effects. Also, it can be seen that there were not significantly associated with self-reported health (Figure 5.2(b)). However, the structured spatial effects in Figure 5.2(a) were found to be significantly associated with self-reported health. In the Xhariep, Lejweleputswa, Thabo Mofutsanyane, Fezile Dabi, Mangaung, Sedibeng and West Rand districts there is a high prevalence of poor health, while in the iLembe, Sisonke and eThekwini districts there is low poor health prevalence.



(a) Structured

(b) Unstructured

Figure 5.2: Predicted posterior odds ratios of the residual spatially structured (a) and unstructured (b) effects with 95% credible intervals based on Model 04

5.4 Structured Additive Regression Models

The structured additive regression (STAR) models provide a unified framework for extending classical models to a more flexible approach. This approach allows for the inclusion of the different type of covariates such as the spatial random effects and nonlinear effects in the linear predictor. The linear predictor in Equation (5.2) only allows the effects of the covariates to be modeled linearly. However, to overcome such constraints, the linear predictor is replaced with an additive linear predictor. Due to Rue et al. (2009) approach, in this section, we replace the formal predictor in Equation (5.2) with a more flexible additive predictor to extend the previous models by accounting for smooth functions of continuous covariates. Thus, these models lead to STAR models. The structured additive predictor is defined as

$$\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \sum_{l=1}^{q} f_l(\boldsymbol{z}_{ijl}) + f_{spat}(s_i), \quad i = 1, \ldots, 52, \tag{5.7}$$

where f_l are nonlinear smooth functions of the continuous covariates $\boldsymbol{z}_{ijl}$ and f_{spat} are functions that caters for the spatial effects of each district or location.

5.4.1 Models specification

The following set of models were examined in order to investigate the linear, spatial and nonlinear effects of generic covariates on self-reported health. The first model is a standard logistic regression model which incorporates fixed effects of categorical covariates and assumes nonlinear effects for age and BMI. This model is given by

$$\text{Model A01} : \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(age_{ij1}) + f_2(bmi_{ij2}).$$

In this model, age and BMI are continuous covariates of an individual. The second model is given by

$$\text{Model A02}: \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(age_{ij1}) + f_2(bmi_{ij2}) + f_{str}(s_i),$$

this model is similar to Model A01, it accounts for fixed effects of categorical covariates, and assumes nonlinear effects of age and BMI, and caters for spatially structured random effects that account for unobserved covariates across the districts or spatial location in general. The third model is similar to Model A02 but instead, this model accounts for spatial unstructured random effects. The model is given by

$$\text{Model A03}: \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(age_{ij1}) + f_2(bmi_{ij2}) + f_{unstr}(s_i),$$

the unstructured heterogeneity caters for unobserved influential covariates that are inherent within the districts. The final model is a structured additive model which incorporates both the spatially structured and spatially unstructured random effects which capture spatial heterogeneity for unobserved influential factors and also accounts for nonlinear effects of age and BMI and the effects of categorical covariates. This model is given by

$$\text{Model A04}: \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + f_1(age_{ij1}) + f_2(bmi_{ij2}) + f_{str}(s_i) + f_{unstr}(s_i).$$

In all the models' formulation in this section, we assumed an independent diffuse prior for the fixed effects $\boldsymbol{\beta} \sim const$, the spatially structured $f_{str}(s_i)$ and spatially unstructured $f_{unstr}(s_i)$ were assumed to follow *intrinsic* conditional autoregressive (iCAR) (Equation (3.21)) and i.i.d Gaussian (Equation (3.24)) distributions respectively. We also assumed that f_1 and f_2 follow a second-order random walk discussed in Section 3.25.

5.4.2 Parameter Estimation

Similar to Section 5.3 the estimation of unknown parameters was carried out using a fully Bayesian (FB) procedure. The latent Gaussian variables of the proposed models is given by $\boldsymbol{\varrho} = \{\{\boldsymbol{\beta}\}, \{f_{str}(\cdot)\}, \{f_{unstr}(\cdot)\}, \{f_1(\cdot)\}, \{f_2(\cdot)\}\}$ with the corresponding hyperparameters $\boldsymbol{\psi} = \{\tau_{str}, \tau_{unstr}, \tau_1, \tau_2\}$. Hence, the posterior distribution is given by

$$p(\boldsymbol{\varrho}, \boldsymbol{\psi}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\varrho}, \boldsymbol{\psi})p(\boldsymbol{\varrho}, \boldsymbol{\psi}). \tag{5.8}$$

The prior specifications were discussed in the previous section and the hyperparameters were assigned conjugate gamma priors as previous, $\tau_1 \sim Gamma(1, 0.00005)$ and $\tau_2 \sim Gamma(1, 0.00005)$. The models were implemented in R using the R-INLA package and all the R codes are also presented in Appendix B.

5.4.3 Application of the spatial models with non-linear effects to NIDS wave 4 data

The posterior odds ratios (PORs) estimates and the corresponding 95% credible intervals (CIs) for all considered models are provided in Table 5.2. The table also provides the results of the model fit statistics. The best fitting model was selected upon the smallest deviance information criterion (DIC). The results for the model fit reveals that the DIC for Model A02 and Model A04 are the smallest and generally the same except for a slight difference. Thus, the results are further interpreted based on Model 04 which incorporates both the spatially structured and unstructured random effects. The results reveal that gender was significantly associated with self-reported health. The odds of reporting poor health among individuals who are male was 0.69 times the odds of reporting poor health for individuals who are female (POR: 0.69, 95% CI: 0.57 to 0.84).

Table 5.2: Parameter estimates of multivariable Bayesian spatial logistic regression models with nonlinear effects.

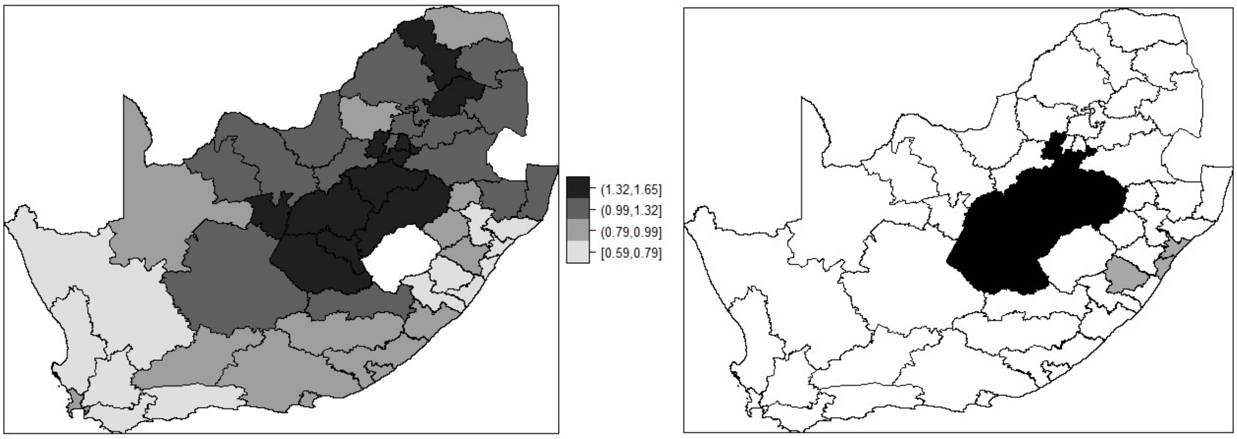
Covariates	**Model A01** **POR (95% CI)**	**Model A02** **POR (95% CI)**	**Model A03** **POR (95% CI)**	**Model A04** **POR (95% CI)**
Gender (ref = Female)				
Male	0.66 (0.55, 0.80)	0.69 (0.57, 0.84)	0.69 (0.57, 0.83)	0.69 (0.57, 0.84)
Race (ref = African)				
Asian/Indian	1.28 (0.61, 2.45)	2.02 (0.94, 4.01)	1.90 (0.88, 3.79)	2.02 (0.94, 4.01)
Coloured	0.53 (0.41, 0.69)	0.71 (0.51, 0.98)	0.63 (0.46, 0.85)	0.71 (0.51, 0.98)
White	1.31 (0.73, 2.24)	1.66 (0.91, 2.88)	1.57 (0.86, 2.72)	1.66 (0.91, 2.88)
Place of residence (ref = Urban informal)				
Rural Formal	0.79 (0.56, 1.11)	0.74 (0.52, 1.06)	0.77 (0.53, 1.10)	0.74 (0.52, 1.06)
Urban Formal	1.10 (0.84, 1.46)	0.95 (0.72, 1.28)	0.99 (0.74, 1.33)	0.95 (0.72, 1.28)
Tribal Authority Areas	0.86 (0.64, 1.16)	0.83 (0.60, 1.15)	0.84 (0.60, 1.16)	0.83 (0.60, 1.15)
Education level (ref = No formal education)				
Primary	0.94 (0.61, 1.43)	0.95 (0.62, 1.45)	0.93 (0.61, 1.43)	0.95 (0.62, 1.45)
Secondary	0.68 (0.47, 0.98)	0.67 (0.46, 0.97)	0.67 (0.46, 0.98)	0.67 (0.46, 0.97)
High	0.46 (0.33, 0.66)	0.45 (0.31, 0.64)	0.45 (0.32, 0.64)	0.45 (0.31, 0.64)
College	0.48 (0.30, 0.77)	0.45 (0.27, 0.72)	0.45 (0.28, 0.73)	0.45 (0.27, 0.72)
Tertiary	0.45 (0.30, 0.67)	0.43 (0.28, 0.65)	0.43 (0.29, 0.66)	0.43 (0.28, 0.65)
Household income (ref = Much below average)				
Below average	1.10 (0.89, 1.36)	1.09 (0.88, 1.35)	1.08 (0.87, 1.35)	1.09 (0.88, 1.35)
Average	1.10 (0.89, 1.35)	1.10 (0.89, 1.36)	1.10 (0.89, 1.36)	1.10 (0.89, 1.36)
Above average	0.90 (0.63, 1.25)	0.87 (0.61, 1.22)	0.88 (0.62, 1.23)	0.87 (0.61, 1.22)
Much above average	0.37 (0.19, 0.64)	0.35 (0.19, 0.62)	0.36 (0.19, 0.64)	0.35 (0.19, 0.62)
Marital status (ref = Not married)				
Widow/Divorced/Seperated	1.02 (0.74, 1.41)	0.96 (0.69, 1.32)	0.97 (0.69, 1.33)	0.96 (0.69, 1.32)
Married/living with partner	1.09 (0.91, 1.30)	1.02 (0.85, 1.23)	1.03 (0.86, 1.24)	1.02 (0.85, 1.23)
Life satisfaction level (ref = Very dissatisfied)				
Dissatisfied	0.73 (0.57, 0.95)	0.74 (0.57, 0.96)	0.74 (0.57, 0.96)	0.74 (0.57, 0.96)
Normal	0.94 (0.73, 1.21)	0.93 (0.72, 1.20)	0.93 (0.72, 1.21)	0.93 (0.72, 1.20)
Satisfied	0.77 (0.58, 1.01)	0.76 (0.57, 1.01)	0.76 (0.57, 1.01)	0.76 (0.57, 1.01)
Very satisfied	0.84 (0.60, 1.16)	0.82 (0.59, 1.15)	0.83 (0.59, 1.15)	0.82 (0.59, 1.15)
Exercise (ref = Never)				
Less than once a week	0.66 (0.47, 0.90)	0.62 (0.44, 0.85)	0.63 (0.45, 0.87)	0.62 (0.44, 0.85)
Once a week	1.11 (0.79, 1.54)	1.07 (0.76, 1.48)	1.08 (0.77, 1.50)	1.07 (0.76, 1.48)
Twice a week	0.86 (0.61, 1.19)	0.84 (0.59, 1.16)	0.85 (0.60, 1.18)	0.84 (0.59, 1.16)
Three or more times a week	0.77 (0.59, 0.98)	0.74 (0.57, 0.95)	0.74 (0.57, 0.95)	0.74 (0.57, 0.95)
Alcohol consumption level (ref = Never drunk alcohol)				
No longer drink	1.68 (1.35, 2.09)	1.60 (1.28, 1.99)	1.61 (1.29, 2.00)	1.60 (1.28, 2.00)
Drink very rarely	1.26 (1.01, 1.56)	1.15 (0.92, 1.43)	1.17 (0.93, 1.45)	1.15 (0.92, 1.43)
Less than once a week	0.82 (0.51, 1.28)	0.75 (0.46, 1.17)	0.77 (0.47, 1.20)	0.75 (0.46, 1.17)
On 1 or 2 days a week	1.37 (1.01, 1.84)	1.30 (0.96, 1.75)	1.31 (0.97, 1.76)	1.30 (0.96, 1.75)
On 3 or 4 days a week	1.36 (0.81, 2.22)	1.27 (0.75, 2.07)	1.30 (0.77, 2.13)	1.27 (0.75, 2.07)
On 5 or 6 days a week	0.65 (0.20, 1.73)	0.59 (0.18, 1.60)	0.59 (0.18, 1.61)	0.59 (0.18, 1.60)
Every day	1.69 (0.66, 3.87)	1.48 (0.58, 3.41)	1.52 (0.59, 3.50)	1.48 (0.58, 3.41)

Table 5.2 *Continues*

Covariates	**Model A01** **POR (95% CI)**	**Model A02** **POR (95% CI)**	**Model A03** **POR (95% CI)**	**Model A04** **POR (95% CI)**
Smokes (ref = No)				
Yes	1.16 (0.99, 1.45)	1.19 (1.00, 1.49)	1.18 (1.00, 1.48)	1.19 (1.01, 1.49)
Type of toilet (ref = None)				
Flush toilet with offsite disposal	0.95 (0.63, 1.45)	0.87 (0.57, 1.33)	0.88 (0.58, 1.36)	0.87 (0.57, 1.33)
Flush toilet with onsite disposal	0.86 (0.58, 1.32)	0.81 (0.53, 1.25)	0.81 (0.54, 1.25)	0.81 (0.53, 1.25)
Bucket toilet	0.86 (0.49, 1.50)	0.78 (0.44, 1.38)	0.78 (0.44, 1.36)	0.78 (0.44, 1.38)
Chemical toilet	1.24 (0.70, 2.18)	1.22 (0.68, 2.17)	1.20 (0.67, 2.13)	1.22 (0.68, 2.17)
Pit latrine with ventilation pipe	0.99 (0.66, 1.50)	1.03 (0.69, 1.58)	1.01 (0.68, 1.55)	1.03 (0.69, 1.58)
Pit latrine without ventilation pipe	0.81 (0.55, 1.22)	0.78 (0.52, 1.18)	0.79 (0.53, 1.20)	0.78 (0.52, 1.18)
Other	2.05 (0.48, 6.98)	1.80 (0.41, 6.21)	1.91 (0.44, 6.54)	1.80 (0.41, 6.21)
Employment status (ref = Unemployed strict)				
Unemployed Discouraged	0.51 (0.20, 1.13)	0.55 (0.21, 1.22)	0.56 (0.22, 1.24)	0.55 (0.21, 1.22)
Not Economically Active	1.46 (1.16, 1.84)	1.49 (1.19, 1.88)	1.50 (1.19, 1.89)	1.49 (1.19, 1.88)
Employed	0.92 (0.74, 1.16)	0.95 (0.75, 1.19)	0.95 (0.75, 1.19)	0.95 (0.75, 1.19)
Was diagnosed with TB? (ref = No)				
Yes	3.06 (2.40, 3.88)	3.20 (2.50, 4.08)	3.12 (2.44, 3.98)	3.20 (2.50, 4.08)
Felt depressed in past week? (ref = Less than 1 day)				
Little of the time (1-2 days)	1.50 (1.27, 1.77)	1.48 (1.25, 1.74)	1.48 (1.25, 1.75)	1.48 (1.25, 1.74)
Moderate amount of time (3-4 days)	2.20 (1.77, 2.71)	2.18 (1.76, 2.70)	2.18 (1.75, 2.69)	2.18 (1.76, 2.70)
All the time (5-7 days)	3.04 (2.15, 4.23)	3.06 (2.16, 4.27)	3.05 (2.15, 4.26)	3.06 (2.16, 4.27)
	Est. (95 % CI)	**Est. (95 % CI)**	**Est. (95 % CI)**	**Est. (95 % CI)**
Additional model parameters				
Spatially structured variation (σ^2_{str})	-	0.18 (0.07, 0.52)	-	0.18 (0.07, 0.52)
Spatially unstructured variation (σ^2_{unstr})	-	-	0.09 (0.04, 0.22)	0.06 (0.02, 0.07)
Age effect (σ^2_{age})	0.04 (0.01, 0.40)	0.05 (0.02, 0.40)	0.05 (0.02, 0.38)	0.05 (0.02, 0.39)
BMI effect (σ^2_{BMI})	0.03 (0.01, 0.23)	0.03 (0.01, 0.19)	0.03 (0.01, 0.21)	0.03 (0.01, 0.19)
Model fit				
DIC	5990.31	5943.94	5947.20	5943.92
$\overline{D}$	5935.21	5866.65	5863.53	5866.64
p_D	55.10	77.30	83.67	77.28

Race was significantly associated with self-reported health. The odds of reporting health for individuals who are coloured were 0.71 times the odds of reporting poor health for individuals who are African (POR: 0.71, 95% CI: 0.51 to 0.98). This justifies that the prevalence of poor health is less for coloured as compared to other race groups. Education was also found to be significantly associated with self-reported health. The odds of reporting poor health among individuals with secondary education were 0.67 (with 95% CI: 0.46 to 0.97) times the odds of reporting poor health for individuals with no formal education. The odds of reporting poor health among individuals with high education were 0.45 times the odds of reporting poor health for individuals with no formal education (POR: 0.45, 95% CI: 0.31 to 0.64). The odds of reporting poor health among individuals with college and tertiary education were respectively 0.45 (with 95% CI: 0.27 to 0.72) and 0.43 (with 95% CI: 0.28 to 0.65) times the odds of reporting poor health for individuals with no formal education. Household income was found to be significantly associated with self-reported health. The odds of reporting poor health for individuals with much above average household income were 0.35 times the odds of reporting poor health for individuals with much below average household income (POR: 0.35, 95% CI: 0.19 to 0.62). Life satisfaction was also found to be significantly associated with self-reported health. The odds of reporting poor health among individuals with dissatisfied life were 0.74 times the odds of reporting poor health for individuals with very dissatisfied life (POR: 0.74, 95% CI: 0.57 to 0.96). The odds of reporting poor health for individuals who exercise less than once a week and three or more times a week were respectively 0.62 and 0.74 times the odds of reporting poor health for individuals who never exercise, their corresponding odds ratios were (POR: 0.62, 95% CI: 0.44 to 0.85) and (POR: 0.74, 95% CI: 0.57 to 0.95) respectively. Alcohol and smoking were found to be significantly associated with self-reported health. The odds of reporting poor health for individuals who no longer drinks alcohol were 1.60 times the odds of reporting poor health for individuals who have never drunk alcohol (POR: 1.60, 95% CI: 1.28 to 2.00). The odds of reporting poor health among individuals who smoke a cigarette was 1.19 times the odds of reporting poor health for individuals who do not smoke a cigarette (POR: 1.19, 95% CI: 1.01 to 1.49). Employment status was found to be significantly associated with self-reported health. The odds of reporting poor health for individuals who are not economically active were 1.49 times the odds of reporting poor health for individuals who are unemployed strict (POR: 1.49, 95% CI: 1.19 to 1.88). Being diagnosed with TB previously was found to be significantly associated with self-reported health. The odds of reporting poor health among individuals who were diagnosed with TB was 3.20 times the odds of reporting poor health for individuals who were not diagnosed with TB (POR: 3.20, 95% CI: 2.50 to 4.08). Depression was also found to be significantly associated with self-reported health. The odds of reporting poor health among individuals who felt depressed for little of the time, a moderate amount of the time and all the time were respectively 1.48 (with 95% CI: 1.25 to 1.74), 2.18 (with 95% CI: 1.76 to 2.70) and 3.06 (with 95% CI: 2.16 to 4.27) times the odds of reporting poor health for individuals who felt depressed in the past week for less than one day. This means the prevalence of poor health is high the higher the depression level.

Figure 5.3 display the total residual spatial effects based on the best fitting model (Model $A04$). Also, shown is the 95% posterior probability map of significance. A similar description of district colors as in Figure 5.1 are used for the maps. There is clear evidence of poor health spatial variations at the district level of South Africa. Figure 5.3(a) reveals that the prevalence of poor health in the districts within central and northern regions of South Africa still remained high. The prevalence of poor health in the districts within western and southern regions also remained low. Figure 5.3(b) depict that the Lejweleputswa, Xhariep, Thabo Mafutsanyane, Fezile Dabi Mangaung, Sedibeng and West Rand districts were found to be significant. Lejweleputswa, Xhariep, Thabo Ma-



(a)

(b)

Figure 5.3: Map of South Africa showing residual total spatial district residual effects estimates (a) and the 95% corresponding map of significance (b) of spatial effect estimates of Model $A04$.

futsanyane, Fezile Dabi Mangaung, Sedibeng and West Rand districts increased the odds of reporting poor health, while the Sisonke, iLembe and eThekwini districts significantly reduced the odds of reporting poor health. Furthermore, it can be observed that there is a high prevalence of poor health in north-eastern regions while less prevalence is recorded in eastern regions.

Figure 5.4 shows the posterior mean estimates and their corresponding 95% CI of self-reported health against age (left) and body mass index (BMI) (right). We assumed a nonlinear relationship between age, BMI and self-reported health. The plots yielded similar shapes as in Figure 4.4, except that the BMI plot is of U shape form more exact. There is a positive linear association between poor health prevalence and individuals age. The odds of reporting poor health increases with age. This is in line with the categorical age. The prevalence of poor health is decreasing to a minimum at BMI between 20-25 then starts increasing again. This is the similar effect as categorical nutrition status.

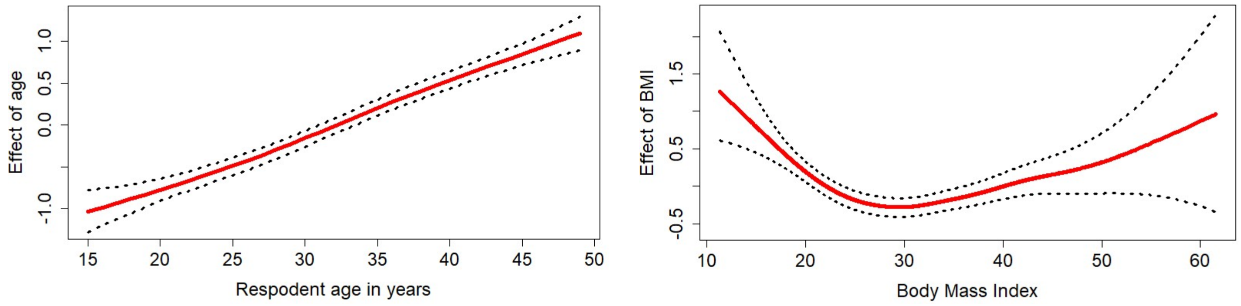


Figure 5.4: Estimated mean (red) of the non-linear effects of individual's age (left) and body mass index (BMI) (right) with 95% credible interval (dotted black lines) of Model $A04$.

The logistic regression model is often used when the response is binary. The assumption under this traditional model is that the predictor is strictly linear. However, the STAR model allows for generic covariates to be added in the predictor in an additive manner. The spatial effects account for unobserved influential factors. The convolution model also showed better fitting models. There was a slight difference between the strictly spatial models and the STAR models. The inclusion of continuous covariates further improved the results. There is evidence of spatial variation of self-reported health in South Africa.

Chapter 6

Spatio-Temporal Modeling of self-reported health in South Africa

A great deal of spatially referenced health datasets is collected over time, leading to an extension of the models from the previous chapters (Chapter 4 and 5) to spatio-temporal modeling.

In this chapter, we introduce models that are an extension to spatial modeling in disease mapping. These models are widely used for modeling disease risk in space and time. We review spatio-temporal models which allow us to understand the change in spatial variation over a time period. Hence, longitudinal data from observations repeatedly made over spatial location and time make it possible to fit such models. In particular, we discuss spatio-temporal modeling of self-reported health among individuals between 15-49 years in South Africa. We fit these models to all the four waves of the NIDS datasets under the Bayesian framework. The Bayesian cumulative logit and logistic regression models for spatio-temporal modeling are discussed and implemented to understand the spatial and temporal variations of poor health prevalence. Basically, we focus on spatial wave-specific modeling under space-time methodology. Furthermore, we investigate several covariate effects on self-reported health.

6.1 Introduction

A frequently used approach to spatial modeling of disease mapping in small areas dates way back to work by Besag et al. (1991) which was extended by Bernardinelli et al. (1995) to include a linear term for space-time interaction, and Knorr-Held (2000) who included a non-parametric spatio-temporal time trend. Waller et al. (1997) proposed nested models, where the spatial main effect is time dependent. However, such models are not of interest in this research. Spatio-temporal models are widely employed in many scientific fields, including disease surveillance studies. The Bayesian hierarchical modeling framework has made it possible to implement these models. These models provide a complex and flexible framework in space and time models, with spatio-temporal interaction being the most important feature. As applied in this research we first review methods based on the Bernardinelli et al. (1995) and Knorr-Held (2000) spatio-temporal framework.

6.2 Review of methods in spatio-temporal modeling

The development of spatio-temporal modeling methods has received robust attention in the field of epidemiology and biostatistics. The aim of these models was to extend the spatial model to consider the temporal dimension. Here we briefly only review two methods which form part of the models proposed in this research.

First, we briefly highlight the approach by Bernardinelli et al. (1995) who proposed a parametric trend space-time model assuming a Poisson distribution. In their approach they defined the log relative risk for area i; $i = 1, \dots, I$ during time t; $t = 1, \dots, T$ to be

$$\log(\theta_{it}) = \eta_{it} = \mu + u_i + v_i + (\beta + \delta_i) \times t, \tag{6.1}$$

where μ is the overall rate intercept, $\phi = u_i + v_i$ are the spatial random effects following the Besag et al. (1991) specifications, β is the global time linear trend effect, and δ_i is an interaction random effect between space and time. The time trend parameters were assigned vague priors to allow the data to reveal the time trend. However, other alternative priors may be used such as autoregressive priors of order 1 denoted as AR(1) (Waller et al., 1997). The i.i.d Gaussian prior was assumed for the interaction random effect δ_i, however other prior specification can be assigned. The specification of priors for the structured and unstructured spatial effects was handled as in the strictly spatial models.

A second approach was developed by Knorr-Held (2000) who modified the previous approach by overcoming the parametric limitation. In this approach, a Binomial distribution was assumed for the number of cases in county i ($i = 1, \ldots, I$) during time t ($t = 1, \ldots, T$) and the log odds was defined as

$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \eta_{it} = \mu + u_i + v_i + \gamma_t + \nu_t + \delta_{it} \tag{6.2}$$

where γ_t and ν_t are temporal random effects which cater for unspecified features of year t, and δ_{it} are interaction effects which capture variation that is not accounted for by the main effects. The u_i and γ_t were assigned *intrinsic* conditional autoregressive (iCAR) and first-order random walk structure, while v_i and ν_t were respectively assigned independent Gaussian priors. The interaction δ_{it} was assumed to have four types of prior implication depending on the spatial effects and temporal effects interaction. For details on these prior type interactions one can refer to Knorr-Held (2000); Blangiardo and Cameletti (2015) among others.

For the two above approaches due to Bernardinelli et al. (1995) and Knorr-Held (2000), parameter estimation was performed under the fully Bayesian (FB) approach with the use of Markov chain Monte Carlo (MCMC) via Gibbs sampling techniques. In our approach, we adopt a similar approach but estimation is carried out using FB and empirical Bayes (EB) for binary and ordinal response outcome respectively. Next, we discuss the methods used in this research.

6.3 Spatio-temporal models in disease mapping

Let y_{ijw} be self-reported health status of individual j in district i: $i = 1, \ldots, 52$ during wave w: $w = 1, 2, 3, 4$. The response outcome variable in this chapter was defined in two natures; the ordinal and the binary response. The ordinal response variable was defined as follows

$$y_{ijw1} = \begin{cases} 1: & \text{excellent} \\ 2: & \text{very good} \\ 3: & \text{good} \\ 4: & \text{fair} \\ 5: & \text{poor} \end{cases}$$

and the binary response version as follows,

$$y_{ijw2} = \begin{cases} 1: & \text{poor} \\ 0: & \text{good} \end{cases}$$

where y_{ijw1} is an ordinal response outcome and y_{ijw2} is a binary response outcome. As mentioned in previous chapters that the commonly used models for ordinal and binary outcomes are the cumulative logit and logistic regression models respectively. Hence we assumed that $y_{ijw1} \sim Multinomial(m_{ijw}, \pi_{ijw})$ and $y_{ijw2} \sim Bernoulli(\pi_{ijw})$, where π_{ijw} are unknown probabilities related to the event probabilities of the models. It is worth noting that both the response outcomes belong to the exponential family of distributions under multivariate and univariate GLMs respectively. The cumulative logit model in this section is denoted by

$$\text{logit}\left[P(y_{ijw} \leq r)\right] = \theta_r - \eta_{ijw}, \quad r = 1, \ldots, k-1, \tag{6.3}$$

while the logistic regression model is given by

$$\text{logit}(\pi_{ijw}) = \beta_0 + \eta_{ijw}. \tag{6.4}$$

The $\eta_{ijw} = \mathbf{x}'_{ijw}\boldsymbol{\beta}$ is the predictor with covariate vector $\mathbf{x} = (x_{ijw1}, \ldots, x_{ijwp})'$, $\boldsymbol{\beta}$ is the vector of regression coefficient given by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, θ_r are cutpoints, and β_0 is the model intercept under the logit model. To allow flexibility we adopt the unified framework of the structured additive regression (STAR) models where the classical predictor can be extended to a more flexible additive predictor. Hence the structured additive predictor can be extended for spatio-temporal modeling as

$$\eta_{ijw} = \mathbf{x}'_{ijw}\boldsymbol{\beta} + f_{spat}(s_i) + f_{wave}(w) + f_{iw}(s_i, w) \tag{6.5}$$

where the functions f_{spat}, f_{wave}, and f_{iw} represent functions appropriate for space, wave and space-wave interaction respectively. One should note that the spatial and temporal terms are independent. The spatial components f_{spat} are decomposed into spatially structured f_{str} and unstructured f_{unstr} effects. Moreover, f_{wave} represent random wave effects which can be modeled as a first-order random walk or AR(1), and $f_{iw}(s_i, w)$ is a space-wave interaction (DiMaggio, 2012).

6.4 Spatio-temporal models

In this section, we propose a series of several models for Bayesian cumulative logit and logistic regression models under spatio-temporal modeling that are considered in this research. The set of models that were employed are:

Model 1 : $\eta_{ijw} = \mathbf{x}'_{ijw}\boldsymbol{\beta} + f_{str}(s_i) + f_{unstr}(s_i)$

Model 2 : $\eta_{ijw} = \mathbf{x}'_{ijw}\boldsymbol{\beta} + f_{str}(s_i) + f_{unstr}(s_i) + \beta_w$

Model 3 : $\eta_{ijw} = \mathbf{x}'_{ijw}\boldsymbol{\beta} + f_{str}(s_i) + f_{unstr}(s_i) + f_{wave}(w)$

$$\text{Model 4}: \eta_{ijw} = \mathbf{x}'_{ijw}\boldsymbol{\beta} + f_{str}(s_i) + f_{unstr}(s_i) + f_{iw}(s_i, w)$$
$$\text{Model 5}: \eta_{ijw} = \mathbf{x}'_{ijw}\boldsymbol{\beta} + f_{str}(s_i) + f_{unstr}(s_i) + \beta_w + f_{iw}(s_i, w)$$
$$\text{Model 6}: \eta_{ijw} = \mathbf{x}'_{ijw}\boldsymbol{\beta} + f_{str}(s_i) + f_{unstr}(s_i) + f_{wave}(w) + f_{iw}(s_i, w)$$
$$\text{Model 7}: \eta_{ijw} = \mathbf{x}'_{ijw}\boldsymbol{\beta} + f_{str}(s_i) + f_{unstr}(s_i) + f_{1wave}(w) + f_{iw}(s_i, w),$$

where in all formulations:

- $\mathbf{x}_{ijw}$ denotes the vector of categorical covariates effects for individual j living in district i during wave w.
- $\boldsymbol{\beta}$ represent a vector of regression coefficients.
- β_w denote the wave-specific fixed effects.
- $f_{str}(s_i)$ and $f_{unstr}(s_i)$ represent the structured and unstructured random effects respectively.
- f_{wave}, f_{1wave} are smooth functions of the temporal random effects.
- $f_{iw}(s_i, w)$ is the spatial-wave interaction effect.

Model 1 only accounts for the spatially structured random effects which account for unobserved influential factors that vary spatially across the districts and spatially unstructured random effects which capture unobserved covariates within districts, and further assume categorical covariates to have a linear effect on self-reported health. This model does not assume any temporal effect. Model 2 is similar to Model 1 but also assumes a linear wave trend captured by β_w. In contrast, Model 3 involves separable space and wave random effects and accounts for the linear effect of categorical covariates. Model 4 is similar to Model 1 but also accounts for space and wave interaction which captures variation that cannot be revealed by the main effects. For Model 5, we assumed linear effects of categorical covariates, spatial random effects, linear wave trend and we assumed space-wave interaction. Model 6 and Model 7 are basically the same but differs in the prior assumptions of the temporal random wave effects $f_{wave}(w)$ and $f_{1wave}(w)$. Furthermore, these two models assume linear effects of categorical covariates, spatial random effects of the locations, and space and wave interaction. In effect, all models assume linear effects of categorical covariates via the term $\mathbf{x}'_{ijw}\boldsymbol{\beta}$.

6.4.1 Prior specifications

In this chapter, two Bayesian approaches were used for the proposed models. The empirical Bayes (EB) via generalized linear mixed model (GLMM) methodology was used for the spatio-temporal cumulative logit models and fully Bayesian approach was used for the spatio-temporal logistic regression models. The fixed effects and linear wave trend were assigned diffuse priors, the spatially structured random effects were modeled with Markov random fields (MRFs) or *intrinsic* conditional autoregressive (iCAR) while the spatially unstructured random effects were assigned i.i.d Gaussian prior. The temporal wave random effects f_{1wave} were modeled by a first-order random walk defined in Section 3.25. However, it is worth noting that different prior specifications for the temporally varying wave random effects f_{wave} were assigned in the models. In the spatio-temporal cumulative logit model, we assumed a first-order autoregressive for f_{wave} and we assigned penalized splines defined in Section 3.8 for the spatio-temporal logistic regression model. Furthermore, the spatial wave-specific effects (interaction) were modeled by independent penalized splines for the cumulative logit model and for the logistic model independent first-order autoregressive model was assumed.

6.4.2 Estimation of Parameters

This research employs two different approaches, we first discuss the procedure of parameter estimation of the spatio-temporal cumulative models. The parameter estimation of the spatio-temporal cumulative models was carried out using the EB approach via GLMM methodology, hence we denote the matrix notation of the parameters to be estimated as

$$\boldsymbol{\eta} = \mathbf{U}\boldsymbol{\gamma} + \mathbf{X}_{str}\boldsymbol{\beta}_{j,str} + \mathbf{X}_{unstr}\boldsymbol{\beta}_{j,unstr} + \mathbf{X}_{wave}\boldsymbol{\beta}_{j,wave} + \mathbf{X}_{iw}\boldsymbol{\beta}_{j,iw} \tag{6.6}$$

where $\boldsymbol{\gamma} = (\theta_1, \ldots, \theta_k, \boldsymbol{\beta}', \beta_w)'$ is the overall vector of regression coefficients, $\mathbf{U}$ is the corresponding design matrix constructed from the covariates $\mathbf{x}'_{ijw}$, and $\mathbf{X}_{str}, \mathbf{X}_{unstr}, \mathbf{X}_{wave}, \mathbf{X}_{iw}$ are appropriate matrices for each spatial, temporal and interaction effect respectively. The elements $\mathbf{X}_{str}, \mathbf{X}_{unstr}, \mathbf{X}_{wave}, \mathbf{X}_{iw}$ and $\boldsymbol{\beta}_{j,str}, \boldsymbol{\beta}_{j,unstr}, \boldsymbol{\beta}_{j,wave}$ are such that $\boldsymbol{f}_j = \mathbf{X}_j\boldsymbol{\beta}_j$. Hence, this equation is of the form of Equation (3.33) and the estimation was carried out using BayesX with the combination to R-software under *R2BayesX* package. The R codes are presented in Appendix A.

We now turn to the estimation procedure for the spatio-temporal logistic regression model. The parameters estimation was done using a fully Bayesian approach. Hence all the unknown parameters are considered to be random variables and are assigned appropriate prior distributions. In the previous section, we discussed such priors and the posterior distribution is given as

$$p(\boldsymbol{\varrho}, \boldsymbol{\psi}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\varrho}, \boldsymbol{\psi})p(\boldsymbol{\varrho}, \boldsymbol{\psi}), \tag{6.7}$$

where $L(\mathbf{y}|\boldsymbol{\varrho}, \boldsymbol{\psi})$ is the likelihood and $p(\boldsymbol{\varrho}, \boldsymbol{\psi})$ are prior distributions of the model. The latent Gaussian field is denoted by $\boldsymbol{\varrho} = \{\{\boldsymbol{\beta}\}, \{\beta_w\}, \{f_{str}(\cdot)\}, \{f_{unstr}(\cdot)\}, \{f_{wave}(\cdot)\}, \{f_{1wave}(\cdot)\}, \{f_{iw}(\cdot)\}\}$ and the corresponding hyperparameters are given by $\boldsymbol{\psi} = \{\tau_{str}, \tau_{unstr}, \tau_{wave}, \tau_{1wave}, \tau_{iw}\}$. All the hyperparameters were assigned conjugate gamma priors $Gamma(1, 0.00005)$. Estimation of parameters was done using R-integrated nested Laplace approximation (INLA) package. The codes are all presented in Appendix B.

6.5 Application to the South Africa wave 1-4 NIDS data.

All the models which were considered in Section 6.4 were then applied to the NIDS wave 1 to 4 datasets. Data description for the NIDS data was described in Section 2.2. The NIDS dataset is a longitudinal data which comprises of categorical and continuous explanatory variables or factors. In addition, it is geo-referenced, thus making it possible to use for understanding spatial variation over time. Moreover, it should be noted that the sample size of the datasets varies across the waves, due to the number of continuing and temporally varying sample members. The sample sizes of each of the NIDS waves dataset is presented in Table F.1. Next, we discuss the results of the model comparison as it is of interest to have the best fitting model for the data under Bayesian analysis.

Table 6.1: Summary of the model fit criterion for model comparison, the AIC, the BIC and the GCV for all the fitted models.

	Models						
Fit criterion	**Model** 1	**Model** 2	**Model** 3	**Model** 4	**Model** 5	**Model** 6	**Model** 7
AIC	106838	106508	106508	105541	105536	105536	105536
BIC	107770	107466	107466	107660	107608	107607	107607
GCV	2.48573	2.47756	2.47756	2.43338	2.43411	2.43412	2.43413

Table 6.1 present the model fit values for the considered spatio-temporal cumulative logit models in Section 6.4. Shown in the table are the AIC, BIC and GCV values. The model with smaller AIC, BIC and GCV values is considered as the best fitting model. Comparing the results for each model, the AIC values for Model 5, Model 6 and Model 7 are the same (AIC = 105536). The BIC favors Model 2 and Model 3, with the same values (BIC = 107466). The GCV value for Model 5, Model 6 and Model 7 are generally the same but favors Model 6 with the smallest value. Therefore, Model 6 is the preferred model based on the majority vote. Thus in what follows, results are presented and interpreted based on Model 6.

Table 6.2: Summary of the model fit criterion for model comparison, the DIC, the mean of deviance and the number of effective parameters for all the fitted models.

	Models						
Fit criterion	**Model** 1	**Model** 2	**Model** 3	**Model** 4	**Model** 5	**Model** 6	**Model** 7
pD	89.14	90.98	90.82	193.81	171.56	171.83	166.63
$\bar{D}$	18519.48	18318.84	18319.01	18124.01	18129.85	18129.71	18128.69
DIC	18608.62	18409.83	18409.84	18317.83	18301.41	18301.55	18295.32

Table 6.2 provide the model fit values for the spatio-temporal logistic regression models considered in Section 6.4. The DIC was used to select the best fitting model, the smaller the DIC value the better the fit. The results reveal that Model 7 is the preferred model, with the DIC value given by (DIC = 18295.32). Thus, results based on Model 7 are presented and interpreted.

Table 6.3 gives posterior odds ratios (PORs) estimates and their corresponding 95% credible intervals (CIs) for the best fitting models mentioned above. The categorical covariates were assumed to have a linear effect on self-reported health. Results of the spatio-temporal cumulative logit model (Model 6) are shown in Table 6.3. Shown are the cumulative POR with their 95% CIs. All the covariates were found to be significantly associated with self-reported health except for race, marital status, and type of toilet facility. We discuss the results for significant covariates only. The results revealed that the odds of reporting poor health seem to increase with age. The odds of reporting poor health is highest for the ages 45-49 years (POR: 2.81, 95% CI: 2.58 to 3.07). The odds of reporting poor among individuals between the age 20-24 years were 1.09 times the odds of reporting poor health for individuals aged between 15-19 years (POR: 1.09, 95% CI: 1.03 to 1.16). The odds of reporting poor among individuals between the age 25-29 years were 1.20 times the odds of reporting poor health for individuals aged between 15-19 years (POR: 1.20, 95% CI: 1.12 to 1.28). The odds of reporting poor health among individuals between the ages 30-34, 35-39 and 40-44 years were respectively 1.38 (with 95% CI: 1.29 to 1.49), 1.68 (with 95% CI: 1.56 to 1.82) and 2.04 (with 95% CI: 1.88 to 2.22) times the odds of reporting poor health for individuals between 15-19 years.

Table 6.3: Posterior odds ratio estimates with corresponding 95% (CI) for the best fitting models.

	Cumulative logit model (Model 6)	**Logistic regression model (Model 7)**
Covariates	**POR (95% CI)**	**POR (95% CI)**
Age group (ref = 15-19)		
20-24	1.09 (1.03, 1.16)	1.36 (1.13, 1.64)
25-29	1.20 (1.12, 1.28)	1.79 (1.48, 2.16)
30-34	1.38 (1.29, 1.49)	3.05 (2.54, 3.66)
35-39	1.68 (1.56, 1.82)	4.36 (3.63, 5.24)
40-44	2.04 (1.88, 2.22)	5.17 (4.29, 6.23)
45-49	2.81 (2.58, 3.07)	7.99 (6.64, 9.62)
Gender (ref = Female)		
Male	0.77 (0.74, 0.81)	0.67 (0.60, 0.74)
Race (ref = African)		
Asian/Indian	1.08 (0.89, 1.31)	1.49 (0.98, 2.22)
Coloured	0.96 (0.88, 1.05)	0.88 (0.74, 1.05)
White	0.98 (0.85, 1.13)	0.95 (0.68, 1.32)
Place of residence (ref = Urban informal)		
Rural formal	0.89 (0.81, 0.98)	0.76 (0.62, 0.93)
Urban formal	0.98 (0.91, 1.07)	0.95 (0.80, 1.12)
Tribal authority Areas	1.02 (0.93, 1.11)	0.88 (0.73, 1.06)
Education level (ref = No education)		
Primary	1.03 (0.90, 1.18)	1.06 (0.87, 1.30)
Secondary	0.80 (0.71, 0.89)	0.89 (0.75, 1.06)
High	0.57 (0.52, 0.64)	0.58 (0.49, 0.69)
College	0.52 (0.45, 0.60)	0.53 (0.40, 0.71)
Tertiary	0.46 (0.40, 0.51)	0.42 (0.34, 0.52)
Household income (ref = Much below average)		
Below average	1.11 (1.05, 1.17)	1.10 (0.98, 1.23)
Average	1.22 (1.15, 1.28)	1.01 (0.90, 1.14)
Above average	1.03 (0.95, 1.12)	0.84 (0.69, 1.01)
Much above average	0.87 (0.78, 0.97)	0.50 (0.36, 0.67)
Marital status (ref = Not married)		
Widow/divorced/seperated	1.05 (0.94, 1.16)	0.97 (0.81, 1.16)
Married/living with partner	0.98 (0.94, 1.03)	0.96 (0.87, 1.06)
Life satisfaction level (ref = Very dissatisfied)		
Dissatisfied	0.86 (0.81, 0.91)	0.77 (0.69, 0.88)
Normal	0.77 (0.72, 0.81)	0.73 (0.64, 0.83)
Satisfied	0.68 (0.64, 0.73)	0.66 (0.56, 0.76)
Very satisfied	0.67 (0.62, 0.73)	0.78 (0.66, 0.93)
Exercise (ref = Never)		
Less than once a week	0.98 (0.92, 1.06)	0.83 (0.70, 0.99)
Once a week	0.95 (0.88, 1.04)	1.08 (0.89, 1.30)
Twice a week	0.94 (0.93, 1.08)	0.98 (0.82, 1.20)
Three or more times a week	0.91 (0.86, 0.97)	0.95 (0.82, 1.10)

Table 6.3 *Continues*

Covariates	**Cumulative logit model (Model 6)** **POR (95% CI)**	**Logistic regression model (Model 7)** **POR (95% CI)**
Alcohol consumption level (ref = Never drunk alcohol)		
No longer drink	1.28 (1.19, 1.36)	1.53 (1.34, 1.74)
Drink very rarely	1.06 (1.02, 1.12)	1.09 (0.96, 1.24)
Less than once a week	1.05 (0.95, 1.17)	1.01 (0.78, 1.26)
On 1 or 2 days a week	1.03 (0.95, 1.12)	0.97 (0.80, 1.16)
On 3 or 4 days a week	1.18 (1.01, 1.38)	1.18 (0.87, 1.58)
On 5 or 6 days a week	1.06 (0.82, 1.39)	0.92 (0.55, 1.50)
Every day	1.23 (0.92, 1.65)	1.01 (0.59, 1.67)
Smokes (ref = No)		
Yes	1.12 (1.06, 1.19)	1.16 (1.03, 1.32)
Type of toilet (ref = None)		
Flush toilet with offsite disposal	0.94 (0.85, 1.04)	1.08 (0.87, 1.33)
Flush toilet with onsite disposal	0.94 (0.85, 1.04)	0.94 (0.76, 1.16)
Bucket toilet	0.88 (0.77, 1.01)	0.92 (0.70, 1.22)
Chemical toilet	1.13 (0.99, 1.28)	0.93 (0.70, 1.22)
Pit latrine with ventilation pipe	0.91 (0.82, 1.01)	0.91 (0.74, 1.11)
Pit latrine without ventilation pipe	0.94 (0.86, 1.03)	1.06 (0.88, 1.28)
Other	1.15 (0.69, 1.92)	0.93 (0.30, 2.44)
Employment status (ref = Unemployed strict)		
Unemployed discouraged	0.90 (0.80, 1.01)	1.01 (0.79, 1.28)
Not economically active	1.08 (1.01, 1.13)	1.29 (1.14, 1.46)
Employed	0.95 (0.90, 1.01)	0.95 (0.84, 1.08)
Nutrition status (ref = Normal)		
Underweight	1.23 (1.13, 1.33)	1.49 (1.25, 1.76)
Overweight/obese	1.02 (0.98, 1.06)	0.90 (0.82, 1.01)
Severe	1.38 (1.22, 1.55)	1.95 (1.55, 2.43)
Was diagnosed with TB? (ref = No)		
Yes	2.64 (2.41, 2.89)	3.46 (3.04, 3.94)
Felt depressed in past week? (ref = Less than 1 day)		
Little of the time (1-2 days)	1.26 (1.21, 1.31)	1.45 (1.32, 1.59)
Moderate amount of time (3-4 days)	1.32 (1.24, 1.40)	1.99 (1.77, 2.24)
All the time (5-7 days)	1.90 (1.70, 2.11)	3.18 (2.68, 3.77)

The odds of reporting poor health for male individuals were 0.77 times the odds of reporting poor health for individuals who are female (POR: 0.77, 95% CI: 0.74 to 0.81). The odds of reporting poor health among individuals staying in rural formal areas were 0.89 times the odds of reporting poor health for individuals living in urban formal areas (POR: 0.89, 95% CI: 0.81 to 0.98). This means the prevalence of poor health is less in rural formal areas compared to urban formal areas. Education was also found to be significantly associated with self-reported health. The odds of reporting poor health seem to decrease as education level increases. The odds of reporting poor health among individuals with secondary education were 0.80 times the odds of reporting poor health for individuals with no education (POR: 0.80, 95% CI: 0.71 to 0.89). The odds of reporting poor health among individuals with high education were 0.57 times the odds of reporting poor health for individuals with no education (POR: 0.57, 95% CI: 0.52 to 0.64). The odds of reporting poor health among individuals with a college education were 0.52 times the odds of reporting poor health for individuals with no education (POR: 0.52, 95% CI: 0.45 to 0.60). Moreover, the odds of reporting poor health among individuals with tertiary education were 0.46 times the odds of reporting poor health for individuals with no education (POR: 0.46, 95% CI: 0.40 to 0.51). The odds of reporting poor health for individuals with below average income were 1.11 times the odds of reporting poor health for individuals with much below average income (POR: 1.11, 95% CI: 1.05 to 1.17). The odds of reporting poor health among individuals with average income were 1.22 times the odds of reporting poor health for individuals with much below average income (POR: 1.22, 95% CI: 1.15 to 1.28). Furthermore, the odds of reporting poor health among individuals with much above average income were 0.87 times the odds of reporting poor health for individuals with much below average income (POR: 0.87, 95% CI: 0.78 to 0.97). Life satisfaction level was found to be significantly associated with self-reported health. The odds of reporting poor health seem to decrease as life satisfaction level increases. The odds of reporting poor health among individuals with dissatisfied life were 0.86 times the odds of reporting poor health for individuals with very dissatisfied life (POR: 0.86, 95% CI: 0.81 to 0.91). The odds of reporting poor health for individuals with normal life were 0.77 times the odds of reporting poor health for individuals with very dissatisfied life (POR: 0.77, 95% CI: 0.72 to 0.81). Furthermore, the odds of reporting poor health for individuals with satisfied and very satisfied life were respectively 0.68 (with 95% CI: 0.64 to 0.73) and 0.67 (with 95% CI: 0.62 to 0.73) times the odds of reporting poor health for individuals with very dissatisfied life. The odds of reporting poor health among individuals who exercise three or more times a week were 0.91 times the odds of reporting poor health for individuals who never exercise (POR: 0.91, 95% CI: 0.82 to 0.97). Alcohol was also found to be significantly associated with self-reported health. The odds of reporting poor health among individuals who no longer drink alcohol were 1.28 times the odds of reporting poor health for individuals who never drunk alcohol (POR: 1.28, 95% CI: 1.19 to 1.36). The odds of reporting poor health for individuals who drink very rarely were 1.06 times the odds of reporting poor health for individuals who never drunk alcohol (POR: 1.06, 95% CI: 1.02 to 1.12). Moreover, The odds of reporting poor health for individuals who drink on 3 or 4 days a week were 1.18 times the odds of reporting poor health for individuals who never drunk alcohol (POR: 1.18, 95% CI: 1.01 to 1.38). The odds of reporting poor health for individuals who smoke a cigarette was 1.12 times the odds of reporting poor health for individuals who do not smoke a cigarette (POR: 1.12, 95% CI: 1.06 to 1.19). The odds of reporting poor health among individuals who are not economically active were 1.08 times the odds of reporting poor health for individuals who are unemployed strict (POR: 1.08, 95% CI: 1.01 to 1.13). Among individuals with underweight and severe nutrition status the odds of reporting poor health were respectively 1.23 and 1.38 times the odds of reporting poor health for individuals with normal nutrition, (POR: 1.23, 95% CI: 1.13 to 1.33) and (POR: 1.38, 95% CI: 1.22 to 1.55) respectively. The odds of reporting poor health for individuals who were diagnosed with TB was 2.64 times the odds of reporting poor health for an individual who was not diagnosed with TB (POR: 2.64, 95% CI: 2.41 to 2.89). Depression was also found to be significantly associated with self-reported health. The results reveal that the odds of reporting poor health increase with depression level. The odds of reporting poor health among individuals who felt depressed in the past week for a little of the time were 1.26 times the odds of reporting poor health for individuals for individuals who felt depressed for less than one day (POR: 1.26, 95% CI: 1.21 to 1.31). The odds of reporting poor health among individuals who felt depressed in the past week for a moderate amount of the time were 1.32 times the odds of reporting poor health for individuals for individuals who felt depressed for less than one day (POR: 1.32, 95% CI: 1.24 to 1.40). Furthermore, the odds of reporting poor health among individuals who felt depressed in the past week all the time were 1.90 times the odds of reporting poor health for individuals for individuals who felt depressed for less than one day (POR: 1.90, 95% CI: 1.70 to 2.11).

Figure 6.1 shows the total residual spatial effects over the four waves based on Model 6. From the figure, districts with dark and dark grey colour show a high association of poor health while grey and light grey colour indicates a low association of poor health. It can be observed that there is a spatial variation in poor health prevalence. Districts within the north-western, central and southern regions had high poor health prevalence. The Siyanda, John Taoli Gaetsewe, Dr. Ruth Segomotsi Mompati, Frances Baard, Lejweleputswa, Xhariep, Mangaung, Chris Hani, Umgungundlovu, Sisonke, Zululand, Umkhanyakude and Uthungulu districts recorded higher poor health prevalence. The districts in north-eastern, south-western and some central regions had low poor health prevalence. The Namakwa, Vhembe, Mopani, Ehlanzeni, Thabo Mofutsanyane, Uthukela

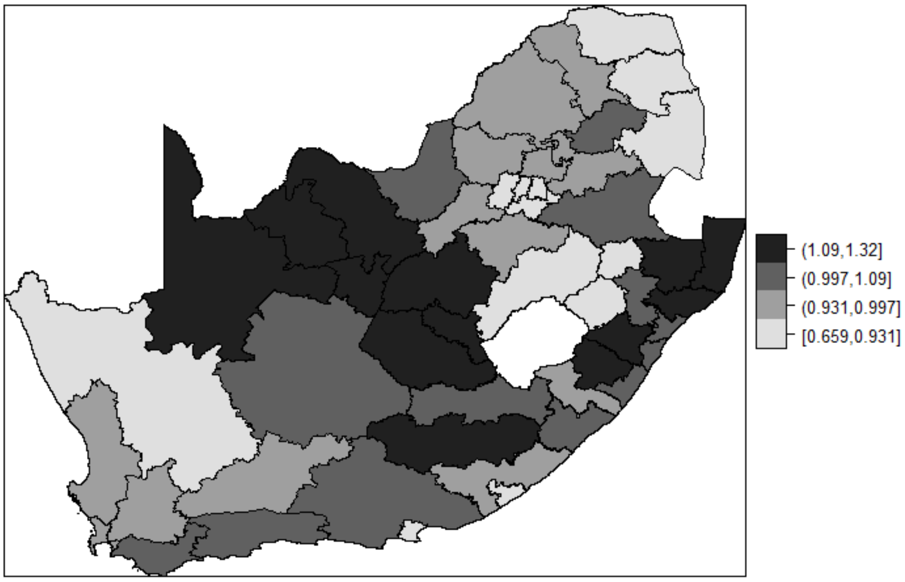


Figure 6.1: Map showing posterior odds estimated district-level residual total spatial effects of self-reported health in South Africa based on a cumulative logit model (Model 6).

and Amajuba districts recorded lower poor health prevalence.

It is of great interest to study how risk factors and disease prevalence changes with time. Here we provide an extensive discussion on the temporal evolution of the space-wave interaction based on Model 6. Figure 6.2 shows the mapped estimated residual spatial effects between wave 1 to 4 in South Africa (SA). The resulting maps are residual spatial effects that represent unobserved spatial factors either not measured in the surveys or abducting the effects of cultural patterns. In all the maps, district colours indicate similar features as in Figure 6.1. The figure shows that the spatial pattern changes much across the study waves period. It is observed that higher concentrations across the wave periods are scattered. High poor health prevalence can be observed in the districts within the northern regions for all the wave periods. Districts in the western regions had high poor

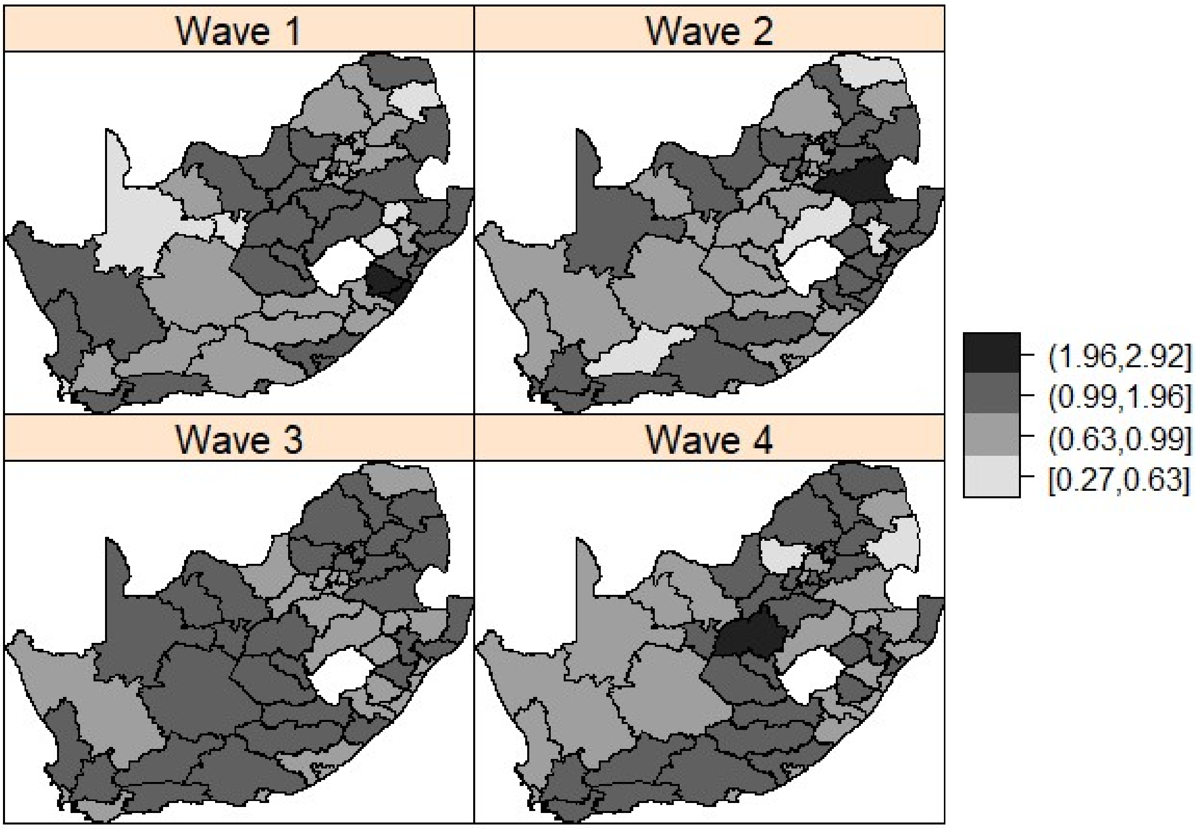


Figure 6.2: Maps showing residual spatial effects of poor self-reported health in South Africa between wave 1 to 4 derived from the spatio-temporal space-wave interaction cumulative logit regression model (Model 6).

health prevalence at the beginning of wave 1 thereafter the poor health prevalence was low between wave 2 to 4. The Namakwa, Siyanda and West Coast districts had reduced odds of reporting poor health across the waves periods. Also, the figure shows that districts within the southern regions start with low poor health prevalence thereafter the poor health prevalence was high between wave 2 to 4. Cacadu, Chris Hani, Overberg, Eden, Central Karoo, and Cape Winelands districts showed increased odds of reporting poor health across the waves period. High poor health prevalence can be seen in the central regions for all the waves except wave 2. The Xhariep, Lejweleputswa, Thabo Mofutsanyane, Fezile Dabi and Mangaung districts recorded high poor health prevalence. Furthermore, the districts within the eastern regions had high poor health prevalence across the wave periods. In wave 1, highest poor health prevalence was recorded in Ugu and Sisonke districts. Poor health prevalence was also high in Umkhanyakude district across the study wave periods.

Figure 6.3 display the temporal wave effects. The figure gives the estimated posterior mean of smooth function and their corresponding 95% CIs. From the figure, there is a decline of the mean wave effect between wave 1 and 2, and then a gradual rise but possibly not the same level by wave 4. This means wave 1 and 2 had low poor health prevalence.

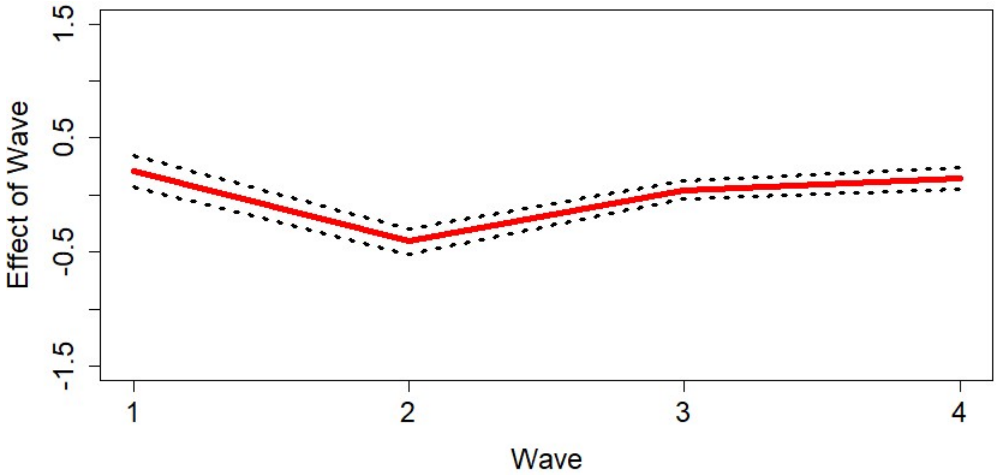


Figure 6.3: Estimated posterior mean (red line) along with the 95% CI (dashed line) of temporal wave random effect for the cumulative logit best fitting model.

Furthermore, wave 2 to 4 had high poor health prevalence. We further discuss the results of the binary response outcome based on Model 7, which was found to be the best fitting model.

Results of the spatio-temporal logistic regression model (Model 7) yielded similar results as Model 6 of the spatio-temporal cumulative logit model, also presented in Table 6.3. All the considered covariates remained associated with self-reported health except for race group, marital status, and type of toilet. The odds of reporting poor health seem to increase with age. The odds of reporting poor health for individuals between 20-24 years were 1.36 times the odds of reporting poor health for individuals between 15-19 years (POR: 1.36, 95% CI: 1.13 to 1.64). The odds of reporting poor health among individuals between 25-29 years were 1.79 times the odds of reporting poor health for individuals between 15-19 years (POR: 1.79, 95% CI: 1.48 to 2.16). The odds of reporting poor health for individuals between 30-34 years were 3.05 times the odds of reporting poor health for individuals between 15-19 years, (POR: 3.05, 95% CI: 2.54 to 3.66). The odds of reporting poor health for individuals between 35-39 years were 4.36 times the odds of reporting poor health for individuals between 15-19 years (POR: 4.36, 95% CI: 3.63 to 5.24). Moreover, the odds of reporting poor health for individuals between 40-44 and 45-49 years were respectively 5.17 and 7.99 times the odds of reporting poor health for individuals between 15-19 years, (POR: 5.17, 95% CI: 4.29 to 6.23) and (POR: 7.99, 95% CI: 6.64 to 9.62) respectively. The odds of reporting poor health for individuals who are male was 0.67 times the odds of reporting poor health for individuals who are female (POR: 0.67, 95% CI: 0.60 to 0.74). The odds of reporting poor health for individuals living in rural formal areas were 0.76 times the odds of reporting poor health for individuals residing in urban informal areas (POR: 0.76, 95% CI: 0.62 to 0.93). The odds of reporting poor health among individuals with high, college, and tertiary education were respectively 0.58 (with 95% CI: 0.49 to 0.69), 0.53 (with 95% CI: 0.40 to 0.71) and 0.42 (with 95% CI: 0.34 to 0.52) times the odds of reporting poor health for individuals with no formal education. The odds of reporting poor health among individuals with much above average household income were 0.50 times the odds of reporting poor health for individuals with much below average household income (POR: 0.50, 95% CI: 0.36 to 0.67). The odds of reporting poor health for individuals with dissatisfied and normal life were respectively 0.77 and 0.73 times the odds of reporting poor health for individuals with very dissatisfied life, (POR: 0.77, 95% CI: 0.69 to 0.88) and (POR: 0.73, 95% CI: 0.64 to 0.83) respectively. Furthermore, the odds of reporting poor health among individuals with satisfied and very satisfied life were respectively 0.66 (with 95% CI: 0.56 to 0.76) and 0.78 (with 95% CI: 0.66 to 0.93) times the odds of reporting poor health for individuals with very dissatisfied life. The odds of reporting poor health for individuals who exercise less than once a week were 0.83 times the odds of reporting poor health for individuals who never exercise (POR: 0.83, 95% CI: 0.70 to 0.99). For individuals who no longer drinks alcohol the odds of reporting poor health were 1.53 times the odds of reporting poor health for individuals who never drunk alcohol (POR: 1.53, 95% CI: 1.34 to 1.74). The odds of reporting poor health among individuals who smoke a cigarette was 1.16 times the odds of reporting poor health for individuals who do not smoke a cigarette (POR: 1.16, 95% CI: 1.03 to 1.32). The odds of reporting poor health among individuals who are not economically active were 1.29 times the odds of reporting poor health for individuals who are unemployed strict, (POR: 1.29, 95% CI: 1.14 to 1.46). The odds of reporting poor health among individuals with underweight and severe nutrition status were respectively 1.49 and 1.95 times the odds of reporting poor health for individuals with normal nutrition status, the corresponding odds ratios (POR: 1.49, 95% CI: 1.25 to 1.76) and (POR: 1.95, 95% CI: 1.55 to 2.43) respectively. The odds of reporting poor health for individuals who were diagnosed with TB was 3.46 times the odds of reporting poor health for an individual who was not diagnosed with TB (POR: 3.46, 95% CI: 3.04 to 3.94). The odds of reporting poor health among individuals who felt depressed the past week for little of the time, a moderate amount of the time and all the time were respectively 1.45, 1.99 and 3.18 times the odds of reporting poor health for individuals who felt depressed for less than one day, (POR: 1.45, 95% CI: 1.32 to 1.59), (POR: 1.99, 95% CI: 1.77 to 2.24) and (POR: 3.18, 95% CI: 2.68 to 3.77) respectively.

Figure 6.4 display the total residual spatial effects over the NIDS four waves based on Model 7. A similar spatial heterogeneity as in Figure 6.1 can be observed. The northwest and central regions yielded higher estimated odds of reporting poor health. The districts in northern and central regions had the highest poor health prevalence while those within the western regions had the lowest poor health prevalence. In contrast to the map in Figure 6.1, the Xhariep, Lejweleputswa, Mangaung, John Taolo Gaetsewe, Siyanda, Frances Baard, and Dr Ruth Segomotsi Mompati districts recorded higher poor health prevalence. The Namakwa, West Coast, Cape Winelands and Central Karoo districts recorded low poor health prevalence.

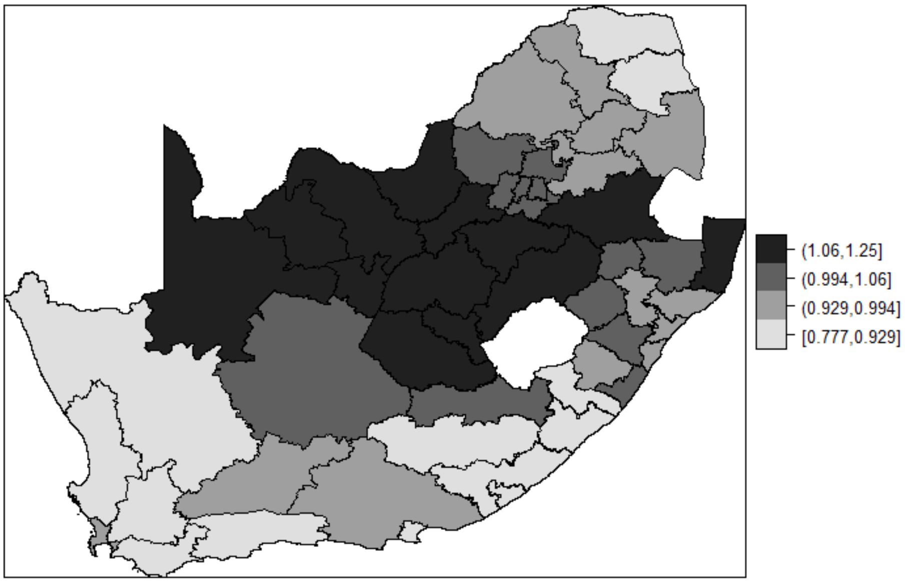


Figure 6.4: Map showing residual total spatial effects on self-reported health in South Africa based on a logistic regression model (Model 7).

Again we are interested in the evolution of the geographic variation of poor self-reported health. The change of spatial effects over wave periods is discussed based on Model 7. Figure 6.5 shows the mapped estimated posterior odds of residual spatial effects for wave 1 to 4 in South Africa (SA). The description of the colors for the districts is the same as mentioned above. This figure yielded similar results to Figure 6.2 but only a slight difference. There is clear evidence of spatial variation across the waves period. Figure

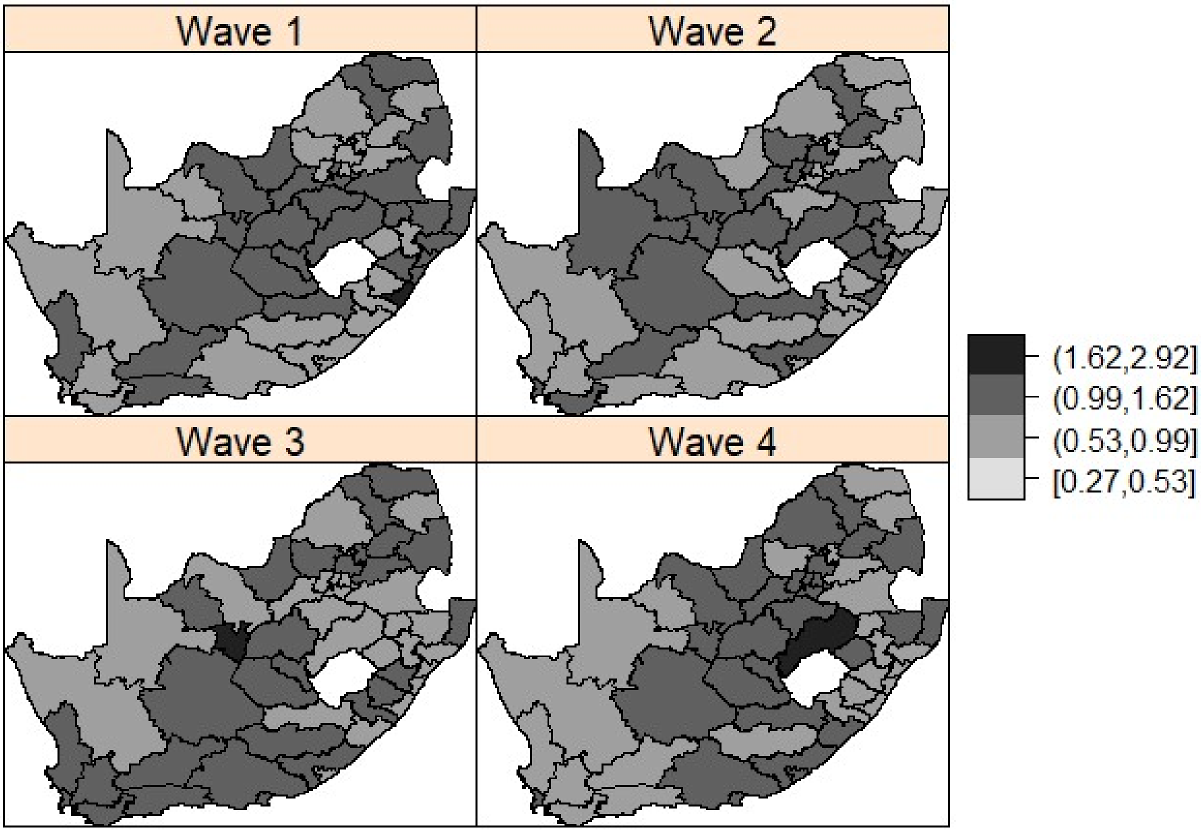


Figure 6.5: Estimated residual spatial effects of self-reported health in South Africa between wave 1 to 4 derived from the spatio-temporal interaction logistic regression model (Model 7).

6.5 show that the high prevalence of poor health across the wave periods can be observed in the districts within the central and southern regions. The Pixely ka Seme, Frances Baard, Lejweleputswa districts had increased odds of reporting poor health across the wave periods. The maps show that the districts within western regions, the prevalence of poor health was low in wave 4 as compared to other waves. Furthermore, the districts within the northern regions show that the pattern of spatial effects is scattered over the wave periods.

Figure 6.6 display a temporal wave effects and their corresponding 95% CIs. The plot provided slightly analogous results as in Figure 6.3, except that this figure shows a decreasing trend between wave 1 and 2, thereafter it becomes uniform. It can be seen that the wave effects are beyond zero at the beginning of wave 1, then below zero thereafter.

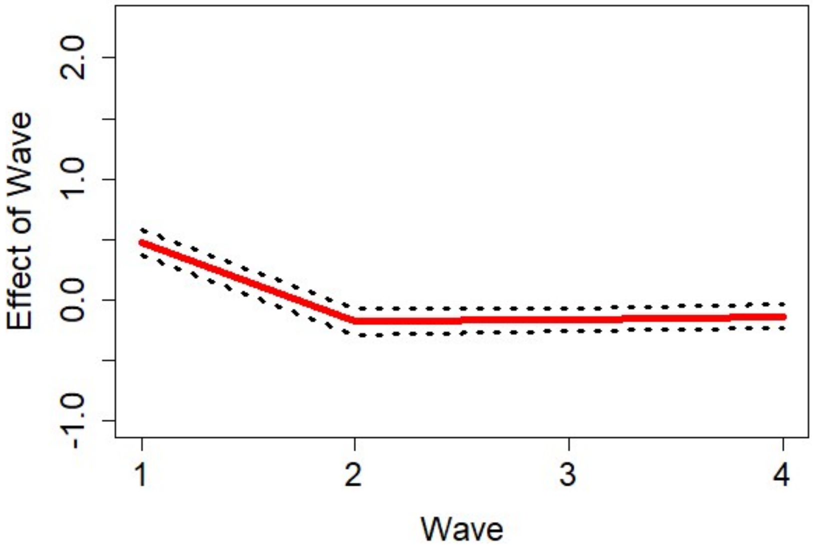


Figure 6.6: Estimated posterior mean (red line) along with the 95% CIs (dashed line) of temporal wave random effect for the logistic regression best fitting model.

Between wave 1 and 2, there was a reduction in poor health prevalence and thereafter remained low. In general, wave effects reduced poor health prevalence between wave 2 to 4.

Spatio-temporal models extend the previous spatial models by including the time effect. The idea behind such models is to investigate how self-reported health changes over the waves periods. The results showed quite interesting trends for both the responses. Health in South Africa is more likely to be improved.

6.6 Categorical interactions

One of the most interesting factors we considered in this research project was to include the interaction terms of the categorical covariates. The primary interest was on the lifestyle categorical covariates, as they are known to greatly influence health. In all the fitted spatial and spatio-temporal models analysis, the interaction between smoking status and alcohol consumption were found not to be statistically significant at 5% level of significance. Literally, none of the interactions that were tested in the analysis were found statistically significant. Therefore, no interaction between the categorical covariates was included in the models.

Chapter 7

Discussion and Conclusion

The aim of this research project is to investigate the determinants and geographic variation of self-reported health using the National Income Dynamics Study (NIDS) adult datasets in South Africa. The objective of this research was to develop and apply suitable statistical models that are used in assessing influential factors and geographical variation of poor self-reported health. It is also to use a unified framework of flexible models within a Bayesian hierarchical modeling in order to understand different types of factors associated with two distinct discrete choice types of self-reported health among individuals between the ages 15 - 49 years in South Africa. The models under consideration are an extension to classical models; this included spatial and spatio-temporal models which were used to identify geographical variation of area-specific effects. Structured additive multinomial cumulative logit and logistic regression models were developed to assess influential factors and districts variation of ordinal and binary self-reported health respectively. Two approaches within Bayesian inference were used in this research, namely the empirical Bayes (EB) via mixed model methodology and fully Bayesian (FB). The former was used in the inferential for the multinomial cumulative logit models while the latter was used in the inferential for logistic regression models. Structured additive modeling paradigm allows for different types of covariates to be added in classical models in an additive manner by borrowing strength from both the parametric and non-parametric models. In particular, we investigated linear, spatial, spatio-temporal and nonlinear effects on self-reported health with applications to the NIDS adult datasets in South Africa using integrated nested Laplace approximation (INLA) and $R2BayesX$ R-packages.

In the exploratory data analysis, we investigated potential covariates that may be associated with self-reported health. The statistical significance of apparent associations between potential covariates and self-reported health were first explored using the chi-square test. Using the chi-square test of independence, the analysis results revealed that age, gender, race, types of residence, education level, household income, marital status, life satisfaction level, exercising level, alcohol consumption, smoking status, type of toilet facilities, employment, nutrition status, Tuberculosis (TB) and depression level were associated with self-reported health. Some of these findings justified similar findings of the study by Reichmann et al. (2009) and Hosseinpoor et al. (2012) with a range of other studies. The proportional odds assumption was tested using the score test. The results show that the proportional odds assumption was insignificant (Table E.1), thus was not violated. Further, the significant covariates were all included to the models of interest.

In Chapter 4 and Chapter 5 we presented the EB and FB approaches respectively to estimate parameters of the spatial and structured additive regression (STAR) modeling of self-reported health in South Africa using wave 4 of NIDS dataset. The models accounted for either structured or unstructured or both spatial random effects. The spatial random effects account for unobserved influential factors that may vary between or within the districts. We assumed a conditional autoregressive (CAR) or independent and normal distributed priors on districts of South Africa. The models' assessment was based on the AIC, BIC, and GCV for the cumulative models while the DIC was used for the logistic regression models. The Bayesian multivariable cumulative logit and logistic regression models which incorporated both the spatially structured and spatially unstructured random effects were better fitting models. The incorporation of these spatial random effects improved the results. These models for both distinct responses yielded similar results except for a slight difference. In the models, it was found that age, gender, education, household income, exercising, alcohol, smoking, employment, nutrition status, TB and depression were significantly associated with self-reported health. The odds of reporting poor health for individuals between 45-49 years who were diagnosed with TB and felt depressed all the time were extreme higher when compared to their counterparts. These findings are accompanied by similar findings of the study by Cau et al. (2016b). However, the results for the household income were contradicting between the two models. The reason for this different result cannot be justified. We speculate that the reason may be due to the two different approaches used in this research. The FB is known to be a better technique than the EB approach (Leyland and Davies, 2005; Bernardinelli et al., 1995). Education and household income are well-known determinants of health. The level of education does not only empower knowledge of an individual but also makes them aware of a healthy living diet. While income level serves as a driving force to better education. The maps of the two above mentioned models showed a significant district level spatial variation (Figure 4.1 and Figure 5.1). Spatial variation was evidence indicating that health is not distributed evenly in South Africa. The variation under both the spatial cumulative logit model and the logistic regression model were almost similar. The dominant feature was that the districts within the central regions had a significantly higher prevalence of poor health. The districts within the south-western regions had lower poor health prevalence, however, the spatial effects of those districts were insignificant.

This research project has also presented flexible approaches used within the Bayesian methods. To further explore influential factors on self-reported health we developed Bayesian STAR models. The STAR models are good with their flexibility to allow the inclusion of generic types of covariates, such as continuous covariates. In the STAR models, the categorical covariates were assumed to have a linear effect on self-reported health. The additional of the age and body mass index (BMI) nonlinear effects did not affect the results of the linear effects. The STAR cumulative logit models and the logistic regression models yielded similar results as of the spatial models which did not account for nonlinear effects. Gender, education, household income, exercise, alcohol consumption, smoking, type of toilet, employment, TB and depression were found to be significantly associated with self-reported health. The odds of reporting poor health were less for individuals who are male. The findings also reveal that those with employment, high education level and not married had lower odds of reporting poor health as compared to their counterparts. The results correspond to similar findings from the study by Wilson et al. (2007). For all the STAR models we assumed that nonlinear effects of age and BMI follow a second order random walk (Rue and Held, 2005; Sørbye and Rue, 2011). We assumed that individuals age and BMI had nonlinear effects on self-reported health in all the STAR models. Both the cumulative logit and logistic regression within the STAR models produced similar plots (Figure 4.4 and Figure 5.4) for age and BMI effects. The age of an individual was found to be linearly associated with self-reported health. The odds of reporting poor health increases as age increases. The explanation for this is that as individuals grow older they are more likely to be exposed to ill health defining conditions such as cardiovascular and chronic diseases. On the other hand, the BMI was found to have a nonlinear relationship with self-reported health. The plot of BMI displayed a U-shaped curve. Poor health prevalence among individuals decreases with BMI up to about 25 kg/m^2 thereafter increased with an increasing BMI. The findings that BMI has a nonlinear effect on self-reported health was expected, given that BMI includes a normal status around 18 to 25 kg/m^2. Thus, poor health prevalence is bound to be high and low for BMI effect increase. The spatial effects maps for these models did not change much from those of spatial models with no nonlinear effects. There were still spatial variations across districts of South Africa. Higher concentrations of poor health prevalence were recorded in the districts within central regions and lower concentrations were in the western regions of South Africa.

Spatio-temporal models are often used when investigating the spatial trends of health outcomes. This research has presented spatio-temporal modeling on self-reported health in South Africa among adults between the ages 15-49 years using wave 1 to 4 of NIDS data collected between 2008 to 2015. Traditional models such as the cumulative logit and logistic regression models account for linearity assumption. However, these models can be extended to spatio-temporal modeling using an additive predictor. The cumulative logit and logistic regression spatio-temporal models yielded similar results. Age, gender, place of residence, education, household income, life satisfaction level, exercise, alcohol, smoking, employment, nutrition status, TB and depression were significantly associated with self-reported health over the four-wave periods. The results demonstrate increasing odds of reporting poor health as individuals age increases. Those who reside in rural formal areas had lower odds of reporting poor health as compared to those living in urban formal areas. This is due to the fact that in urban areas there are many factories which promote poor health such as air pollution which affects the health of an individual. Higher levels of education and much above average household income had lower odds of poor health. This is because higher education and higher income are driving factors of healthy living, thus promotes better health than individuals who are deprived of such. The districts within the central and north-western regions had the highest poor health prevalence. This may suggest lack of good education, toilet facilities, water source, healthcare institutes or higher HIV prevalence. The spatio-temporal trends effects were found to vary over the waves periods. Over the four waves, higher poor health prevalence was found in the districts within northern, central and southern regions. Lowest poor health prevalence was recorded in the districts within the western regions. We also estimated wave temporal effects on self-reported health. The plots showed a decreasing trend from wave 1 to wave 2 thereafter it was stationary. This may suggest a reduction of poor health prevalence over the four waves.

Every research has its own limitations. In this research, a major limitation was the number of waves available for us to estimate the temporal trends of self-reported health. This issue limited this research not to investigate the trend of self-reported health during the early years, more especially during the course of the HIV epidemic. Another limitation of this research was the administration area level which ends at the district level. It limits us not to compare the area levels such as district and municipal levels, thus to focus on those areas in a more reliable manner and also advice policymakers to focus their interventions on such relevant areas.

This research project used EB and FB approaches within the Bayesian hierarchical modeling to model adults self-reported health in South Africa. Also, we used the Bayesian structured additive approach to model the determinants of poor health. We showed that the prevalence of poor health can be modeled and mapped using these different approaches since the data was geo-referenced and collected repeatedly over time, hence suitable for spatio-temporal modeling. The findings of this research suggest that the improvement of health in South Africa is likely to be established. Recommendations from this research are that, the commission of social determinants of health, Healthy People 2020 and Sustainable Development Goals (SDGs) policy proponents must focus on improving individuals education, income inequality, and gender inequality by empowering women's position in development programmes. The creation of more jobs by the Government for the unemployed individuals will have an impact on improving wealth and decreasing depression. On the other hand, healthy living programs should be emphasized on social media and TV programmes in order to increase health awareness. Improvement could also be achieved by evaluation of health programmes to promote vaccine usage for communicable diseases such as TB and other diseases. In particular, the findings of this research imply that the main focus of these interventions should be in the districts within the northern, central and southern regions of South Africa.

There is a large space for further research on this study. This research provided modeling based on the conditional autoregressive (CAR) models. Future research may consider other prior distributions for the spatial random effects, such as the two-dimensional P-spline. A FB approach for modeling the different nature of self-reported health response would be of interest for future research. We also hope to consider the relationship between self-reported health and HIV status. Furthermore, Bayesian models for joint disease mapping should be considered. The main aim of joint disease mapping is to simultaneously investigate the determinants of health with other diseases that may influence health, such as Mortality, TB, HIV, and Malnutrition. In addition, we also hope to consider spatially varying coefficient for categorical covariates in order to understand how each covariate is distributed spatially.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* JohnWiley & Sons, Inc, Hoboken, New Jersey.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Angus, S, D. and Robert, T. (2015). People In Sub-Saharan Africa Rate Their Health And Health Care Among Lowest In World. *Health Aff (Millwood)*, 34(3):519 – 527.

Anita, R. (2013). Social determinants of health. University Lecture: Retrieved from https://www.alpbach.org/wp-content/uploads/2013/08/Forum2013_Rieder1.pdf (2018/05/31).

Ataguba, J. E.-O., Day, C., and McIntyre, D. (2015). Explaining the role of the social determinants of health on health inequality in south africa. *Global Health Action*, 8(1):28865. PMID: 28156737.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition.* Crc Press, London & New York NY, 2^{nd} edition.

Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf, N. (2009). BayesX: Software for Bayesian inference in structured additive regression models. http://www.uni-goettingen.de/de/bayesx/550513.html.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995). Bayesian analysis of space?time variation in disease risk. *Statistics in medicine*, 14(21-22):2433–2443.

Bernardinelli, L. and Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Mediciine*, 11(8):983–1007.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.

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

Best, N., Richardson, S., and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical methods in medical research*, 14(1):35–59.

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

Browning, C. R., Cagney, K. A., and Wen, M. (2003). Explaining variation in health status across space and time: implications for racial and ethnic disparities in self-rated health. *Social science & medicine*, 57(7):1221–1235.

Brunello, G., Fort, M., Schneeweis, N., and Winter-Ebmer, R. (2016). The causal effect of education on health: what is the role of health behaviors? *Health economics*, 25(3):314–336.

Cabrera-Barona, P. (2017). Influence of Urban Multi-Criteria Deprivation and Spatial Accessibility to Healthcare on Self-Reported Health. *Urban Science*, 1(2):11.

Cau, B. M., Falcão, J., and Arnaldo, C. (2016a). Determinants of poor self-rated health among adults in urban Mozambique. *BMC public health*, 16(1):856.

Cau, B. M., Falco, J., and Arnaldo, C. (2016b). Determinants of poor self-rated health among adults in urban Mozambique. *BMC Public Health*, 16(856):856.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681.

Congdon, P. (2005). *Bayesian models for categorical data.* Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 1^{st} edition.

Cramm, J. M. and Nieboer, A. P. (2011). The influence of social capital and socio-economic conditions on self-rated health among residents of an economically and health-deprived south african township. *International Journal for Equity in Health*, 10(1):51.

Cressie, N. (1993). *Statistics for Spatial Data.* Wiley-Interscience, Canada, revised edition.

Devine, O. J. and Louis, T. A. (1994). A constrained empirical Bayes estimator for incidence rates in areas with small populations. *Statistics in Medicine*, 13(11):1119–1133.

DiMaggio, C. (2012). Spatial Epidemiology Notes: applications and vignettes in R.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 11(2):89–102.

Ender, P. (2010). Collin. *Stata command to compute collinearity diagnostics. UCLA: Academic Technology Services, Statistical Consulting Group. http://www.ats.ucla.edu/stat/stata/ado/analysis/. Accessed*, 17.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space–time data: a Bayesian perspective. *Statistica Sinica*, 14:731–761.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models.* Springer Science & Business Media, New York.

Fayissa, B. and Gutema, P. (2005). The Determinants of Health Status in Sub-Saharan Africa (SSA). *The American Economist*, 49(2):60–66.

Gill, J. (2014). *Bayesian methods: A Social and Behavioral Sciences Approach.* CRC press, London & New York NY, 3 edition.

Gill, P. J. (2000). *Generalized Linear Models: A Unified Approach (Quantitative Applications in the Social Sciences).* Quantitative Applications in the Social Sciences, series no. 07-134. Sage Publications, Inc, London, 1 edition.

Gomez-Rubio, V., Bivand, R. S., and Rue, H. (2014). Spatial models using laplace approximation methods. In *Handbook of regional science*, pages 1401–1417. Springer.

Grut, L., Mji, G., Braathen, S. H., and Ingstad, B. (2012). Accessing community health services: challenges faced by poor people with disabilities in a rural community in South Africa. *African Journal of Disability*, 1(1):1–7.

Gutirrez-Pea, E. (1997). Moments for the canonical parameter of an exponential family under a conjugate distribution. *Environmental and Ecological Statistics*, 84:727–732.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall/CRC, New York.

Hosseinpoor, A. R., Williams, J. S., Amin, A., De Carvalho, I. A., Beard, J., Boerma, T., Kowal, P., Naidoo, N., and Chatterji, S. (2012). Social determinants of self-reported health in women and men: understanding the role of gender in population health. *PloS one*, 7(4):e34799.

Househam, K. C. (2010). Africa's burden of disease: The University of Cape Town Sub-Saharan Africa Centre for Chronic Disease. *SAMJ*, 100(2):94 – 95.

Johanson, R. (2001). *Education and Health in Sub-Saharan Africa: A Review of Sector-wide Approaches.* Washington, District of Columbia: World Bank.

Kammann, E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):1–18.

Kandala, N. B., Lang, S., Klasen, S., and Fahrmeir, L. (2001). Semiparametric analysis of the socio-demographic and spatial determinants of undernutrition in two african countries. *Research in Official Statistics*, 245:81–100.

Kawachi, I., Kennedy, B. P., and Glass, R. (1999). Social capital and self-rated health: a contextual analysis. *American journal of public health*, 89(8):1187–1193.

Kistemann, T., Dangendorf, F., and Schweikart, J. (2002). New perspectives on the use of Geographical Information Systems (GIS) in environmental health sciences. *International journal of hygiene and environmental health*, 205(3):169–181.

Kneib, T. (2006). *Mixed model based inference in structured additive regression*. PhD thesis, lmu.

Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space–time data: A mixed model approach. *Biometrics*, 62(1):109–118.

Kneib, T., Müller, J., and Hothorn, T. (2008). Spatial smoothing techniques for the assessment of habitat suitability. *Environmental and Ecological Statistics*, 15(3):343–364.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space–time variation in disease risk. *Statistics in medicine*, 19(17-18):2555–2567.

Komro, K. A., Burris, S., and Wagenaar, A. C. (2014). Social determinants of child health: concepts and measures for future research. *Health Behavior and Policy Review*, 1(6):432–445.

Lamarca, G. A., do C Leal, M., Sheiham, A., and Vettore, M. V. (2013). The association of neighbourhood and individual social capital with consistent self-rated health: a longitudinal study in Brazilian pregnant and postpartum women. *BMC pregnancy and childbirth*, 13(1):1.

Lang, S. and Brezger, A. (2000). Bayesx - software for bayesian inference based on markov chain monte carlo simulation techniques. collaborative research center 386, discussion paper 187. http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1577-9.

Lau, Y. K. and Ataguba, J. E. (2015). Investigating the relationship between self-rated health and social capital in South Africa: a multilevel panel data analysis. *BMC public health*, 15(1):266.

Lawson, A. B. (2008). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC press, Boca Raton & London & New York.

Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press, London & New York NY.

Leibbrandt, M., Woolard, I., and de Villiers, L. (2009). Methodology: Report on NIDS Wave 1, National Income Dynamics Study. Technical Report 1, Southern Africa Labour and Development Research Unit, University of Cape Town.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In Halloran, M. E. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–191, New York, NY. Springer New York.

Leyland, A. H. and Davies, C. A. (2005). Empirical Bayes methods for disease mapping. *Statistical Methods in Medical Research*, 14(1):17–34.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society: Series b (statistical methodology)*, 61(2):381–400.

Lock, K. (2000). Health impact assessment. *British Medical Journal*, 320(7246):1395–1398.

Ma, J., Mitchell, G., Dong, G., and Zhang, W. (2017). Inequality in Beijing: A spatial multilevel analysis of perceived environmental hazard and self-rated health. *Annals of the American Association of Geographers*, 107(1):109–129.

Manor, O., Matthews, S., and Power, C. (2000). Dichotomous or categorical response? Analysing self-rated health and lifetime social class. *International Journal of Epidemiology*, 29(1):149–157.

Mariella, L. and Tarantino, M. (2010). Spatial temporal conditional auto-regressive model: A new autoregressive matrix. *Austrian Journal of Statistics*, 39(3):223–244.

Marshall, R. J. (1991). Mapping Disease and Mortality Rates Using Empirical Bayes Estimators. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40(2):283–294.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society, Series B (Methodological)*, (42):109–142.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models.* Chapman & Hall, London.

McGranahan, G. and Martine, G. (2012). *Urbanization and Development: Policy lessons from the BRICS experience.* Routledge, London & New York.

Mirowsky, J. and Ross, C. E. (2013). *Education, social status, and health.* ALDINE DE GRUYTER, New York.

Moodley, J. and Ross, E. (2015). Inequities in health outcomes and access to health care in South Africa: a comparison between persons with and without disabilities. *Disability & Society*, 30(4):630–644.

Morris, C. and Normand, S. (1992). Hierarchical models for combining information and for meta-analyses. *Bayesian statistics*, 4:321–344.

Motala, S., Ngandu, S., Mti, S., Arends, F., Winnaar, L., Khalema, E., Makiwane, M., Ndinda, C., Moolman, B., Maluleke, T., et al. (2015). *Millennium Development Goals: country report 2015*. Statistics South Africa, Pretoria.

Murata, N., Yoshizawa, S., and Amari, S.-i. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5(6):865–872.

Ngwira, A. and Kazembe, L. (2016). Analysis of severity of childhood anemia in Malawi: a Bayesian ordered categories model. *Open Access Medical Statistics*, 2016(6):9–20.

OECD (2013). *Health at a Glance 2013: OECD Indicators*. OECD publishing. http://dx.doi.org/10.1787/health_glance-2013-en.

Phillips, L. J., Hammock, R. L., and Blanton, J. M. (2005). PEER REVIEWED: Predictors of Self-rated Health Status Among Texas Residents. *Preventing Chronic Disease*, 2(4).

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reichmann, W. M., Katz, J. N., Kessler, C. L., Jordan, J. M., and Losina, E. (2009). Determinants of self-reported health status in a population-based sample of persons with radiographic knee osteoarthritis. *Arthritis Care & Research*, 61(8):1046–1053.

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

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

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, 11(4):735–757.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Sen, A. (2002). Health: perception versus observation. *British Medical Journal*, 324(7342):860–861.

Sørbye, S. H. and Rue, H. (2011). Simultaneous credible bands for latent Gaussian models. *Scandinavian Journal of Statistics*, 38(4):712–725.

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

StataCorp (2015). Stata statistical software: Release 14. College Station, TX: StataCorp LP. http://www.stata.com.

Statistics South Africa (2015). *GENDER SERIES VOLUME II: Education, 2004-2014* (rep. no. 2). Retrieved from http://www.statssa.gov.za/?p=5933.

Subramanian, S. V., Huijts, T., and Avendano, M. (2010). Self-reported health assessments in the 2002 World Health Survey: how do they correlate with education? *Bulletin of the World Health Organization*, 88(2):131–138.

Sudipto, B., Gelfand Alan, E., and Carlin Bradley, P. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC Press, London & New York NY.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. Suri-Kagaku (Mathematical Sciences). 153:12–18. In Japanese.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86.

Tutz, G. (2003). Generalized semiparametrically structured ordinal models. *Biometrics*, 59(2):263–273.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical association*, 92(438):607–617.

Weimann, A., Dai, D., and Oni, T. (2016). A cross-sectional and spatial analysis of the prevalence of multimorbidity and its association with socioeconomic disadvantage in south africa: a comparison between 2008 and 2012. *Social Science & Medicine*, 163:144–156.

WHO, Sambo, L. G., Harris, M., and Beveridge, M. (2014). *The health of the people: what works: the African regional health report 2014*. World Health Organization, Regional Office for Africa. http://www.aho.afro.who.int/en/publication/1786/african-regional-health-report-2014-health-people-what-works.

Wilson, K., Elliott, S. J., Eyles, J. D., and Keller-Olaman, S. J. (2007). Factors affecting change over time in self-reported health. *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, 98(2):154–158.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, New York.

World Health Organization. (2008). *The Global Burden of Disease: 2004 Update*. Geneva: World Health Organization.

World Health Organization. (2013). *World health statistics 2013*. World Health Organization.

World Health Organization (2015). Tuberculosis country profile: TB burden estimates and country-reported TB data, South Africa. Retrieve from https://www.who.int/tb/country/data/profiles/en/.

World Health Organization. (2016). *Atlas of African Health Statistics 2016: health situ-*

ation analysis of the African Region. World Health Organization, Regional Office for Africa. http://www.who.int/iris/handle/10665/206547.

Wu, S., Wang, R., Zhao, Y., Ma, X., Wu, M., et al. (2013). The relationship between self-rated health and objective health status: a population-based study. *BMC public health*, 13(1):320.

Appendix A

R Codes of ordinal response

```
rm(list=ls())
# Packages required
library("MASS")
library("lattice")
library("ctv")
library("sp")
library(maptools)
library(rgdal)
library(spdep)
require(RColorBrewer)
require(ztable)
library(foreign)
library(R2BayesX)
#Loading map file
samap <- read.bnd("samap.bnd")
#Loading data file
dat <- read.csv("Adults.csv",sep = ",",header = T)
attach(dat)
dat <- na.omit(dat) #removing missing observations
dat[!complete.cases(dat),] #checking for completeness
dat$health <- ordered(dat$health) #ordering the response variable
############# Models ############################################
```

```
ctr<-bayesx.control(model.name ="OrdRes",outfile="C:/OrdRes",family="cumlogit",
method="REML") #BayesX estimation properties
#Specifying models formula
m1<-health~age_grp+gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+
alcohol+smokes+toilet+employment+maln+tb_diag+depressed
m2<-health~age_grp+gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+
alcohol+smokes+toilet+employment+maln+tb_diag+depressed+sx(district11,bs="mrf",map=samap)
m3<-health~age_grp+gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+
alcohol+smokes+toilet+employment+maln+tb_diag+depressed+sx(district11,bs="re")
m4<-health~age_grp+gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+
alcohol+smokes+toilet+employment+maln+tb_diag+depressed+sx(district11,bs="mrf",map=samap)+
sx(district11,bs="re")
#Running the models estimation
sm1 <- bayesx(m1,data = dat,control = ctr)
sm2 <- bayesx(m2,data = dat,control = ctr)
sm3 <- bayesx(m3,data = dat,control = ctr)
sm4 <- bayesx(m4,data = dat,control = ctr)

########## Spatial effects ########################################
sa.graph <- readOGR("District_Municipalities_2016.shp")
mrf <- sm4$effects$`sx(district11):mrf`[,c(2,3,7)]
re <- sm4$effects$`sx(district11):re`[,c(2,3,7)]
tot <- mrf+re
sa.graph$tot <- exp(tot$Estimate)
summary(sa.graph$tot)
sa.graph$totcuts <- cut(sa.graph$tot,breaks=c(0.42,0.71,0.99,2.30,3.70),include.lowest=T)

############ Map ################################################
spplot(sa.graph,"totcuts",col.regions=gray(3.5:0.5/4))
```

```
############ Non linear effects of age and BMI ##################
ctr<-bayesx.control(model.name ="OrdRes",outfile="C:/OrdRes",family="cumlogit",method=
"REML") #BayesX estimation properties
#Specifying models formula
m1<-health~gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+alcohol+
smokes+toilet+employment+tb_diag+depressed+sx(age,bs="rw2")+sx(BMI,bs="rw2")
m2<-health~gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+alcohol+
smokes+toilet+employment+tb_diag+depressed+sx(district11,bs="mrf",map=samap)+
sx(age,bs="rw2")+sx(BMI,bs="rw2")
m3<-health~gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+alcohol+
smokes+toilet+employment+tb_diag+depressed+sx(district11,bs="re")+sx(age,bs="rw2")+
sx(BMI,bs="rw2")
m4<-health~gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+alcohol+
smokes+toilet+employment+tb_diag+depressed+sx(district11,bs="mrf",map=samap)+
sx(district11,bs="re")+sx(age,bs="rw2")+sx(BMI,bs="rw2")
#Running the models estimation
sm1 <- bayesx(m1,data = dat,control = ctr)
sm2 <- bayesx(m2,data = dat,control = ctr)
sm3 <- bayesx(m3,data = dat,control = ctr)
sm4 <- bayesx(m4,data = dat,control = ctr)

########## Spatial effects ########################################
sa.graph <- readOGR("District_Municipalities_2016.shp")
mrf <- sm4$effects$‘sx(district11):mrf‘[,c(2,3,7)]
re <- sm4$effects$‘sx(district11):re‘[,c(2,3,7)]
tot <- mrf+re
sa.graph$tot <- exp(tot$Estimate)
sa.graph$totcuts <- cut(sa.graph$tot,breaks = c(0.43,0.72,0.99,2.30,3.65),
include.lowest = T)
```

```
########## Map ######################################################
spplot(sa.graph,"totcuts",col.regions=gray(3.5:0.5/4))

########## Plots ####################################################
#BMI
plot(sm4$effects$‘sx(BMI)‘[,1],sm4$effects$‘sx(BMI)‘[,2],type="l",col="red",ylim=c(-1,2.30)
,lwd=5,lty=1,xlab="Body Mass Index",ylab="Effect of BMI",cex.lab=1.5, cex.axis=1.5)
par(new=TRUE)
plot(sm4$effects$‘sx(BMI)‘[,1],sm4$effects$‘sx(BMI)‘[,3],ann=FALSE,axes=FALSE,type="l",
lwd=3,ylim=c(-1,2.30),lty=21)
par(new=TRUE)
plot(sm4$effects$‘sx(BMI)‘[,1],sm4$effects$‘sx(BMI)‘[,7],ann=FALSE,axes=FALSE,type="l",
lwd=3,ylim=c(-1,2.30),lty=21)
par(mfrow=c(1,1))
#Age
plot(sm4$effects$‘sx(age)‘[,1],sm4$effects$‘sx(age)‘[,2],type="l",col="red",
ylim=c(-1.26,1.5),lwd=5,lty=1,xlab="Respodent age in years",ylab="Effect of age",
cex.lab=1.5,cex.axis=1.5)
par(new=TRUE)
plot(sm4$effects$‘sx(age)‘[,1],sm4$effects$‘sx(age)‘[,3],ann=FALSE,axes=FALSE,type="l",
lwd=3,ylim=c(-1.26,1.5),lty=21)
par(new=TRUE)
plot(sm4$effects$‘sx(age)‘[,1],sm4$effects$‘sx(age)‘[,7],ann=FALSE,axes=FALSE,type="l",
lwd=3,ylim=c(-1.26,1.5),lty=21)
par(mfrow=c(1,1))

########### Spatio temporal #######################################
rm(list=ls())
samap <- read.bnd("samap.bnd")
dat <- read.csv("Adults.csv",sep = ",",header = T)
```

```
attach(dat)
dat$health_ord <- ordered(dat$health_ord)
dat$wave1 <- factor(dat$wave)
dat$distrct1 <- dat$distrct11
ctr<-bayesx.control(model.name="OrdResSpatio",outfile="C:/OrdResSpatio",family="cumlogit",
method="REML") #BayesX estimation properties
#Specifying models formula
m1<-health_ord~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+sx(distrct11,bs="mrf",map=samap)+
sx(distrct11,bs="re") #Spatial model
m2<-health_ord~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+sx(distrct11,bs="mrf",map=samap)+
sx(distrct11,bs="re")+as.factor(wave) #Linear trend model
m3<-health_ord~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+sx(distrct11,bs="mrf",map=samap)+
sx(distrct11,bs="re")+sx(wave,bs="ps") #Simple trend model with one random time effect
m4<-health_ord~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+sx(distrct11,bs="mrf",map=samap)+
sx(distrct11,bs="re")+r(distrct1,bs="re",map = samap,by = wave1) #Interaction only
#Interaction + linear trend
m5<-health_ord~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+sx(distrct11,bs="mrf",map=samap)+
sx(distrct11,bs="re")+r(distrct1,bs="re",map = samap,by = wave1)+as.factor(wave)
#Interaction with one random time effect (p-spline)
m6<-health_ord~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+sx(distrct11,bs="mrf",map=samap)+
sx(distrct11,bs="re")+sx(wave)+r(distrct1,bs="re",map = samap,by = wave1)
#Interaction with one random time effect (random walk of order one)
m7<- health_ord ~ age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+sx(distrct11,bs="mrf",map=samap)+
sx(distrct11,bs="re")+r(wave,bs="rw1")+r(distrct1,bs="re",map = samap,by = wave1)

#Running the models estimation
sm1 <- bayesx(m1,data = dat,control = ctr)
sm2 <- bayesx(m2,data = dat,control = ctr)
sm3 <- bayesx(m3,data = dat,control = ctr)
sm4 <- bayesx(m4,data = dat,control = ctr)
sm5 <- bayesx(m5,data = dat,control = ctr)
sm6 <- bayesx(m6,data = dat,control = ctr)
sm7 <- bayesx(m7,data = dat,control = ctr)

########## Spatial effects ########################################
wav1 <- exp(sm6$effects$‘sx(distrct1):wave11:re‘[,2])
wav2 <- exp(sm6$effects$‘sx(distrct1):wave12:re‘[,2])
wav3 <- exp(sm6$effects$‘sx(distrct1):wave13:re‘[,2])
wav4 <- exp(sm6$effects$‘sx(distrct1):wave14:re‘[,2])
cuts <- c(0.27,0.63,0.99,1.96,2.92)
sa.graph <- readOGR("District_Municipalities_2016.shp") #Loading map
sa.graph$wav1 <- wav1
sa.graph$wav2 <- wav2
sa.graph$wav3 <- wav3
sa.graph$wav4 <- wav4
sa.graph$wav1 <- cut(sa.graph$wav1,breaks = cuts,include.lowest = TRUE)
sa.graph$wav2 <- cut(sa.graph$wav2,breaks = cuts,include.lowest = TRUE)
sa.graph$wav3 <- cut(sa.graph$wav3,breaks = cuts,include.lowest = TRUE)
sa.graph$wav4 <- cut(sa.graph$wav4,breaks = cuts,include.lowest = TRUE)

########## Maps ############################################
spplot(sa.graph,c("wav3","wav4","wav1","wav2"),names.attr=c("Wave 3","Wave 4","Wave 1",
"Wave 2"),col.regions=gray(3.5:0.5/4),par.settings=list(fontsize=list(text=15)),xlab = "",
ylab = "",scales=list(draw = FALSE))


########## Plot ##########################################################

####Time effect###

plot(sm6$effects$‘sx(wave)‘[,1],sm6$effects$‘sx(wave)‘[,2],type="l",col="red",ylim=c(-1.50,

1.50),lwd=5,lty=1,xlab="Wave",ylab="Effect of Wave",cex.lab=1.5, cex.axis=1.5,xaxt = ’n’)

axis(1,at = seq(1,4,1),lty = 1,cex.lab=1.5, cex.axis=1.5)

par(new=T)

plot(sm6$effects$‘sx(wave)‘[,1],sm6$effects$‘sx(wave)‘[,3],ann=F,axes=F,type="l",lwd=3,

ylim=c(-1.50,1.50),lty=21)

par(new=T)

plot(sm6$effects$‘sx(wave)‘[,1],sm6$effects$‘sx(wave)‘[,7],ann=F,axes=F,type="l",lwd=3,

ylim=c(-1.50,1.50),lty=21)

par(mfrow=c(1,1))
```

Appendix B

R Codes for binary response models

```
rm(list=ls())

# Packages required

library("MASS")

library("lattice")

library("ctv")

library("sp")

library(maptools)

library(rgdal)

library(spdep)

require(INLA)

require(RColorBrewer)

library(ztable)

sa.graph <- readOGR("District_Municipalities_2016.shp")

adjsa <-poly2nb(sa.graph)#Creates adjacency for sa

nb2INLA("sa.graph",adjsa)

#Loading data

dat <- read.csv("Adults.csv",sep = ",",header = T)

attach(dat)

#head(dat,5)

#tail(dat,5)

dim(dat)

dat <- na.omit(dat) #removing missing observations
```

```
dat[!complete.cases(dat),] #checking for completeness
####################################################################################
#districts duplicates
dat$district1 <- dat$district11
dat$district2 <- dat$district11
###### Models formulation and estimation ###########################################
formula011<-health_bin~age_grp+gender+race+res2001+edu+income+marital_stat+satisfied_cat+
exercise+alcohol+smokes+toilet+employment+maln+tb_diag+depressed
res011<-inla(formula011,family="binomial",data=dat,control.compute=list(dic=T,cpo=T))
formula012<-health_bin~age_grp+gender+race+res2001+edu+income+marital_stat+satisfied_cat+
exercise+alcohol+smokes+toilet+employment+maln+tb_diag+depressed+
f(district11,model="besag", graph="sa.graph")
res012<-inla(formula012,family="binomial",data=dat,control.compute = list(dic=T,cpo=T))
formula013<-health_bin~age_grp+gender+race+res2001+edu+income+marital_stat+satisfied_cat+
exercise+alcohol+smokes+toilet+employment+maln+tb_diag+depressed+
f(district11,model="iid")
res013<-inla(formula013,family="binomial",data=dat,control.compute=list(dic=T,cpo=T))
formula014<-health_bin~age_grp+gender+race+res2001+edu+income+marital_stat+satisfied_cat+
exercise+alcohol+smokes+toilet+employment+maln+tb_diag+depressed+
f(district1,model="besag", graph="sa.graph")+f(district2,model="iid")
res014<-inla(formula014,family="binomial",data=dat,control.compute=list(dic=T,cpo=T))

########## Spatial effects ################################
spatial014 <- res014$summary.random$district1
spatial014re <- res014$summary.random$district2
tot <- spatial014$"0.5quant"+spatial014re$"0.5quant"
sa.graph$r <- exp(spatial014$"0.5quant"+spatial014re$"0.5quant")
sa.graph$tot <- cut(sa.graph$r , breaks = c(0.60,0.79,0.99,1.30,1.61) ,include.lowest = T)

########## Map ############################################
```

```
spplot(sa.graph,"tot",col.regions=gray(3.5:0.5/4))


############ Non linear effects of age and BMI #######################################
###### Models formulation and estimation ###############################
formula011<-health_bin˜gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+
alcohol+smokes+toilet+employment+tb_diag+depressed+f(age,model="rw2")+f(BMI,model="rw2")
res011<-inla(formula011,family="binomial",data=dat,control.compute=list(dic=TRUE,cpo=T))
formula012<-health_bin˜gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+
alcohol+smokes+toilet+employment+tb_diag+depressed+f(district11,model="besag",graph=
"sa.graph")+f(age,model="rw2")+f(BMI,model="rw2")
res012 <- inla(formula012,family= "binomial",data=dat,control.compute=list(dic=T,cpo=T))
formula013<-health_bin˜gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+
alcohol+smokes+toilet+employment+tb_diag+depressed+f(district11,model="iid")+f(age,model=
"rw2")+f(BMI,model="rw2")
res013 <- inla(formula013, family = "binomial",data=dat,control.compute= list(dic=T,cpo=T))
formula014<-health_bin˜gender+race+res2001+edu+income+marital_stat+satisfied_cat+exercise+
alcohol+smokes+toilet+employment+tb_diag+depressed+f(district1,model="besag", graph=
"sa.graph")+f(district2,model="iid")+f(age,model="rw2")+f(BMI,model="rw2")
res014 <- inla(formula014, family = "binomial",data=dat,control.compute =list(dic=T,cpo=T))

########## Spatial effects ################################
r<-exp(res014$summary.random$district1$‘0.5quant‘+
res014$summary.random$district2$‘0.5quant‘)
sa.graph$r <- r
sa.graph$tot<-cut(sa.graph$r,breaks=c(0.59,0.79,0.99,1.32,1.65),include.lowest=TRUE)



########## Map ####################################################
spplot(sa.graph,"tot",col.regions=gray(3.5:0.5/4))
```

```
########## Plots ###################################################
#BMI
plot(res014$summary.random$BMI$ID,res014$summary.random$BMI$mean,type="l",col="red",ylim=
c(-0.60,2.30),lwd=5,lty=1,xlab="Body Mass Index",ylab="Effect of BMI",cex.lab=1.5,
cex.axis=1.5)
par(new=T)
plot(res014$summary.random$BMI$ID,res014$summary.random$BMI$‘0.025quant‘,ann=F,axes=F,type=
"l",lwd=3,ylim=c(-0.60,2.30),lty=21)
par(new=T)
plot(res014$summary.random$BMI$ID,res014$summary.random$BMI$‘0.975quant‘,ann=F,axes=F,type=
"l",lwd=3,ylim=c(-0.60,2.30),lty=21)
par(mfrow=c(1,1))
#Age
plot(res014$summary.random$age$ID,res014$summary.random$age$mean,type="l",col="red",ylim=
c(-1.26,1.32),lwd=5,lty=1,xlab="Respodent age in years",ylab="Effect of age",cex.lab=1.5,
cex.axis=1.5)
par(new=T)
plot(res014$summary.random$age$ID,res014$summary.random$age$‘0.025quant‘,ann=F,axes=F,type=
"l",lwd=3,ylim=c(-1.26,1.32),lty=21)
par(new=T)
plot(res014$summary.random$age$ID,res014$summary.random$age$‘0.975quant‘,ann=F,axes=F,type=
"l",lwd=3,ylim=c(-1.26,1.32),lty=21)
par(mfrow=c(1,1))

########### Spatio temporal ######################################
rm(list=ls())
sa.graph <- readOGR("District_Municipalities_2016.shp")
adjsa <-poly2nb(sa.graph)#Creates adjacency for sa
nb2INLA("sa.graph",adjsa)
dat <- read.csv("Adults.csv",sep = ",",header = T)
```

```
attach(dat)
##################################################################################
dat$distrct1 <- dat$distrct11
dat$distrct2 <- dat$distrct11
dat$wave1 <- dat$wave
dat$wave2 <- dat$wave

###### Models formulation and estimation ##########################################
formula1<-health_bin~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+f(distrct1,model="besag",graph="
sa.graph")+f(distrct2,model="iid",graph="sa.graph") #Spatial model
mod1<-inla(formula1,family="binomial",data=dat,control.compute=list(dic=T,cpo=T),verbose=F)
formula2<-health_bin~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+f(distrct1,model="besag",graph="
sa.graph")+f(distrct2,model="iid",graph="sa.graph")+as.factor(wave) #Linear trend model
mod2<-inla(formula2,family="binomial",data=dat,control.compute=list(dic=T,cpo=T),verbose=F)
#Simple trend model with one random time effect
formula3<-health_bin~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+f(distrct1,model="besag",graph="
sa.graph")+f(distrct2,model="iid",graph="sa.graph")+f(wave,model = "ar1")
mod3<-inla(formula3,family="binomial",data=dat,control.compute=list(dic=T,cpo=T),verbose=F)
formula4<-health_bin~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+f(distrct1,model="besag",graph=
"sa.graph")+f(distrct2,model="iid",graph="sa.graph")+f(distrct11,model="iid",group=wave,
control.group=list(model="ar1"),adjust.for.con.comp = FALSE) #Interaction only
mod4<-inla(formula4,family="binomial",data=dat,control.compute=list(dic=T,cpo=T),verbose=F)

#Interaction only + linear trend
formula5<-health_bin~age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+f(distrct1,model="besag",graph="
sa.graph")+f(distrct2,model="iid",graph="sa.graph")+f(distrct11,model="iid",group=wave,
control.group=list(model="ar1"),adjust.for.con.comp = FALSE)+as.factor(wave)
mod5<-inla(formula5,family="binomial",data=dat,control.compute=list(dic=T,cpo=T),verbose=F)
#Interaction with one random time effect (ar1)
formula6<-health_bin˜age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+f(distrct1,model="besag",graph="
sa.graph")+f(distrct2,model="iid",graph="sa.graph")+f(wave,model = "ar1")+f(distrct11,
model="iid",group=wave,control.group=list(model="ar1"),adjust.for.con.comp=FALSE)
mod6<-inla(formula6,family="binomial",data=dat,control.compute=list(dic=T,cpo=T),verbose=F)
#Interaction with one random time effect (random walk)
formula7<-health_bin˜age_group+gender+race+geo2001+edu+income+marital_stat+satisf+exercise+
alcohol+smokes+toilet+empl_stat+maln+TB+depressed+f(distrct1,model="besag",graph="
sa.graph")+f(distrct2,model="iid",graph="sa.graph")+f(wave,model="rw1")+f(distrct11,model="
iid",group = wave,control.group=list(model="ar1"),adjust.for.con.comp = FALSE)
mod7<-inla(formula7,family="binomial",data=dat,control.compute=list(dic=T,cpo=T),verbose=F)

########## Spatial effects ################################

wav1 <- exp(mod6$summary.random$distrct11[1:52,5])
wav2 <- exp(mod6$summary.random$distrct11[53:104,5])
wav3 <- exp(mod6$summary.random$distrct11[105:156,5])
wav4 <- exp(mod6$summary.random$distrct11[157:208,5])

sa.graph$wav1 <- wav1
sa.graph$wav2 <- wav2
sa.graph$wav3 <- wav3
sa.graph$wav4 <- wav4

cuts <- c(0.58,0.79,0.99,1.50,2.20)
sa.graph$wav1 <- cut(sa.graph$wav1,breaks = cuts,include.lowest = TRUE)
sa.graph$wav2 <- cut(sa.graph$wav2,breaks = cuts,include.lowest = TRUE)
sa.graph$wav3 <- cut(sa.graph$wav3,breaks = cuts,include.lowest = TRUE)
sa.graph$wav4 <- cut(sa.graph$wav4,breaks = cuts,include.lowest = TRUE)

########## Maps ############################################
spplot(sa.graph,c("wav3","wav4","wav1","wav2"),names.attr=c("Wave 3","Wave 4","Wave 1"
,"Wave 2"),col.regions=gray(3.5:0.5/4),par.settings=list(fontsize=list(text=15)),xlab="",
ylab="",scales=list(draw = FALSE))

########## Plots ##################################################
#### Time effect ###
plot(mod6$summary.random$wave$ID,mod6$summary.random$wave$‘0.5quant‘,type="l",col="red",
ylim=c(-1.0,2.30),lwd=5,lty=1,xlab="Wave",ylab="Effect of Wave",cex.lab=1.5,cex.axis=1.5,
xaxt = ’n’)
axis(1,at = seq(1,4,1),lty = 1,cex.lab=1.5, cex.axis=1.5)
par(new=TRUE)
plot(mod6$summary.random$wave$ID,mod6$summary.random$wave$‘0.025quant‘,ann=F,axes=F,
type="l",lwd=3,ylim=c(-1.0,2.30),lty=21)
par(new=T)
plot(mod6$summary.random$wave$ID,mod6$summary.random$wave$‘0.975quant‘,ann=F,axes=F,
type="l",lwd=3,ylim=c(-1.0,2.30),lty=21)
par(mfrow=c(1,1))
```

Appendix C

Study area

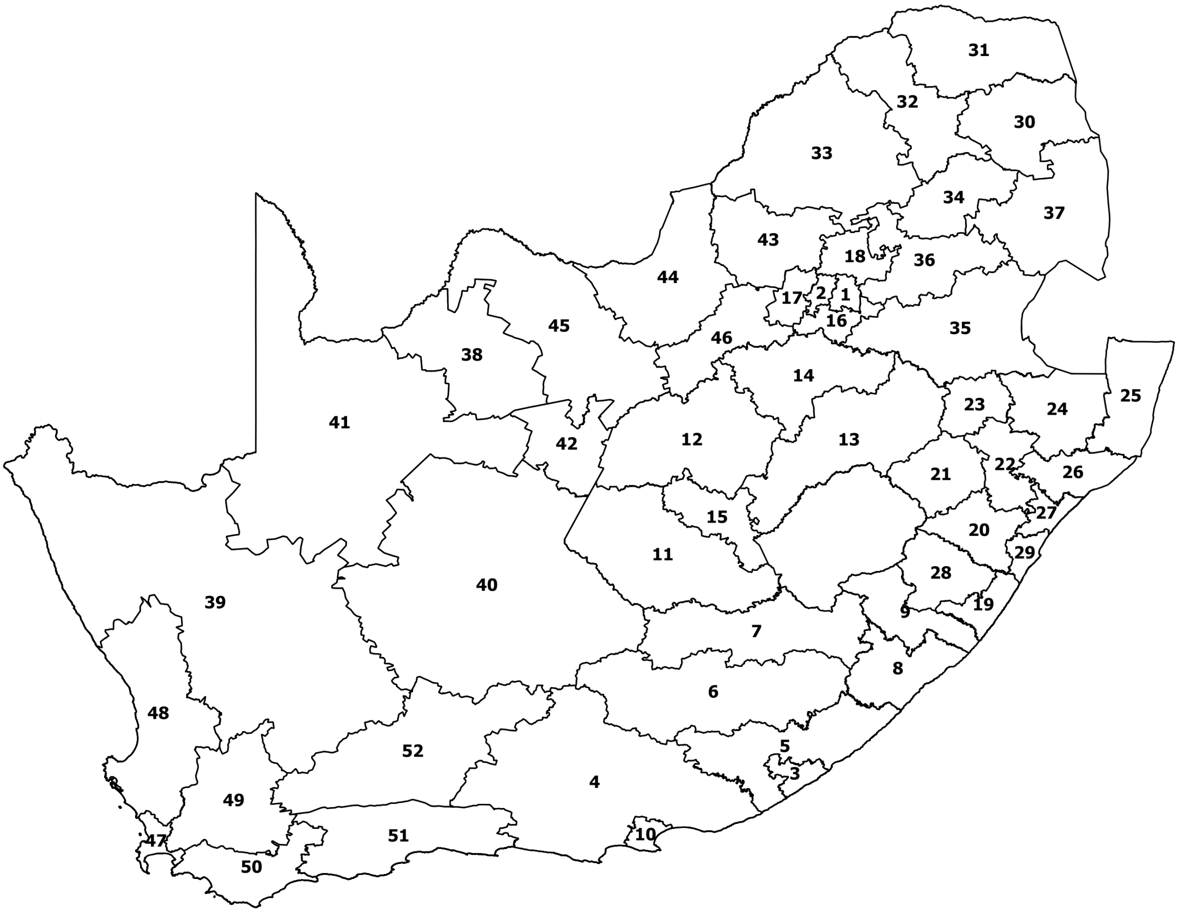


Figure C.1: Map of South Africa showing district municipalities within provinces.

Table C.1 present the district municipality names along with corresponding codes displayed in the South African map in Figure C.1. The region which is not numbered is the Lesotho country.

Table C.1: Districts municipality of South Africa and their corresponding geographic codes used in NIDS data.

District	Code	District	Code	District	Code
Ekurhuleni	1	Ugu	19	Ehlanzeni	37
City of Johannesburg	2	Umgungundlovu	20	John Taolo Gaetsewe	38
Buffalo City	3	Uthukela	21	Namakwa	39
Cacadu	4	Umzinyathi	22	Pixley ka Seme	40
Amathole	5	Amajuba	23	Siyanda	41
Chris Hani	6	Zululand	24	Frances Baard	42
Joe Gqabi	7	Umkhanyakude	25	Bojanala	43
O.R.Tambo	8	Uthungulu	26	Ngaka Modiri Molema	44
Alfred Nzo	9	iLembe	27	Dr Ruth Segomotsi Mompati	45
Nelson Mandela Bay	10	Sisonke	28	Dr Kenneth Kaunda	46
Xhariep	11	eThekwini	29	City of Cape Town	47
Lejweleputswa	12	Mopani	30	West Coast	48
Thabo Mofutsanyane	13	Vhembe	31	Cape Winelands	49
Fezile Dabi	14	Capricorn	32	Overberg	50
Mangaung	15	Waterberg	33	Eden	51
Sedibeng	16	Great sekhukhune	34	Central Karoo	52
West Rand	17	Gert Sibande	35		
City of Tshwane	18	Nkangala	36		

Appendix D

Multicollinearity results for the NIDS wave 4 dataset.

Table D.1: Multicollinearity results of all the considered covariates for the Bayesian multivariable spatial models.

Covariates	VIF	$\sqrt{\text{VIF}}$	Tolerance	R^2	Standard Error
Age group	1.72	1.31	0.5819	0.4181	0.0100541
Gender	1.37	1.17	0.7295	0.2705	0.0345057
Race	1.18	1.09	0.8472	0.1528	0.0322828
Place of residence	1.21	1.10	0.8271	0.1729	0.0204496
Marital status	1.40	1.19	0.7118	0.2882	0.0199674
Household income	1.06	1.03	0.9474	0.0526	0.0150253
Life satisfaction level	1.08	1.04	0.9237	0.0763	0.0133721
Alcohol consumption level	1.41	1.19	0.7098	0.2902	0.0122051
Exercise	1.13	1.06	0.8827	0.1173	0.0101177
Smokes	1.45	1.20	0.6889	0.3111	0.0455896
Type of toilet	1.36	1.16	0.7368	0.2632	0.0092655
Employment status	1.34	1.16	0.7462	0.2538	0.0122931
Nutrition status	1.21	1.10	0.8297	0.1703	0.0158066
Previously diagnosed with TB	1.02	1.01	0.9810	0.0190	0.0814654
Felt depressed in the past week?	1.05	1.02	0.9537	0.0463	0.0196538

Multicollinearity is a state which occurs when a combination of two or more covariate are highly correlated. When it occurs it becomes a problem, as the estimates will be imprecise and results to wider standard errors. The common method to check for multicollinearity is the variance inflation factor (VIF). In literature it is argued that maximum level of the VIF that does not raise concern about multicollinearity should be 5 or 10. The multicollinearity was tested using `collin` command (Ender, 2010) in STATA 14. The results reveal that no VIF values were close to both the maximum levels for all the covariates in Table D.1. Hence, all the covariates were included in the models.

Appendix E

Proportional odds assumption of the cumulative logit model

One of the underlying assumption in the cumulative logit (proportional odds) model is that the effects of any covariates included in the model are identical across the different thresholds, thus this is termed the assumption of proportional odds or the parallel regression assumption.

Table E.1: Testing for the proportional odds assumption.

Model	**AIC**	**Log-likelihood**	**Difference in log-likelihood** $(2(l_M - l_C))$	**df**	**p-value**
Cumulative logit	39169	-19533.00	14.00	9	0.1151
Multinomial logit	39172	-19526.00			

In order to have a substantial interpretation of the results from the cumulative logit model, the validity of the proportional odds assumption was tested. Table E.1 presents the results of the proportional odds assumption test based on the likelihood ratio test. The results reveal that the assumptions were not violated ($p = 0.1151$), since the p-value is greater than 0.05. Furthermore, it can be seen that the cumulative logit model has the smaller Akaike information criterion (AIC) (39169) as compared to the multinomial logit model.

Appendix F

NIDS data sample distribution of the fours waves for all the spatio-temporal analysis.

Table F.1: Sample sizes of respondents for the four NIDS waves in South Africa

Wave	**1**	**2**	**3**	**4**	**Total**
Sub-sample (n)	7462	6348	13161	15707	42678

Table F.1 presents the sample size of the NIDS waves in South Africa. As shown in the table, wave 3 and 4 accounts for a large sample size compared to the other two waves. This may be due to most respondents joining the study in recent years.