UNIVERSITY OF
KWAZULU-NATAL

# Bayesian spatial joint and spatial-temporal disease modeling with application to HIV, HSV-2 and Malaria using case studies from Kenya and Angola respectively.

by

Okango Elphas Luchemo

**A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy in Statistics in the**

**College of Agriculture, Engineering and Science**
**School of Mathematics, Statistics and Computer Science**

March 2017

**Dedication**

To God the Almighty.

To my Parents Mr. Livingstone A. Okango and Mrs. Judith Okango

# Declaration

This research work was carried out under the supervision of Prof. H. Mwambi and Dr. O. Ngesa at the School of Mathematics, Statistics and Computer Science, University of Kwazulu Natal, Pietermaritzburg campus. The work represents an original work by the author and it has not been submitted in any form for any degree or any other qualification at this University or any other institution. Due acknowledgement is given where works of others have been quoted.

| | |
|---|---|
| Author: Elphas L. Okango | Date |
| Supervisor: Prof. Henry Mwambi | Date |
| Co-supervisor: Dr. Oscar Ngesa | Date |

# List of Journal Publications and Conference papers

1. Okango Elphas, Henry Mwambi, and Oscar Ngesa. "Spatial modeling of HIV and HSV-2 among women in Kenya with spatially varying coefficients." BMC Public Health 16.1 (2016): 1.

2. Okango Elphas, Henry Mwambi, and Oscar Ngesa, Thomas Achia. "Semi-Parametric Spatial Joint Modeling of HIV and HSV-2 among Women in Kenya." PloS one 10.8 (2015): e0135212.

3. Okango Elphas, Henry Mwambi, and Oscar Ngesa, "Semi-Parametric Spatial Joint modelling of HIV and HSV-2 among Women in Kenya with spatially varying coefficients". The South African Statistical association's 57th Annual conference, Department of Statistics, University of Pretoria, 29th November 2015-2nd December 2015.

4. Okango Elphas, Henry Mwambi, and Oscar Ngesa, "Spatial modeling of HIV and HSV-2 among women in Kenya with spatially varying coefficients".ASSAf-TWAS ROSSA Young Scientists Conference 2015, Birchwood Hotel and O.R Tambo Conference center on 16-18 September 2015.

5. Okango Elphas, Henry Mwambi, and Oscar Ngesa, "Relaxing the linearity assumption in Spatial Joint modelling with application to HIV and HSV-2 in Kenya". Strathmore International Mathematics Conference, Nairobi, Kenya. 3 - 7 August,2015

6. Okango Elphas, Henry Mwambi, and Oscar Ngesa, "Semi-Parametric Spatial Joint Modeling of HIV and HSV-2 among Women in Kenya". 1st sub-Saharan Africa conference on spatial and spatial temporal statistics, School of Public health, University of Witwatersrand, Johannesburg, 17-24th November 2014

7. Okango Elphas, Artemisa Lima, Henry Mwambi, and Oscar Ngesa, “Spatio-temporal modeling of Malaria among children under the age of 5 in Angola”. Submitted

8. Okango Elphas, Henry Mwambi, and Oscar Ngesa, “Relaxing some limiting assumptions in disease mapping with applications”. Submitted

**University of KwaZulu Natal**

# *Abstract*

**College of Agriculture, Engineering and Science**
**School of Mathematics, Statistics and Computer Science**

**Doctor of Philosophy**

**by Okango Elphas Luchemo**

In this thesis we develop and extend existing statistical models for spatial disease modeling and apply them to HIV, HSV-2 and malaria data. The availability of geo-referenced data and free software has seen many disease mapping models developed and applied in epidemiology, public health, agriculture and ecology among other areas. In chapter 1 we provide a background and developments in the field of disease mapping. We present in brief some limiting assumptions and how recent developments have tried to relax them. Chapter 2 introduces a model; the semi-parametric joint model to model HIV and HSV-2. The semi-parametric joint model performed better than the single models in terms of DIC. The limiting linearity assumption was relaxed by using the penalized regression splines for the continuous covariate age. The main focus of chapter 3 was to develop a model that relaxes the stationarity assumption. This was achieved by allowing the effects of the covariates to vary spatially by using the conditional autoregressive model. This new model performed better than the stationary models.

In chapter 4 we introduce a spatial temporal spatially varying covariate model. In this model, the covariates were allowed to vary both spatially and temporally. We fit this model to the Angolan malaria data. The fifth chapter presents a review of various assumptions in spatial disease modeling and improvements for some

limiting assumptions such as the normality assumption on random effects and linearity assumption on the covariates. We use the non-parametric spatial model approach to relax the limiting normality assumption. The last part of chapter 5 involves developing a joint spatially varying model (an extension of the spatially varying coefficient model in chapter 3) and fitting it to the HIV and HSV-2 data. Chapter six of the study provides the overview of the thesis, the conclusion and presents areas of further studies.

# *Acknowledgements*

Praise God from whom all blessings flow, to God the almighty glory and honor. I would like to thank my supervisors Prof. Henry Mwambi and Dr. Oscar Ngesa for their guidance encouragement and their positive words throughout my work. I would like to thank the Kisii Universiy for giving me a study leave that enabled me undertake my PhD. Special thanks to Prof. Mailutha for being such a good friend. I am indebted to Dr. Oscar Ngesa, Mr Humphreys M, Prof. Uppal, Prof Odhiambo, Ms. Jane Akinyi, Prof Mwita, Dr. Mwema, Dr Y. Dawood, Prof. Kihoro and other teaching and non-teaching staff for their assistance during my masters program at JKUAT. My parents Mr. Livingstone A Okango and Judith Okango who exchanged their comfort for our our success. My brothers and sisters, my many friends Preston, Ong'ayo, Linda, Mary, Ongata, and many others. Thanks to my friends and colleagues at the University of Kwazulu-Natal for their time and any assistance. Thanks to Lima, Kalinda, Orakwelu, Malinzi, Sbongakonke, Makweti, Abdallah, Jesca, Christel, Bev, Keli, Lucas, Catherine, and the SDASM-PMB family.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **AMIS** | **A**ngola **M**alaria **I**ndicator**S**urvey |
| **CAR** | **C**onditional **A**utoregressive |
| **DIC** | **D**eviance **I**nformation **C**riterion |
| **DP** | **D**irichlet **P**rior |
| **HIV** | **H**uman **I**mmuno-Deficiency Virus |
| **HSV-2** | **H**erpes **S**implex Virus -**T**ype 2 |
| **INLA** | **I**nla **N**ested **L**aplace **A**pproximation |
| **KAIS** | **K**enya **A**IDS **I**ndicator **S**urvey |
| **MCAR** | **M**ultivariate **C**onditional **A**utoregressive **M**odel |
| **MCMC** | **M**arkov **c**hain **M**onte **C**arlo |
| **MMMC** | **M**ultiple **M**embership **M**ultiple **C**lassification |
| **MPT** | **M**ultivariate **P**olya **T**rees |

# Chapter 1

# General Introduction

## 1.1 Overview

Many fields of study, for example public health, epidemiology, agriculture and ecology of late have vast amounts of geo-referenced data. This geo-referencing is usually by point referencing, i.e. latitudes and longitudes or areal referencing i.e. districts, counties, states, provinces and other administrative units. The availability of such data has necessitated the development and application of spatial statistical methods in analysis of geographically correlated data. This thesis focuses on hierarchical modeling of binary data and in particular the modeling of Human Immuno-Deficiency Virus (HIV), Herpes Simplex Virus type 2 (HSV-2) and malaria prevalence and mortality data. In the introductory part of this thesis, we highlight some key concepts in spatial modeling and subsequently expound on the development of spatial models, their limitations and offer alternative solutions and methods to overcome some of these limitations.

## 1.2 Disease Mapping

Disease mapping may be defined as the estimation and presentation of summary measures of health outcomes by geographical location [Rezaeian et al., 2007]. This is usually in order to;

- Generate hypotheses about a disease and its risk factors.
- Describe the geographical variation of diseases and/or risk factors.
- Generate disease atlases and maps.
- Detect disease clustering.

Detection of disease clustering, understanding of geographical variations of diseases and risk factors and accurate hypotheses about a disease and its risk factors can be used by policy makers when making decisions on public health resource allocation, assessment of inequalities and informing on tailor-made intervention strategies. In the overview that follows, we only discuss the applications for discrete variation of disease and the type of spatial data considered is called areal data (lattice data). There exist other types of spatial data namely point referenced data (geostatistical data) and point pattern data. Modeling of geostatistical data types is discussed extensively by Cressie [1992] while Diggle et al. [1998] discuss extensively on modeling of the point pattern data types.

## 1.3 Disease mapping models

Due to availability of vast amounts of geo-referenced data and free statistical software, many disease mapping models have been developed and implemented. These models usually arise from the generalized linear model. Let $y_{ij}$ be the disease status of an individual $j$ in region $i$, with $y_{ij} = 0$, if the individual tests negative for the disease and 1 otherwise. Thus a region, location or similar structure contains a cluster of observations. The disease status $y_{ij}$ is modeled as a Bernoulli random variable in the generalized linear model (GLM) context. Covariates or predictors may be included in this model and they may either be categorical or continuous. Categorical covariates are usually assumed to have linear effects on the response variable while the continuous covariates may not necessarily be linearly related to the response variable.

Often and realistically, covariates may not account for all the variation observed in the response variable and hence the GLM may be extended to include random effects leading to a generalized linear mixed model (GLMM) [Lawson et al., 2003]. The simplest form of random effects model introduces an additional parameter $\mu_i$ into the linear predictor for each response unit $i$. The $\mu_i^{'}s$ are assumed independent and exchangeable and are usually modeled with a zero mean normal distribution with unknown variance $\sigma^2$. The model is specified in a hierarchical form with two stages. The observed statuses in a cluster are conditionally independent given the values of the random effects and in the second stage the distribution of the random effects is specified. These random effects may be correlated or uncorrelated. Correlated random effects can be introduced through a spatial covariance

matrix by making the random effects form a single vector following an appropriate distribution with a specified mean and a spatial variance-covariance matrix. The spatial variance-covariance matrix is made up of parametric functions defining the covariance structure based on any two units of study. In the case of geostatistical data, the spatial covariance between two observations is dictated by the distance between the two observations [Cressie, 1992; Diggle et al., 1998; Waller and Gotway, 2004], while in the case of lattice data, neighbourhoods can be specified based on sharing a border, the distance between the centroids of any pair of regions or a combination of these two.

The multivariate Gaussian distribution is the most commonly used distribution for the random effects [Gaetan et al., 2010; Sherman, 2011; Waller and Gotway, 2004]. The use of this assumption is mainly because of its computational simplicity. The argument against the normality assumption is that some random effects may exhibit skewness, fat-tailness, multimodality e.t.c. and this may obscure some important features of between subjects and within subjects variations. Many studies have tried relaxing the normality assumption by use of both parametric and non-parametric distributions. The generalized Gaussian distribution (GGD) was employed by Ngesa et al. [2014a] to relax the normality assumption. They showed that it can produce better results when the normality assumption is violated due to high or low peakedness in the data. Skew distributions have also been effectively used in modeling heterogeneous data with asymmetry features. These distributions include the skew-t and the skew normal distributions [Azzalini, 1985]. Bayesian nonparametric spatial modeling approaches for disease

incidence and prevalence data include the use of the Dirichlet distribution by Ferguson [1973] or its variation and the polya trees processes by Lavine [1992] for random effects distribution.

The earliest spatially structured prior distribution for random effects was introduced by Clayton and Kaldor [1987]. Using the empirical Bayes approach, they estimated the relative risk of a region as a tradeoff between local data and a weighted average of observations in the neighborhood of that region. A fully Bayesian counterpart to the Clayton and Kaldor [1987] approach, the conditional autoregressive model (CAR), was introduced by [Besag et al., 1991]. They implemented their model using Markov chain Monte Carlo (MCMC) algorithms. This specification gives an alternative to using the multivariate Gaussian models. Here, the conditional distribution of the random effects in a region given all the others is a weighted average of all the other random effects. A number of studies have used different weighting schemes. These weighting schemes are either fixed or data-driven. Besag et al. [1991] assigned the weights based on whether a pair of regions shared a boundary or not; if the regions share a boundary, the weight assigned is 1, otherwise the weight assigned in 0. Best et al. [2001] used distance-based spatial weights however the adjacency based model performed better than the distance-based model according to the Deviance Information Criterion (DIC). Earnest et al. [2007] found considerable differences in the smoothing properties of the CAR model, depending on the neighborhood structure specified. This in turn had an effect on their models' ability to predict the observed risk in an area. These results have significant implications for all researchers using CAR models,

since the neighborhood weight matrices chosen may markedly influence a study's findings. Lu et al. [2007] developed a Bayesian hierarchical model that permits the estimation of the spatial neighborhood structure. Their approach allowed the data and other observed covariate information to help in the degree and the nature of spatial smoothing.

The random effects can be split into two to capture both clustering and heterogeneity by including both the spatially structured random effects (clustering) and the spatially unstructured random effects (heterogeneity) through a convolution model. The choice between the clustering and the heterogeneity model depends on the prior belief one has about the scope of dominant risk determinants. However the prior weight needs to be assigned fairly to the structured and unstructured components to avoid either global over smoothing (clustering) or local over smoothing (heterogeneity). This can be achieved by ensuring that the standard deviation of the conditional distribution of the spatially structured random effects is 0.7 times that of the spatially unstructured random effects Bernardinelli et al. [1995] however this conclusion is still open for debate. Other formulations of the convolution model include using only one random intercept but splitting its variance covariance matrix into spatial and non-spatial components, with a parameter controlling the spatial dependency [Leroux et al., 2000]. Other alternatives include a parametric bootstrap approach due to MacNab and Dean [2000] and the hidden Markov approach by Green and Richardson [2002] but these they are not commonly used.

When diseases share common risk factors, it could be more precise to model them

together. The joint modeling of disease outcomes within a spatial statistical context may provide more insight on the interaction of diseases both at individual and at regional level. The most common approaches for handling multiple disease modeling are the multivariate CAR approach suggested by Carlin and Banerjee [2003]; Gelfand and Vounatsou [2003], the shared component approach by MacNab [2010] and the multiple membership multiple classification (MMMC) approach, see [Browne et al., 2001].

In the shared component model for jointly modeling of two diseases, one component is relevant to the two diseases while another is specific to one of the diseases. These two components account for the unobserved spatial variables that affect disease risk and are not captured by the systematic component via the covariates. There is a slight difference between the MCAR and the MMMC approach and this lies in the way the spatial correlation is achieved. In the MCAR approach, the spatial correlation is achieved through a variance structure while in MMMC, the spatial correlation is achieved through a multiple membership relationship and that the neighborhood random effects are not independent.

It would be of great epidemiological importance if disease risks are observed both in space and in time. This may help bring out how the effects of the risk factors change with time and in space, unmasking of endemic regions and periods and bringing out both new risk factors and those that no longer are at play as far as the disease prevalence or incidence is concerned. With this information, policy makers may be informed as to whether their intervention strategies are working over time and whether different approaches need to be considered. Spatio-temporal

models have been developed to aid in passive surveillance of diseases in space and time. Bernardinelli et al. [1995] introduced a spatio-temporal model for count data. They assumed a Poisson GLM model with the linear predictors containing separate terms for space and time as well as space and time interaction effects allowing for different temporal trends in different regions. In their spatio-temporal model, Waller et al. [1997] used the CAR model developed by Besag et al. [1991] for their spatio-temporal model. Their model allowed each time period to have a separate spatial and non-spatial random effect. Other spatio-temporal models are discussed in [MacNab and Dean, 2002; Sun et al., 2000].

Other than introducing the spatial dependence via the random effects, coefficients can be allowed to vary through the spatial domains leading to spatially varying coefficients. This allows the relationship between responses and covariates to vary by region in the spatial domain. Assunçao [2003] allowed the covariates to vary spatially by assigning its regression parameters the Bayesian autoregressive (BAR), simultaneous autoregressive (SAR) or the conditional autoregressive (CAR) model. Other studies that allowed the covariates to vary by region in the spatial domain are by Hastie and Tibshirani [1990] and by Gelfand et al. [2003].

So far we have discussed how spatial dependence can be introduced via the random effects and the spatially varying coefficients. Spatial dependence can also be introduced via the observations. The leading models for introducing the spatial dependence via the observations are the auto Poisson models [Besag, 1974; Griffith, 2001] and the autologistic models [Hoeting et al., 2000].

The most common approach for doing inference for latent Gaussian models has been via the MCMC algorithms although other methods have been employed including the expectation propagation (EP) by Minka [2001] and variational Bayes method [Hinton and Van Camp, 1993]. The downside of the MCMC algorithm is usually computational time and convergence [Rue et al., 2009]. The integrated nested Laplace approximation (INLA) by Rue et al. [2009] is a new inference technique that has been widely used by many researchers giving similar results to the MCMC but in a shorter time. The INLA approach basically involves three steps; the posterior of the hyper-parameters given the data are approximated and these are used to determine the grid of hyper-parameter values. The second step involves approximating the posterior marginal distributions given the data and the hyper-parameter values on the grid. Lastly a numerical integration of the product of the two approximations is done to obtain the posterior marginal of interest. This approach saves computational time as compared to the MCMC which directly samples from the joint posterior distribution.

## 1.4 Statement of the problem

In the recent past, disease mapping has gained traction and this has been largely due to the availability of geo-referenced data and free software. Most of the models developed for disease mapping have given impressive results. These models have however made some assumptions that may be limiting hence lead to inaccurate results and interpretations. These models normally use random effects which are split into spatial and non-spatial components. The normality assumption is usually

used for the non-spatial component. This assumption may not always be true as data may exhibit fat tailness, skewness, multimodality or high or low peakedness. There is need to consider models that allow flexibility in the normal random effects. Other than using a parametric assumption for the random effects, non-parametric models can be used instead. In this work we employ the non-parametric modeling of the random effects using the mixture of dirichlet and the mixture of polya trees models. This thesis also considers the various methods of relaxing the linearity assumption on covariate effects including the use of the penalized regression splines and the random walk model.

The assumption that the same stimulus provokes the same response across the study region may also be misleading. For spatial processes, the same stimulus may provoke different responses across the study area. This may be due to attitudes, cultures, preferences, climate among other reasons which are localized within an area. It is therefore reasonable to assume that the effects of the individual risk factors vary from place to place. This can be achieved by allowing the regression coefficients to vary across space.

When several diseases share common risk factors it may be more effective to model the diseases jointly. By pooling the available data from different disease sources, there are gains in precision and efficiency of estimates, especially in rare diseases. In this study we develop a model for jointly modeling HIV and HSV-2 and later extend this model to also allow the coefficients to vary spatially.

The effect of disease and particularly the effect of the intervention strategies put in place by policy can be observed over time. It would also be of great importance

to observe how the effects of a particular risk factor evolve both over time and in space. In this study we explore these effects by developing a spatio-temporal spatially varying coefficient model and apply it to the Angolan malaria data.

## 1.5 Main Objective

The main objective of this study is to extend and develop disease mapping models for lattice data.

## 1.6 Specific Objectives

The specific objectives are;

- to develop semi parametric-joint disease models and apply them to the Kenya HIV and HSV-2 women data.
- to relax the stationarity assumption of covariates and apply the spatially varying covariates model to the data.
- to review the limitations of some disease mapping models and give alternatives.
- to extend the existing spatial models to spatio-temporal models and also allow covariates to vary both in space and time and apply the model to malaria prevalence data.
- to develop a joint spatially varying coefficient model.

## 1.7 Dissertation Outline

In this thesis we develop models and methods for spatial and spatio-temporal analysis of diseases. This thesis is divided into six chapters. In chapter 1 we provide an introduction to the study, discussing the developments in disease mapping. Chapters 2-5 represents full research articles that have been published in peer-reviewed journals or submitted. Chapter 6 provides the overall conclusions. The contents of the chapters are:

**Chapter 1**: This chapter contains the introduction of this thesis with discussions on developments in disease mapping for both single and multiple diseases and the objectives of the study.

**Chapter 2**: Here, we develop models for joint modeling of HIV and HSV-2 prevalence among women in Kenya using the 2007 Kenya AIDS indicator survey data (KAIS). The joint disease model is found to perform better than the single disease model.

**Chapter 3**: In this chapter, we develop a spatial model that introduces spatial dependence through both the coefficients and the random effects, thereby relaxing the stationarity assumption of the covariates. The new model performs better than the stationary models. The model is applied to modeling HIV and HSV-2 among women in Kenya.

**Chapter 4**: In this chapter we develop a model that allows the coefficients to vary both in space and time. The spatio-temporal model is fitted to the Angolan malaria data for children under the age of 5 years. The chapter also discusses

the inference techniques and developments on the spatial-temporal modeling of disease.

**Chapter 5**: This chapter reviews the limiting assumptions and the areas that still need improvements in disease mapping. The chapter gives alternatives to some of the limiting assumptions with applications using the KAIS data. In this chapter we also develop a joint spatially varying coefficient model and apply it in modeling HIV and HSV-2 among women in Kenya.

**Chapter 6**: This chapter gives a summary of the thesis. We summarize the findings, the conclusions and highlight some topics that need further research.

# Chapter 2

# Semi-Parametric spatial Joint modeling of HIV and HSV-2 among women in Kenya

## 2.1 Introduction

According to the world health organization (WHO), more than 1 million people acquire sexually transmitted infections (STI) daily. The WHO report of 2013 indicates that more than 530 million people (about 7.5%) have the virus that causes genital herpes or the herpes simplex virus type 2 (HSV-2) [WHO, 2013]. It was estimated that out of these, 123.7 million or 23% resided in sub-Saharan Africa, among whom 63% were women [Looker et al., 2008]. HSV-2 prevalence in the age group 15-49 in sub-Saharan Africa region ranges from 30% to 80% among

women and from 10% to 50% among men [Weiss, 2004]. People living with HIV were estimated to be about 35 million by the end of 2013 with 2.1 million new infections [UNAIDS, 2013]. HSV-2 is associated with a two to threefold increased risk of HIV acquisition and an up to fivefold increased risk of HIV transmission per sexual act, and may account for 40% to 60% of new HIV infections in populations where HSV-2 has a high prevalence Looker et al. [2008], hence modeling these two diseases jointly may provide more insights on how these two diseases relate in Kenya. STIs can have serious consequences beyond the immediate impact of the infection itself, through mother-to-child transmission (MTCT) of infections and chronic diseases. Drug resistance is a major threat to reducing the impact of STIs worldwide [WHO, 2013].

Many studies have focused on monitoring HIV and HSV-2 trends in a country and comparison between countries using national averages [Mishra et al., 2007]. These averages, though important, can hide the HIV and HSV-2 prevalence variability among administration units of a country and hence intervention strategies rolled out at national levels may not be effective at the administration level.

The national HIV and HSV-2 prevalence rates in 2002 in Kenya within the adult population (15-64 years) were estimated to be as high as 5.6% and 7.1% respectively as reported by NASCOP [2012], with a wide gender and geographical variation. The HIV prevalence among women was 6.9% while among men was 4.4%. The North Eastern region had HIV prevalence of as low as 2.1% while regions around Lake Victoria and the Western regions had prevalence ranging from between 13%-25% [NASCOP, 2007]. The Kenya National AIDS and STI Control

Program (NASCOP) in their Kenya AIDS Indicator Survey (KAIS) 2007 report stated that age had non-linear relationships with HIV and HSV-2 prevalence. This is consistent with several studies which have shown that HIV and HSV-2 prevalence by age have a non-linear relationship assuming an inverted U shape [Ghebremichael et al., 2009; NASCOP, 2007]. HIV prevalence increases with age until it plateaus at between ages 25-35, then starts decreasing with increasing age. HSV-2 prevalence increases with age up to between ages 35-45 then begins to decline with increasing age.

Several studies have assumed that all the covariates in the study have a linear relationship with the response variable. This linear relationship may not hold for all variables as in our case age, which has a non-linear relationship with the response variable. Our objective is to perform a spatial joint modeling which allows for studying of the relationship between diseases and also between regions under study and at the same time captures this nonlinear relationship. We extend the spatial semi parametric model based on penalized regression spline proposed in previous studies such as Ngesa et al. [2014b] to model HIV and HSV-2 jointly among women in Kenya.

## 2.2 Methods

### 2.2.1 Data

The data for this study was obtained from the Kenya AIDS Indicator Survey (KAIS) which was carried out by the Kenyan government with financial support from the United States President's Emergency Plan for AIDS Relief (PEPFAR) and the United nations (UN). The main aim of the survey was to obtain high quality data on the prevalence of HIV and Sexually Transmitted Infections (STI) among adults and to assess the knowledge of HIV and STIs in the population.

The sampling frame for KAIS was the National Sample Survey and Evaluation Program IV (NASSEP IV). It consisted of 1800 clusters comprising of 1260 rural and 540 urban clusters; of these, 294 rural and 141 urban clusters were sampled for KAIS. The overall design for KAIS 2007 was a stratified, two-stage cluster sampling design. The first stage involved selecting clusters from NASSEP IV, and the second stage involved the selection of household for KAIS with equal probability in the urban-rural strata within the districts. A sample of 415 clusters and 10,375 households were systematically selected for KAIS. A uniform sample of 25 households per cluster was selected using an equal probability systematic sampling method. The multilevel structure of the data in our analysis was accounted for through the random effects to model within and between county variability.

The survey was twofold: A household questionnaire was used to collect the characteristics of the living environment and an individual questionnaire to collect information on demographic characteristics and the knowledge of HIV and STIs

on men and women aged 15-64 years. A representative sample of households and individuals was selected from eight provinces in the country. Each individual was asked for consent to provide a venous blood sample for HIV and HSV-2 testing. More information on the survey methodologies used in collecting the data is found in the final KAIS, 2007 report [NASCOP, 2007]. This study uses the 2007 data even though a new round of KAIS, 2012 NASCOP [2012] has been done. The final release of this new study had not been made, hence the data was not available for use. This study uses the women's data from the KAIS, 2007 survey. Information from 4864 women, aged 15-64 years who had provided venous blood for HIV and HSV-2 testing and also had full covariate information was used in the analysis. In the data, age was captured as both categorical and continuous while all other covariates were categorical. An initial exploratory data analysis was carried out using a univariate standard logistic regression model to determine the association of each single covariate with the outcome variable (HIV and HSV-2 status). These variables were categorized into four groups, namely: demographic, social, biological and behavioral.

From this initial analysis, education level, age at first sex, perceived risk, partners in the last one year, marital status, place of residence, STI status in the last one year and age of the respondent were found to be associated with HIV and HSV-2 infection. The results are contained in tables 2.1 and 2.2. It was also established that age had a non-linear effect on HIV and HSV-2 infection, hence its continuous form (mean=33.31, SD=10.87) was used in the subsequent analysis.

TABLE 2.1: Exploratory data analysis for HIV

| Variable | p-value | Unadjusted OR |
|---|---|---|
| **Demographic characteristics** | | |
| Place of residence (Ref Rural) | | 1 |
| Urban | 0.001 | 0.749(0.635,0.884) |
| Age (Ref 15-19) | 0.001 | 1 |
| 20-24 | 0.000 | 2.825(1.982,4.026) |
| 25-29 | 0.000 | 3.055(2.133,4.375) |
| 30-34 | 0.000 | 4.656(3.276,6.618) |
| 35-39 | 0.000 | 3.682(2.544,5.328) |
| 40-44 | 0.000 | 2.796(1.869,4.181) |
| 45-49 | 0.000 | 2.783(1.858,4.169) |
| 50-54 | 0.000 | 2.347(1.490,3.696) |
| 55-59 | 0.294 | 1.352(0.770,2.375) |
| 60-64 | 0.173 | 0.487(0.173,1.371) |
| **Social Characteristics** | | |
| Wealth Quantile (ref poorest) | 0.525 | 1 |
| Second | 0.652 | 1.058(0.827,1.353) |
| Middle | 0.392 | 0.896(0.696,1.153) |
| Fourth | 0.564 | 1.074(0.843,1.369) |
| Richest | 0.592 | 0.938(0.741,1.186) |
| Media access(Ref No) | | 1 |
| Yes | 0.257 | 0.913(0.781,1.068) |
| Education level (Ref none) | 0.000 | 1 |
| Primary | 0.386 | 1.078(0.910,1.276) |
| Secondary | 0.574 | 0.929(0.720,1.200) |
| Higher | 0.000 | 0.451(0.303,0.671) |
| Marital status(Ref Married, 1 partner) | 0.000 | 1 |
| Married, +2partners | 0.001 | 1.536(1.192,1.980) |
| Divorced/separated | 0.000 | 2.503(1.960,3.197) |
| Widowed | 0.000 | 3.301(2.645, 4.120) |
| Never married | 0.000 | 0.647(0.510,0.820) |
| Perceived-Risk(Ref No risk) | 0.000 | 1 |
| Small Risk | 0.000 | 0.325(0.231,0 .457) |
| Moderate Risk | 0.000 | 0.447(0.335,0.597) |
| Great Risk | 0.574 | 0.916(0.676,1.242) |
| Age-first-sex(Ref Never had sex) | 0.000 | 1 |
| Under 11 | 0.000 | 8.524(3.569,20.358) |
| Between 12-14 | 0.000 | 10.162(5.774, 7.885) |

TABLE 2.1: Exploratory data analysis for HIV (continued)

| | | |
|---|---|---|
| Between 15-17 | 0.000 | 8.636(5.034,14.817) |
| Over 18 | 0.000 | 4.870(2.833,8.371) |
| **Biological characteristics** | | |
| Had STI(Ref Yes) | | 1 |
| No | 0.000 | 0.406(0.277,0.597) |
| Ever given birth(Ref Yes) | | 1 |
| No | 0.061 | 0.405(0.316,0.519) |
| **Behavioral Characteristics** | | |
| Partners in last 1 year (Ref No partner) | 0.000 | 1 |
| 1 partner | 0.034 | 1.021(0.314,0.812) |
| 2 partners | 0.665 | 1.232(0.771,3.433) |
| 3 or more partners | 0.999 | 2.455(1.759,11.233) |
| Travel away (didn't stay away) | 0.029 | 1 |
| Stayed away 1-2 times | 0.015 | 1.241(1.042,1.477) |
| Stayed away 3-5 times | 0.006 | 1.362(1.092,1.698) |
| Stayed away 6-10 times | 0.451 | 1.170(0.778,1.761) |
| Stayed away >11 times | 0.748 | 0.894(0.451,1.772) |

TABLE 2.2: Exploratory data analysis for HSV-2

| Variable | P-Value | Unadjusted OR |
|---|---|---|
| **Demographic characteristics** | | |
| Place of residence (Ref Rural) | | 1 |
| Urban | 0.001 | 0.823(0.746,0.907) |
| Age (Ref 15-19) | 0.000 | 1 |
| 20-24 | 0.000 | 2.745(2.254,3.343) |
| 25-29 | 0.000 | 4.374(3.591,5.329) |
| 30-34 | 0.000 | 6.794(5.559,8.303) |
| 35-39 | 0.000 | 8.299(6.739,10.220) |
| 40-44 | 0.000 | 9.389(7.538,11.694) |
| 45-49 | 0.000 | 8.641(6.936,10.765) |
| 50-54 | 0.000 | 8.378(6.592,10.649) |
| 55-59 | 0.294 | 8.661(6.720,11.162) |
| 60-64 | 0.173 | 5.751(4.279,7.729) |
| **Social Characteristics** | | |
| Wealth Quantile (ref poorest) | 0.051 | 1 |
| Second | 0.011 | 1.199(1.042,1.381) |
| Middle | 0.466 | 1.053(.916,1.212) |
| Fourth | 0.001 | 1.279(1.113,1.469) |
| Richest | 0.569 | 1.039(0.910,1.186) |
| Media access(Ref No) | | 1 |
| Yes | 0.821 | 1.010(0.924,1.104) |

TABLE 2.2: Exploratory data analysis for HSV-2 (continued)

| | | |
|---|---|---|
| Education level (Ref none) | 0.000 | 1 |
| Primary | 0.000 | 0.814(0.738,0.898) |
| Secondary | 0.000 | 0.704(0.610,0.813) |
| Higher | 0.000 | 0.457(0.381,0.548) |
| Marital status(Ref Married, 1 partner) | 0.000 | 1 |
| Married, +2partners | 0.000 | 2.381(2.042,2.778) |
| Divorced/separated | 0.000 | 1.904(1.607,2.256) |
| Widowed | 0.000 | 3.238(2.719,3.857) |
| Never married | 0.000 | 0.292(0.257,0.333) |
| Perceived-Risk(Ref No risk) | 0.000 | 1 |
| Small Risk | 0.000 | 0.452(0.371,0.551) |
| Moderate Risk | 0.000 | 0.581(0.483,0.699) |
| Great Risk | 0.675 | 0.957(0.778,1.177) |
| Age-first-sex(Ref Never had sex) | 0.000 | 1 |
| Under 11 | 0.000 | 12.572(7.554,20.922) |
| Between 12-14 | 0.000 | 18.384(13.685,24.697) |
| Between 15-17 | 0.000 | 15.053(11.477,19.743) |
| Over 18 | 0.000 | 9.797(7.487,12.818) |
| **Biological characteristics** | | |
| Had STI(Ref Yes) | | 1 |
| No | 0.000 | 0.556(0.407,0.760) |
| Ever given birth(Ref Yes) | | 1 |
| No | 0.000 | 0.187(0.163,0.215) |
| **Behavioral Characteristics** | | |
| Partners in last 1 year (Ref No partner) | 0.009 | 1 |
| 1 partner | 0.802 | 0.990(0.873,1.276) |
| 2 partners | 0.831 | 1.108(1.925,6.294) |
| 3 or more partners | 0.938 | 0.535(0.699,1.434) |
| Travel away (didn't stay away) | 0.000 | 1 |
| Stayed away 1-2 times | 0.000 | 1.251(1.133,1.380) |
| Stayed away 3-5 times | 0.000 | 1.468(1.289,1.672) |
| Stayed away 6-10 times | 0.017 | 1.324(1.052,1.665) |
| Stayed away >11 times | 0.198 | 1.258(0.887,1.786) |

## 2.3 Statistical Model

A univariate standard logistic model was used to test the association of each single covariate with the outcome variable (HIV and HSV-2 status).The association was considered significant at 5% significance level. These results are shown in Tables 2.1 and 2.2 .

Let $y_{ijk}$ be the disease $k$ status (0 or 1), $k = 1$ for HIV and $k = 2$ for HSV-2, for individual $j$ in county $i$:$i = 1, 2, \ldots, 46$. In this notation $y_{ij1} = 1$ if individual $j$ in county $i$ is HIV positive and zero otherwise and $y_{ij2} = 1$ if individual $j$ in county $i$ is HSV-2 positive and zero otherwise. This study assumes the dependent variable $y_{ijk}$ is Bernoulli distributed, i.e. $y_{ijk}|p_{ijk} \backsim Bernoulli(p_{ijk})$.

The vector $X_{ijk}$=$(x_{ij1}, x_{ij2}, \ldots, x_{ijp})'$ contains $p$ continuous predictors and

$$W_{ijk} = (w_{ij1}, w_{ij2}, \ldots, w_{ijr})'$$

contains $r$ categorical predictors with the first component accounting for the intercept. In this study, $p = 1$ since we only have one continuous variable, age and $r = 8$.

The unknown $E(y_{ijk}) = p_{ijk}$ relates to the predictors as follows:

$$h(p_{ij1}) = X^T\beta_1 + W^T\gamma_1, \text{ for HIV} \tag{2.1}$$

and

$$h(p_{ij2}) = X^T\beta_2 + W^T\gamma_2, \text{ for HSV-2} \tag{2.2}$$

Where $h(.)$ is the logit link function in both models, $\beta_1$ and $\beta_2$ are $p$ dimensional vector of regression coefficients for the continuous predictors, and $\gamma_1$ and $\gamma_2$ are $r$ dimensional vectors of regression coefficients for the categorical predictors. An extension to a semi parametric model utilizing the penalized regression spline approach and convolution model was employed in order to cater for both the non-linear effects of the continuous covariates and the spatial autocorrelation in the data.

The penalized regression spline approach relaxes the highly restrictive linear predictor by a more flexible semi-parametric predictor, defined as:

$$h(p_{ij1}) = \sum_{t=1}^{p} f_t(x_{ijt}) + f_{spat}(S_{i1}) + W^T\gamma_1, \text{ for HIV} \tag{2.3}$$

and

$$h(p_{ij2}) = \sum_{t=1}^{p} f_t(x_{ijt}) + f_{spat}(S_{i2}) + W^T\gamma_2, \text{for HSV-2} \tag{2.4}$$

The function $f_t(.)$ is a non-linear twice differentiable smooth function for the continuous covariate and $f_{spat}(S_i)$ is a factor that caters for the spatial effects of each county. This study utilized the convolution model which assumes that the spatial effect can be decomposed into two components: spatially structured and spatially unstructured components i.e. $f_{spat}(S_{ik}) = f_{str}(S_{ik}) + f_{unstr}(S_{ik})$, $k = 1, 2$ [Manda and Leyland, 2007; Ngesa et al., 2014a]. The spatially unstructured random effects cover the unobserved covariates that are inherent within the counties or the correlation within the counties e.g. common cultural practices, climate, cultures etc.

while the spatially structured random effect accounts for any unobserved covariates which vary spatially across the counties, this is called spatial autocorrelation and it is technically defined as the dependence due to geographical proximity. The final model is expressed as:

$$h(p_{ijk}) = \sum_{K=1}^{K} f_t(x_{ijt}) + f_{str}(S_{ik}) + f_{unstr}(S_{ik}) + W^T\gamma_k, \tag{2.5}$$

with $k = 1$ for HIV and $k = 2$ for HSV-2.

## 2.4 Parameter estimation

This study used a full Bayesian approach in estimation and parameters were assigned appropriate prior distributions as will be discussed in section 2.6 dedicated to prior distributions.

## 2.5 The Penalized regression spline

Several studies have discussed extensively the methods for estimating the smooth function $f_t(.)$ [Fahrmeir and Tutz, 2001; Hastie et al., 2001]. In this study we utilize the penalized regression splines proposed by Eilers and Marx [1996]. Here, the assumption is that the effect of the continuous covariates can be approximated using the polynomial spline. The assumption is that the smooth function $f_t(.)$ can be estimated by a spline of degree $l$ with $k$ equally spaced knots, $x_{p,min} = \Psi_{p1} <$

$\Psi_{p2} < \ldots < \Psi_{pk-1} < \Psi_{pk} = x_{p,max}$ giving:

$$f(x,\theta) = \phi_0 + \phi_1 x + \ldots \phi_p x^l + \sum_{k=1}^{K} b_k (x - \Psi_k)^l{}_+ \tag{2.6}$$

where, $\theta$=$(\phi_0, \phi_1, \ldots \phi_p, b_1, b_2, \ldots b_K)'$ and $(\Lambda - \Omega)_+$ is equal to $(\Lambda - \Omega)_+$ if $(\Lambda - \Omega)$ is positive and zero otherwise.

This study uses a quadratic spline ($l = 2$) with 20 knots to ensure flexibility and takes the $k^{th}$ knot to be defined as the sample quantile of the continuous predictors obtained by the probability equal to $\frac{k}{K+1}$. Green and Silverman [1993] suggested a roughness penalty $-\frac{1}{2}\lambda \int_{xmin}^{xmax} [f''(x)]^2 dx$ imposed in the log-likelihood to avoid getting a smooth function which "wiggles" too much, yielding the penalized log-likelihood function given by: $L = l(y, \theta, \gamma) - \frac{1}{2}\lambda \int_{xmin}^{xmax} [f''(x)]^2 dx$, where $\lambda$ dictates the balance between flexibility and smoothness.

## 2.6 Prior distributions

The nearest neighbor multivariate Gaussian Markov random field (GMRF) is used as a prior distribution for the spatially structured effects $\boldsymbol{f_{(srt)}(S_i)}$=$(f_{(str)}(S_{i1}), f_{(str)}(S_{i2}))^T$. This is specified as: $f_{(srt)}(S_i, S_j) \backsim MCAR(1, \sum)$ where, $\sum$ is the covariance matrix inducing correlation.

### 2.6.1 Multivariate Conditional Autoregressive (MCAR) Model

The development of the multivariate model is based on Mardia [1988] extension of Besag [1974] results to a multivariate setting. Mardia [1988] showed conditions under with the conditional multivariate distributions uniquely determine the corresponding multivariate joint pdf. Using these results Carlin and Banerjee [2003] developed the MCAR as follows. Let $\Phi^T = (\phi_1^T, \phi_2^T, \ldots, \phi_p^T)$, where each $\phi_i$ is an $n \times 1$ vector. Then $\Phi$ is an $np \times 1$ vector. Also $\Phi$ have a multivariate Gaussian distribution with mean 0 and dispersion matrix B, written as

$$P(\phi_i|\phi_i) = (2\pi)^{\frac{-np}{2}} |B|^{\frac{1}{2}} exp\left\{-\frac{1}{2}\Phi^T B \Phi\right\} \tag{2.7}$$

B is an $np \times np$ symmetric and positive definite matrix. It is informative to look at B as a $p \times p$ block matrix with $n \times n$ block $B_{ij}$. The full conditional distribution are given by

$$P(\phi_i|\phi_{-i}) \propto exp\left[-\frac{1}{2}\left(\phi_i - B_{ii}^{-1}\sum_{j\neq i}(-B_{ij})\phi_j\right)^T B_{ii}\left(\phi_i - B_{ii}^{-1}\sum_{j\neq i}(-B_{ij})\phi_j\right)\right] \tag{2.8}$$

This implies that $\phi_i|\phi_{-i} \sim N_n(B_{ii}^{-1}\sum_{j\neq i}(-B_{ij})\phi_j, B_{ii}^{-1})$. The full conditional probability density functions are

$$P(\phi_i|\phi_{-i}) = N_n\left(\sum_{j\neq i} C_{ij}\phi_j, \sum\nolimits_i\right), i = 1, 2, \ldots, p \tag{2.9}$$

where $\sum_i$ and $C_{ij}$ are $n \times n$ matrix analogues of $c_{ij}$ are $\sigma_i^2$ defined in. The matrix $\sum_i$ is also symmetric and positive definite in Appendix 2. We now write $\sum_i$ and

$C_{ij}$ in terms of B, the precision matrix of the joint distribution as $C_{ij} = -B_{ii}^{-1}B_{ij}$ and $\sum_i = B_{ii}^{-1}$. If we set $\sum$ to be a block diagonal matrix with $\sum_i$ blocks and C as a partitioned matrix with blocks $C_{ij}$ and $C_{ii} = 0_{n\times n}$, then

$$B = \sum^{-1}(I - C) \tag{2.10}$$

A propriety parameter $\alpha$ can be added into precision matrix in equation 2.9 to yield

$$B = \sum^{-1}(I - \alpha C) \tag{2.11}$$

For B to be symmetric then a condition to satisfy this is that $C_{ij}\sum_i = \sum_i C_{ij}^T$. Carlin and Banerjee [2003] denoted this distribution by $MCAR(C, \sum)$.

The unstructured spatial effects were assumed to follow a Multivariate Gaussian prior i.e. $f_{unstr}(S_i, S_j)|\tau^2_{unstr} \backsim MVN(0, \tau^2_{unstr}I)$, where $I$ is the identity matrix.

Inverse gamma distributions were assigned to the variance hyper parameters as:

$$\tau^2_{str} \backsim IG(0.0001, 0.0001)$$

and

$$\tau^2_{unstr} \backsim IG(0.0001, 0.0001)$$

The fixed effects coefficients were given the following prior distributions:

$$\phi_0, \phi_1, \ldots \phi_p \backsim N(0, 10^6), \lambda_1, \lambda_2, \ldots \lambda_r \backsim N(0, 10^6), b_k \backsim N(0, {\tau_b}^2)$$

and

$$\tau_b^2 \backsim IG(0.0001, 0.0001), \beta_1, \beta_2 \backsim N(0.01, 0.01) \text{ being the intercepts}$$

## 2.7 Posterior Distribution

The posterior distribution is obtained by updating the prior distribution with the observed data and hence it is the distribution of the parameters after observing the data. This posterior distribution is what gives samples for Bayesian inference. The Markov chain Monte Carlo (MCMC) method overcomes the problem of high dimensionality as it allows for direct sampling from this posterior distribution repeatedly and estimates such as the mean and median are calculated from these sample data summaries. Assuming conditional independence between the response variable and the hyper parameters, the posterior distribution for the Bernoulli model is given by:

$$\begin{aligned} P_{post}(\phi, \lambda, b, \tau^2|y) &\propto L(y|\phi, \lambda, b, \tau^2) P_{pri}(\phi, \lambda, b, \tau^2) \\ &= \prod_i \prod_j L(y_{ij}|\theta, \lambda, \tau^2) \prod_{k=1}^{p} [P(b_k|\tau^2{}_k) P(\tau^2{}_k)] \times \\ &\qquad \prod_{j=1}^{r} [P(\gamma_j|\tau^2{}_j) P(\tau^2{}_j)] \times \\ &P(f_{str}|\tau^2{}_{str}) P(\tau^2{}_{str}) P(f_{unstr}|\tau^2{}_{unstr}) P(\tau^2{}_{unstr}) \end{aligned} \tag{2.12}$$

All the analyses in this study were carried out using WinBUGS 14 [Spiegelhalter et al., 2007]. In the implementation, 20,000 Markov chain Monte Carlo (MCMC)

iterations for each model was run, with the initial 10,000 discarded to carter for the burn in period. The 10,000 iterations left were used for assessing the convergence of the MCMC and parameter estimation.

## 2.8 Model Diagnostics

The models were compared using the deviance information criterion (DIC) suggested by Spiegelhalter et al. [2002]. The best fitting model is one with the smallest DIC. The DIC value is obtained as: $DIC = \overline{D}(\theta) + pD$, where $\overline{D}$ is the posterior mean of the deviance that measures the goodness of fit while $pD$ gives the effective number of parameters in the model which penalizes for complexity of the model. In using the DIC, low values of $\overline{D}$ indicate a better fit while small values of $pD$ indicate model parsimony. One challenge with the DIC is, how big the difference in DIC values of two competing models needs to be in order to declare one model as being better than the other is not well defined. Studies have shown that a difference of 3 in DIC values between two models cannot be distinguished while a difference of between 3 and 7 can be weakly differentiated [Kazembe et al., 2008; Spiegelhalter et al., 2002].

## 2.9 Data Analysis

This study investigated four sets of models in order to get an insight on the effect of the covariates, the unobserved effects on the distribution and relationship between

HIV and HSV-2 in Kenya based on the female data. Studies have discussed these classes of models and their advantages over classical models [Hastie and Tibshirani, 1995; Wand et al., 2001].

$$\text{Model 1}: \quad logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma \text{ for HSV-2}$$

$$\text{Model 2}: \quad logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma + f_{unstr}(S_{i1}) \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma + f_{unstr}(S_{i2}) \text{ for HSV-2}$$

$$\text{Model 3}: \quad logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma + f_{str}(S_{i1}) \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma + f_{str}(S_{i2}) \text{ for HSV-2}$$

$$\text{Model 4}: \quad logit(\rho_{ij1}) = \beta_{01} + f(age) + W_{ij}^T\gamma + f_{unstr}(S_{i1}) + f_{str}(S_{i1}) \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W_{ij}^T\gamma + f_{unstr}(S_{i2}) + f_{str}(S_{i2}) \text{ for HSV-2}$$

**Model 1**: This is a model of fixed categorical covariates which are assumed to have linear effects on the response variable namely, education level, age at first sex, perceived risk, partners in the last one year, marital status, place of residence, STI status in the past one year, number of times one had stayed away from home in the past one year and one continuous covariate, age, modeled with a non-linear

smooth function. Results from [Johnson and Way, 2006; Mishra et al., 2007] supports modeling age with a non-linear smoothing prior. Model 1 does not take into account the spatially structured and the spatially unstructured random effects and the two diseases are modeled independently.

**Model 2**: This is an additive model that assumes linear effects of the categorical covariates listed in model 1 above, non-linear effect of the continuous covariate age and spatially unstructured random effects which cover the unobserved covariates that are inherent within the counties. Here the joint modeling is initiated by the multivariate normal distribution.

**Model 3**: This model explores the effect of the linear covariates listed in model 1 above, non-linear covariate age and spatially structured random effect which accounts for any unobserved covariates which vary spatially among counties. The joint modeling is initiated by the multivariate conditional autoregressive model.

**Model 4**: Examines the nonlinear effects of age, linear effects of the categorical covariates and a convolution of spatially structured and spatially unstructured random effects, and the joint modeling is initiated by both the multivariate normal distribution and the multivariate conditional autoregressive model.

TABLE 2.3: Nesting nature of the models under study

| Model | Nonlinear effect of age | Linear effects of categorical covariates | Spatially unstructured random effects | Spatially structured random effects |
|---|---|---|---|---|
| $M_1$ | ✓ | ✓ | - | - |
| $M_2$ | ✓ | ✓ | ✓ | - |
| $M_3$ | ✓ | ✓ | - | ✓ |
| $M_4$ | ✓ | ✓ | ✓ | ✓ |

## 2.10 Results

### 2.10.1 Model assessment and comparison

Table 2.3 gives the nesting nature of the models under study. Model 1 basically examines the linear and nonlinear effects of the covariates, model 2 extends model 1 to include spatially unstructured random effects, model 3 extends model 1 to include spatially structured random effects and finally model 4 is model 1 plus both structured and unstructured random effects.

Table 2.4 presents model diagnostics for the four fitted models. The model with the smallest DIC provides a better fit. However studies have reported that a difference of 3 in DIC between two models cannot be distinguished while a difference of between 3 and 7 can be weakly differentiated [Kazembe et al., 2008; Spiegelhalter et al., 2002]. This implies therefore that model 2 and model 4 are indistinguishable since the difference in their DIC is less than 3. We therefore present and discuss results based on model 4 as it captures both spatially structured and unstructured random effects i.e. provides more information than model 2. However it does point to the fact that unobserved effects are accounted for more by the unstructured random effects than the structured random effects.

TABLE 2.4: Nesting nature of the models under study

| | Model1 | | Model2 | | Model3 | | Model4 | |
|---|---|---|---|---|---|---|---|---|
| | HIV | HSV-2 | HIV | HSV-2 | HIV | HSV-2 | HIV | HSV-2 |
| Individual pD | 23.425 | 25.424 | 32.755 | 56.869 | 43.211 | 57.869 | 43.149 | 58.133 |
| Individual $\overline{D}(\theta)$ | 2447.41 | 6040.86 | 2319.64 | 5732.09 | 2312.85 | 5733.05 | 2308.84 | 5733.01 |
| Individual DIC | 2470.83 | 6066.29 | 2252.40 | 2252.40 | 2356.06 | 5790.91 | 2351.99 | 5791.14 |
| Total DIC | 8537.27 | | 8141.36 | | 8146.97 | | 8143.13 | |

## 2.11 Fixed effects

Table 2.5 gives the posterior estimates of the odds ratios (OR) and their corresponding 95% credible intervals (CI) for the categorical covariates which were assumed to have linear effects under the logit model on HIV and HSV-2 statuses based on model 4 statuses.

Place of residence, marital status, education level, perceived risk, age at first sex, number of partners in the last year, if an individual had STI in the last 12 months and the number of times an individual had stayed away from home in the last one year were found to be significantly associated with HIV and HSV-2 infection statuses.

## 2.12 HIV

Place of residence (urban/rural) was found to be associated with HIV infection among women. The odds of HIV infection among women staying in urban areas was 1.592 times as likely as that of women living in rural areas (OR: 1.592, 95% CI: 1.116 to 2.211). Marital status was also significantly associated with HIV infection. The odds of HIV infection among divorced/separated women was 1.78

Table 2.5: Odds ratios based on Model 4

| Covariates | HIV | HSV-2 |
|---|---|---|
| **Demographic characteristics** | | |
| Place of residence (Ref Rural) | 1 | 1 |
| Urban | 1.592(1.116,2.211) | |
| **Social Characteristics** | | |
| Marital status(Ref Married, 1 partner) | 1 | 1 |
| Married, +2partners | 0.923(0.623,1.320) | 1.934(1.532,2.427) |
| Divorced/separated | 2.780(1.810,4.091) | 2.504(1.818,3.365) |
| Widowed | 4.603(2.598,7.477) | 3.110(1.856,5.000) |
| Never married | 1.376(0.891,2.016) | 0.991(0.763,1.275) |
| Perceived-Risk(Ref No risk) | 1 | 1 |
| Small Risk | 0.493(0.315,0.722) | 0.665(0.511,0.835) |
| Moderate Risk | 0.536(0.363,0.754) | 0.705(0.549,0.869) |
| Great Risk | 0.873(0.590,1.239) | 0.955(0.729,1.201) |
| Age-first-sex(ref Over 18) | 1 | 1 |
| Under 11 | 2.702(0.846,6.095) | 2.196(0.966,4.342) |
| Between 12-14 | 1.691(1.153,2.393) | 2.055(1.604,2.575) |
| Between 15-17 | 1.407(1.063,1.851) | 1.610(1.373,1.866) |
| Stay away(ref >11 times)) | 1 | 1 |
| Didn't stay away | 1.282(0.514,2.594) | 1.220(0.712,2.046) |
| 1-2 times | 1.179(0.474,2.351) | 1.290(0.754,2.194) |
| 3-5 times | 1.725(0.681,3.469) | 1.437(0.838,2.472) |
| 6-10 times | 1.368(0.461,3.039) | 1.232(0.668,2.176) |
| Education(ref Higher) | 1 | 1 |
| Primary | 2.168(1.260,3.715) | 2.072(1.581,2.666) |
| Secondary | 2.343(1.274,4.086) | 1.808(1.346,2.383) |
| **Behavioral Characteristics** | | |
| Partners in last 1 year(3 or more)) | 1 | 1 |
| 1 Partners | 1.283(0.235,5.762) | 1.896(0.411,6.478) |
| 2 Partners | 1.992(0.323,8.993) | 2.528(0.507,8.682) |
| **Biological Characteristics** | | |
| STI(ref no) | 1 | 1 |
| Yes | 1.570(0.842,2.611) | 1.382(0.916,1.995) |
| Random effects | | |
| Spatially unstructured ($\tau_{unstr}$) | 0.143(0.000,0.645) | 0.167(0.012,0.533) |
| Spatially unstructured ($\tau_{str}$) | 0.141(0.024,0.982) | 0.159(0.412,1.323) |
| Spline Coefficients ($\tau_b$) | 5674(1003,7554) | 7683(870.8,9356) |
| Correlation (HIV-HSV-2) | 0.683(0.386,0.871) | |

times higher than women who were married with one partner (OR: 2.78, 95% CI: 1.81 to 4.091). Women who had never been married were found to be 1.376 as likely to be HIV positive as women who were married with one partner (OR: 1.376, 95%

CI: 0.8911 to 2.016), though not significant. Widowed women were 3.603 times more likely to be HIV positive than women who were married to one partner (OR: 4.603, 95% CI: 2.598 to 7.477). Those women who had some perceived risk of HIV infection (small risk, Moderate risk, Great risk) were less likely to be HIV positive than those who had no perceived risk. Age at first sex is negatively associated with HIV infection. The likelihood for HIV was higher for those women who had had their first sex before age 11 as compared to those who had had their first sex after age 18, but this was not significant as indicated by the odds ratio and its corresponding credible interval (OR: 2.702, 95% CI: 0.8462 to 6.095). The chance of testing positive for HIV was 0.691 times higher for women who had had their first sex between ages 12-14 years than those who had their first intercourse after age 18 ( OR: 1.691, 95% CI: 1.153 to 2.393). Education level was also found to be associated with HIV infection. Those women with no education were 1.425 times more likely to test positive for HIV than those with higher education (OR: 2.425, 95% CI: 1.425 to 4.199). The chance of HIV infection was lowest among women with higher education. Individuals who contracted an STI in the last 12 months were found to be 1.57 times as likely to test positive for HIV as those who had not (OR: 1.57, 95% CI: 0.8439 to 2.611).

## 2.13 HSV-2

Place of residence (urban/Rural) was found to be associated with HSV-2 infection. Women who resided in urban locations were 1.904 times as likely to test HSV-2 positive as those residing in rural areas (OR: 1.904, 95% CI: 1.549 to 2.313).

Marital status was also found to be associated with HSV-2 infection among women. The odds of testing positive for HSV-2 was 0.9912 times as less likely for those women who were never married as for those who were married with one partner (OR: 0.991, 95% CI: 0.763 to 1.275). Women who were married with more than one partner were 1.934 times as likely to test positive for HSV-2 as those who were married with one partner (OR: 1.934, 95% CI: 1.532 to 2.427). Divorced/separated women were 1.504 times more likely to test positive for HSV-2 than those women who were married with one partner (OR: 2.504, 95% CI: 1.818 to 3.365). Widowed women were most likely to test positive for HSV-2. Widowed women were 3.11 as likely to test positive for HSV-2 as those women who were married with one partner (OR: 3.11, 95% CI: 1.856 to 5.000). HSV-2 infection is positively associated with perceived risk. The chance of testing positive increased with increasing perceived risk. However, women who had some perceived risk were less likely to test positive for HSV-2 as compared to those who felt they had no risk. Women who perceived great risk of infection were 0.955 as less likely to test positive for HSV-2 as those who felt no risk at all, although this was not significant (OR: 0.955, 95% CI: 0.729 to 1.201). The likelihood of infection on women that had a perception of moderate risk was 0.705 as less likely as for those women who felt not at risk (OR: 0.7051, 95% CI: 0.549 to 0.869). Women who had their first intercourse below age 11 years were 1.196 times more likely to test positive for HSV-2 than those who had their first intercourse after age 18. The odds of women who had had their first sexual intercourse between ages 12 and 14 to be infected with HSV-2 were 2.055 times as high as those who had engaged in their first intercourse after age 18.

HSV-2 infection is negatively related with education. The likelihood of HSV-2 infection was 1.184 higher for those with no education as compared with those who had attained higher education, (OR: 2.184, 95% CI: 1.662 to 2.851). Women who had primary education were 1.072 times more likely to test positive for HSV-2 than those with higher education (OR: 2.072, 95% CI: 1.581 to 2.666). Another finding of this study is that those women with higher education qualification were less likely to test positive for both HIV and HSV-2.

## 2.14 Nonlinear effects of age

Figures 2.1 and 2.2 show the nonlinear association between age of an individual and HIV infection and age of an individual and HSV-2 infection. The figures give the posterior mean of the smooth function and their corresponding 95% CI. From the figures it is evident that there is a nonlinear relationship between age and HIV and HSV-2 infection. An assumption of linear relationship would have led to miss leading results and subsequently wrong interpretations. The chance of HIV infection increases with age up to a maximum age of about 30 years then starts declining with increase in age.

For HSV-2, the likelihood of infection increases with age up to a maximum age of about 40 years then starts to decline thereafter with increasing age. The results depict that the prevalence of HIV picks earlier in age than HSV-2. Early age at first sex often times leads to individuals developing risky sexual behaviors like having multiple partners and not using protection as the individual grows older increasing the chances of getting HIV or HSV-2 with increasing age. HIV and HSV-2 prevalence also increases with age from between age 15 and 30 as this is the time the youth is in risky behavior such as unprotected sex and having multiple partners. HIV and HSV-2 prevalence stagnates at 30 and 40 respectively before dropping and this could be assumed to be the age where women have either settled in marriage and are practicing safe sexual relationships or are becoming less active sexually hence the declining prevalence.

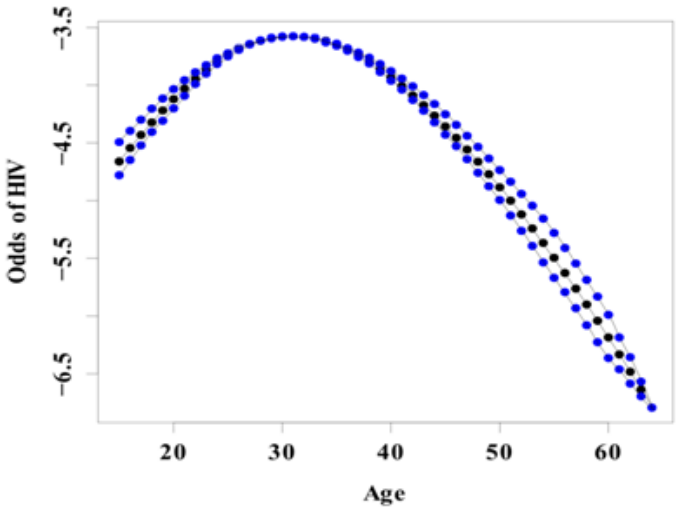



FIGURE 2.1: Estimated mean of the Nonlinear effect of age (in black) on HIV infection and the corresponding 95% credible interval(blue)

.

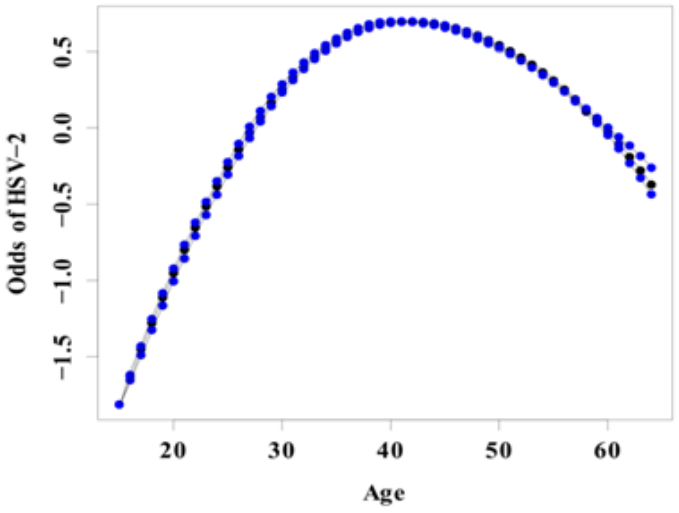



FIGURE 2.2: Estimated mean of the Nonlinear effect of age (in black) on HSV-2 infection and the corresponding 95% credible interval(blue)

.

## 2.15 Joint Spatial effects

We present spatial effects based on model 4. These are shown in Figures 2.3.

From the figures, counties with dark blue shading show high association of HIV

and HSV-2 infection while light blue shading indicates low association of HIV and HSV-2 infection. The figures show spatial variation of HIV and HSV-2. From Figure 2.3, counties in the Western and around Lake Victoria regions had high HIV prevalence. Counties in the North Eastern region had low HIV prevalence. Siaya, Homabay, Migori and Kisumu counties recorded the highest HIV prevelance. Figure 2.3 also shows that Siaya, Homabay, Migori, Kisumu and Turkana counties recorded the highest HSV-2 prevalence. HSV-2 prevalence was higher than HIV prevalence and more spread than HIV.

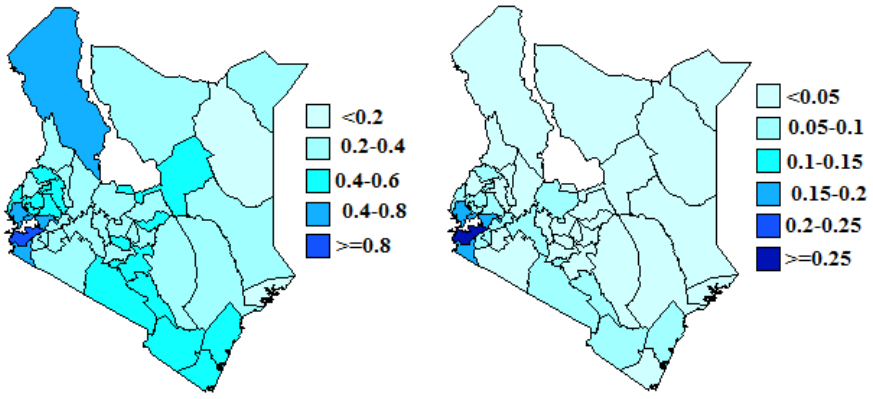



FIGURE 2.3: Residual spatial effect of county on HIV (on the left) and HSV-2 (on the right)

## 2.16 Discussion

This study utilized a full Bayesian approach to perform a semi-parametric spatial joint modeling of HIV and HSV-2 in Kenya. In particular, we used these methods to analyze the regional variation, risk factors of HIV and HSV-2 and the association between HIV and HSV-2. The works of Eilers and Marx [1996] on the B-splines, their construction and the penalized likelihood and that of Carroll and Ruppert [2003] on semi-parametric regression provided a basis for this study. In particular we modeled the non-linear effects using the penalized regression splines, in a semi parametric model paradigm, allowing for spatial variation in the response variables. The linearity assumption between the response variable and the covariates is limiting, unrealistic and can lead to misleading results in many situations. Semi parametric models are more flexible as they combine both parametric and semi parametric models hence enriching the standard parametric model by exploring the non-parametric domain while still keeping intact the linear structure [Besag and Kooperberg, 1995]. This flexibility improves the accuracy of the model and hence the results.

Age was found to have a non-linear effect on both HIV and HSV-2. i.e. an inverted "U" shape. The likelihood of HIV infection among women increases with age up to about age 30 then reduces thereafter with increasing age. On the other hand the likelihood of HSV-2 infection increases with age up to about age 40 and then starts declining with age. These findings were consistent with other studies [Johnson and Way, 2006]. The late peaking of HSV-2 could be attributed to its late detection as they have mild to no symptoms at all or their symptoms may be mistaken

for other conditions. This carries with it a negative public health implication in that, this is the age when the youth is most active and more willing to take risks. High prevalence in this age group implies high number of new infections and hence curbing HIV and HSV-2 becomes more difficult. Strategies to delaying age at first sex, practicing responsible sexual behavior will help reduce the prevalence of these two diseases.

The spatial effects in the model are modeled using a Gaussian Markov Random Field (GMRF) while the spatially unstructured random effects are modeled using a zero mean Gaussian process [Besag et al., 1991; Kazembe et al., 2008]. Bayesian and non-Bayesian methods have been proposed for joint disease modeling [Knorr-Held and Best, 2001; Langford et al., 1999]. The maximum likelihood (frequentist) approaches are not viable for these models due to the high complexity and intractability, hence the Bayesian inference, utilizing the MCMC techniques is highly favored [Ngesa et al., 2014b]. The computational limitations of the frequentist approach makes the Bayesian approach through the MCMC algorithm more appealing as it is less cumbersome to implement. Bayesian approaches allow for complex and flexible hierarchical modeling while providing more reliable estimates and predictions for many realistic epidemiological problems. While parameters are estimated similarly under the two methods, random effects variance estimates are generally attenuated under the frequentist approach compared to the Bayesian approach [Manda and Leyland, 2007].

Place of residence was found to be significantly associated with HIV and HSV-2 infection among women when controlled for other covariates. Women in urban

areas were more likely to be HIV and HSV-2 positive than women living in rural areas. Many studies have reported the effect of place of residence on HIV infection but with mixed conclusions [Johnson and Way, 2006; Kleinschmidt et al., 2007]. From our study, HSV-2 infection was also more prevalent in Turkana County which is mostly rural when considering the prevalence at county levels. These findings could be used to inform area specific approaches and campaign strategies to help curb the prevalence of these two diseases.

Marital status was also significantly associated with HIV and HSV-2 infection. Women who had been married before and then divorced, separated or widowed were more likely to test positive for HIV and HSV-2 than those who were married with one partner or never been married. Widowed women were more likely to test positive for HIV and HSV-2 than those who were married with one partner. This could be attributed to wife inheritance. Wife inheritance is a widespread cultural practice in sub-Saharan Africa that increases the risk of HIV acquisition and transmission [Amornkul et al., 2009; Kenya, 1997]. The life expectancy of females is higher than that of males in most cases and countries, with the gap between sexes steadying at 5 since 1990 WHO [2014], this in effect means that it is more likely that a man will die leaving behind his HIV/HSV-2 infected wife, and if she accepts to be inherited, she will pass it to her inheritor who will acquire the disease and pass it to the wife before dying and leaving them. These two widows will then be inherited by other individuals and the chain goes on. In most cases these inheritors engage in concurrent sex and are polygamous with some having more than 2 wives.

This study also found that age at first sex was negatively associated with HIV and HSV-2 infection. Those who had had their first sexual contact before age 11 were more likely to test positive for HIV and HSV-2 than those who had had their first intercourse after age 18. Other studies have found similar results [Ghebremichael et al., 2009]. This knowledge can help in designing of prevention programs not only aiming at delaying the age at first sex but also addressing the factors leading to early sexual practices.

Women who had had STI in the last 12 months were also more likely to test positive for HIV and HSV-2. This has been documented in various studies [Cohen, 1998; Røttingen et al., 2001]. Education level was found to be inversely related to HIV and HSV-2 infection. Those who had attained higher education qualification were less likely to test positive for HIV and HSV-2. This is consistent with other studies which reported similar results [Burgoyne and Drummond, 2008]. The introduction of free primary education and the subsequent subsidizing of secondary education is hoped to increase the number of people attaining higher education level [Adrienne and Mbiti, 2012].

HIV and HSV-2 infection were also found to be highly spatially correlated, and this was significant: (OR: 0.683, 95% CI: 0.386 to 0.871). This means counties with high HSV-2 prevalence had a high HIV prevalence too.

Spatial effects in the model account for unobserved variables that represent those variables that vary spatially. Identifying high prevalence areas and the relationship between HIV and HSV-2 can provide more insight that can be useful in coming up with tailor made campaigns and prevention strategies for specific regions. There

was evidence of spatial variation of HIV and HSV-2 infection among counties. The highest prevalence rate for HIV was observed in Western part of the country and around Lake Victoria likewise highest prevalence for HSV-2 was observed in Western region, around Lake Victoria and Turkana. Availability of free software like R and WinBUGS makes the establishing and testing epidemiological hypothesis easier and the implementation of these complex models cheaper.

The major limitation for this study was that the data used for county estimation was collected when the country was still based on the old administrative units (provinces) however these new administrative units (counties) were formed by combining several districts together. This made it easy for the county where an individual belongs to be allocated easily since each district belongs to only one county. The knots used in the penalized spline regression were assumed to be fixed and were calculated as quantiles from the continuous variable age. A more flexible analysis can allow the knots to be data driven [DiMatteo et al., 2001]. Another limitation for this study is that the data used for this study are from 2007 survey. A more recent KAIS survey has been conducted although it had not yet been made public by the time this study was carried out. The models introduced in this study can be replicated in other countries with similar data. Future work can also allow for time trends to exploit subsequent surveys that collect data on the two infections.

# Chapter 3

# Spatial Modeling of HIV and HSV-2 Among Women in Kenya with Spatially Varying Coefficients

## 3.1 Introduction

The World Health Organization (WHO) places at more than 1 million, the number of people who acquire sexually transmitted infections (STI) daily. By 2013 more than 530 million (about 7.5%) had the virus that causes genital herpes or the herpes simplex virus type 2 (HSV-2) [WHO, 2013]. Out of these, it is estimated that about 123.7 million or 23% resided in sub-Saharan Africa, among whom 63%

were women [Looker et al., 2008]. HSV-2 prevalence in the age group 15-49 in the sub-Saharan Africa region ranges from 30% to 80% among women and from 10% to 50% among men [Weiss, 2004]. There were about 35 million individuals living with HIV in sub-Saharan Africa by the end of 2013 with 2.1 million new infections [UNAIDS, 2013]. HSV-2 is associated with a two to three-fold increased risk of HIV acquisition and an up to five-fold increased risk of HIV transmission per-sexual act, and may account for 40% to 60% of new HIV infections in populations where HSV-2 has a high prevalence [Looker et al., 2008].

HIV and HSV-2 share common risk factors e.g. education level, place of residence, and age among others. Therefore understanding the spatial distribution, the dynamics and the underlying factors that propagate the spread of these diseases will help in ultimately winning the war against them. STIs can have serious consequences beyond the immediate impact of the infection itself, through mother-to-child transmission (MTCT) of infections and chronic diseases. Drug resistance is a major threat to reducing the impact of STIs worldwide [WHO, 2013].

The national HIV and HSV-2 prevalence rates in Kenya within the adult population (15-64 years) were estimated to be as high as 5.6% in men and 7.1% in women NASCOP [2012], with a wide gender and geographical variation. The North Eastern region had HIV prevalence of as low as 2.1% while regions around Lake Victoria and the Western region had prevalence ranging from between 13%-25% [NASCOP, 2007]. HIV and HSV-2 prevalence by age have a non-linear relationship assuming an inverted U shape [Ghebremichael et al., 2009; NASCOP, 2007]. HIV prevalence increases with age until it plateaus at between ages 25-35, then starts decreasing

with increasing age. HSV-2 prevalence increases with age up to between ages 35-45 then begins to decline with increasing age.

In the conventional generalized linear regression models applied to spatial data, many studies have assumed stationarity in that the same stimulus of a disease predictor provokes the same response in all parts of the study region [Hastie and Tibshirani, 1995; Mishra et al., 2007]. The models developed in chapter 2 ride on this assumption. This assumption may be highly restrictive for spatial processes. This may be as a result of sampling variation, intrinsically different relationships across space e.g. attitudes, cultures, preferences and model misspecification. Some of these causes are inherent effects that need to be taken care of in the modeling process. It is therefore realistic to assume that the regression coefficients vary across space [Fotheringham et al., 2003]. The results obtained in chapter 2 may be enriched by further investigating the effects of the covariates as one moves across space. The issue of spatial non-stationarity can be addressed by allowing the relationships we are measuring to vary over space through the geographically weighted regression (GWR) model where the weights applied to observations in a series of locally weighted regression models across the study area are determined by a spatial kernel function as suggested by Fotheringham et al. [2003], or the Bayesian spatially varying coefficients process (BSVCP), where spatially varying coefficients are modeled as a multivariate spatial process [Wheeler and Waller, 2009]. In the BSVCP model as discussed by Assunçao [2003], the covariates are allowed to vary spatially by assigning its coefficients the Bayesian autoregressive (BAR), simultaneous autoregressive (SAR) or the conditional autoregressive

(CAR) model [Assunçao, 2003]. Assunçao [2003] applied the BSVCP to model agricultural development in Brazil. The model showed significant regional differences in agricultural development [Assunçao et al., 1998]. Evidence of spatially varying parameters, even against strong prior belief on the absence of such variation, can be indicative of spatial differences of database collection procedures e.g. large differences on underreporting rates [Assunçao, 2003]. Several studies that use the linear predictor class of models including both the general and generalized linear models assume that all the covariates in the study have a linear relationship with the response variable. This linear relationship may not hold for all variables as in our case; age, which has a non-linear relationship with the response variable. Our objective is to perform a spatial modeling analysis while relaxing the stationarity and the linearity assumption by respectively employing the BSVCP and the random walk model of order 2 to model HIV and HSV-2 among women in Kenya.

## 3.2 Methods

### 3.2.1 Data

The data for this study was obtained from the Kenya AIDS Indicator Survey (KAIS) which was carried out by the Kenyan government with financial support from the United States President's Emergency Plan for AIDS Relief (PEPFAR) and the United Nations (UN). The main aim of the survey was to obtain high quality data on the prevalence of HIV and Sexually Transmitted Infections (STI) among adults and to assess the knowledge of HIV and STIs in the population.

The sampling frame for KAIS was the National Sample Survey and Evaluation Programme IV (NASSEP IV). It consisted of 1800 clusters comprising 1260 rural and 540 urban clusters; of these, 294 rural and 141 urban clusters were sampled for KAIS. The overall design for KAIS 2007 was a stratified, two-stage cluster sampling design. The first stage involved selecting clusters from NASSEP IV, and the second stage involved the selection of households for KAIS with equal probability in the urban-rural strata within the districts. A sample of 415 clusters and 10,375 households were systematically selected for KAIS. A uniform sample of 25 households per cluster was selected using an equal probability systematic sampling method. The survey was twofold: A household questionnaire was used to collect the characteristics of the living environment and an individual questionnaire to collect information on demographic characteristics and the knowledge of HIV and STIs on men and women aged 15-64 years. A representative sample of households and individuals was selected from eight provinces in the country. Each individual was asked for consent to provide a venous blood sample for HIV and HSV-2 testing. More information on survey methodologies used in collecting the data is found in the final KAIS, 2007 report NASCOP [2007]. This study uses the 2007 data even though a new round of KAIS, 2012 NASCOP [2012] has been done. The final release of this new data had not been made hence the data was not available for use. This study uses the women's data from the KAIS, 2007 survey. Information from 4864 women, aged 15-64 years who had provided venous blood for HIV and HSV-2 testing and also had full covariate information was used in the analysis. In the data, age was captured as both categorical and continuous while all other covariates were categorical. Readers are directed to the KAIS, 2007

report NASCOP [2007] for more information. An initial exploratory data analysis was carried out using a univariate standard logistic regression model to determine the association of each single covariate with the outcome variable (HIV and HSV-2 status). These variables were categorized into four groups, namely: demographic, social, biological and behavioral [Ngesa et al., 2014b].

From this initial analysis, education level, age at first sex, perceived risk, partners in the last one year, marital status, place of residence, STI status in the last one year and age of the respondent were found to be associated with HIV and HSV-2 infection. The choice of covariates to be included in the model is vital for inference. By first fitting a standard logistics regression we assume, although not universally the case, that ignoring spatial correlation leads to spurious significance hence one is unlikely to miss any important covariates. There are other methods that could be used for choosing the covariates to be included in the model. The Bayesian model averaging (BMA) averages over all the possible combinations of the covariates and ranks the models by their Bayesian Information Criterion (BIC) and the covariates in order of importance by giving their posterior inclusion probability. The combination with the lowest BIC can then be used in the analysis, or the covariates with the highest PIP in the best model can be used.

## 3.3 Statistical Model

The covariates were tested for significance by fitting a univariate standard logistic model between each single covariate with the outcome variables (HIV and HSV-2

status). The association was considered significant at 5% significance level.

Let $y_{ijk}$ be the disease $k$ status (0 or 1), $k = 1$ for HIV and $k = 2$ for HSV-2, for individual $j$ in county $i$: $i = 1, 2, \ldots, 46$, $y_{ij1} = 1$ if individual $j$ in county $i$ is HIV positive and zero otherwise and $y_{ij2} = 1$ if individual $j$ in county $i$ is HSV-2 positive and zero otherwise. This study assumes the dependent variables $y_{ij1}$ and $y_{ij2}$ are univariate Bernoulli distributed, i.e. $y_{ij1}|p_{ij1} \backsim Bernoulli(p_{ij1})$ and $y_{ij2}|p_{ij2} \backsim Bernoulli(p_{ij2})$

The continuous predictors are contained in the vector $X_{ijk}$=$(x_{ij1}, x_{ij2}, \ldots, x_{ijp})'$ while $W_{ijk}$=$(w_{ij1}, w_{ij2}, \ldots, w_{ijr})'$ contains $r$ categorical predictors with the first component accounting for intercept. In this study, $p = 1(age)$ and $r = 8$. The unknown $E(y_{ijk}) = p_{ijk}$ relates to the predictors as follows:

$$h(p_{ij1}) = X^T\beta_1 + W^T\gamma_1, \text{ for HIV} \tag{3.1}$$

and

$$h(p_{ij2}) = X^T\beta_2 + W^T\gamma_2, \text{ for HSV-2} \tag{3.2}$$

Where $h(.)$ is the logit link function, $\beta_1$ and $\beta_2$ are $p$ dimensional vector of regression coefficients for the continuous predictors, and $\gamma_1$ and $\gamma_2$ are $r$ dimensional vector of regression coefficients for the categorical predictors. A random walk model of order 2 (RW2) and a convolution model were employed in order to cater

for both the non-linear effects of the continuous covariates and the spatial autocorrelation in the data.

$$h(p_{ij1}) = \sum_{t=1}^{p} f_t(x_{ijt}) + f_{spat}(S_{i1}) + W^T\gamma_1, \text{ for HIV} \tag{3.3}$$

and

$$h(p_{ij2}) = \sum_{t=1}^{p} f_t(x_{ijt}) + f_{spat}(S_{i2}) + W^T\gamma_2, \text{ for HSV-2} \tag{3.4}$$

The function $f_t(.)$ is a non-linear twice differentiable smooth function for the continuous covariate and $f_{spat}(S_{ik})$ is a factor that caters for the spatial effects of each county. This study utilized the convoluted spatial structure which assumes that the spatial effect can be decomposed into two components: spatially structured and spatially unstructured i.e. $f_{spat}(S_{ik}) = f_{str}(S_{ik}) + f_{unstr}(S_{ik})$, $k = 1, 2$ [Manda and Leyland, 2007; Ngesa et al., 2014b]. The spatially unstructured random effects cover the unobserved covariates that are inherent within the counties or the correlation within the counties e.g. common cultural practices, climate, cultures etc. while the spatially structured random effect accounts for any unobserved covariates which vary spatially among counties. This is called spatial autocorrelation and it is technically defined as the dependence due to geographical proximity. Thus the final model is expressed as:

$$h(p_{ijk}) = \sum_{t=1}^{p} f_t(x_{ijt}) + f_{str}(S_{ik}) + f_{unstr}(S_{ik}) + W^T\gamma_k, \tag{3.5}$$

with $k = 1$ for HIV and $k = 2$ HSV-2

## 3.4 Parameter Estimation

This study used a full Bayesian estimation approach where parameters were assigned prior distributions as will be discussed in section 3.7 .

## 3.5 Non-linear effects

Several studies have discussed extensively the methods for estimating the smooth function $f_t(.)$ [Eilers and Marx, 1996; Fahrmeir and Tutz, 2001]. The penalized regression splines model proposed by Eilers and Marx [1996] for example is commonly used. Here, the assumption is that the effect of the continuous covariates can be approximated using the polynomial spline. They assumed that the smooth function $f_t(.)$ can be estimated by a spline of degree $l$ with $k$ equally spaced knots, $x_{p,min} = \Psi_{p1} < \Psi_{p2} < \ldots < \Psi_{pk-1} < \Psi_{pk} = x_{p,max}$. Many studies have explored the relationships between the Gaussian Markov Random Fields (GMRF) and smoothing splines [Fahrmeir and Knorr-Held, 1997; Fahrmeir and Wagenpfeil, 1996]. In this study we used the random walk model for estimating the smooth function $f_t(.)$. This is briefly discussed in Appendix 1.

## 3.6 Spatially Varying Coefficients

As stated before, many studies have been done with the assumption that the relationship between the explanatory variable and the response variables in a regression model are constant across the study region [Hastie and Tibshirani, 1995;

Mishra et al., 2007]. This assumption is unrealistic for spatial processes as factors such as sampling variation, different relationships across space e.g. attitudes, preferences, culture etc. contribute to a different response to the same stimuli as one moves across space. Two competing spatially varying models are the GWR and the BSVCP. The GWR addresses this by estimating $\boldsymbol{\beta}'\boldsymbol{s}$ by the weighted least squares method, where more emphasis in terms of weights are placed on the observations which are close to location $i$ , since it is assumed that the observations close to $i$ exert more influence on the parameter estimates at location $i$ than those farther away as noted by Fotheringham et al. [2003]. The weighting schemes can be fixed or adaptive. In the fixed scheme, observations that are within some distance $d$ are given the weight of 1 while those farther away beyond some distance $d$ from location $i$ are given a weight of zero. Under the adaptive scheme, weights inside some radius $d$ are made to decrease monotonically to zero as the radius increases. In this study we used the BSVCP (Appendix 2) model to relax the stationarity assumption. The covariates are allowed to vary spatially by assigning its coefficients the conditional autoregressive (CAR) model [Assunçao, 2003].

## 3.7 Priors for the spatial components

The prior for the structured random effects was defined to follow the CAR model while for the unstructured random effects, the independently and identically distributed normal distribution.

## 3.8 Posterior distribution

This is the distribution of the parameters after observing the data. The posterior distribution is obtained by updating the prior distribution with the observed data. Since our study is fully Bayesian, inference is made by sampling from this posterior distribution. Markov Chain Monte Carlo (MCMC) is the most common approach to do inference for latent Gaussian models however this method is slow and performs poorly when applied to such models [Rue et al., 2009]. This poor performance is usually due to the fact that convergence of the chain takes a long time and at times it may be difficult to identify and prove that the chain has converged. The other reason why MCMC performs poorly when applied to such models is the speed at which the chain explores the target equilibrium distribution. This is referred to as mixing. It is desirable to have rapid mixing and so therefore have fast convergence. The Integrated Nested Laplace (INLA) criterion is a relatively new technique developed to circumvent these shortfalls [Rue et al., 2009]. The posterior distribution for the latent Gaussian model is:

$$\pi(x,\theta|y) \propto \pi(\theta)\pi(x|\theta)\prod_{i\in I}\pi(y_i|x_i,\theta)$$

$$\propto \pi(\theta)|Q(\theta)|^{\frac{n}{2}}exp\left(-\frac{1}{2}x^TQ(\theta)x+\sum_{i\in I}log\pi(y_i|x_i,\theta)\right), \tag{3.6}$$

Where $x$ is the class of latent fields, $\theta$ is the set of hyper parameters and $y$ is the data. In the INLA approach, the posterior marginals of interest are:

$$\pi(x_i|y) = \int \pi(x_i|\theta, y)\pi(\theta|y)\, d\theta \tag{3.7}$$

and

$$\pi(\theta_j|y) = \int \pi(\theta|y)\, d\theta_{-j}, \tag{3.8}$$

and these are used to construct the nested approximations:

$$\widetilde{\pi}(x_i|y) = \int \widetilde{\pi}(x_i|\theta, y)\widetilde{\pi}(\theta|y)\, d\theta \tag{3.9}$$

and

$$\widetilde{\pi}(\theta_j|y) = \int \widetilde{\pi}(\theta|y)\, d\theta_{-j} \tag{3.10}$$

The analysis in this study were carried out using the R software with the INLA package.

## 3.9 Model Diagnostics

The models were compared using the deviance information criterion (DIC) suggested by [Spiegelhalter et al., 2002]. The best fitting model is one with the smallest DIC. The DIC value is obtained as: $DIC = \overline{D}(\theta) + pD$, where $\overline{D}$ is the posterior mean of the deviance that measures the goodness of fit while $pD$ gives the effective number of parameters in the model which penalizes for complexity

of the model. In DIC, low values of $\overline{D}$ indicate a better fit while small values of $pD$ indicate model parsimony. One challenge with the DIC is, how big the difference in DIC values of two competing models needs to be in order to declare one model as being better than the other is not well defined. Studies have shown that a difference of 3 in DIC between two models cannot be distinguished while a difference of between 3 and 7 can be weakly differentiated [Kazembe et al., 2008; Spiegelhalter et al., 2002].

## 3.10 Data Analysis

The following sets of models were investigated in order to understand the effect of the observed covariates and unobserved effects on the distribution of HIV and HSV-2 in Kenya among the female population

$$\text{Model 1} : logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma \text{ for HSV-2}$$

$$\text{Model 2} : logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma + f_{unstr}(S_{i1}) \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma + f_{unstr}(S_{i2}) \text{ for HSV-2}$$

$$\text{Model 3} : logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma + f_{str}(S_{i1}) \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma + f_{str}(S_{i2}) \text{ for HSV-2}$$

$$\text{Model 4}: logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma + f_{unstr}(S_{i1}) + f_{str}(S_{i1}) \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma + f_{unstr}(S_{i2}) + f_{str}(S_{i2}) \text{ for HSV-2}$$

**Model 1**: This is a model of fixed categorical covariates which are assumed to have linear effects on the response variable namely, education level, age at first sex, perceived risk, partners in the last one year, marital status, place of residence, STI status in the past one year, number of times one had stayed away from home in the past one year and one continuous covariate, age, modeled with a non-linear smooth function: the RW2 model. Model 1 does not take into account the spatially structured and the spatially unstructured random effects and the two diseases are modeled independently.

**Model 2**: This is an additive model that assumes linear effects of the categorical covariates listed in model 1 above, non-linear effect of the continuous covariate age and spatially unstructured random effect which caters for the unobserved covariates that are inherent within the counties specified by the identically and independently distributed (iid) normal distribution.

**Model 3**: This model explores the effect of the linear covariates listed in model 1 above, non-linear covariate age and spatially structured random effect which accounts for any unobserved covariates which vary spatially among counties, specified by the CAR model.

**Model 4**: Examines the effects of the nonlinear effects of age, linear effects of the categorical covariates and a convolution of spatially structured and spatially

unstructured random effect, specified by the CAR model and the iid normal distribution respectively.

**Models 5-8** are similar to models 1-4 respectively, the only difference is that the regression coefficients $\gamma$ in these models are assumed to vary spatially and are assigned CAR priors.

## 3.11 Results

### 3.11.1 Model assessment and comparison

TABLE 3.1: Stationary model

| | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|---|---|
| | HIV | HSV-2 | HIV | HSV-2 | HIV | HSV-2 | HIV | HSV-2 |
| pD | 12.83 | 13.71 | 38.50 | 51.28 | 38.84 | 51.17 | 38.47 | 51.28 |
| $\overline{D}(\theta)$ | 2509.47 | 6202.836 | 2366.25 | 5827.92 | 2367.05 | 5827.87 | 2366.24 | 5827.90 |
| Total DIC | 2522.30 | 6216.54 | 2404.75 | 5879.20 | 2405.89 | 5879.04 | 2404.71 | 5879.18 |

TABLE 3.2: Spatially Varying coefficients

| | Model 5 | | Model 6 | | Model 7 | | Model 8 | |
|---|---|---|---|---|---|---|---|---|
| | HIV | HSV-2 | HIV | HSV-2 | HIV | HSV-2 | HIV | HSV-2 |
| pD | 32.43 | 61.70 | 38.68 | 69.58 | 39.05 | 68.57 | 38.58 | 69.34 |
| $\overline{D}(\theta)$ | 2430.02 | 5932.32 | 2365.98 | 5773.91 | 2365.77 | 5779.05 | 2365.80 | 5773.84 |
| Total DIC | 2462.45 | 5994.02 | 2404.66 | 5843.49 | 2404.82 | 5847.62 | 2404.38 | 5843.17 |

Table 3.1 shows the DICs for the four separately fitted models for HIV and HSV-2. These four models were assumed to have stationary coefficients. Table 3.2 shows the DICs for the four separate models with spatially varying coefficients. The model with the smallest DIC provides the best fit. Studies have however reported that two models with a difference of 3 or less in DIC are indistinguishable, while a

difference of between 3 and 7 suggests that the two models are weakly distinguishable [Spiegelhalter et al., 2002]. From the tables, all the spatially varying models have a lower DIC as compared with their corresponding stationary models. For HIV, Spatially varying coefficient models 6, 7, 8 are not significantly different form each other and from the corresponding stationary model counterparts as the difference in DIC is less than 3. This suggests that the covariates for HIV do not vary significantly across space. For HSV-2, the spatially varying models are significantly better than the stationary models since they have significantly lower DICs. This suggests that the covariates may provoke different responses across space for HSV-2. Spatially varying model 8 provided the best fit for HSV-2.

We therefore present and discuss the results based on model 8 for both HIV and HSV-2, which allows the covariates to vary spatially by the CAR model and also captures the structured and the unstructured random effects.

### 3.11.2 Spatially Varying Effects

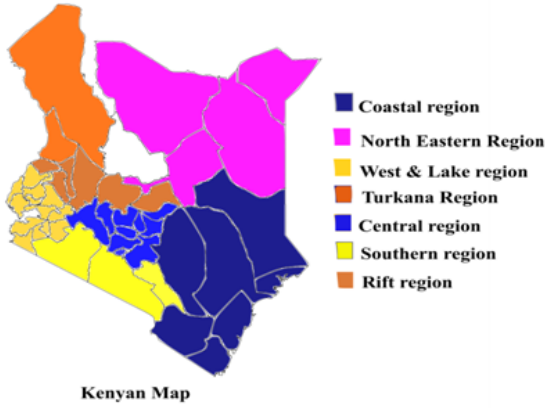



FIGURE 3.1: A map of geographical regions of Kenya

The DIC values indicate that the SVC models are better than the stationary ones, especially for the HSV-2 model. The choropleth maps show the varying effects of each covariate across space. Figure 3.1 shows the map of Kenya. Kenya is positioned on the equator on Africa's East Coast. The administration units in Kenya were provinces before changing to counties after the 2010 promulgation of the constitution. There are 47 counties in Kenya but this study discusses results from 46 counties as the KAIS 2007 was not conducted in Samburu County due to insecurity.

### 3.11.3 Spatially Varying Effects

### 3.11.4 HIV

Though the SVC models for HIV provided almost the same fit as their stationary counterparts since their DICs were almost equal, the choropleth maps suggest that the effects of some of the covariates indeed do vary across space. The effect of education on HIV prevalence among women was more in the North Eastern, Coastal, Southern regions and parts of Central region indicated by the yellow to orange shading in the choropleth map in Figures 3.2 and 3.3. Age at first sex also had a greater effect in those parts where education had greater effects as compared with the other parts of the country suggesting a correlation between education and age at first sex. The effect of number of partners had in the last one year was almost the same across the country except for some parts of West, Lake and Central region, where the effect was greater indicated by yellow/orange

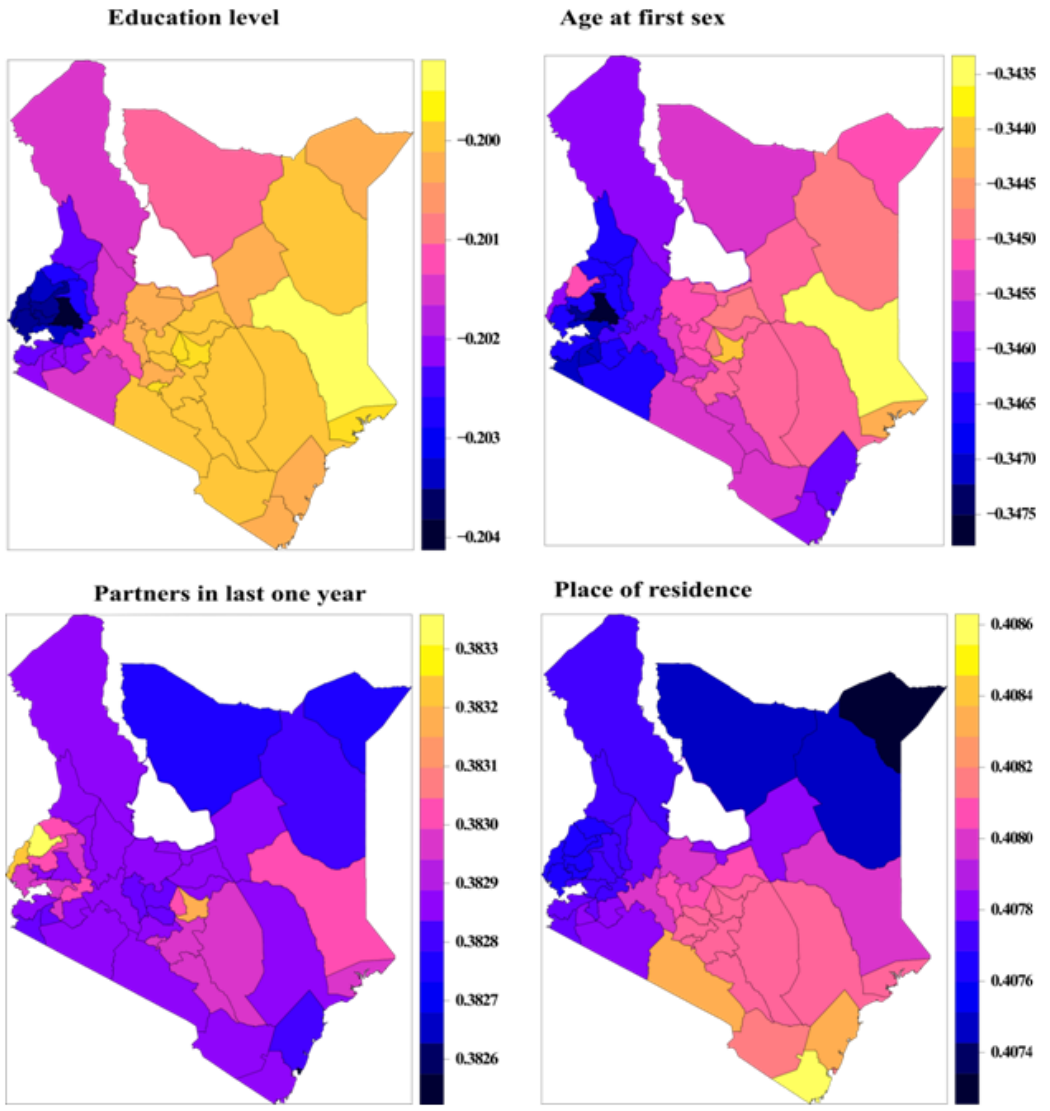



FIGURE 3.2: Figures of spatially varying effects of covariates on HIV status

shading on the choropleth map in Figures 3.2 and 3.3. The effect of frequency of travel away was also evident in the North Eastern, Coastal and Southern regions and parts of Central region while that of marital status was dominant in the Lake region.

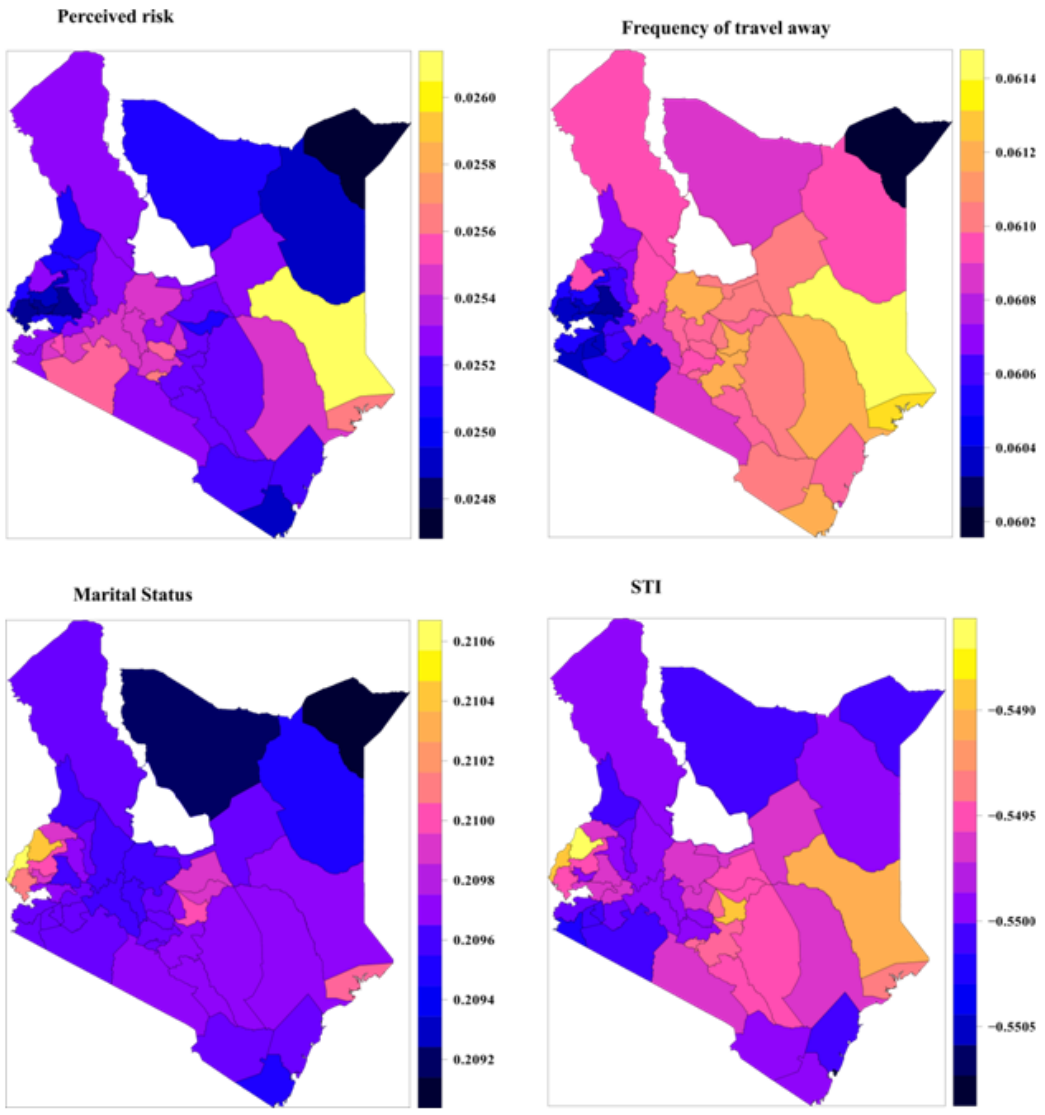



FIGURE 3.3: Figures of spatially varying effects of covariates on HIV status (continued)

### 3.11.5 HSV-2

The effect of education on HSV-2 status was lower in North Eastern and parts of Rift region than most of the other parts of the country shown by the blue shading on the map in Figures 3.4 and 3.5. Age at first sex also had a greater bearing in the Costal and some parts of North Eastern, parts of Rift and West and Lake regions (pink/yellow shading) suggesting either early marriages or child prostitution. The highest rates of arranged marriages among adolescent girls in

Kenya are found in Northeastern (73 percent), Rift Valley (22 percent), and Coast (21 percent) provinces [CBS, 2004]. A study by the University of Chicago in Kenya and Zambia found that among 15-to-19 year old girls who are sexually active, being married increased their chance of HIV and other STIs by more than 75 percent. This is due to the fact that most of these young marrieds were more likely to be in a polygamous union [Clark, 2004]. The number of partners had in the last one year had more effect on HSV-2 status in the West and Lake regions and some parts of the Central and Southern regions depicted by yellow shading on Figures 3.4 and 3.5, while the number of partners had in the last one year had less effect in the regions with blue shading. The effect of place of residence (rural/urban) also varied spatially. The effects were higher in the West and Lake regions, Southern and parts of Central and Coastal and Rift regions depicted by yellow shading on Figure 3.4 and 3.5. The urban environment is very different from the rural one. For counties near the capital, say those in central region etc., the effect of place of residence vary spatially substantially. In remote counties, the difference between urban and rural environment are almost indistinguishable and therefore the expected result.

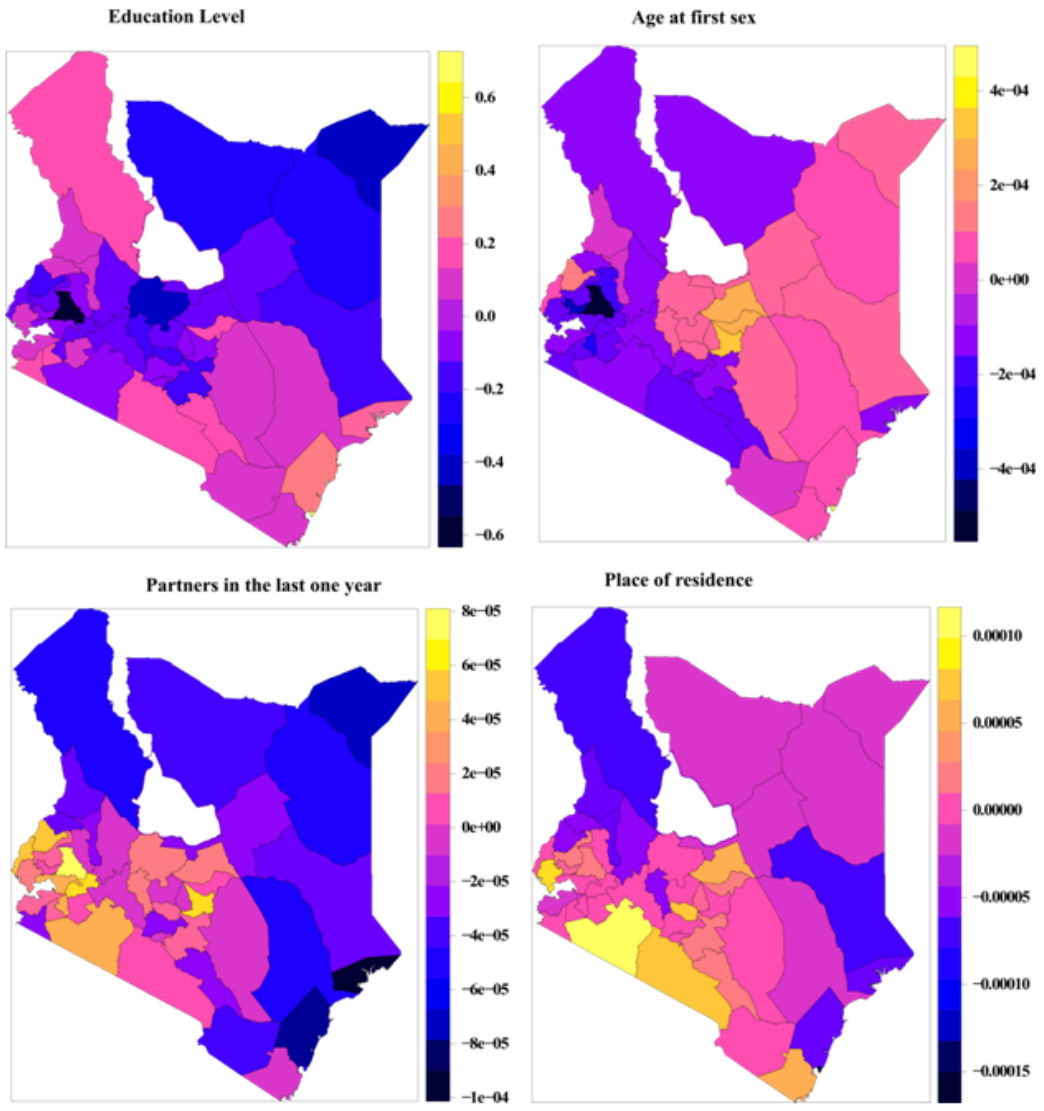



FIGURE 3.4: Figures of spatially varying effects of covariates on HSV-2 status

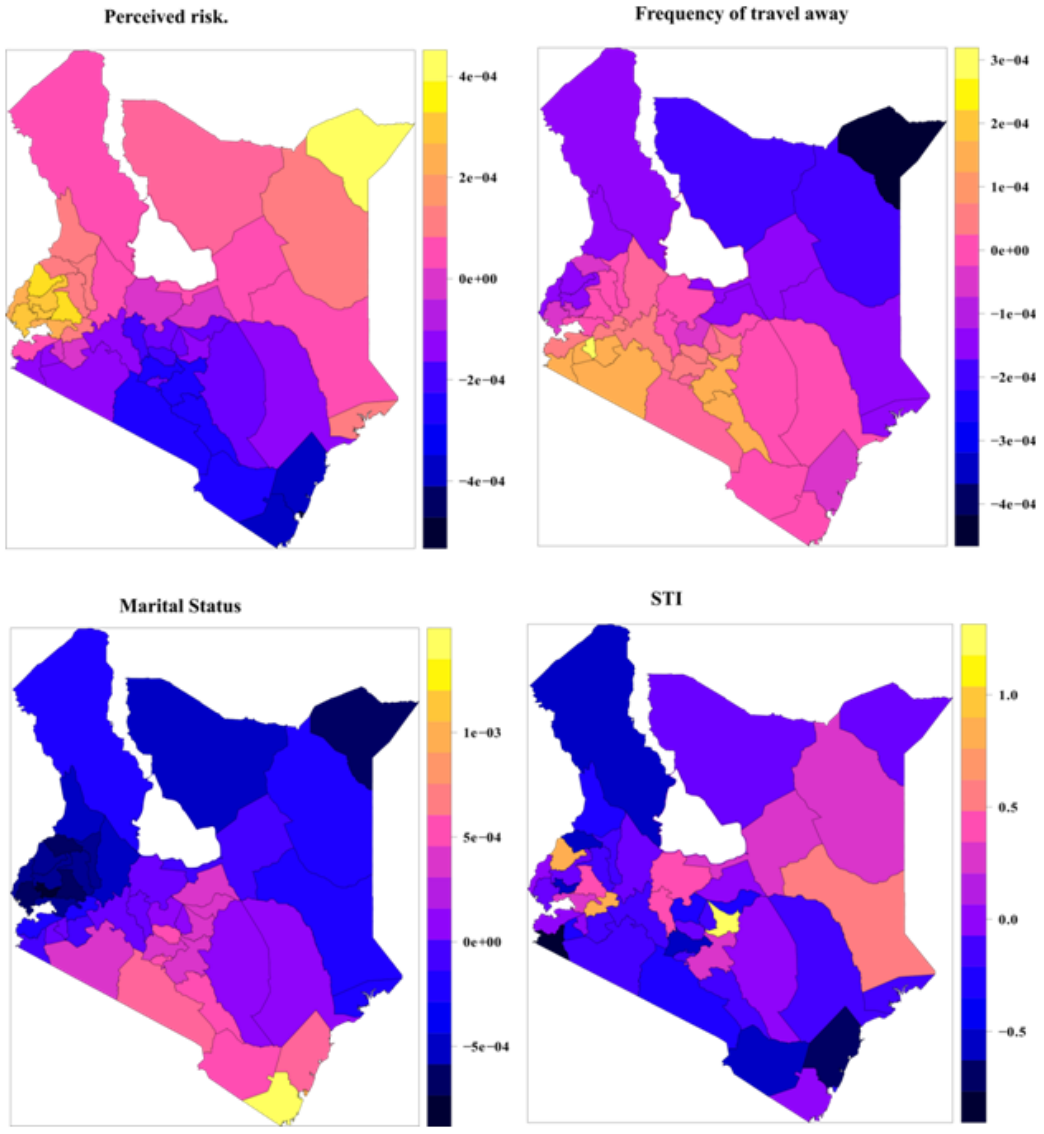



FIGURE 3.5: Figures of spatially varying effects of covariates on HSV-2 status (continued)

The spatial effects based on model 4 indicate that HIV prevalence varies spatially with areas in the Central, West and Lake regions recording the highest prevalence. HIV prevalence is lowest in the North Eastern region (shown by blue shading on Figure 3.6) with some significant prevalence in some parts of the Coastal region. On the other hand , HSV-2 prevalence is also highest in the West and Lake regions, but also generally high across the country as shown in the yellow/orange shading on the choropleth map in Figure 3.6. Most regions with high HSV-2 prevalence had also a high HIV prevalence. Identifying the effects of individual covariates on each area can go a long way in informing strategies to deal with HIV and HSV-2 prevalence.

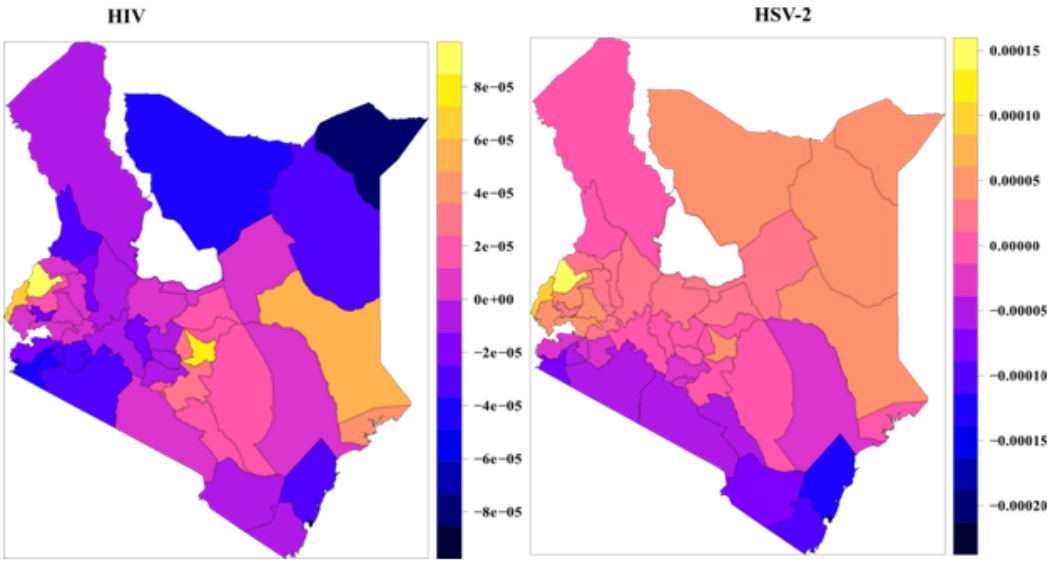



FIGURE 3.6: Figure of spatial effects of HIV and HSV-2

### 3.11.6 The Non-linear effect of age

Figure 3.7 shows the nonlinear association between age of an individual and HIV infection and age of an individual and HSV-2 infection. The Figures give the

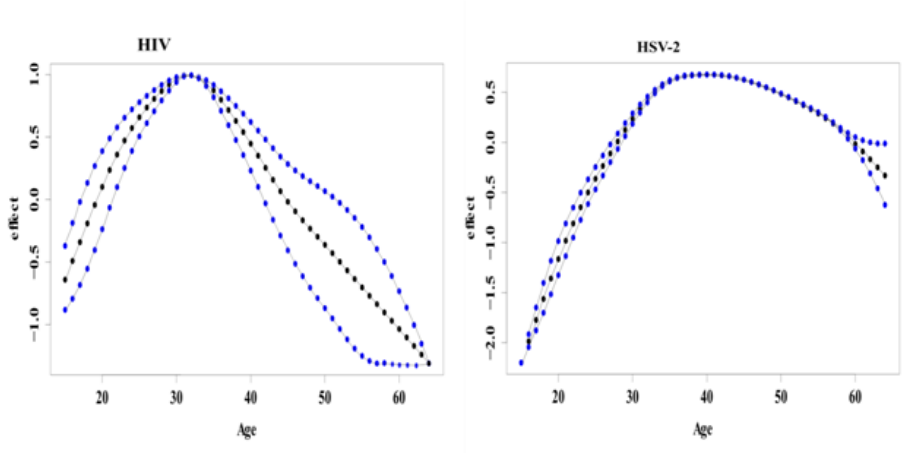



FIGURE 3.7: Figure of non-linear effect of age on HIV and HSV-2

posterior mean of the smooth function and their corresponding 95% CI. From the Figures it is evident that there is a nonlinear relationship between age and HIV and HSV-2 infection. An assumption of linear relationship would have led to misleading results and subsequently wrong interpretations. The chance of HIV infection increases with age up to an optimum age of about 30 years then starts declining with increase in age. For HSV-2, the likelihood of infection increases with age up to an optimum age of about 40 years then starts to decline thereafter with increasing age. The results depict that the prevalence of HIV peaks earlier in age than HSV-2.

### 3.11.7 Discussion

This study found that the effect of the covariates on HIV and HSV-2 prevalence varied spatially, although the spatially varying HIV model was not much different from the stationary one. This could be due to bias introduced by deletion of cases. A stationarity assumption would therefore have masked these varying effects. The models developed in chapter 2 only showed the blanket or countrywide effect of

the covariates on HIV and HSV-2 prevalence. The BSVCP models developed in chapter 3 however shows the region (county) specific effects of each covariate. The major strength of the spatially varying model is that it is able to unmask the effect of each covariate on HIV and HSV-2 prevalence in each region. A study by Assunçao [2003] using the BSVP to model agricultural development in Brazil showed significant regional differences in agricultural development. Age at first sex had the greatest effect on HSV-2 prevalence in the Central and parts of Rift region and more effect on HIV prevalence in the Coastal, North Eastern and Central regions. This may suggest either early marriages, child prostitution or teenage sex. Intervention strategies geared towards delaying the age at first sex, stopping childhood prostitution or early marriages can be put in place in these regions. The number of partners had in the last one year had more effect on HSV-2 status in the West and Lake regions and some parts of the Central region. Residents in these regions can be educated on faithfulness, use of protection and/or abstinence. Place of residence had more effect on HSV-2 prevalence in the Southern, parts of Central, West, Lake and Coastal regions.Various studies have documented that education level is inversely related to HIV and HSV-2 infection [Burgoyne and Drummond, 2009; Cohen, 1998]. Education level provoked more response in HIV prevalence in the North Eastern, Coastal, Southern and parts of Central region. In the Coastal region where tourism is rife, vices such as child prostitution and drug abuse can greatly contribute to the prevalence of HIV and HSV-2. Education can not only detract an individual from activities that can lead to a high probability of acquiring HIV and/or HSV-2, but also make them aware of the safe practices. The effects of frequency of travel away on HIV prevalence was dominant in Coastal, Central and

Rift regions, with some parts of North Eastern region having a near zero effect while for HSV-2 prevalence, the effect was dominant in the West and Lake regions and some parts of Central and Rift region. This shows that frequency of travel away has different effects across the regions suggesting that women in the Coastal, Central and Rift regions travel away from their homes/regions more than women from the rest of the country. This may imply that these women engage in risky behaviors when they travel away and/or their spouses engage in risky behaviors when their partners have traveled away. Frequency of travel away also has different effects on HIV and HSV-2. Since its effect on HSV-2 is dominant in West and Lake region, this could mean that the regions visited by these women have high HSV-2 prevalence and the same applies for HIV. The 2011-12 Tanzanian HIV/AIDS and malaria indicator survey found that women who traveled away from home five or more times in a year were twice likely to be infected with HIV(STIs) compared to women who did not travel [TACAIDS, 2013]. This could be due to the fact that these women are more likely to engage in risky sexual behaviors when they are away from home. The effect of marital status on HIV prevalence was dominant in the West and the Lake region. This could be attributed to traditional practices such as wife inheritance which is rife in these regions. Wife inheritance is a widespread cultural practice in sub-Saharan Africa that increases the risk of HIV acquisition and transmission [Amornkul et al., 2009; Kenya, 1997].

Age was found to have a non-linear effect on both HIV and HSV-2. i.e. an inverted "U" shape. The likelihood of HIV infection among women increases with age up to about age 30 then reduces thereafter with increasing age. On

the other hand the likelihood of HSV-2 infection increases with age up to about age 40 and then starts declining with age. These findings were consistent with other studies [Amornkul et al., 2009]. Spatial effects in the model account for unobserved variables that represent those variables that vary spatially. Identifying high prevalence areas and the relationship between HIV and HSV-2 can provide more insight that can be useful in coming up with campaigns and prevention strategies for specific regions. There was evidence of spatial variation of HIV and HSV-2 infection among counties. HIV prevalence was lowest in the North Eastern region with some significantly high prevalence in some parts of the Coastal, Central, Western and lake regions. HSV-2 prevalence was highest in the West and Lake regions, but generally high across the country. Identifying the effects of individual covariates on each region will help in informing region specific strategies to deal with HIV and HSV-2 prevalence.

The spatially varying coefficient model has a huge epidemiological implication. With limited resources such as funds, time and personnel, intervention strategies may be tailor made for specific regions instead of rolling out blanket intervention strategies. More emphasis for example can be put in delaying the age at first sex in those regions where the effect of age at first sex on HIV and HSV-2 was great etc. Areas where individuals engage in sexual activities with multiple partners can for example be targeted with intervention strategies tailored to either help these individuals stick to one partner or educate them on the use of protection rather than addressing issues that do not contribute much to the prevalence of HIV and HSV-2 in that particular area thereby wasting valuable resources.

**Appendix 1: The Random walk model**

Random walk (RW) models can be used as priors to derive the discretized Bayesian smoothing spline estimator [Speckman and Sun, 2003]. The Random walk was made spatially adaptive by introducing local smoothing parameters into the models [Lang et al., 2002; Yue et al., 2012]. The random walk model of order 2 (RW2) for the Gaussian vector $X = (x_1 \ldots, x_n)$ is constructed assuming independent second-order increments:

$\Delta^2 x_i = x_i - 2x_{i+1} + x_{i+2} \sim N(0, \tau^{-1})$. The density of $X$ is derived from its n-2 second-order increments as:

$$\pi(X|\tau) \propto \tau^{(n-2)/2} exp\left\{-\tau/2 \sum (\Delta^2 x_i)^2\right\}$$

The term $x_i - 2x_{i+1} + x_{i+2}$ can be interpreted as an estimate of the second-order derivative of a continuous time function $x(t)$ at $t = i$ using the values of $x(t)$ at $t = i, i+1, i+2$ [Lindgren and Rue, 2008]. The RW2 model is quite flexible due to its invariance to addition of a linear trend, and also computationally convenient due to its Markov properties i.e. $\pi(x_i|x_{-i}) = \pi(x_i|x_{i-2}, x_{i+1}, x_{i+2})$ for $2 < i < n-2$. RW2 is also a GMRF for which efficient numerical methods for sparse matrices in place of Markov chain Monte Carlo algorithms exists [Rue, 2001; Rue and Held, 2005].

**Appendix 2: The Bayesian Spatially Varying Coefficient Process (BSVCP)**

The specification of the BSVCP is in a hierarchical manner. The first stage is to specify the distribution of the data conditional on unknown parameters, and the

second stage is specifying these unknown parameters conditional on other parameters.

The SVCP model is:

$$y_{ijk}|p_{ijk} \backsim Bernoulli(p_{ijk})$$

$$h(p_{ij1}) = X^T\beta_k + W^T\gamma_k$$

The prior distribution for the regression coefficients is given by Wheeler and Waller [2009] as:

$$\left[\gamma|\mu_\gamma, \sum\nolimits_\gamma\right] = N(1_{n\times 1} \otimes \mu_\gamma, \sum\nolimits_\gamma) \tag{3.11}$$

Where: $\mu_\gamma = (\mu_{\gamma 0}, \mu_{\gamma 1}, \ldots, \mu_{\gamma_p})^T$ is the vector of means of the regression coefficients corresponding to each of $p$ explanatory variables. Spatial dependence is taken into account through the covariance $\sum_\gamma$. This is achieved by specifying the priors for $\boldsymbol{\gamma}'s$ as an areal unit model e.g. the conditional autoregressive model (CAR) or the spatial autoregressive model (SAR) as shown in Banerjee et al. [2014] or a geostatistical approach, where a parametric distance-based covariance function is specified [Wheeler and Waller, 2009]. Our focus is on the aerial unit model and in particular we assume the CAR priors for the $\gamma's$.

**Conditional autoregressive (CAR) Model**

Consider a vector $\phi = (\phi_1, \ldots, \phi_p)$ of $p$ components that follows a multivariate Gaussian distribution with mean 0 and $B$ as the inverse of the dispersion matrix, so that $B$ is a $p \times p$ symmetric and positive definite matrix. The density for $\phi$ is

given by:

$$p(\phi) = (2\pi)^{p/2}|B|^{1/2}exp(-\frac{1}{2}\phi^T B\phi)$$

For the CAR model, the conditional distribution of a particular component given the remaining components is considered. In terms of the elements of the matrix $B = (b_{ij})$, from the normal theory, $\phi_i$ has a full conditional distribution;

$$p(\phi_i|\phi_{-i}) \propto exp\left(-\frac{1}{2}b_{ii}\left(\phi_i - \sum_{j\neq i}\frac{-b_{ij}}{b_{ii}}\phi_j\right)^2\right)$$

which is normally distributed i.e.

$$\phi_i|\phi_{-i} \backsim N\left(\sum_{j\neq i}\frac{-b_{ij}}{b_{ii}}\phi_j, \frac{1}{b_{ii}}\right) \tag{3.12}$$

Mardia [1988] showed the conditions under which the full conditional distributions specified above uniquely define a full joint distribution.

We let $c_{ij} = \frac{-b_{ij}}{b_{ii}}$ and $b_{ii} = \frac{1}{{\sigma^2}_i}$ and form a matrix $C$ with $c_{ii} = 0$ and $c_{ij} = -\frac{-b_{ij}}{b_{ii}}$, and another matrix $M = Diag({\sigma_i}^2)$ and $M^{-1} = Diag(b_{ii})$. The inverse of the dispersion matrix, B is then related to C and M as:

$$B = M^{-1}(I - C). \tag{3.13}$$

$I$ is the identity matrix and the joint distribution of $\phi$ is $MVN(0, M^{-1}(I-C))$. C and M must be modeled properly to ensure the symmetry of B, and this is achieved by conditioning ${c_{ij}\sigma_j}^2 = {c_{ji}\sigma_i}^2$. The C matrix is also specified to show relationship

between neighbors. The elements of matrix C are defined as $c_{ii} = 0$ and $c_{ij} = \frac{1}{m_i}$ as in Besag [1974], if $j$ is adjacent to $i$ and zero otherwise. This is a commonly used adjacency matrix for lattice data. Here, $m_i$ represent the number of neighbors of region $i$, define another matrix $W$ to hold the adjacency structure, where, $w_{ij} = 1$ if region $i$ and region $j$ are neighbors and zero otherwise. Then, $C = W_s$ where $W_s = diag(\frac{1}{m_i})W$. i.e. $W_s$ is a scaled adjacency matrix, the $i^{th}$ row being scaled by the number of neighbors of region $i$. The above expressions for the elements of C and M translate to the following specifications for inverse covariance matrix B: $b_{ii} = \lambda m_i$, and $b_{ij} = -\lambda$ if $j$ is adjacent to $i$ and 0 otherwise. Thus B is symmetric and it can be expressed as $B = \lambda(Diag(m_i) - C)$. The expression $M^{-1}(I - C)$ has a positive definite structure for the conditional distribution to give rise to a valid probability distribution function (pdf). The definition of the adjacency matrix above leads to an improper joint pdf. This is overcome by introducing a parameter into the precision matrix B, to give:

$$B = M^{-1}(I - \alpha C). \tag{3.14}$$

If $|\alpha| < 1$ then the matrix $M^{-1}(I - \alpha C)$ is diagonally dominant and symmetric. Symmetric and diagonally dominant matrices are positive definite [Harville, 1997].

# Chapter 4

# Spatio-temporal modeling of Malaria among children under the age of 5 in Angola

## 4.1 Introduction

About half of the world's population is currently at risk of malaria [WHO, 2016]. When infected, the most vulnerable to the disease are young children, pregnant women and non-immune travelers from malaria-free areas. Between 2000 and 2015, malaria incidence among populations at risk (the rate of new cases) fell by 37% globally. In that same period, malaria death rates among populations at risk fell by 60% globally among all age groups and by 65% among children under 5. Sub-Saharan Africa however carries a disproportionately big share of the global

malaria burden. In 2015, the region was home to 88% of malaria cases and 90% of malaria deaths [WHO, 2016].

Malaria is a major public health problem in Angola. It represents close to 35 percent of the demand for curative care, 20 percent of hospital admissions, 40 percent of perinatal deaths, and 25 percent of maternal mortality [AMIS, 2012]. As at 2005, it was estimated that malaria accounted for 35 percent of overall mortality in children under five, 60 percent of hospital admissions of children under five, and 10 percent of hospital admissions of pregnant women [Ruebush et al., 2005]. In 2008, Angola was divided into 3 regions according to endemicity level; the hyperendemic, mesoendemic stable and mesoendemic unstable as shown in Figure 4.1. The hyperendemic region comprises 6 provinces to the North-Eastern part of the country. The transmission period in this region is all year round, with highest transmission rates occurring between November and January. About 28% of the population was at risk in the hyperendemic region during the transmission period. The mesoendemic stable region covers 8 provinces in central Angola. The highest transmission period in this region is November to May while the lowest transmission period is July to October. Slightly more than half of the population (55%) was at risk in the mesoendemic stable region between November and May. The mesoendemic unstable region covers the Southern 4 provinces with 17% of the population at risk. Low transmission period for this region is from May to December [AMIS, 2007].

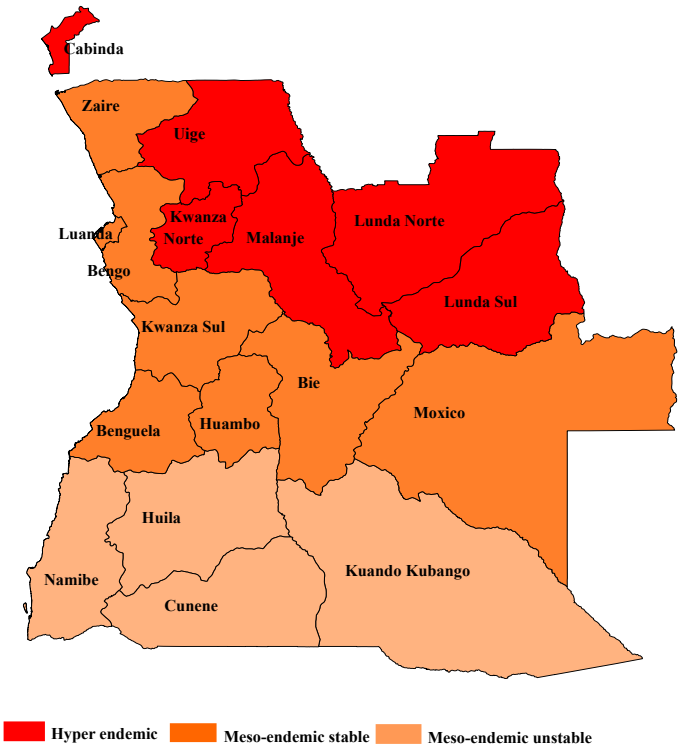



FIGURE 4.1: Figure of malaria endemicity in Angola

The clues of disease etiology can be unmasked by studying its risk factors in space and time. In particular, it would be of great epidemiological consequence if the effects of these risk factors can be observed in both space (from administrative unit to administrative unit) and time (from time period to time period). This may help bring out the way the effects of the risk factors change in time, unmask endemic regions and periods, unravel both new and less effective risk factors with time and hence help inform the policy makers on the effects of the intervention strategies laid down and whether different approaches need to be used.

In this study, we perform a spatial temporal modeling of malaria in Angola using the 2006-2007 and 2011 Angola malaria indicator survey (AMIS) data. We then extend this method to allow the effects of the covariates to vary both in space

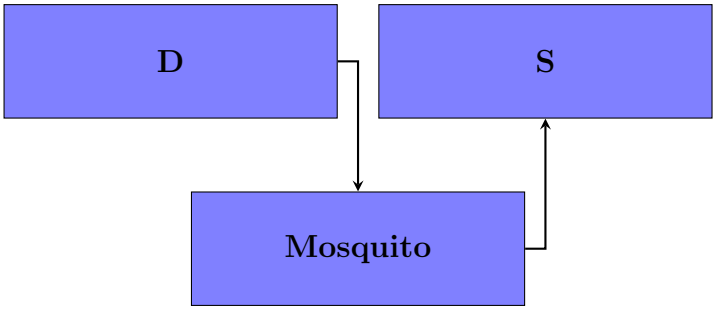
using the conditional autoregressive model and in time using the autoregressive model of order one.

## 4.2 Methods

### 4.2.1 Study area

Angola is located on the southwest of Africa and is the sixth largest country in Africa with an area of about 1,246,620 square kilometers. It shares borders with the Republic of Congo-Brazzaville on the North, Democratic of Republic of Congo on the Northeast, Zambia on the East and Namibia on the South. The republic of Angola is divided into 18 provinces with 164 municipalities. The tropical climatic conditions in Angola makes it a good breeding ground for mosquitoes which are responsible for transmitting malaria. The transmission from an infected individual (D) to a susceptible individual (S) is depicted in the simple diagram below.

FIGURE 4.2: Transmission of Malaria



## 4.3 Data

The data for this study was obtained from the Angola Malaria Indicator Survey (AMIS) 2006-2007 and 2011 [AMIS, 2007, 2012]. In particular, this study uses the children's data in the age group 0-5 years. Information from 2,310 children from the 2006-2007 survey and 3,432 children from the 2011 survey, who provided venous blood for malaria testing and also had full covariate information, was used. This information was obtained with the consent and assistance of the parent/guardian of the child. The variables of interest that were subsequently used in the study were place of residence (urban/rural), province, wealth index, gender, whether a child sleeps under a mosquito net and age of the child which was captured as continuous. Readers are directed to the AMIS [2007, 2012] report for more information about the data.

## 4.4 Statistical model

### 4.4.1 The Spatio-temporal model

Spatio-temporal disease mapping has become an important tool in passive surveillance of diseases. Understanding how disease risks and prevalence and/or incidence vary over time may provide information that may be of great epidemiological significance. Spatio-temporal models are extensions of the basic spatial models by simply including a linear or a non-parametric trend in time, time space, time-covariate and time-space-covariate interactions.

### 4.4.2 The model

Let $y_{ijt}$ be the malaria status for place of a given child $j$ at time $t$ in province $i$: $i = 1, 2, \ldots, 18$ such that $y_{ijt} = 1$ if child $j$ in province $i$ has malaria at time $t$ and zero otherwise. This study assumes the dependent variable $y_{ijt}$ is univariate Bernoulli distributed, i.e. $y_{ijt}|p_{ijt} \backsim Bernoulli(p_{ijt})$.

The unknown mean response namely $E(y_{ijt}) = p_{ijt}$ may relate to the predictors as follows:

$$h(p_{ijt}) = intercept + \underbrace{C + S + T}_{\text{main effects}} + \underbrace{CS + CT + ST + CST}_{\text{interactions terms}}$$

Where,

The function $h(.)$ is a logit link function, the intercept term gives the initial amount of risk shared by all individuals, provinces and time. The main effects $C, S$ and $T$ represents the covariate, spatial and temporal effects respectively. The second order interaction terms $CS, CT, ST$ represents contribution to the risk due to a combination of main effects that cannot be explained additively by main effects, $CST$ represents the covariate-space-time interaction [López-Quilez and Munoz, 2009].

## 4.5 Structure of the effects

### 4.5.1 The covariate effect

The effects of the covariates on the response variable can be either linear or non-linear. Depending on the type of variable, linearity is usually achieved by assigning the regression coefficients flat priors, usually flat normal distribution. Some covariates may be stratified into several categories and included as fixed effects or as structured random effects. There are several ways of allowing some covariates to have a non-linear effect on the response variable. A class of models called the generalized additive models were introduced by Hastie and Tibshirani [1990] which replaces each linear term in the additive logistic regression by a more general functional form $f_t(.)$. Several studies have discussed extensively the methods for estimating the smooth function $f_t(.)$. Green and Silverman [1993] used penalization and splines to model the smooth function $f_t(.)$. Some other methods include the p-spline method by Lang and Brezger [2004], continuous indexed spline models by Wahba [1978], Gaussian processes by O'Hagan and Kingman [1978], the penalized regression splines method proposed by Eilers and Marx [1996] and the random walk models [Cleveland et al., 1992; Fahrmeir and Tutz, 2001].

### 4.5.2 The Spatial effects

Spatial effects are introduced in the model as random effects. Spatial effects are of two types, namely spatially structured random effects which cater for the unobserved covariates which vary spatially across (clustering) the study areas and spatially unstructured random effects which cater for the unobserved covariates inherent within (heterogeneity) the study areas. Some models incorporate a convolution of both structured and the unstructured random effects.

For the unstructured random effects, the spatial effects are assumed to be sampled from a normal distribution with 0 mean and a precision $\tau$ i.e. $\mu_i \backsim N(0, \tau)$.

The mean of the structured effect $v_i$ is allowed to depend on the neighboring $V_j's$ through the Gaussian conditionally autoregressive (CAR) distribution given by Hinton and Van Camp [1993] as.

$$v_i|v_j \backsim N\left(\frac{\sum_{j\neq i} C_{ij}v_j}{\sum_{j\neq i} C_{ij}}, \frac{1}{\tau_{CAR}\sum_{j\neq i} C_{ij}}\right),\ i = 1, \ldots, I \tag{4.1}$$

By convention $c_{ii}$ is set to zero for all $i$ so that no region is its own neighbor while $C_{ij} = 1$ if region $j$ is adjacent to region $i$, and $C_{ij} = 0$ otherwise. Other weighting options to adjacency-based weighing system also exist but are less widely applied. Best et al. [2001] used distance-based spatial weights however the adjacency based model performed better than the distance-based model based on the DIC. Earnest et al. [2007] found considerable differences in the smoothing properties of the CAR model, depending on the type of neighbors specified. This in turn had an effect on their models' ability to predict the observed risk in an area. These

results have significant implications for all researchers using CAR models, since the neighborhood weight matrices chosen may markedly influence a study's findings. Other studies have also allowed the weights to be data driven i.e. the weights are estimated from the data [Lu et al., 2007].

The choice between the clustering and the heterogeneity model depends on the prior belief one has about the scope of risk determinants. Risk determinants exceeding the limits of one or more regions leads to clustering since they include similar risk values in neighboring regions, while when the scope of the risk determinants is smaller than a region's size we have heterogeneity [López-Quilez and Munoz, 2009]. The risk associated with a region can be broken down as the sum of the heterogeneity and a clustering effect and hence the spatial effect $S_i$ is given by: $S_i = \mu_i + v_i$.

Spatial effects can also be modeled by a two dimensional splines. In particular penalized splines by Eilers and Marx [1996] was used by Currie et al. [2006] to smooth both Gaussian and non-Gaussian data.

## 4.6 Temporal effect T

The temporal effects are frequently modeled as structured random effects, ensuring that contiguous periods are likely to be similar, but allowing for flexible shapes in the evolution curve, especially when long periods of time are being considered. Many studies have modeled the temporal effects with the autoregressive processes (AR) for example by Mabaso et al. [2006], first and second order random walks

(RW1, RW2) by Lindgren and Rue [2008] or the splines. For equally spaced time points, the difference between the two approaches is that RW1 is the limiting case of AR1 where the parameter goes to 1. Further AR1 is stationary, RW1 is only intrinsically stationary.

Other studies have stratified time into a few blocks of time and modeled the effect as a fixed effect, thus estimating the effect of each block independently from others.

## 4.7 Covariate interactions CS, CT, CST

Most studies usually assume independence between the covariates and the spatial and the temporal effects. It is however realistic that the effect of the covariate will vary both spatially and temporally and hence the interactions. In chapter 3 we examined the covariate space interaction and in this study we explore the covariate space and time interaction. Sun et al. [2000] stratified age into four groups and assumed that each age group could present a different evolution pattern in mortality rates due to the disease they were modeling. They incorporated the age group as a fixed-effect covariate and modeled the covariate-time interaction terms as linear functions of time with slope depending on age group i.e.$C_kT_j = age_kt_j$.

## 4.8 Spatio-temporal interaction ST

This is the key aspect of spatio-temporal models. There exist many possibilities of spatio-temporal interactions however there is no accepted standard that functions well. Knorr-Held [1999] discussed four possible types of interactions between spatial and temporal random effects;

- **Type I interaction:**

  This can be thought of as independent unobserved covariates for each combination of region and period, hence without any structure. This interaction is a global space-time heterogeneity effect, and is usually modeled as white noise. This is the simplest way of implementing a spatio-temporal interaction allowing the data to show if there is a need for further investigation

- **Type II interaction:**

  In this type of interaction, each region has a specific evolution structure that is independent of that in the neighboring regions. The form of the evolution structure may be as many as there are the forms of the temporal main effect itself. This type of interaction is suitable for modeling factors affecting specific regions and inducing deviations from the global trend.

- **Type III interaction:**

  This interaction can be assumed to have a spatial structure for each period, independent of adjacent periods (its neighbors in time). This is usually modeled with a CAR distribution for each period. Such an interaction could

represent situations where an unobserved regional factor is affecting an area containing two or more adjacent regions, but not persistent in time.

- **Type IV interaction:**

  This type of interaction arises when deviations from the global trends are assumed to be correlated with their neighbors both in space and time. This interaction can model hidden factors whose effects exceed the limits of one or more regions and also persistent for more than one period of time.

## 4.9 Spatial temporal model

In this section we employ two models namely; a spatial model (model 1) and a spatio-temporal model (model 2). The spatio-temporal model captures covariate-space-time interaction. The models used are;

$$\text{Model 1} \quad : \quad logit(p_{ijt}) = \beta_{01} + f(age) + W_{ijt}^{T}\gamma + v_i \text{ Spatial model}$$

$$\text{Model 2} \quad : \quad logit(p_{ijt}) = \beta_{01} + f(age_{jt}) + W_{ijt}^{T}\lambda z_t + v_i \text{ Spatio-temporal model}$$

where;

- $\beta_{01}$ is the intercept representing the logit prevalence rate when all covariates have a zero value.

- $f(age)$: This represents a function of age. Age was captured as continuous and it is assumed to have a non-linear effect on malaria prevalence.

- $w_{ijt}^{T}$: This represents the vector of categorical covariates effects for child $j$ living in province $i$ at time $t$ in both models 1 and 2.

The component $\gamma$ represent the regression coefficients for the spatial model and they are assigned the flat normal distribution priors while $\lambda$ and $z_t$ represents the space and time dependent regression coefficients modeled using the conditional autoregressive model (CAR discussed in appendix 1) and the autoregressive model of order 1 (AR1) respectively. The component $w\lambda z$ represents the covariate, space and time interaction. The AR1 model for a Gaussian vector $X = (X_1, X_2 \ldots, X_n)$ is defined as;

$$\begin{aligned} x_1 &\sim N\left(0, \left(\tau\left(1-\rho^2\right)\right)^{-1}\right) \\ x_i &= \rho x_{i-1} + \epsilon_i; \epsilon_i \sim N(0, \tau^{-1}), i = 2, \ldots, n \\ |\rho| &< 1 \end{aligned} \tag{4.2}$$

- $v_i$ Represents both the structured and the unstructured spatial random effects

**Model 1**: This model does not cater for any interaction i.e. covariate-time, space-time, space-covariate or space-covariate-time. It is a model of continuous covariate age modeled with a random walk model of order 2 as discussed by Lindgren and Rue [2008] which is assumed to have a non-linear effect on malaria prevalence, categorical covariates which are assumed to have a linear effect on malaria status

and the structured random effects modeled using the conditional autoregressive model (CAR).

**Model 2**: This model caters for space-covariate-time interaction. Malaria is modeled as a non-linear function of age using the random walk model of order 2, while the rest of the covariates are modeled as functions of time and space using the autoregressive process of order 1 and the conditional autoregressive model respectively. The model incorporates the structured random effects which cater for any unobserved covariates which vary spatially across the provinces. There are various formulations of the spatial temporal models that do exist with no particular convention one. In fact it is also impossible to justify any single formulation of the spatio-temporal models on empirical grounds. Our study for the purpose of demonstration adapts models 1 and 2 and warn that these models are not in themselves final.

## 4.10 Priors for the parameters and spatial components

A non-informative normal distribution prior was used for the fixed effects while a random walk model of order 2 was used for the continuous covariate age. The temporal effects were modeled by a first order autoregressive process allowing for correlation between the two time periods and the provinces [Mabaso et al., 2006]. The spatial components prior was the CAR model for the structured random effects.

## 4.11 Posterior distribution

This is the distribution of the parameters after observing the data. The posterior distribution is obtained by updating the prior distribution with the observed data. Since our study is fully Bayesian, inference is made by sampling from this posterior distribution. There exist several approximation schemes for latent Gaussian models common of which includes Markov Chain Monte Carlo (MCMC) approach, variational Bayes (VB) methodology, the expectation-propagation (EP) approach and the Integrated Nested Laplace Approximation (INLA). The posterior distribution for the latent Gaussian model is:

$$\pi(x,\theta|y) \propto \pi(\theta)\pi(x|\theta)\prod_{i\in I}\pi(y_i|x_i,\theta)$$

$$\propto \pi(\theta)|Q(\theta)|^{\frac{n}{2}}exp\left(-\frac{1}{2}x^TQ(\theta)x+\sum_{i\in I}log\pi(y_i|x_i,\theta)\right), \tag{4.3}$$

Where $x$ is the class of latent fields, $\theta$ is the set of hyper parameters and $y$ is the data.

## 4.12 The variational Bayes approach

This approach was developed in the machine learning literature as discussed in Hinton and Van Camp [1993] and has provided numerous promising results in areas like hidden Markov models, mixture models, graphical models and state space models among others. For a posterior distribution $\pi(x,\theta|y)$ of a generic

Bayesian model, with observation $y$, latent variable $x$ and hyperparameter $\theta$, the VB method uses as an approximation the joint density $q(x, \theta)$ that minimizes the Kullback-Leibler contrast of $\pi(x, \theta|y)$ with respect to $q(x, \theta)$. This minimization is subject to some constraints on $q(x, \theta)$, most commonly $q(x, \theta) = q_x(x)q_\theta(\theta)$. The VB approximated density $q(x, \theta)$ does not capture the dependence between $x$ and $\theta$ although one hopes that its marginals (of $x$ and $\theta$) approximate well the posterior marginals. The solution of this minimization problem is approached through an iterative, expectation-maximization like algorithms.

## 4.13 Expectation Propagation (EP)

The EP method of approximation as discussed by Minka [2001] was developed and studied mainly in the machine learning applications. The advantage of the EP method over say the Laplace methods of approximation is that they yield approximations that are more accurate [Kuss and Rasmussen, 2005; Minka, 2001] with a computational cost and that they can be applied in cases where the Laplace method is out of question e.g. when the log-posterior is not twice differentiable [Seeger, 2008].

## 4.14 INLA versus MCMC

Markov Chain Monte Carlo (MCMC) is the most common approach to do inference for latent Gaussian models however this method is slow and performs poorly

when applied to such models [Rue et al., 2009]. The Integrated Nested Laplace Approximation (INLA) method is a relatively new technique developed by Rue et al. [2009] to circumvent these shortfalls. In most cases, similar results can be obtained by both INLA and MCMC however there are some differences in the way the posterior distribution is estimated. The MCMC algorithm samples directly from the joint posterior distribution while INLA uses a closed form expression to express the marginal posterior distribution. The greatest advantage of the INLA approach over the MCMC approach is the saving on computational time as INLA does in seconds and minutes what MCMC does in hours or even days with almost similar results. The INLA approach can be summarized into three steps i.e. a) Approximating the posterior of the hyper-parameters given the data and use this to determine the grid of hyper-parameter values, b) Approximating the posterior marginal distributions given the data and the hyper-parameter values on the grid and c) Numerically integrating the product of the two approximations to obtain the posterior marginal of interest. In the INLA approach, the posterior marginals of interest are:

$$\pi(x_i|y) = \int \pi(x_i|\theta, y)\pi(\theta|y)\, d\theta \tag{4.4}$$

and

$$\pi(\theta_j|y) = \int \pi(\theta|y)\, d\theta_{-j} \tag{4.5}$$

these are used to construct the nested approximations:

$$\widetilde{\pi}(x_i|y) = \int \widetilde{\pi}(x_i|\theta, y)\widetilde{\pi}(\theta|y)\, d\theta \tag{4.6}$$

and

$$\widetilde{\pi}(\theta_j|y) = \int \widetilde{\pi}(\theta|y)\, d\theta_{-j} \tag{4.7}$$

The notation $\theta_{-j}$ is used to denote that the integration is done over all the components of $\theta$ except $\theta_j$.

The analysis in this study were carried out using the R software with the INLA package.

## 4.15 Results

TABLE 4.1: The estimated odds ratios with the corresponding 95% credible intervals for the spatial model.

| Covariates | Year 2006-2007 | 2011 |
|---|---|---|
| **Residence (ref rural)** | 1 | 1 |
| Urban | 0.212(0.101,0.442) | 0.104(0.059,5.667) |
| **Have net (ref No)** | 1 | 1 |
| Yes | 0.738(0.566,0.961) | 0.651(0.504,0.847) |
| **Wealth index (Poorest)** | 1 | 1 |
| Poorer | 0.806(0.601,0.93) | 0.731(0.961,1.854) |
| Middle | 0.449(0.286,0.695) | 0.773(0.904,1.854) |
| Richier | 0.326(0.174,0.596) | 0.558(0.393,0.875) |
| Richest | 0.377(0.189,0.729) | 0.441(0.270,0.646) |
| **Gender (ref Male)** | 1 | 1 |
| Female | 1.023(0.821,1.275) | 0.974(0.615,1.026) |

Table 4.1 gives posterior estimates of the odds ratios and their corresponding 95% credible intervals (CI) for the spatial model. The categorical covariates were assumed to have linear effects on malaria prevalence and all showed significant relationship (using CI).

The odds of having malaria in the 2006/2007 and 2011 periods for children living in urban areas is significantly lower than the odds of having malaria for those who

reside in the rural area, (OR 0.212, 95%CI: 0.101 to 0.442) and (OR 0.104, 95% CI: 0.059 to 5.667) respectively. In addition, the odds of malaria for urbanite children was lower in the 2011 data than the 2006-2007 data. Studies have established that urbanization affects some species of mosquitoes in the environment, diversity, numbers, survival rates, infection rates and the frequency with which they bite people [Hay et al., 2005]. Children who had nets were also less likely to contract malaria compared to those who did not have nets in 2006/07 and 2011, (0.738, 95% CI: 0.566 to 0.961) and (0.651 95% CI: 0.504 to 0.837) respectively the lower odds in 2011 could indicate higher or increased use of nets over time. Studies have shown that households with bed nets had a lower chance of infection [Yusuf et al., 2010].This is for the obvious reason that one is protected from mosquito bites which is the key transmission route. Wealth index had a linear relationship with malaria; the wealthier a child's family is the less likely the child is to contract malaria. A study by Njau et al. [2006], found that people from the better-off stratum were significantly less likely to be parasitaemic, and significantly more likely to obtain antimalarials than those in the middle or poor stratum. The better treatment obtained by the better off led them to spend two to three times more than the middle and poor had spent. This could explain why the rich and the richest stratum fair better as far as malaria prevalence is concerned. The interesting observation is that although the conclusion about wealth index was consistent in the two study periods, the odd of malaria were generally higher in 2011 than in 2006/07. The effect of a child's gender was found to be insignificant in malaria prevalence.

### 4.15.1 The effect of age

In this study we had assumed that age had a non-linear effect on malaria prevalence on children under the age of 5. However our study revealed that age had a positive linear relationship with malaria prevalence. Figure 4.3 shows the effects of age on prevalence of malaria. The likelihood of malaria infection increases with increasing age. This increasing prevalence of malaria with increasing age could be as a result of nutrient deficiency in their diet which may render them vulnerable to malaria than those below 12 months who were still likely to be exclusively breast-fed. This is because it has been observed that micronutrient deficiencies among infants could impinge on their immune system, increasing their risk for malaria [Nyarko and Cobblah, 2014; Pérez-Escamilla et al., 2009].

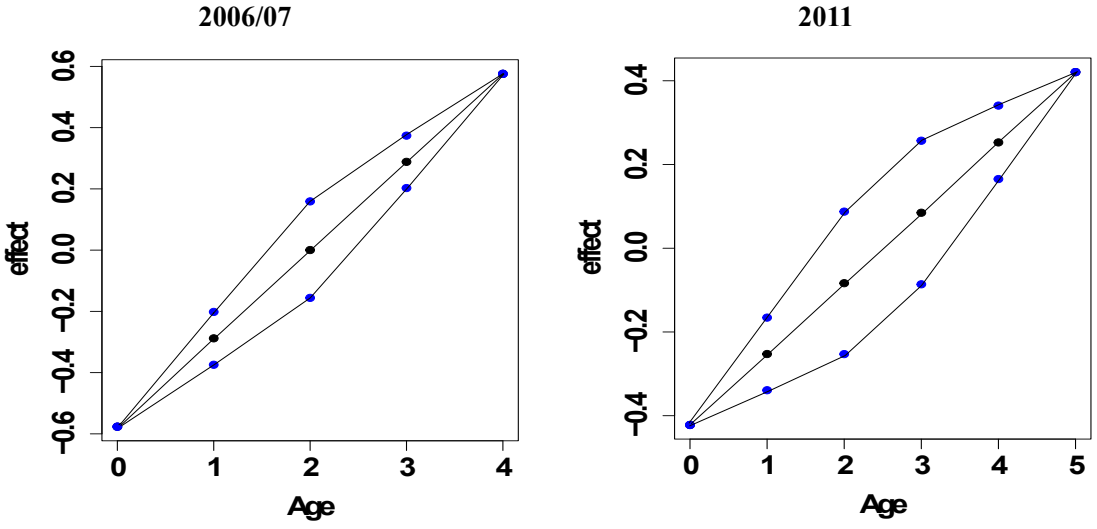



FIGURE 4.3: The linear effects of age on malaria for children aged 0-5 years in the years 2006/07 (left panel) and 2011 (right panel) in Angola.

## 4.16 Spatio-temporal effects

It is of great importance to study how risk factors and disease prevalence evolve in time. We discuss in brief the spatio-temporal evolution from model 2 of some few risk factors namely place of residence, whether a child slept under a bed net and wealth index.

### 4.16.1 Place of residence

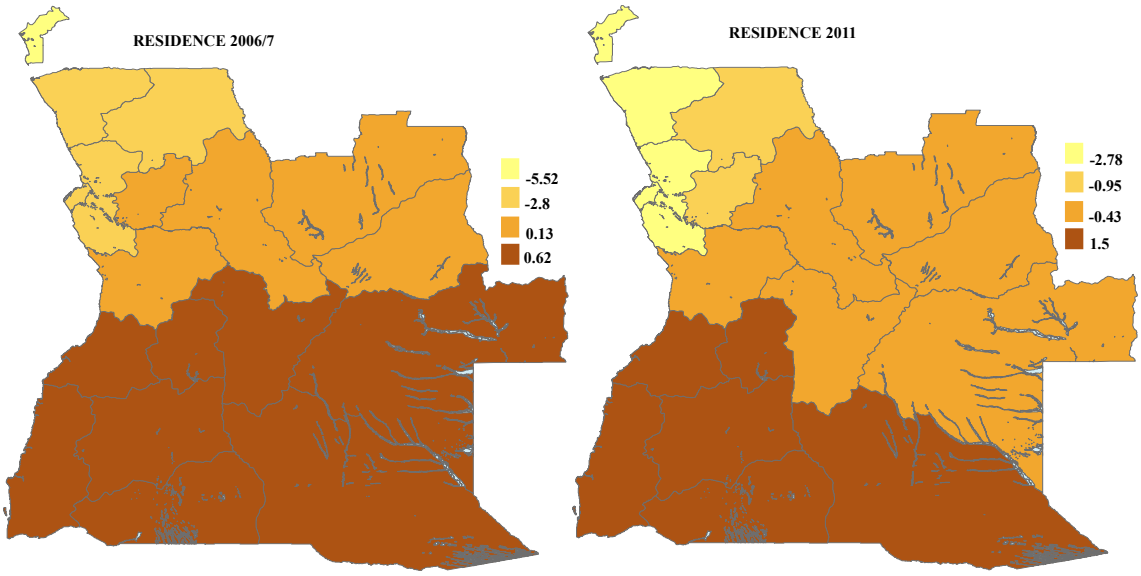



FIGURE 4.4: Effect of place of residence on malaria prevalence

From figure 4.4, the regions are almost clustered as per malaria endemicity as depicted in figure 1. For 2006/07, the mesoendemic unstable and the hyperendemic regions had the highest prevalence rates. The 2006/07 study was done between November 2006 and April 2007 when the mesoendemic- unstable region experiences the highest transmission rates [AMIS, 2007]. For 2011, the study was conducted between January and May 2011 which is also the highest transmission period for mesonedemic unstable regions [AMIS, 2012]. The hyperendemic regions experience

a high transmission rate all year round with highest transmission rates occurring between November and January.

### 4.16.2 Mosquito Nets

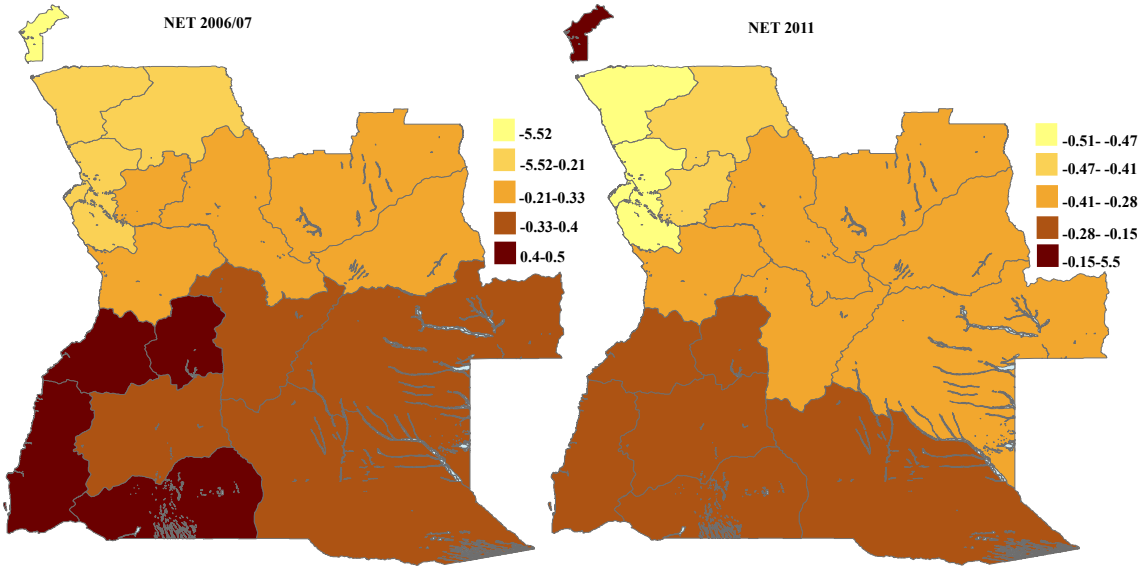



FIGURE 4.5: Effect of Mosquito nets on malaria prevalence

Figure 4.5 depicts the effects of mosquito net in Angola in 2006/07 and 2011. The effect of mosquito nets was high in the hyper endemic and the meso-endemic unstable region. This coincided with the period in time where malaria transmission was highest in these regions in November 2006 and April 2007 and January and May 2011. This suggests awareness in the part of the individuals on the times they are most vulnerable and on the methods of reducing their vulnerability. The choropleth maps depict an increased effect of bed nets on malaria prevalence in 2011 as compared to 2006/07.

### 4.16.3 Wealth Index

As discussed earlier, wealth has a negative linear relationship with malaria prevalence. Figure 4.6 shows the effect of wealth on malaria prevalence. The effect of wealth was higher around Luanda, Bengo and Zaire which are more urbanized and hence have higher standards of living. The effect of wealth was least in the South eastern part of the country.

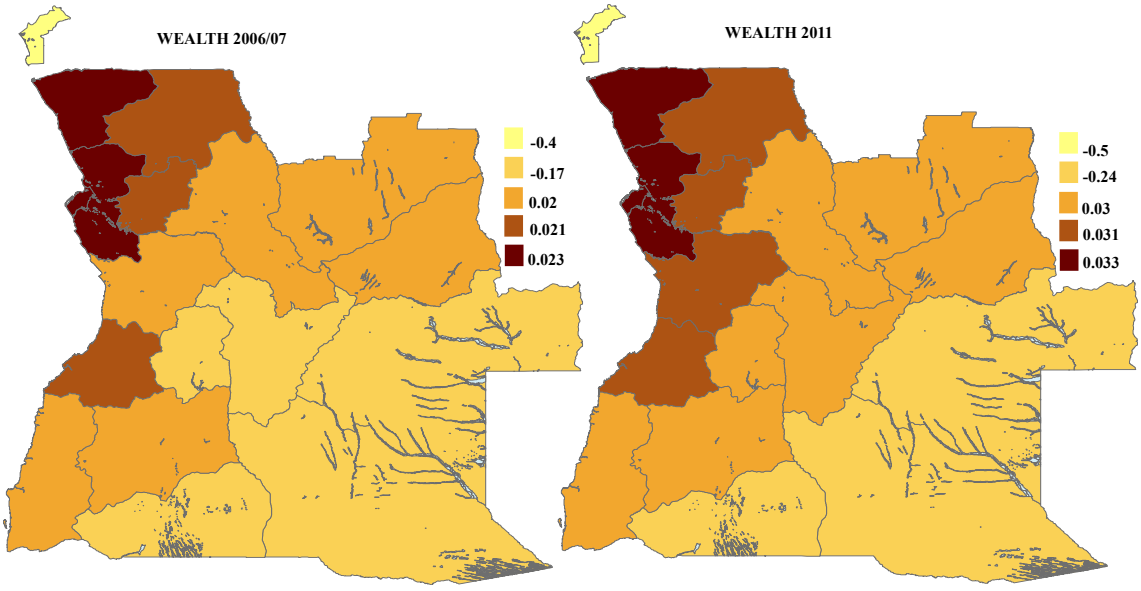



FIGURE 4.6: Effect of wealth on malaria prevalence Spatio-temporal distribution of malaria

Figure 4.7 shows the distribution of malaria in years 2006/07 and 2011. The prevalence of malaria was high in the Northern parts of the country both in the year 2006.07 and 2011. It is noted that the Northern parts are hyper-endemic malaria region. The highest prevalence at the time of both studies was in Zaire province. The choropleth maps show that the prevalence of malaria was lower in 2011 than in 2006/07. This temporal decline is a positive trend showing that

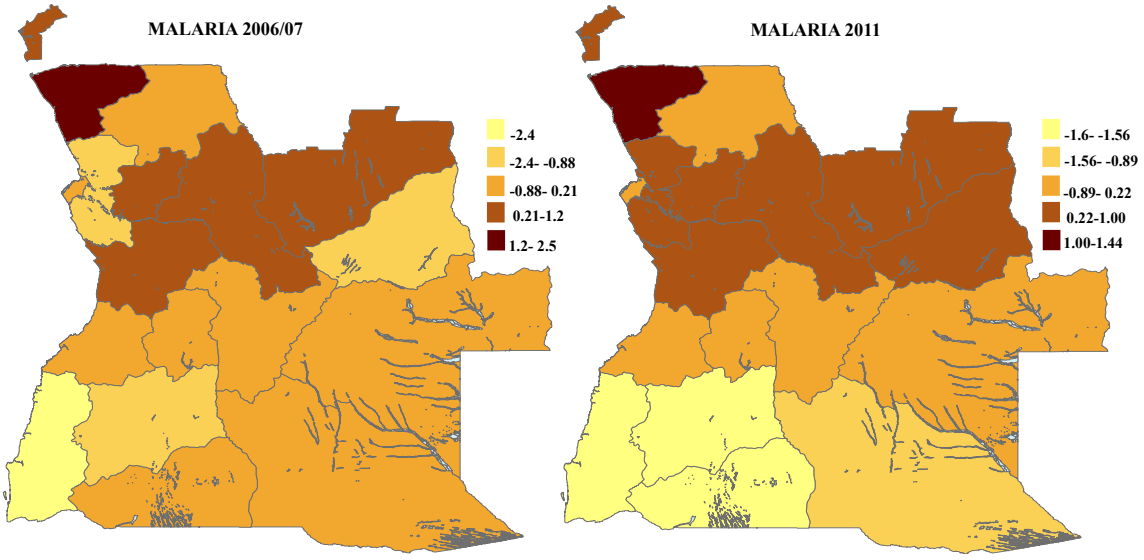



FIGURE 4.7: Overall spatio-temporal distribution of malaria in Angola

the interventions in place could be yielding a desired result. Thus in this analysis result it is clear that both spatially varying and temporal effects are present.

## 4.17 Discussion

This study utilized a full Bayesian approach to perform a spatial and spatio-temporal modeling of malaria prevalence in Angola. We assumed that all the covariates had a linear effect on malaria prevalence except age which was initially assumed to have a non-linear relationship and was modeled using random walk 2 model [Lindgren and Rue, 2008]. The study however found that age had a positive linear relationship with malaria prevalence (Figure 2). Children under 1 year are most likely to be breast feeding and therefore have strong immune systems. As the years progress they are weaned and breast feeding is reduced or stopped altogether and this reduces their immunity making them more vulnerable to malaria. A study by UNICEF reported a high prevalence of stunting among children under 5 years

old in Angola and the reliance on traditional healers or untrained nursing staff to take care of their children by mothers who live in rural areas which serves to worsen the health of the children [UNICEF, 2013].

Wealth was found to have a negative linear relationship with malaria prevalence. People from a well off group economically were significantly less likely to be parasitaemic, and significantly more likely to obtain antimalarials than those in the middle or poor group [Yusuf et al., 2010]. This can be due to the fact that the wealthier in the society can afford better quality medical care and also the fact that they are better placed to afford mitigations against malaria such as treated bed nets, sprays and so on.

Children living in urban areas were less likely to contract malaria than those living in urban areas. The reason for this could be three-fold: One is the fact that generally those individuals living in urban areas are more endowed economically as compared to their rural counterparts. The other factor is that urbanization affects some species of mosquitoes in the environment, diversity, numbers, survival rates, infection rates and the frequency with which they bite people Hay et al. [2005] and hence reducing the prevalence in the urban regions. It is also almost always the case that urban areas have better health care facilities than rural areas. Children who had bed nets were less likely to get malaria than those who were without. This is consistent with other studies [Yusuf et al., 2010]. Insecticide treated nets (ITNs) are very important tools for controlling malaria in Africa [UNICEF, 2013]. This is due to the fact that the principal malaria vectors, the *Giles Anopheles gambiae* and *Anopheles funestus species* complexes as reported in White [1974]

primarily feed indoors at night [Pates and Curtis, 2005]. Bed nets may therefore reduce exposure or mosquito contact and hence transmission.

Spatial effects in the model account for unobserved variables that represent those variables that vary spatially. Identifying high prevalence areas may help in informing intervention strategies for those regions. Figure 4.7 shows the spatial temporal distribution of malaria in the periods 2006/07 and 2011. These patterns are important in that they show how malaria prevalence is changing and may help unmask new patterns and risk factors. The prevalence of malaria was high in the hype-endemic regions of the country both in the year 2006/07 and 2011. The highest prevalence at the time of both studies was in Zaire province. The maps depicts a reduction in malaria prevalence in 2011 as compared to the year 2006/07. This is important to policy makers as it informs that the interventions in place are working as supported by a sound statistical analysis of the data.

# Chapter 5

# Relaxing some limiting assumptions in disease mapping with application

## 5.1 Introduction

The field of disease mapping has gained traction in the recent past. Several reviews on disease mapping have been done by Manda [2011], Clayton and Bernardinelli [1992] and Wakefield [2007] among others. Disease mapping is useful in describing geographical variation of diseases, generation of atlases for diseases and identification of disease clusters. The models developed for modeling disease outcomes however have made limiting assumptions that may lead to less meaningful results

and subsequent interpretations. Some of these assumptions include normality/-parametric distribution assumption for random effects, linearity assumption for covariates, stationarity assumption for covariates. Single modeling of diseases has also been done by many studies but it would be more informative if diseases are modeled jointly if they share common risk factors. In this study we discuss these assumptions and relax the normality, stationarity and linearity assumptions and extend the single disease modeling to multiple disease modeling. We use both the mixture of Dirichlet process (MDP) and the multivariate mixture of Polya trees (MMPT) to relax the normality (parametric) assumption on the random effects, the spatially varying model to relax the stationarity assumption while simultaneously modeling HIV and HSV-2 jointly as a real example using the multivariate normal distribution and the multivariate conditional autoregressive model and using the penalized regression splines to relax the linearity assumption. We also discuss the use of multiple membership multiple classification (MMMC) model for joint modeling.

## 5.2 Data

The data for this study was obtained from the Kenya AIDS Indicator Survey 2007 (KAIS 2007) NASCOP [2007] which was carried out by the Kenyan government with financial support from the United States President's Emergency Plan for AIDS Relief (PEPFAR) and the United Nation (UN). The main aim of the survey was to obtain a high quality data on the prevalence of HIV and Sexually Transmitted Infections (STI) among adults and to assess the knowledge of HIV and

STIs in the population. The data reports the disease (HIV and HSV-2) status for an individual in 46 counties covering the whole country including covariate information.

## 5.3 Normality

Most studies have used the normality assumption on the spatially unstructured random effects. The use of this assumption is mainly because of its computational simplicity. The argument against the normality assumption is that some random effects may exhibit skewness, fat-tailness, multimodality e.t.c. and this may obscure some important features of between subjects and within cluster, area, subject etc.variations depending on the application area. A number of studies have tried to address the issue of normality assumption on the random effects. Ngesa et al. [2014a] employed a generalized Gaussian distribution (GGD) and showed that it can produce better results when the normality assumption is violated due to high or low peakedness in the data. The generalized Gaussian distribution allows the distributional assumption to be dictated by the data itself in the case where the random effects are truly normal.The GGD however assumes that the random effects are symmetric and this assumption may sometimes be wrong. Skew distributions have emerged as an effective tool in modeling heterogeneous data with asymmetry features. These distributions include the univariate and multivariate skew normal [Azzalini, 1985, 1986, 2005]. Many studies have employed the skew-normal distribution and its modifications for random effects. Hosseini et al. [2011] used a closed skew normal (CSN) and they found that in addition

to admitting skewness to the normal distribution, the CSN possesses some desirable properties similar to those of the normal distribution in that it is closed under marginalization, conditioning and linear transformations (full column or row rank) [Dominguez-Molina et al., 2003]. Komárek and Lesaffre [2008] replaced the normal prior model in GLMM with a penalized Gaussian mixture distribution. They assumed the random effects $b_1, \ldots, b_N$ were independently and identically distributed with a density $g(b)$. They standardized the random effects by a vector of unknown scale parameters and modeled the shape of the density of these standardized random effects using the penalized splines method. Some studies have also used Bayesian nonparametric spatial modeling approaches for disease incidence data. This allows for data driven deviations from the normality assumption for the spatial random effects. Many studies have employed a form of Dirichlet process prior or its modifications to model the random effects non-parametrically. Kleinman and Ibrahim [1998] used a mixture Dirichlet process (MDP) structure for their model on marker data from an AIDS study. They found that the MDP model is useful in generalized linear mixed models where inferences are sensitive to distributional assumptions on the random effects.

## 5.4 Mixture of Dirichlet Process and Polya tree processes for random effects

In the generalized linear mixed models (GLMM), unobservable variables are taken into account via the random effects. The GLMM is a hierarchical model that has

a structure that uses the first stage modeling in the observed outcomes and the second stage mostly involves an exchangeable prior distribution on the unobservables (random effects), which parametrize the distribution of the observables. A problem now arises concerning the form of the random effects distribution. Many studies have used the normality assumption for random effects. This is usually because of computational simplicity. The information about the distribution of the random effects is usually unavailable, and this might lead to poor parameter estimates when the distribution is miss-specified [Walker and Mallick, 1997]. Estimates of the covariate effects may also show changes in sign and magnitude depending on the form of the random effect distribution [Heckman and Singer, 1984; Laird, 1978]. Ferreira and Garcia [2001] also note that asymptotic unbiasedness for estimates of random effects variance depends on the form of the random effect distribution. The assumption that the random effects arise from a known parametric distribution is therefore not always correct.

Though not widely used in practice, a more flexible approach would be to model the random effects terms non-parametrically. This flexibility can be introduced when the random effects distribution is drawn from a large class of distributions. Such a large class can be formed by using nonparametric approaches to model the random effect distribution. In this section we demonstrate the use of Dirichlet process mixture of normals and the Polya tree processes for random effects.

## 5.5 Mixture of Dirichlet processes(MDP) Model

The Dirichlet process (DP) is a stochastic process used in Bayesian nonparametric models of data. It is a distribution over distributions implying that each draw from a Dirichlet process is itself a distribution. The Dirichlet distribution is a multivariate generalization of a beta distribution and was introduced by [Ferguson, 1973]. Although the DP is a simple and computationally tractable prior for an unknown distribution, it produces distributions that are discrete with probability 1, making it unsuitable for density modeling. This can be avoided by convoluting the distribution with some continuous kernel, or more generally, by using a DP to define a mixture distribution with infinitely many components, each of some simple parametric form [Jara et al., 2012]. This approach is called DP mixtures (DPM) [Escobar and West, 1995]. A random vector $X = (x_1, x_2 \ldots, x_n)$ is said to have a Dirichlet distribution with parameter $\alpha = (\alpha_1, \ldots, \alpha_n) > 0$: $X \backsim Dir(\alpha)$ if its density function is given by:

$$f(x, \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{n} x_i^{\alpha_i - 1} I(X \in S), \tag{5.1}$$

as given by Johnson et al. [2002], where $B(\alpha)$ is a generalized beta function of $\alpha$.

$I(X \in S)$ is an indicator function that $X$ is in the probability simplex $S$. The probability simplex $S$ is such that $S = \{X \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^{n} X_i = 1\}$.

For the MDP, suppose an $n_i \times 1$ random vector $x_i$ has a univariate normal distribution indexed by the $w \times 1$ vector $\theta_i$, $i = 1, 2 \ldots, n$. Suppose also that the $\theta_i$

themselves have a prior distribution with known hyperparameters $\varphi_0$. Thus

$$\begin{aligned} Stage1 &: [x_i|\theta_i] \backsim D_{n_i}(h_1(\theta_i)) \\ Stage2 &: [\theta_i|\varphi_0] \backsim D_w(h_2(\varphi_0)) \end{aligned}$$

where $D_s(.)$ is a generic label for an s-dimensional parametric multivariate distribution and $h_1$ and $h_2$ are functions of $\theta_i$ and $\varphi_0$ respectively [Kleinman and Ibrahim, 1998]. The MDP model (Escobar and West [1995]; MacEachern [1994]) removes the assumption of a parametric prior at the second stage and replaces it with a general distribution $G$ which in turn has a Dirichlet process prior as discussed in Ferguson [1973], i.e. suppose the random vector $x_i$ has a univariate normal distribution indexed with unknown mean $\theta_i$ and a known variance ${\sigma_x}^2$. Suppose also that each $\theta_i$ has also a univariate normal distribution. Then the classical specification is:

$$\begin{aligned} Stage1 &: [x_i|\theta_i, \sigma_x^2] \backsim N(\theta_i, {\sigma_x}^2) \\ Stage2 &: [\theta_i|\mu, \varphi_0] \backsim N(\mu, {\sigma_0}^2) \end{aligned}$$

The MDP model removes the normality assumption in the second stage resulting in;

$$\begin{aligned} Stage1 &: [x_i|\theta_i, \sigma_x^2] \backsim N(\theta_i, {\sigma_x}^2) \\ Stage2 &: \theta_i \backsim G \\ Stage3 &: [G|\alpha, \varphi_0] \backsim DP(\alpha \bullet G_0(h_2(\varphi_0))) \end{aligned}$$

where $G_0$ is a $w$ dimension parametric distribution and $\alpha$ is a positive scalar.

## 5.6 The statistical model

Let $y_{ij}$ be the HIV status (0 or 1), for individual $j$ in county $i$:$i = 1, 2, \ldots, 46$. We assume that the dependent variable $y_{ij}$ is Bernoulli distributed i.e. $y_{ij}|p_{ij} \backsim Bernoulli(p_{ij})$. The unknown $E(y_{ij}) = p_{ij}$ relates to the predictors as follows:

$$h(p_{ij}) = X^T{}_{ij}\beta + u_i \tag{5.2}$$

where the vector $X^T{}_{ij}$ contains $p$ categorical predictors with the first component accounting for intercept and $u_i$ represents the conditional area specific random effects, $h(.)$, a logit link function, $\boldsymbol{\beta}$ is a $p$ dimensional vector of regression coefficients for the categorical predictors. In this study we use $p = 8$ categorical variables for demonstration purposes.

## 5.7 Prior distribution

The fixed effects are assigned a flat normal distribution i.e. $\beta \backsim N_p(\mu_0, \sum_0)$ while the prior for the random effects are specified as follows;

$$u_i|G \backsim G$$

$$G|\alpha, G_0 \backsim DP(\alpha G_0)$$

Where $G_0 = N(\mu|\lambda, \sum)$. The independent hyperpriors are given as in Jara et al. [2012];

$$\begin{aligned} \alpha|\alpha_0, b_0 &\backsim Gamma(\alpha_0, b_0) \\ \lambda|\lambda_b, S_b &\backsim N(\lambda_b, S_b) \\ \sum|\nu_0, T &\backsim IW(\nu_0, T) \end{aligned} \tag{5.3}$$

## 5.8 The Mixture of Multivariate Polya Trees (MMPT) prior for random effects

Instead of convolving the distribution with some continuous kernel, or more generally, by using a DP to define a mixture distribution with infinitely many components, each of some simple parametric form as stated above, one can consider Bayesian parametric models which admit continuous distributions. An example of such a model is the Polya trees that can be viewed as generalizations of the DP [Ferguson, 1973]. Polya trees (PT) and mixture of Polya trees (MPT) provide a highly flexible non parametric alternative to the traditional parametric and Dirichlet process mixture process. One downside of the DP and the MDP priors is that they suffer from intractability in some settings due to the discreteness of the DP [Johnson and Christensen, 1989].

## 5.9 Definition

Let $E = \{0, 1\}$, $E^0 = \oslash$, $E^m$ be the m-fold product $E \times E \times E \times \ldots \times E$, $E^* = \bigcup_0^\infty E^m$ and $E^N$ be the set of infinite sequences of elements of $E$. Let $\Omega$ be a separable measurable space,$\pi_0 = \Omega$ and $\Pi = \{\pi_m; m = 0, 1, \ldots\}$ be a separating binary tree of partitions of $\Omega$; that is, let $\pi_0, \pi_1, \ldots$ be a sequence of partitions such that $\bigcup_0^\infty \pi_m$ generates the measurable sets and such that every $B \in \pi_{m+1}$ is obtained by splitting some $B^{'} \in \pi_m$ into two pieces. Let $B_\oslash = \Omega$ and, for all $\varepsilon = \varepsilon_1, \ldots \varepsilon_m \in E^*$, let $B_{\varepsilon 0}$ and $B_{\varepsilon 1}$ be the two pieces into which $B_\varepsilon$ is split. Degenerate splits are permitted, for example, $B_\epsilon = B_{\epsilon 0} \bigcup \oslash$.

A random probability measure $\rho$ on $\Omega$ is said to have a Polya tree distribution, or Polya tree prior with parameters $\Pi$ and $A$ i.e. $\rho \sim PT(\Pi, A)$, if there exists non-negative numbers $A = \{\alpha_\in : \varepsilon \in E^*\}$ and random variables $y = \{Y_\varepsilon : \varepsilon \in E^*\}$ such that;

1. All the random variables $y$ are independent;

2. For every $\varepsilon \in E^*$, $Y_\varepsilon$ has a Beta distribution with parameters $\alpha_{\epsilon 0}$ and $\alpha_{\varepsilon 1}$

3. For every $m = 1, 2, \ldots$ and every $\varepsilon \in E^m$ ,

$$\rho\left(B_{\varepsilon_1 \ldots \varepsilon_m}\right) = \left(\prod_{j=1;\epsilon_j=0}^{m} Y_{\varepsilon_1 \ldots \varepsilon_{j-1}}\right)\left(\prod_{j=1;\epsilon_j=1}^{m} \left(1 - Y_{\varepsilon_1 \ldots \varepsilon_{j-1}}\right)\right) \tag{5.4}$$

where the first term in the product is interpreted as $Y_\oslash$ and the second as $1 - Y_\oslash$. The random variables $\Theta, \Theta_2, \ldots$ are said to be samples from $\rho$ if, given $\rho$, they are i.i.d. with distribution $\rho$. The $Y_\varepsilon^{'}s$ are interpreted as follows: $Y_\oslash$ and $1 - Y_\oslash$ are

respectively the probabilities that $\Theta_i \in B_0$ and $\Theta_i \in B_1$, and $Y_\epsilon$ and $1 - Y_\epsilon$ are the conditional probabilities that $\Theta_i \in B_{\varepsilon 0}$ and $\Theta_i \in B_{\varepsilon 1}$ given that $\Theta_i \in B_\varepsilon$ [Lavine, 1992].

### 5.9.1 The Model

As discussed in the MDP model above,

$$h(p) = X^T \beta + u_i$$

where we suppress the indicators for simplicity. As adopted from Jara et al. [2012], the prior distributions are as follows; the fixed effects are assigned flat normal distribution prior $\beta \backsim N_p(\gamma_0, \sum_0)$ and the random effects $u_i$ the MMPT prior i.e.

$$u_i | G \backsim G$$

$$G | \alpha, \lambda, \sum, O \backsim PT^M(\prod^{\lambda, \sum, O} A)$$

where $O$ is an orthogonal matrix defining the decomposition of the centering covariance matrix. As in Hanson [2012], the PT prior is centered around the $N_d(\lambda, \sum)$ distribution. Jara et al. [2012] considered the class of partitions $\prod^{\lambda, \sum, O}$ where partitions starts with base sets that are cartesian products of intervals obtained as quantiles from the standard normal distribution. A multivariate location-scale transformation $\theta = \lambda + \sum^{1/2} Z$ is applied to each base set yielding the final sets where $\sum^{1/2} = T'O'$, being the unique upper triangular Cholesky matrix of $\sum$. The family $A = \{\alpha_e : e \in E^*\}$, where $E^* = \bigcap_{m=0}^{M} E_d^m$ with $E_d$ and $E_d^m$ are the $d$

fold product of $E = \{0, 1\}$ and the $m$ fold of product $E_d$ respectively. The family is specified as $\alpha_{e1} \dots e_m = \alpha m^2$. To complete the model specification, independent hyperpriors are assumed i.e. i.e.

$$\begin{aligned} \alpha_0, b_0 &\backsim Gamma(\alpha_0, b_0) \\ \lambda|\lambda_b, S_b &\backsim N(\lambda_b, S_b) \\ \sum|\nu_0, T &\backsim IW(\nu_0, T) \qquad O \backsim Haar(q) \end{aligned} \tag{5.5}$$

where Haar() denotes the prior distribution (measure) for the orthogonal matrix defining the decomposition of the centering covariance matrix.

### 5.9.2 Haar Measure

Suppose that $M$ is a compact metric space (such as a sphere in $\mathbb{R}^n$), and that $G$ is a group of isometries of $M$.

1. There exists a Borel probability measure $\mu$ on $M$ which is invariant under $G$. That is, $\mu(S) = \mu(gS)$ for all $g \in G$ and $S \subset M$.

2. If $G$ is transitive, then the Haar measure is unique. Here, "transitive" means for all $x, y \in X$, there exists $g \in G$ such that $gx = y$. Where $x$ and $y$ are any arbitrary pair [Milman and Schechtman, 2009].

For the normal model the random effects are assigned the normal priors.

## 5.10 Posterior distribution

This is obtained by updating the prior distribution with the observed data. The posterior distribution gives the sample for Bayesian inference. We employ the Markov chain Monte Carlo (MCMC) for direct sampling from this posterior distribution hence overcoming the problem of high dimensionality. The downside of MCMC is that it is slow. The analyses for this study were carried out using R packages; DPpackage and MCMCglmm. The DPpackage was used to fit the MDP and the MMPT models for random effects while the MCMCglmm was used to fit the GLMM with normal random effects.

## 5.11 Model Diagnostics

Spiegelhalter et al. [2002] suggested the deviance information criterion (DIC) for model diagnostics and this was used for this study to compare model fit. The DIC value is obtained as: $DIC = \overline{D}(\theta) + pD$, where $\overline{D}$ is the posterior mean of the deviance that measures the goodness of fit while $pD$ gives the effective number of parameters in the model which penalizes for complexity of the model. In DIC, low values of $\overline{D}$ indicate a better fit while small values of $pD$ indicate model parsimony. The best fitting model is one with the smallest DIC. Studies have shown that a difference of 3 in DIC between two models cannot be distinguished while a difference of between 3 and 7 can be weakly differentiated [Spiegelhalter et al., 2002]. The DIC has its limitations including the fact that $pD$ is not invariant to reparameterization, for example different values of $pD$ are obtained if parameterization

is done in terms of $\sigma$ or $\log(\sigma)$, even if the priors on each were mathematically equivalent. The $pD$ component can also sometimes be negative if the posterior of $\theta$ is very non-normal and so $\overline{\theta}$ does not provide a very good estimate of $\theta$. Another limitation of the DIC is lack of consistency. However the main aim of DIC is in optimizing the short-term prediction of a particular type and not trying to identify a 'true' model. The DIC is also not based on a proper predictive criterion as it uses plug-in predictions $P(Y^{rep}|\widehat{\theta})$ rather than the full predictive distributions given by $P(Y^{rep}) = \int P(Y^{rep}|\theta)P(\theta|y)\, d\theta$ which would provide invariance to reparameterization as noted by Spiegelhalter et al. [2002]. Celeux et al. [2006] showed in their context of mixture models that DIC is not based on a universal principle that could lead to a procedure that was both computationally practical and generically applicable and therefore has a weak theoretical justification. Some studies have tried to address these issues. Spiegelhalter et al. [2014] suggests patching up the $pD$ so that instead of using the posterior mean of the stochastic parents of $\theta$ i.e. if there are stochastic nodes $\varphi$ such that $\theta = f(\varphi)$, then $D(\tilde{\theta}) = D\{f(\tilde{\varphi})\}$, it would be better to use the posterior mean of an appropriate function of the 'direct parameters' to give the plug-in deviance. Stochastic nodes are variables that are given a distribution, they may be parents or children (parameters of other distributions or both). Stochastic nodes may be observed in which case they are data , or may be unobserved and hence be parameters. Gelman et al. [2014] suggested an alternative measure of complexity denoted as $p_v$; Suppose that one has a non-hierarchical model with a weak prior, so that $D(\theta) \approx D(\overline{\theta}) + {\chi^2}_k$, where $E[{\chi^2}_k] = k$ the true number of parameters then $E[D(\theta)] \approx D(\overline{\theta}) + k$ so $pD \approx k$ and $V\{D(\theta)\} \approx 2k$. Thus, with negligible prior information, half the variance of the deviance is an

estimate of the number of free parameters in the model. This estimate generally turns out to be remarkably robust and accurate, and this has suggested the use of $p_v = V(D)/2$ as an estimate of the effective number of parameters in a model in more general situations with informative prior information. The posterior distribution of the deviance is not affected by equivalent reparameterizations, and so $p_v$ will be invariant to reparameterization see Spiegelhalter et al. [2014].

## 5.12 Results

TABLE 5.1: Model comparison statistics for the tree models

| | $\overline{D}$ | $\widehat{D}$ | $pD$ | $DIC$ |
|---|---|---|---|---|
| Normal-Model | 2417.54 | 2370.41 | 47.13 | 2464.67 |
| MDP4 | 2436.36 | 2402.97 | 33.39 | 2469.75 |
| MMPT | 2403.00 | 2403.00 | 0 | 2403.00 |

Table 5.1 shows the DICs for the three separately fitted models for HIV. The model with the smallest DIC provides the best fit. From the table, the MMPT model provided the best fit since it had the smallest DIC. From Table 5.2, all the

TABLE 5.2: Parameter estimates for risk factors of HIV and their corresponding 95% credible intervals from the tree candidate models

| | NORMAL-MODEL | MDP | MMPT |
|---|---|---|---|
| (Intercept) | -1.350(-2.449,-0.206) | -1.420(-2.782,-0.113) | -1.350(-1.467,-1.227) |
| Education level | -0.095(-0.272,-0.009) | -0.099(-0.236,0.039) | -0.100(-0.100,-0.100) |
| Age at first sex | -0.319(-0.453,-0.193) | -0.320(-0.468,-0.160) | -0.320(-0.320,-0.320) |
| Perceived risk | 0.026(-0.121,0.137) | 0.032(-0.092,0.154) | 0.032(0.032,0.032) |
| Partners had last 1yr | 0.479(0.083,1.033) | 0.487(0.029,0.937) | 0.483(0.483,0.483) |
| Residence | 0.460(0.146,0.814) | 0.460(0.108,0.796) | 0.440(0.440,0.440) |
| Freq of travel away | 0.054(-0.038,0.154) | 0.055(-0.060,0.169) | 0.05790.057,0.057) |
| Marital status | 0.168(0.093,0.273) | 0.151(0.072,0.228) | 0.149(0.149,0.149) |
| STI in last 1yr | -0.601(-2.121,-0.014) | -0.591(-1.113,-0.031) | -0.611(-0.611,-0.611) |

three models provided almost equal parameter estimates.

From the MMPT model, the number of partners an individual had in the last one year had a strong positive effect on HIV status. Education level, age at first sex and whether an individual had contracted a sexually transmitted infection (STI) in the last one year had negative effects on HIV. As discussed earlier, estimates of the covariate effects may also show changes in sign and magnitude depending on the form of the random effect distribution. From the table, the changes in magnitude are negligible while the sign remains the same on the three choices of random effects. The multivariate mixture of Polya trees provides a highly flexible non parametric alternative to the traditional parametric and Dirichlet process mixture process. It is worth noting that the results from the three models i.e. the normal, MDP and MMPT are almost similar suggesting that the random effect distribution is approximately normal. In such a case these complex models are not warranted. However when the random effects distribution departs from the normal distribution, it would be prudent to employ other distributions such as the MDP, MMPT etc.

## 5.13 Linearity

The assumption that all covariates have a linear relationship with the response variable may also be limiting. The assumption places a parametric constraint on the shape of the exposure-response relationship and disallows the adjustments for cyclical patterns in cofounders. This linear relationship is not necessarily true for all covariates as some may have a non-linear relationship with the response variable. Hastie and Tibshirani [1990] introduced a class of models called the

generalized additive models where they replaced each linear term in the additive logistic regression by a more general functional form $f_t(.)$.

Several studies have discussed extensively the methods for estimating the smooth function $f_t(.)$. Green and Silverman [1993] used the penalization and splines to model the smooth function $f_t(.)$. Some other methods include the p-spline method by Lang and Brezger [2004], continuous indexed spline models by Wahba [1978], Gaussian processes by O'Hagan and Kingman [1978], the penalized regression splines method proposed by Eilers and Marx [1996] and the random walk models [Cleveland et al., 1992; Fahrmeir and Tutz, 2001].

In the section that follows we discuss the development of the joint spatially varying coefficient model.

## 5.14 Stationarity

The assumption that the relationship effect between the explanatory variable and the response variables in a regression model are constant across the study region may be highly restrictive for spatial processes as factors such as sampling variation, different relationships across space e.g. attitudes, preferences, culture etc. may contribute to a different response to the same stimuli as one moves across space. Two competing spatially varying models are geographically weighted regression (GWR) and the Bayesian spatially varying coefficient process (BSVCP). The GWR addresses this by estimating the coefficients $\beta's$ by the weighted least squares method, where more emphasis in terms of weights are placed on the observations

which are close to location $i$, since it is assumed that the observations close to $i$ exert more influence on the parameter estimates at location $i$ than those farther away [Fotheringham et al., 2003]. The weighting schemes can be fixed or adaptive. In the fixed distance scheme, observations that are within some distance $d$ are given the weight of 1 while those farther away beyond some distance $d$ from location $i$ are given a weight of zero, while in the adaptive scheme, weights of observations inside some radius $d$ are made to decrease monotonically to zero as the radius increases. In the BSVCP model, the covariates are allowed to vary spatially by assigning its coefficients the conditional autoregressive (CAR) model, the Bayesian autoregressive (BAR) or the simultaneous autoregressive (SAR) [Assunçao, 2003].

## 5.15 Joint modeling

Univariate disease mapping has been common in many studies. Many diseases however share common risk factors. Herpes simplex virus type-2 is for example associated with a two to threefold increased risk of HIV acquisition and an up to fivefold increased risk of HIV transmission per-sexual act, and may account for 40% to 60% of new HIV infections in populations where HSV-2 has a high prevalence [Looker et al., 2008]. The joint modeling of two or more diseases across a geographical area to estimate relative risks is of both methodological and epidemiological importance. By pooling all the available data from different disease sources, there are gains in precision and efficiency of estimates especially in rare diseases [Dabney and Wakefield, 2005]. Joint modeling of diseases other than being useful in helping to identify disease specific risk factors also provides estimates and inferences on

the pairwise and cross-covariances between the risks of disease outcomes [Dabney and Wakefield, 2005; Manda et al., 2012]. The joint modeling is usually initiated via the random effects by a number of possible approaches among them the multivariate normal distribution (MVN), the multivariate conditional autoregressive model (MCAR) and the multiple membership multiple classification (MMMC) approaches. The random effects can be decomposed into structured random effects which accounts for any unobserved covariates which vary spatially across the regions and unstructured random effects which caters for the unobserved covariates that are inherent within the regions under study.

Suppose that $y_{ij1}$ and $y_{ij2}$ represents the disease 1 and 2 status respectively of individual $j$ living in country $i$. We assume that the dependent variable $y_{ijk}$ follows a Bernoulli distribution, i.e. $y_{ijk}|p_{ijk} \backsim (p_{ijk})$. The unknown

$$h(p_{ij1}) = X^T\beta_1 + W^T{}_{ij1}\gamma_1 + u_{i1} + v_{i1}, \text{ for disease 1} \tag{5.6}$$

and

$$h(p_{ij2}) = X^T\beta_2 + W^T{}_{ij2}\gamma_2 + u_{i2} + v_{i2}, \text{ for disease 2} \tag{5.7}$$

where the vector $X_{ijk} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})'$ contains $p$ continuous predictors and $W_{ijk} = (w_{ij1}, w_{ij2}, \ldots, w_{ijr})'$ contains $r$ categorical predictors with the first component accounting for the intercept, $u_i$ and $v_i$ represents the unstructured and the structured random effects respectively. A bivariate model to measure risk for the two diseases can be imposed via $u_{ik}$ and $v_{ik}$ or both $u_i = (u_{i1}, u_{i2})^T$

and $v_i = (v_{i1}, v_{i2})^T$. The unstructured random effects $u_i = (u_{i1}, u_{i2})^T$ are assigned a bivariate normal distribution with covariance matrix, $\sum_u$ to allow for correlation between the disease risks, $u_i \backsim MVN_2(0, \sum_u)$, where $\sum_{u11} = \sigma^2_{u1}$. Similarly, for the spatially structured terms $v_i = (v_{i1}, v_{i2})^T$, either the bivariate intrinsic conditional autoregressive model (ICAR) or the MMMC model is used. For the bivariate ICAR model, the structured terms are assigned a bivariate normal distribution, $v_i \backsim MVN_2(\overline{V_i}, \sum_v)$, where $\overline{V_i}$ is the mean vector: $\overline{V_i} = (\sum_{i\in\Theta_i} v_{i1}/m_i, \sum_{i\in\Theta_i} v_{i2}/m_i)^T$, where $\Theta_i$ is the set of neighbors, $m_i$ is the number of neighbours of area $i$ and $\sum_v$ is the covariance matrix of $v_i = (v_{i1}, v_{i2})^T$. The conditional variance for $v_{i1}$ and $v_{i2}$ respectively are $\sum_{v11} = \sigma^2_{v1}/m_i$ and $\sum_{v22} = \sigma^2_{v2}/m_i$. Manda et al. [2012] applied the MMMC models to spatial epidemiology models. They used two classifications: an area classification capturing the non-spatial variation (classification level 2) and a neighbor classification (classification level 3) to capture effects due to neighboring areas. They used the following notations where the superscript represents the classification levels:

$b_i = \sum_{i\neq j} W_{ij} u_j^{(3)}$ where $W_{ij}$ is the weighting factor that relates area $i$ to each of the neighbor $j$ in the neighborhood set $\Theta_i$ and $u_j^{(3)}$ is the effect of area $j$ on area $i$ weighted by $W_{ij}$ while the non-spatial random effects $u^{(2)}_{area(i)}$ are assigned independent normal distributions, $u^{(2)}_{area(i)} \backsim N(0, \sigma^2_{u(2)})$ and areas in the classification set $\Theta_i$ have random effects $u_j^{(3)} \backsim N(0, \sigma^2_{u(3)})$. The standard choice of the weighting function is similar to that of the MCAR (CAR) model i.e. $W_{ij} = \frac{1}{m_i}$ where $m_i$ is the number of neighbors implying that the more the neighbors an area has the more precision is for that area. The difference between the MCAR

and the MMMC is that for MCAR, the spatial correlation is achieved through a variance structure while for MMMC the spatial correlation is achieved through a multiple membership relationship and that the neighborhood random effects are not independent.

## 5.16 Application

In the application, we relax the stationarity assumption using the CAR model while at the same time modeling HIV and HSV-2 jointly using the MVN and the MCAR models.

We compared four models in assessing the effect of the covariates across the counties in Kenya, the unobserved effects on the distribution and relationship between HIV and HSV-2 in Kenya based on the female data.

$$\text{Model 1}: \quad logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma \text{ for HSV-2}$$

$$\text{Model 2}: \quad logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma + f_{unstr}(S_{i1}) \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma + f_{unstr}(S_{i2}) \text{ for HSV-2}$$

$$\text{Model 3}: \quad logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma + f_{str}(S_{i1}) \text{ for HIV}$$

$$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma + f_{str}(S_{i2}) \text{ for HSV-2}$$

Model 4 : $logit(\rho_{ij1}) = \beta_{01} + f(age) + W^T\gamma + f_{unstr}(S_{i1}) + f_{str}(S_{i1})$ for HIV

$logit(\rho_{ij2}) = \beta_{02} + f(age) + W^T\gamma + f_{unstr}(S_{i2}) + f_{str}(S_{i2})$ for HSV-2

**Model 1**: This is a model of the eight categorical covariates that were allowed to vary spatially and one continuous covariate, age, modeled with a non-linear smooth function. This model does not take into account the spatially structured and the spatially unstructured random effects and the two diseases are modeled independently.

**Model 2**: This is an additive model that assumes non stationarity for the categorical covariates, non-linear effect of the continuous covariate; age and spatially unstructured random effects which cover the unobserved covariates that are inherent within the counties. The joint modeling here is initiated by the multivariate normal distribution.

**Model 3**: This model explores the non stationarity effect of the categorical, non-linear covariate age and spatially structured random effects which accounts for any unobserved covariates which vary spatially among counties. The joint modeling in model 3 is initiated by the multivariate conditional autoregressive model.

**Model 4**: Examines the nonlinear effect of age, spatially varying effects of the categorical covariates and a convolution of spatially structured and spatially unstructured random effects, and the joint modeling is initiated by both the multivariate normal distribution and the multivariate conditional autoregressive model.

## 5.17 Priors for the parameters

The hyper parameters were assigned the inverse gamma distributions as: $\tau^2_{str} \backsim IG(0.0001, 0.0001)$ and $\tau^2_{unstr} \backsim IG(0.0001, 0.0001)$. The coefficients were given the following prior distributions $\phi_0, \phi_1, \ldots \phi_p \backsim N(0, 10^6)$, $\lambda_1, \lambda_2, \ldots \lambda_r \backsim N(0, 10^6)$, $b_k \backsim N(0, \tau_b^2)$ and $\tau_b^2 \backsim IG(0.0001, 0.0001)$, and the intercepts: $\beta_1, \beta_2 \backsim N(0.01, 0.01)$ [Ngesa et al., 2014b].

## 5.18 Posterior Distribution

The posterior distribution gives the sample draws for Bayesian inference. It is obtained by updating the prior distribution with the observed data. We employ the Markov chain Monte Carlo (MCMC) for direct sampling from this posterior distribution hence overcoming the problem of high dimensionality. The downside of MCMC is that it is slow, the integrated nested laplace approximation (INLA) Rue et al. [2009] is among the methods that have been developed to circumvent this problem. The quality of the integrated nested Laplace approximations can be assessed using simulation studies.

If we assume conditional independence between the response variable and the hyper parameters, the posterior distribution for the Bernoulli model is given by:

$$
\begin{aligned}
P_{post}(\phi, \lambda, b, \tau^2|y) &\propto L(y|\phi, b, \tau^2) P_{pri}(\phi, \lambda, b, \tau^2) \qquad (5.8)\\
&= \prod_i \prod_j L(y_{ij}|\theta, \lambda, \tau^2) \prod_{k=1}^{p} [P(b_k|{\tau^2}_k) P({\tau^2}_k)] \times \\
&\qquad \prod_{j=1}^{r} [P(\gamma_j|{\tau^2}_j) P({\tau^2}_j)] \times \\
&P(f_{str}|{\tau^2}_{str}) P({\tau^2}_{str}) P(f_{unstr}|{\tau^2}_{unstr}) P({\tau^2}_{unstr})
\end{aligned}
$$

## 5.19 Results

Table 5.3 gives the model diagnostics for the 4 fitted models. Model 2 has the least total DIC value and hence provided the best fit. Model 2 has the least individual DIC value for HIV while model 4 has the least individual DIC value for HSV-2. The subsequent discussions and results are therefore based on the best fitting model 2 in adherence to model parsimony. The model allows the covariates to vary spatially by the CAR model and also captures the unstructured random effects.

TABLE 5.3: Model Comparison

| | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|---|---|
| | HIV | HSV-2 | HIV | HSV-2 | HIV | HSV-2 | HIV | HSV-2 |
| Individual pD | 39.809 | 72.266 | 32.543 | 67.945 | 35.127 | 71.887 | 36.304 | 72.38 |
| Individual $\overline{D}(\theta)$ | 2358.47 | 5761.97 | 2354.89 | 5755.24 | 2360.86 | 5755.49 | 2358.21 | 5748.82 |
| Individual DIC | 2398.28 | 5834.23 | 2387.43 | 5823.18 | 2395.99 | 5827.38 | 2394.51 | 5821.2 |
| Total DIC | | 8232.52 | | 8210.61 | | 8223.36 | | 8215.72 |

### 5.19.1 Model assessment and comparison

## 5.20 Joint modeling

### 5.20.1 Joint Spatially Varying Effects

The results of the spatially varying coefficients are presented in choropleth maps and the correlation between the effect of these varying coefficients on HIV and HSV-2 prevalence are discussed. In particular the discussions are based on some of those covariates whose effects on HIV and HSV-2 prevalence have strong positive correlation. Figure 5.1 shows the map of Kenya divided into various regions.

### 5.20.2 Age at first sex

Figure 5.2 depicts clearly that the effect of age at first sex on HIV and HSV-2 prevalence varies spatially, with greater effects in the Central, some parts of North Eastern, Lake and Rift valley regions. Age at first sex is negatively associated with HIV and HSV-2 prevalence [MacEachern, 1994]. There is a greater chance of

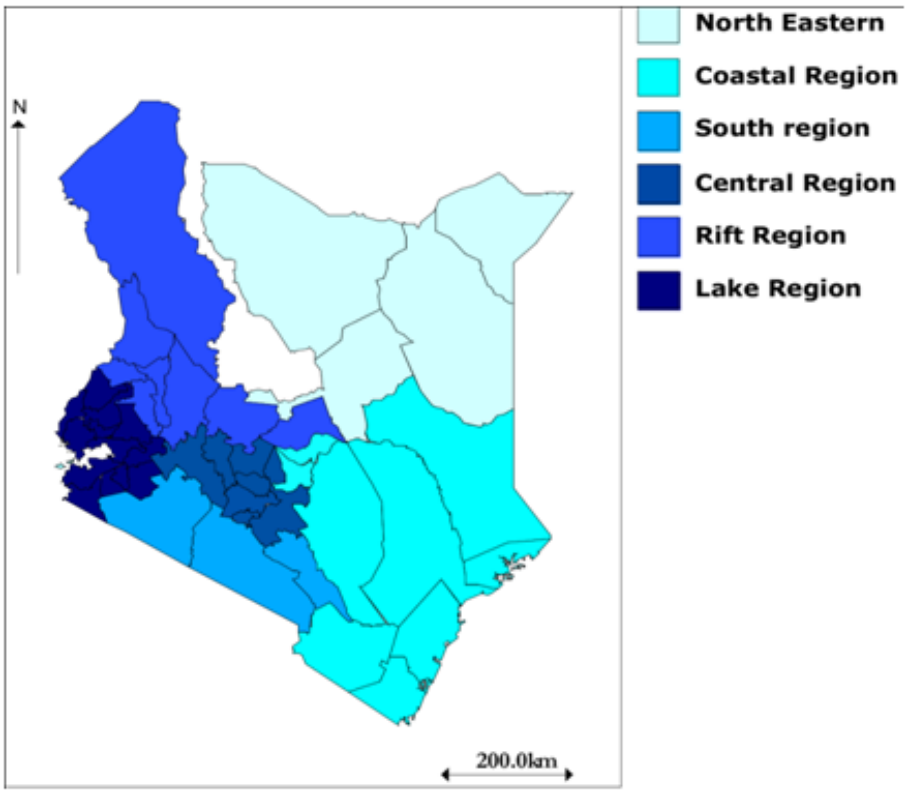



FIGURE 5.1: The map of Kenya

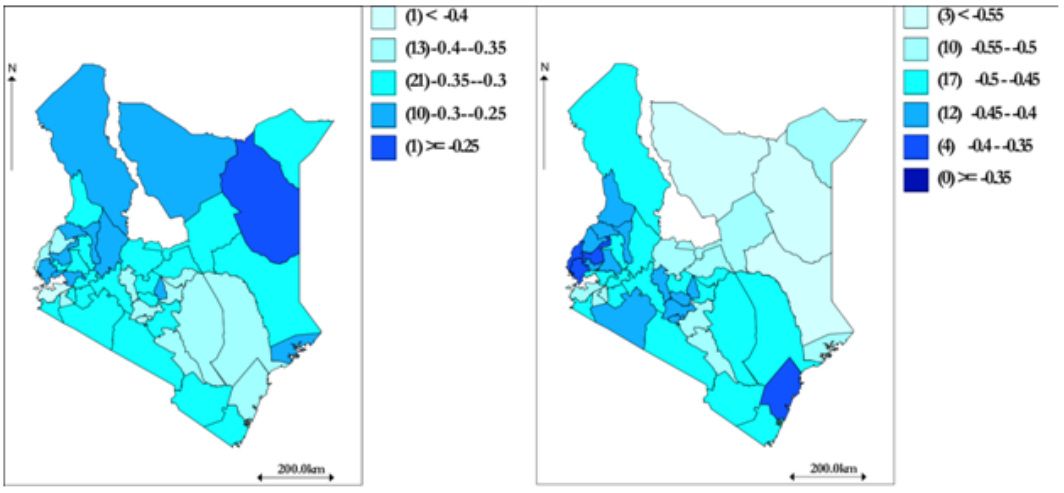



FIGURE 5.2: The effect of age at first sex on HIV (left panel) and HSV-2 prevalence (right panel)

contracting HIV or HSV-2 for those individuals who had their first sexual intercourse at an earlier age than those who had at an older age. There was a strong correlation of 0.6073 between the effect of age at first sex on HIV and HSV-2 prevalence implying that age at first sex has similar effects on both HIV and HSV-2

prevalence.

### 5.20.3 Education Level

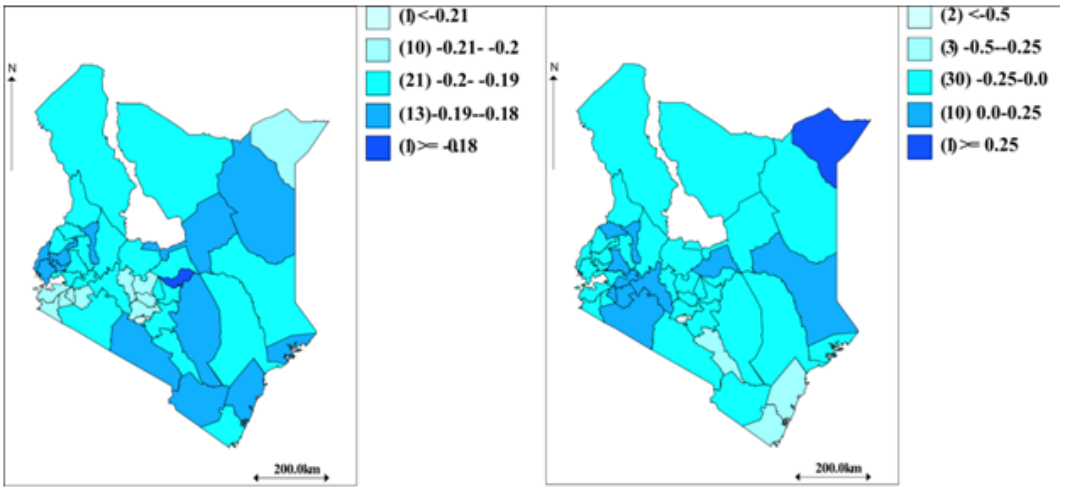



FIGURE 5.3: Effect of Education level on the prevalence of HIV (left panel) and HSV-2 (right panel)

The effect of education level on HIV and HSV-2 prevalence was high throughout the country. The effect of education level was greater in areas with darker shading decreasing to the areas with lighter shading. Women with higher education qualifications were less likely to test positive for HIV and HSV-2 than those with lower education qualification [MacEachern, 1994]. The effect of education level on HIV and HSV-2 were highly correlated: 0.765. Place of residence is associated with HIV and HSV-2 infection [MacEachern, 1994]. Urban residents were more likely to test positive for both HIV and HSV-2 than the rural residents. The effect of place of residence on HIV and HSV-2 prevalence were positively correlated: 0.541. The effect of place of residence on HIV prevalence was the least in the North Eastern region and greatest in some parts of Coastal and Lake region. The effect on HSV-2 was more in the Lake, Coastal and some parts of the Southern region.

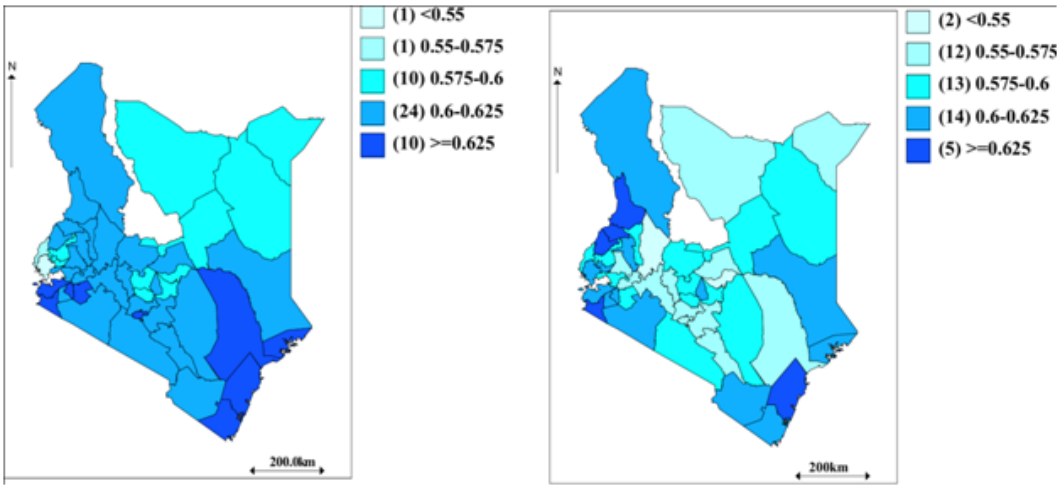



FIGURE 5.4: Effect of place of residence on HIV and HSV-2 status

## 5.21 Joint Spatial effects

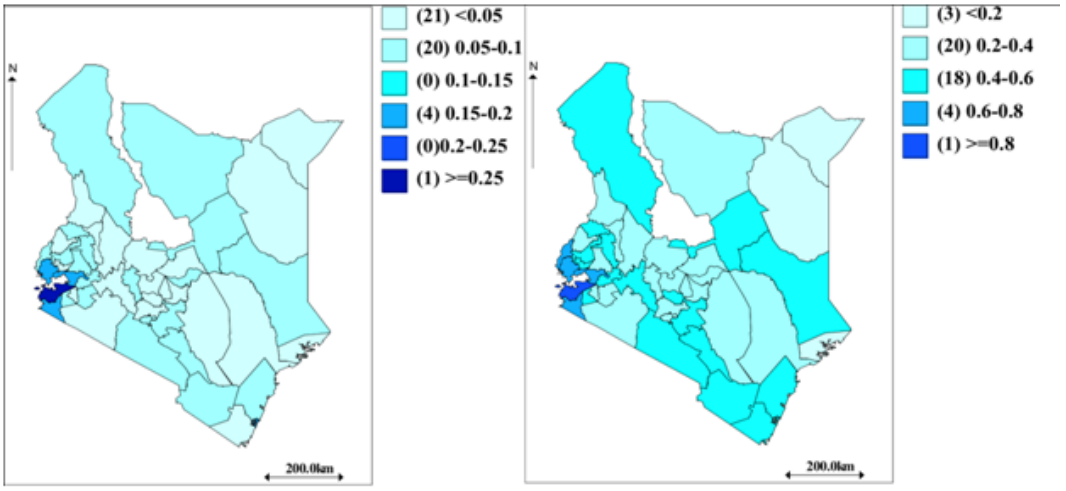



FIGURE 5.5: Residual spatial effect of County on HIV (left panel) and HSV-2 (right panel)

Figures 5.5 show a high association between HIV and HSV-2. There was a strong positive correlation of 0.854 between HIV and HSV-2. Those counties that registered a high HSV-2 prevalence also had high HIV prevalence. HSV-2 is associated with a two- to three-fold increased risk of HIV acquisition and an up to five-fold increased risk of HIV transmission per-sexual act, and may account for 40% to 60% of new HIV infections in populations where HSV-2 has a high prevalence [Looker

et al., 2008]. The HSV-2 map depicts a generally high HSV-2 prevalence across the country putting a large population at risk. The joint modeling has unmasked this strong association and therefore interventions can be put in place in order to help curb HSV-2 in areas where it has high prevalence as this would in turn result in reducing the HIV incidence.

## 5.22 Discussion

In this study we outlined some of the limiting assumptions in disease mapping and reviewed the methods other studies have employed in trying to relax them. The normality assumption does not necessarily hold as some random effects may exhibit skewness, fat-tailness, multimodality e.t.c. This may obscure some important features of between subjects and within subjects' variations. The idea of also using normal random effects to model say binomial or count data may also be limiting. Linearity assumption is also not always true for all covariates. Age for example has been found to have a non-linear relationship with HIV and HSV-2 infection. Other studies have made stationarity assumption in that one stimulus e.g. education, provokes the same response in all the regions under study and this is also quite restrictive. Responses to stimuli may vary from region to region due to aspects like culture, preferences and attitudes. The spatial joint modeling may also unmask the hidden association between multiple disease outcomes such as HIV and HSV-2. We relaxed the stationarity assumption using the conditional autoregressive model, the linearity assumption by using the penalized regression splines method, the normality assumption by using the mixture of Dirichlet and the

multivariate Polya trees prior and initiated the joint modeling of HIV and HSV-2 using the multivariate conditional autoregressive model and the multivariate normal distribution. We also discussed the multiple membership multiple classification approach as another viable method of initiating joint modeling. This study utilized a fully Bayesian approach. The analyses were done using WinBUGS for joint modeling, relaxing the linearity and the stationarity assumption, R package DPpackage for fitting the MDP and the MMPT random effects and MCMCglmm for the normal random effects model.

The multivariate mixture of Polya trees provide a highly flexible non parametric alternative to the traditional parametric and Dirichlet process mixture process as one of the downside of the DP and the MDP priors is that they suffer from intractability in some settings due to the discreteness of the DP. The three models i.e. normal, MDP and MMPT gave almost equal parameter estimates. The MMPT model however provided the best fitting model with more accurate results on the account of credible intervals.

The covariates used in this study were found to vary spatially hence a stationarity assumption would have led to less meaningful results hence interpretations. Age at first sex had greater effects on HIV and HSV-2 prevalence in the Central, some parts of North Eastern, Lake and Rift valley regions. The effect of place of residence on HIV prevalence was the least in the North Eastern region and greatest in some parts of Coastal and Lake region. The effect on HSV-2 was more in the Lake, Coastal and some parts of the Southern region. These findings have huge epidemiological implication. With limited funds, intervention strategies

may be tailor made for specific regions instead of rolling out blanket intervention strategies. More emphasis for example can be put in delaying the age at fist sex in those regions where the effect of age at first sex on HIV and HSV-2 was greater and so on. Modeling of diseases jointly while at the same time allowing the covariates to vary spatially can unmask underlying patterns and changing covariate effects both spatially and on the diseases.

The joint modeling unmasked a strong association between HIV and HSV-2. In particular, there was a strong positive correlation of 0.8540 between HIV and HSV-2. Those counties that registered a high HSV-2 prevalence also had high HIV prevalence. This simply implies that curbing HSV-2 would in turn result in curbing HIV as HSV-2 is associated with a two- to three-fold increased risk of HIV acquisition and an up to five-fold increased risk of HIV transmission per-sexual act, and may account for 40% to 60% of new HIV infections in populations where HSV-2 has a high prevalence [Fahrmeir and Tutz, 2001].

This study found that age had a non-linear effect on HIV prevalence. An assumption of linear relationship in this case therefore would have led to misleading results and subsequently wrong interpretations. The chance of HIV infection increases with age up to an optimum age of about 30 years then starts declining with increase in age. For HSV-2, the likelihood of infection increases with age up to an optimum age of about 40 years then starts to decline thereafter with increasing age. The results depict that the prevalence of HIV picks earlier in age than HSV-2. These results are important in knowing the target age group where more effort can be directed to help curb the spread of HIV and HSV-2.

Further work could be dedicated into exploring other models for relaxing the normality assumption so as to allow for skewness, fat tailness. Extending the non-parametric models such as the MMPT and MDP to be used for joint modeling or to replace conditional and multivariate conditional autoregressive (CAR and MCAR) models could provide interesting results as the both the CAR and the MCAR models are derived from the normal and the multivariate normal distributions respectively. There exists other ways of relaxing the linearity assumption and further studies can be geared towards finding out which models work best for particular circumstances. The model diagnostic tool: the DIC suffers some shortfalls. These shortfalls include negative values of $pD$ especially when the posterior of $\theta$ is highly non-normal which makes no sense, lack of consistency and therefore better tools for model diagnostics need to be developed. It would be of interest to explore how the effects of covariates evolve jointly in space and time and future study may incorporate a time effect on the joint spatially varying model. There are also some gray areas that need to be addressed to enhance the accuracy of spatial models. The leading tools for inference are MCMC and INLA. These methods cannot be trusted to give reliable inference without careful tuning and diagnostic checking. The choice of models for regression coefficients and prior distributions also affects the reliability of the results.

# Chapter 6

# Discussion, Conclusion and Future Research

In this thesis we developed and extended existing statistical models for spatial disease modeling. We applied these models to HIV, HSV-2 and malaria. These models only catered for areal (lattice) data, Geostatitsical and point pattern data were not considered in this study. In particular we used the KAIS 2007 and the AMIS 2006/07 and 2011 data.

In Chapter 2 we introduced a semi-parametric joint model to model HIV and HSV-2. This model was the best fitting model in terms of DIC. The joint model provided insight on the interaction between HIV and HSV-2. Areas with high HSV-2 prevalence had also high HIV prevalence. The models introduced in this chapter are applicable when one or more covariate has a non-linear relationship

with the response variable and also when one wants to jointly model two or more disease outcomes.

Chapter 3 introduces a model that relaxes the liming assumption of stationarity. In this model, the effects of the covariates are allowed to vary spatially by assigning their coefficients the CAR model. This new model provided a better fit than the stationary model. It may be that one wants to observe effects of the covariates on disease in each location and/or whether their intervention strategies are working or not. The BSVC model is a good candidate for this kind of analysis.

In Chapter 4 we introduce a spatial temporal spatially varying model. The covariates were allowed to vary both spatially and temporally. We fit this model to the Angolan malaria data. If data is collected over a period of time, a spatial temporal model or its variations would help in providing insight on how the disease and/or covariates effects vary over time.

Chapter 5 presents a review of various assumptions in spatial disease modeling. We provide alternatives for some limiting assumptions such as normality and linearity assumption. A joint spatially varying model is also developed allowing the effect of HIV and HSV-2 to jointly vary spatially. The applications of the models developed and discussed in this study are problem specific and their choice mostly would depend on the objective of the researcher and/or the nature of the data.

This study is not exhaustive in that it does not address all the issues in spatial disease modeling. There exist a number of areas that need further research. So far most studies have modeled the covariate effect of the categorical covariates in a

linear fashion by normal prior. A non-parametric fashion for modeling these effects can be considered in further studies. In spatial analysis, the boundary problem or effect may interfere with the accurate estimation of the statistical parameter. This is particularly true when boundaries cut off say at the border of two countries. The assumption has always been that the effect across the boundary is zero. This assumption is not necessarily true as an outbreak in the neighboring country maybe the reason of high prevalence at the border points. Further research may consider a model that takes into account the boundary effect.

This study considered complete case analysis. Individuals with missing entries were completely excluded from analysis. Further research could either use missing data techniques or incorporate the sampling weights to account for this deletion, a task impossible for this study as the weights were based on different administrative units (provincial) instead of counties.

There are considerable differences in the smoothing properties of the CAR model, depending on the type of neighbors specified. This has significant implications on the users of the CAR models since the neighborhood weight matrices chosen may markedly influence a study's findings. Further studies should look into the best neighborhood structure for the CAR models.

Non-parametric models used for random effects in this study performed better than their parametric counterparts. A non-parametric spatial model can be developed and used instead of CAR model. The standard method for measuring model fit in Bayesian analysis and in many cases disease mapping has been the deviance

information criterion (DIC). DIC has some limitations among them lack of consistency, $pD$ the effective number of parameters in the model which penalizes for complexity of the model is not invariant to reparameterization. Further research could explore other model diagnostic techniques or focus on improving the DIC.

# Appendix A

# WinBUGS Codes for chapter Two Models

```
############################Multivariate CAR model###################
model
{
#spline
for(i in 1: N)
{
for(l in 1:degree+1)
{
X[i,l]<-pow(Age[i],l-1)
}
}
for(i in 1: N)
```

```
{
for(k in 1:20)
{
u[i,k]<-(Age[i]-knot[k])*step(Age[i]-knot[k])
Z[i,k]<-pow(u[i,k],degree)
}
}
#likelihood
for(i in 1: N)
{
###None=0,primary=1,secondary=2,higher=3##
D.education1[i]<-equals(education[i],0)
D.education2[i]<-equals(education[i],1)
D.education3[i]<-equals(education[i],2)

###married,1partner=1,married,+2partners=2,divorced/seperated=3#,widowed=4,
nevermarried=5######
D.Married1[i]<-equals(Married[i],2)
D.Married2[i]<-equals(Married[i],3)
D.Married3[i]<-equals(Married[i],4)
D.Married4[i]<-equals(Married[i],5)

###Norisk=0,smallrisk=1,moderaterisk=2,greatrisk=3#####
D.Perceived2[i]<-equals(Perceived[i],1)
```

```
D.Perceived3[i]<-equals(Perceived[i],2)

D.Perceived4[i]<-equals(Perceived[i],3)


###Neverhadsex=0,under11=1,between12-14=2,between15-17=3,over18=4####

D.AgeatFirst1[i]<-equals(AgeatF[i],1)

D.AgeatFirst2[i]<-equals(AgeatF[i],2)

D.AgeatFirst3[i]<-equals(AgeatF[i],3)


###STI###

###Yes=1,No=2###

D.STI[i]<-equals(STI[i],1)


###Didn'tstayaway=0,stayedaway1-2times=1,stayedaway3-5times=2,

stayedaway6-##10times=3,stayedaway>11times=4###

D.Stayaway1[i]<-equals(Stayaway[i],0)

D.Stayaway2[i]<-equals(Stayaway[i],1)

D.Stayaway3[i]<-equals(Stayaway[i],2)

D.Stayaway4[i]<-equals(Stayaway[i],3)


##Urban##

##No=0,Yes=1###

D.Urban[i]<-equals(Urban[i],1)


##partners last one year##
```

```
###No partner=0,1partner=1,2partners=2,3ormorepartners=3###

D.Partner1[i]<-equals(Partner[i],1)

D.Partner2[i]<-equals(Partner[i],2)


#for HIV

hiv[i]~dbern(p1[i])

p1[i]<-min(1,max(0,PHIV[i]))


logit(PHIV[i])<-beta1+edu1[1]*D.education1[i]+edu1[2]*D.education2[i]+

edu1[3]*D.education3[i]+Marrd1[2]*D.Married1[i]+Marrd1[3]*D.Married2[i]+

Marrd1[4]*D.Married3[i]+Marrd1[5]*D.Married4[i]+

Perc1[2]*D.Perceived2[i]+Perc1[3]*D.Perceived3[i]+Perc1[4]*D.Perceived4[i]+

AgeF1[1]*D.AgeatFirst1[i]+AgeF1[2]*D.AgeatFirst2[i]+AgeF1[3]*D.AgeatFirst3[i]+

STI1*D.STI[i]+Stay1[1]*D.Stayaway1[i]+Stay1[2]*D.Stayaway2[i]+

Stay1[3]*D.Stayaway3[i]+Stay1[4]*D.Stayaway4[i]+urb1*D.Urban[i]+

Partn1[1]*D.Partner1[i]+Partn1[2]*D.Partner2[i]+

S[1,county[i]]+U[county[i],1]+spline1[i]

spline1[i]<-inprod(b1[ ], Z[i, ])+inprod(betaS1[ ], X[i, ])


#for herpes

herpes[i]~dbern(p2[i])

p2[i]<-min(1,max(0,PHRP[i]))

logit(PHRP[i])<-beta2+edu2[1]*D.education1[i]+edu2[2]*D.education2[i]+

edu2[3]*D.education3[i]+Marrd2[2]*D.Married1[i]+Marrd2[3]*D.Married2[i]+
```

```
Marrd2[4]*D.Married3[i]+Marrd2[5]*D.Married4[i]+

Perc2[2]*D.Perceived2[i]+Perc2[3]*D.Perceived3[i]+Perc2[4]*D.Perceived4[i]+

AgeF2[1]*D.AgeatFirst1[i]+AgeF2[2]*D.AgeatFirst2[i]+

AgeF2[3]*D.AgeatFirst3[i]+STI2*D.STI[i]+Stay2[1]*D.Stayaway1[i]+

Stay2[2]*D.Stayaway2[i]+Stay2[3]*D.Stayaway3[i]+

Stay2[4]*D.Stayaway4[i]+urb2*D.Urban[i]+

Partn2[1]*D.Partner1[i]+Partn2[2]*D.Partner2[i]+

S[2,county[i]]+U[county[i],2]+spline2[i]


spline2[i]<-inprod(b2[ ], Z[i, ])+inprod(betaS2[ ], X[i, ])

}

#Herpes

edu2[4]<-0

Marrd2[1]<-0

Perc2[1]<-0

AgeF2[4]<-0

Stay2[5]<-0

Partn2[3]<-0


edu1[4]<-0

Marrd1[1]<-0

Perc1[1]<-0

AgeF1[4]<-0

Stay1[5]<-0
```

```
Partn1[3]<-0

#priors

beta2~dnorm(0.01,0.01)

STI2~dnorm(0.01,0.01)

urb2~dnorm(0.01,0.01)


beta1~dnorm(0.01,0.01)

STI1~dnorm(0.01,0.01)

urb1~dnorm(0.01,0.01)

#Education coefficients

for(j in 1: 3)

{

edu1[j]~dnorm(0.01,0.01)

edu2[j]~dnorm(0.01,0.01)

}

#Married coefficients

for(m in 2:5)

{

Marrd1[m]~dnorm(0.01,0.01)

Marrd2[m]~dnorm(0.01,0.01)

}

#perceived risk

for(k in 2:4)

{
```

```
Perc1[k]~dnorm(0.01,0.01)

Perc2[k]~dnorm(0.01,0.01)

}

#age at first sex coefficients

for(g in 1:3 )

{

AgeF1[g]~dnorm(0.01,0.01)

AgeF2[g]~dnorm(0.01,0.01)

}

#stay away coeff

for(t in 1:4)

{

Stay1[t]~dnorm(0.01,0.01)

Stay2[t]~dnorm(0.01,0.01)

}

for(g in 1:2)

{

Partn1[g]~dnorm(0.01,0.01)

Partn2[g]~dnorm(0.01,0.01)

}

for(l in 1:degree+1)

{

betaS1[l]~dnorm(0,0.0001)

}
```

```
#priorsplines

for(k in 1:20)

{

b1[k]~dnorm(0,taub1 )

}

taub1~dgamma(1000,0.001)

for(l in 1:degree+1)

{

betaS2[l]~dnorm(0,0.0001)

}

#priorsplines

for(k in 1:20)

{

b2[k]~dnorm(0,taub2 )

}

taub2~dgamma(1000,0.001)

#ODDS ratios

#Education coefficients

for(j in 1: 4)

{

ORedu1[j]<-exp(edu1[j])

ORedu2[j]<-exp(edu2[j])

}
```

```
#married coefficient

for(m in 1:5)

{

ORMarrd1[m]<-exp(Marrd1[m])

ORMarrd2[m]<-exp(Marrd2[m])

}

#Perceived risk

for(k in 1:4)

{

ORPerc1[k]<-exp(Perc1[k])

ORPerc2[k]<-exp(Perc2[k])

}

#Age at first sex

for(g in 1:4)

{

ORAgeF1[g]<-exp(AgeF1[g])

ORAgeF2[g]<-exp(AgeF2[g])

}

#stay away coefficients

for(t in 1:5)

{

ORStay1[t]<-exp(Stay1[t])

ORStay2[t]<-exp(Stay2[t])

}
```

```
for(g in 1:2)

{

ORPartn1[g]<-exp(Partn1[g])

ORPartn2[g]<-exp(Partn2[g])

}

ORSTI2<-exp(STI2)

ORurb2<-exp(urb2)

ORSTI1<-exp(STI1)

ORurb1<-exp(urb1)

# MVCAR prior

S[1:Ndiseases, 1 : Nareas] ~ mv.car(adj[], weights[], num[], omega[ , ])

for (i in 1:sumNumNeigh)

{

weights[i] <- 1

}

R[1,1] <- 3

R[1,2] <- 0

R[2,1] <- 0

R[2,2] <- 2

# Precision matrix of MVCAR

omega[1 : Ndiseases, 1 : Ndiseases] ~ dwish(R[ , ], Ndiseases)

# Covariance matrix of MVCAR

sigma2[1 : Ndiseases, 1 : Ndiseases] <- inverse(omega[ , ])

# conditional SD of S[1, ] (HIV)
```

```
sigma[1] <- sqrt(sigma2[1, 1])

# conditional SD of S[2,] (HSV-2)

sigma[2] <- sqrt(sigma2[2, 2])

# within-area conditional correlation

corr <- sigma2[1, 2] / (sigma[1] * sigma[2])

# between HIV and HSV-2.

mean1 <- mean(S[1,])

mean2 <- mean(S[2,])

for(j in 1: 46)

{

S1[j]<-S[1,j]

S2[j]<-S[2,j]

}

#prior

for(i in 1: N)

{

for(j in 1: 46)

{

PH[j,i]<-(PHIV[i])*(equals(county[i],j))

PHPS[j,i]<-(PHRP[i])*(equals(county[i],j))

}

}

for(j in 1: 46)

{
```

```
for(i in 1: N)
{
count[j,i]<-equals(county[i],j)
}
number[j]<-sum(count[j,])
PCHV[j]<-sum(PH[j,])/number[j]
PCHPS[j]<-sum(PHPS[j,])/number[j]
}
#unstructured prior
for(i in 1:Nareas)
{
U[i, 1:Ndiseases] ~ dmnorm(zero[], tau[ , ])
}
# Precision matrix of MV Normal
tau[1:Ndiseases, 1:Ndiseases] ~ dwish(Q[ , ], Ndiseases)
# Covariance matrix of MV Normal
sigma2.U[1:2, 1:2] <- inverse(tau[ , ])
sigma.U[1] <- sqrt(sigma2.U[1, 1])
sigma.U[2] <- sqrt(sigma2.U[2, 2])
# within-area correlation between unstructured component of variation in HIV
and
# HSV-2
corr.U <- sigma2.U[1, 2] / (sigma.U[1] * sigma.U[2])
# within-area conditional correlation between total random effect
```

```
# (i.e. spatial + unstructured components) for HIV and for HSV-2

corr.sum <- (sigma2[1, 2] + sigma2.U[1, 2]) /

(sqrt(sigma2[1, 1] + sigma2.U[1, 1]) * sqrt(sigma2[2, 2] + sigma2.U[2, 2]))

}

#Data

#INITIALS
```

# Appendix B

# R Codes for chapter Three

# Models

```
# Packages required

library("MASS")

library("lattice")

library("ctv")

library("sp")

library(maptools)

library(rgdal)

require(RColorBrewer)

library(spdep)

require(INLA)

ken_data<-read.csv("C:/Users/okango/Desktop/shpfileoscar/

femaleNotMissGLM1.csv",header=T,sep=",")
```

```
head(ken_data)

attach(ken_data)


ken_data$county1<-ken_data$countyX

ken_data$county2<-ken_data$countyX

ken_data$county3<-ken_data$countyX

ken_data$county4<-ken_data$countyX

ken_data$county5<-ken_data$countyX

ken_data$county6<-ken_data$countyX

ken_data$county7<-ken_data$countyX

ken_data$county8<-ken_data$countyX

ken_data$county9<-ken_data$countyX


ken.graph<- readShapePoly("C:/Users/okango/Desktop/shpfileoscar/

ken_hds_test.shp")

plot(ken.graph)

adjken<-poly2nb(ken.graph)#Creates adjacency for ken

adjken

nb2INLA("ken.graph",adjken) #INLA graph file #spdep command

#unstructured

###HIV###

formula<-HIV~educationlevel+age_first_sex+perceived_Risk+

partners_last_1yr+Urban+Freq_of_travel_away+MaritalStatusX+STI+
```

```
f(age,model="rw2")+f(countyX, model="iid")
result0<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result0)

###Herpes###
formula<-herpes~educationlevel+age_first_sex+perceived_Risk+
partners_last_1yr+Urban+Freq_of_travel_away+MaritalStatusX+
STI+f(age,model="rw2")+f(countyX, model="iid")
result01<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result01)
#Structured
formula<-HIV~educationlevel+age_first_sex+perceived_Risk+
partners_last_1yr+Urban+Freq_of_travel_away+
MaritalStatusX+STI+f(age,model="rw2")+
f(countyX, model="besag",graph.file="ken.graph")
result00<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result00)

##Herpes###
formula<-herpes~educationlevel+age_first_sex+perceived_Risk+
partners_last_1yr+Urban+Freq_of_travel_away+MaritalStatusX+
```

```
STI+f(age,model="rw2")+f(countyX, model="besag",graph.file="ken.graph")
result001<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result001)

###structured-unstructured###
formula<-HIV~educationlevel+age_first_sex+perceived_Risk+
partners_last_1yr+Urban+Freq_of_travel_away+
MaritalStatusX+STI+f(age,model="rw2")+
f(countyX, model="besag",graph.file="ken.graph")+f(county1, model="iid")
result000<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result000)

###Herpes###
formula<-herpes~educationlevel+age_first_sex+perceived_Risk+
partners_last_1yr+Urban+Freq_of_travel_away+MaritalStatusX+
STI+f(age,model="rw2")+f(countyX, model="besag",graph.file="ken.graph")+
f(county1, model="iid")
result0001<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result0001)

##SVCUnstructured##
```

```
formula<-HIV~f(county1,educationlevel,model="besag",graph="ken.graph",
constr=FALSE)+
f(county2,age_first_sex,model="besag",graph="ken.graph",constr=FALSE)+
f(county3,perceived_Risk,model="besag",graph="ken.graph",constr=FALSE)+
f(county4,partners_last_1yr,model="besag",graph="ken.graph",constr=FALSE)+
f(county5,Urban,model="besag",graph="ken.graph",constr=FALSE)+
f(county6,Freq_of_travel_away,model="besag",graph="ken.graph",constr=FALSE)+
f(county7,MaritalStatusX,model="besag",graph="ken.graph",constr=FALSE)+
f(county8,STI,model="besag",graph="ken.graph",constr=FALSE)+f(age,model="rw2")+
f(countyX, model="iid")
result1<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,mlik=TRUE,cpo=TRUE))
summary(result1)


###Herpes##
formula<-herpes~f(county1,educationlevel,model="besag",graph="ken.graph",
constr=FALSE)+
f(county2,age_first_sex,model="besag",graph="ken.graph",constr=FALSE)+
f(county3,perceived_Risk,model="besag",graph="ken.graph",constr=FALSE)+
f(county4,partners_last_1yr,model="besag",graph="ken.graph",constr=FALSE)+
f(county5,Urban,model="besag",graph="ken.graph",constr=FALSE)+
f(county6,Freq_of_travel_away,model="besag",graph="ken.graph",constr=FALSE)+
f(county7,MaritalStatusX,model="besag",graph="ken.graph",constr=FALSE)+
f(county8,STI,model="besag",graph="ken.graph",constr=FALSE)+f(age,model="rw2")+
```

```
f(countyX, model="iid")
result11<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,mlik=TRUE,cpo=TRUE))
summary(result11)

##SVCStructured###
formula<-HIV~f(county1,educationlevel,model="besag",graph="ken.graph",
constr=FALSE)+
f(county2,age_first_sex,model="besag",graph="ken.graph",constr=FALSE)+
f(county3,perceived_Risk,model="besag",graph="ken.graph",constr=FALSE)+
f(county4,partners_last_1yr,model="besag",graph="ken.graph",constr=FALSE)+
f(county5,Urban,model="besag",graph="ken.graph",constr=FALSE)+
f(county6,Freq_of_travel_away,model="besag",graph="ken.graph",constr=FALSE)+
f(county7,MaritalStatusX,model="besag",graph="ken.graph",constr=FALSE)+
f(county8,STI,model="besag",graph="ken.graph",constr=FALSE)+f(age,model="rw2")+
f(countyX, model="besag",graph="ken.graph")
result2<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result2)

##Herpes##
formula<-herpes~f(county1,educationlevel,model="besag",graph="ken.graph",
constr=FALSE)+
f(county2,age_first_sex,model="besag",graph="ken.graph",constr=FALSE)+
```

```
f(county3,perceived_Risk,model="besag",graph="ken.graph",constr=FALSE)+
f(county4,partners_last_1yr,model="besag",graph="ken.graph",constr=FALSE)+
f(county5,Urban,model="besag",graph="ken.graph",constr=FALSE)+
f(county6,Freq_of_travel_away,model="besag",graph="ken.graph",constr=FALSE)+
f(county7,MaritalStatusX,model="besag",graph="ken.graph",constr=FALSE)+
f(county8,STI,model="besag",graph="ken.graph",constr=FALSE)+f(age,model="rw2")+
f(countyX, model="besag",graph="ken.graph")
result22<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result22)


##SVCStructuredunstructured##
formula<-HIV~f(county1,educationlevel,model="besag",graph="ken.graph",
constr=FALSE)+
f(county2,age_first_sex,model="besag",graph="ken.graph",constr=FALSE)+
f(county3,perceived_Risk,model="besag",graph="ken.graph",constr=FALSE)+
f(county4,partners_last_1yr,model="besag",graph="ken.graph",constr=FALSE)+
f(county5,Urban,model="besag",graph="ken.graph",constr=FALSE)+
f(county6,Freq_of_travel_away,model="besag",graph="ken.graph",constr=FALSE)+
f(county7,MaritalStatusX,model="besag",graph="ken.graph",constr=FALSE)+
f(county8,STI,model="besag",graph="ken.graph",constr=FALSE)+f(age,model="rw2")+
f(county9, model="besag",graph="ken.graph")+f(countyX, model="iid")
result3<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
```

```
summary(result3)


###Herpes###

formula<-herpes~f(county1,educationlevel,model="besag",graph="ken.graph",

constr=FALSE)+

f(county2,age_first_sex,model="besag",graph="ken.graph",constr=FALSE)+

f(county3,perceived_Risk,model="besag",graph="ken.graph",constr=FALSE)+

f(county4,partners_last_1yr,model="besag",graph="ken.graph",constr=FALSE)+

f(county5,Urban,model="besag",graph="ken.graph",constr=FALSE)+

f(county6,Freq_of_travel_away,model="besag",graph="ken.graph",constr=FALSE)+

f(county7,MaritalStatusX,model="besag",graph="ken.graph",constr=FALSE)+

f(county8,STI,model="besag",graph="ken.graph",constr=FALSE)+f(age,model="rw2")+

f(county9, model="besag",graph="ken.graph")+f(countyX, model="iid")

result33<-inla(formula,family="binomial",data=ken_data,control.compute=

list(dic=TRUE,mlik=TRUE,cpo=TRUE))

summary(result33)


##SVC-Besag##

formula<-HIV~f(county1,educationlevel,model="besag",

graph="ken.graph")+

f(county2,age_first_sex,model="besag",graph="ken.graph")+

f(county3,perceived_Risk,model="besag",graph="ken.graph")+

f(county4,partners_last_1yr,model="besag",graph="ken.graph")+

f(county5,Urban,model="besag",graph="ken.graph")+
```

```
f(county6,Freq_of_travel_away,model="besag",graph="ken.graph")+
f(county7,MaritalStatusX,model="besag",graph="ken.graph")+
f(county8,STI,model="besag",graph="ken.graph")+f(age,model="rw2")
result4<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result4)


##herpes##
formula<-herpes~f(county1,educationlevel,model="besag",
graph="ken.graph")+
f(county2,age_first_sex,model="besag",graph="ken.graph")+
f(county3,perceived_Risk,model="besag",graph="ken.graph")+
f(county4,partners_last_1yr,model="besag",graph="ken.graph")+
f(county5,Urban,model="besag",graph="ken.graph")+
f(county6,Freq_of_travel_away,model="besag",graph="ken.graph")+
f(county7,MaritalStatusX,model="besag",graph="ken.graph")+
f(county8,STI,model="besag",graph="ken.graph")+f(age,model="rw2")
result44<-inla(formula,family="binomial",data=ken_data,control.compute=
list(dic=TRUE,mlik=TRUE,cpo=TRUE))
summary(result44)


###RESEULTS_FOR_STRUNSTUCTURED###
Mod4<-result3$summary.random$age
Mod4
```

```
Mod44<-result3$summary.random

Mod44


?spplot


###NON-LINEAR EFFECTS##

##AGE ON HIV###


plot(result3$summary.random$age$ID, result3$summary.random$age[,5],

xlab="Age",ylab="effect")

par(new=TRUE)

plot(result3$summary.random$age$ID, result3$summary.random$age[,4],

col="blue",ann=FALSE, axes=F)

par(new=TRUE)

plot(result3$summary.random$age$ID, result3$summary.random$age[,6],

col="blue",ann=FALSE,axes=F)


###AGE ON HSV-2####


plot(result33$summary.random$age$ID, result33$summary.random$age[,5],

xlab="Age",ylab="effect")

par(new=TRUE)

plot(result33$summary.random$age$ID, result33$summary.random$age[,4],

col="blue",ann=FALSE, axes=F)
```

```
par(new=TRUE)

plot(result33$summary.random$age$ID, result33$summary.random$age[,6],

col="blue",ann=FALSE,axes=F)


###################MAPS#######

###structured###

structured_Spatial_Effect<-result3$summary.random$county9

structured_Spatial_Effect

#UNSTR<-rbind(Unstructured_Spatial_Effect,0)

ken.graph$NUNSTR<-structured_Spatial_Effect$"0.5quant"

spplot(ken.graph,"NUNSTR",   col.regions=bpy.colors(20))


##HIV####

##education##

education<-result3$summary.random$county1

education

ken.graph$educ<-education$"0.5quant"

spplot(ken.graph,"educ",   col.regions=bpy.colors(20))


##age-at-first-sex##

Firstsex<-result3$summary.random$county2

ken.graph$First<-Firstsex$"0.5quant"

spplot(ken.graph,"First",   col.regions=bpy.colors(20))
```

```
###Perceived-Risk##

Perceived<-result3$summary.random$county3

ken.graph$risk<-Perceived$"0.5quant"

spplot(ken.graph,"risk",   col.regions=bpy.colors(20))


###Partners-last-one-year##

partners<-result3$summary.random$county4

ken.graph$part<-partners$"0.5quant"

spplot(ken.graph,"part",   col.regions=bpy.colors(20))


##Urban###

urban<-result3$summary.random$county5

ken.graph$urb<-urban$"0.5quant"

spplot(ken.graph,"urb",   col.regions=bpy.colors(20))


##Travel--away##

travel<-result3$summary.random$county6

ken.graph$trav<-travel$"0.5quant"

spplot(ken.graph,"trav",   col.regions=bpy.colors(20))


##Marital##

marital<-result3$summary.random$county7

ken.graph$marry<-marital$"0.5quant"

spplot(ken.graph,"marry",   col.regions=bpy.colors(20))
```

```
##STI##

sti<-result3$summary.random$county8

ken.graph$st<-sti$"0.5quant"

spplot(ken.graph,"st",   col.regions=bpy.colors(20))


###HSV-2###


##education##

education<-result44$summary.random$county1

education

ken.graph$educ<-education$"0.5quant"

spplot(ken.graph,"educ",   col.regions=bpy.colors(20))


##age-at-first-sex##

Firstsex<-result44$summary.random$county2

ken.graph$First<-Firstsex$"0.5quant"

spplot(ken.graph,"First",   col.regions=bpy.colors(20))


###Perceived-Risk##

Perceived<-result44$summary.random$county3

ken.graph$risk<-Perceived$"0.5quant"

spplot(ken.graph,"risk",   col.regions=bpy.colors(20))
```

```
###Partners-last-one-year##

partners<-result44$summary.random$county4

ken.graph$part<-partners$"0.5quant"

spplot(ken.graph,"part",   col.regions=bpy.colors(20))


##Urban###

urban<-result44$summary.random$county5

ken.graph$urb<-urban$"0.5quant"

spplot(ken.graph,"urb",   col.regions=bpy.colors(20))


##Travel--away##

travel<-result44$summary.random$county6

ken.graph$trav<-travel$"0.5quant"

spplot(ken.graph,"trav",   col.regions=bpy.colors(20))


##Marital##

marital<-result44$summary.random$county7

ken.graph$marry<-marital$"0.5quant"

spplot(ken.graph,"marry",   col.regions=bpy.colors(20))


##STI##

sti<-result44$summary.random$county8

ken.graph$st<-sti$"0.5quant"

spplot(ken.graph,"st",   col.regions=bpy.colors(20))
```

```
####HIV-SVC-structured Map###

SVCstructured_Spatial_Effect<-result3$summary.random$county9

SVCstructured_Spatial_Effect

#SVCSTR<-rbind(SVCstructured_Spatial_Effect,0)

ken.graph$SVCNSTR<-SVCstructured_Spatial_Effect$"0.5quant"

spplot(ken.graph,"SVCNSTR")


####HSV-SVC-structured Map###

HSVCstructured_Spatial_Effect<-result33$summary.random$county9

HSVCstructured_Spatial_Effect

#SVCSTR<-rbind(SVCstructured_Spatial_Effect,0)

ken.graph$HSVCNSTR<-HSVCstructured_Spatial_Effect$"0.5quant"

spplot(ken.graph,"HSVCNSTR")
```

# Appendix C

# R Codes for chapter Four Models

```
#ANGOLA INLA R CODE

rm(list=ls())

# Packages required

library("MASS")

library("lattice")

library("ctv")

library("sp")

library(maptools)

library(rgdal)

library(spdep)

require(INLA)

ang.graph<- readRDS("C:/Users/okango/Desktop/shpfileoscar/ANGO.rds")

#angola<-read.csv("C:/Elphasphd/aminata/data/NEWangola.csv",header=T,sep=",")

#angola07<-read.csv("C:/Elphasphd/aminata/data/maldata07.csv",header=T,sep=",")
```

```
#head(angola07)

#tail(angola07)

#angola07$prov1=angola07$Province

#angola07$prov2=angola07$Province

#angola07<-read.csv("C:/Elphasphd/aminata/data/maldata12.csv",header=T,sep=",")

plot(ang.graph)

adjang<-poly2nb(ang.graph)#Creates adjacency for ken

adjang

nb2INLA("ang.graph",adjang)


formula<-Malaria~1+as.factor(Residence)+as.factor(Net)+as.factor(Wealth)+

as.factor(Gender)+f(Age,model="rw2")+

f(prov1,model="besag",graph="ang.graph",adjust.for.con.comp = FALSE)+

f(prov2,model="iid",graph="ang.graph",adjust.for.con.comp = FALSE)

result0<-inla(formula,family="binomial",data=angola07,control.compute=

list(dic=TRUE,cpo=TRUE))

summary(result0)


#####NON LINEAR EFFECT OF AGE####

age=result0$summary.random$Age

age

plot(result0$summary.random$Age$ID, result0$summary.random$Age[,5],

xlab="Age",ylab="effect")

par(new=TRUE)
```

```
plot(result0$summary.random$Age$ID, result0$summary.random$Age[,4],
col="blue",ann=FALSE, axes=F)
par(new=TRUE)
plot(result0$summary.random$Age$ID, result0$summary.random$Age[,6],
col="blue",ann=FALSE,axes=F)

#####SAPATIAL EFFECTS####
sapt07=result0$summary.random$prov1
sapt07

sapt207=result0$summary.random$prov2
sapt207

#####2012#####
angola12<-read.csv("C:/Elphasphd/aminata/data/maldata12.csv",header=T,sep=",")
head(angola12)
tail(angola12)
angola12$prov11=angola12$Province
angola12$prov22=angola12$Province
formula<-Malaria~1+as.factor(Residence)+as.factor(Net)+as.factor(Wealth)+
as.factor(Gender)+f(Age,model="rw2")+
f(prov11,model="besag",graph="ang.graph",adjust.for.con.comp = FALSE)+
f(prov22,model="iid",graph="ang.graph",adjust.for.con.comp = FALSE)
result1<-inla(formula,family="binomial",data=angola12,control.compute=
```

```
list(dic=TRUE,cpo=TRUE))

summary(result1)


#####NON LINEAR EFFECT OF AGE####

age=result1$summary.random$Age

age

plot(result1$summary.random$Age$ID, result1$summary.random$Age[,5],

xlab="Age",ylab="effect")

par(new=TRUE)

plot(result1$summary.random$Age$ID, result1$summary.random$Age[,4],

col="blue",ann=FALSE, axes=F)

par(new=TRUE)

plot(result1$summary.random$Age$ID, result1$summary.random$Age[,6],

col="blue",ann=FALSE,axes=F)


#####SAPATIAL EFFECTS####

sapt11=result1$summary.random$prov11

sapt11

sapt211=result1$summary.random$prov22

sapt211


##########spatio-temporal#############


#angola<-read.csv("C:/Elphasphd/aminata/data/NEWangola.csv",
```

```
header=T,sep=",")

angola<-read.csv("C:/Elphasphd/aminata/data/ANGOLNEW.csv",

header=T,sep=",")

head(angola)

angola$prov3=angola$Province

angola$prov33=angola$Province

angola$prov4=angola$Province

angola$prov5=angola$Province

angola$prov6=angola$Province

angola$prov7=angola$Province

angola$prov8=angola$Province


#formula<-Malaria~f(Province,Net,model="besag",graph="

ang.graph",adjust.for.con.comp = FALSE,group=time,control.group=

list(model="ar1"))

#formula<-Malaria~f(Province,Wealth,model=

"besag",graph="ang.graph",adjust.for.con.comp = FALSE,group=time,control.group=

list(model="ar1"))

#mm=table(angola$Net)

#mm

#formula=Malaria~f(Net,model="besag",graph="ang.graph",adjust.for.con.comp =

FALSE,group=time,control.group=list(model="ar1"))+

f(Wealth,model="besag",graph="ang.graph",adjust.for.con.comp =

FALSE,group=time,control.group=list(model="ar1"))+
```

```
f(Residence,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Gender,model="besag",graph="ang.graph",adjust.for.con.comp =
 FALSE,group=time,control.group=list(model="ar1"))+
f(Age,model="rw2")+f(prov3,model="besag",graph="ang.graph",group=
time,adjust.for.con.comp = FALSE)+
f(prov33,model="iid",graph="ang.graph",group=time,adjust.for.con.comp =FALSE)
#formula<-Malaria~f(Residence,model="besag",graph=
"ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Net,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Wealth,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Gender,model="besag",graph="ang.graph",adjust.for.con.comp =
 FALSE,group=time,control.group=list(model="ar1"))+
f(Age,model="rw2")+f(prov3,model="besag",graph="ang.graph",adjust.for.con.comp =
f(prov33,model="iid",graph="ang.graph",adjust.for.con.comp = FALSE)
#formula=Malaria~f(Net,model="besag",graph="ang.graph",adjust.for.con.comp =
 FALSE,group=time,control.group=list(model="ar1"))+
f(Wealth,model="besag",graph="ang.graph",adjust.for.con.comp =
 FALSE,group=time,control.group=list(model="ar1"))+
f(Residence,model="besag",graph="ang.graph",adjust.for.con.comp =
 FALSE,group=time,control.group=list(model="ar1"))+
```

```
f(Gender,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+f(Age,model="rw2")+
f(prov3,model="besag",graph="ang.graph",group=time,adjust.for.con.comp = FALSE)+
f(time,model="ar1")
formula=Malaria~f(Net,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Wealth,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Residence,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Gender,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Age,model="rw2")+f(prov3,model="besag",graph="ang.graph",
group=time,adjust.for.con.comp = FALSE)


formula=Malaria~f(Net,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Wealth,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Residence,model="besag",graph="ang.graph",adjust.for.con.comp =
FALSE,group=time,control.group=list(model="ar1"))+
f(Age,model="rw2")+f(prov3,model="besag",graph="ang.graph",
group=time,adjust.for.con.comp = FALSE)
result3<-inla(formula,family="binomial",
```

```
data=angola,control.compute=list(dic=TRUE,cpo=TRUE))

summary(result3)


####################Results###################


nnet=result3$summary.random$Net

nnet

WWEALTH=result3$summary.random$Wealth

WWEALTH

RESD=result3$summary.random$Residence

RESD

GEN=result3$summary.random$Gender

GEN

Spastruc=result3$summary.random$prov3

Spastruc

AGE=result3$summary.random$Age

AGE

dd<-result3$summary.random$Residence

ddd=dd$"0.5quant"

ddd

mm=ddd[1:18]

mm

zz=ddd[19:36]

ang.graph$welt<-zz
```

```
spplot(ang.graph,"welt",   col.regions=bpy.colors(20))
```

# Appendix D

# WinBUGS and R Codes for chapter Five Models

```
require(DPpackage)

require(MCMCglmm)

ken_data<-read.csv("C:/Users/okango/Desktop/shpfileoscar/

femaleNotMissGLM1.csv",

header=T,sep=",")

head(ken_data)

ID<-1:4864

ken_data$ID<-ID

head(ken_data)

attach(ken_data)

prior <- list(G = list(G1 = list(V = 1,nu = 0.002)))

mod1<-MCMCglmm(HIV~educationlevel+age_first_sex+perceived_Risk+
```

```
partners_last_1yr+Urban+Freq_of_travel_away+

MaritalStatusX+STI,random=~countyX,family = "categorical",data =

ken_data, prior = prior, verbose = FALSE,

pr=T,burnin=5000,nitt = 25000,thin=20)

summary(mod1)


M1dic<-mod1$DIC

M1dic

Dev<-mod1$Deviance

Dev

mean(Dev)

dhat<-mod1$Dhat

dhat


# Prior information


beta0<-rep(0,8)

beta0

Sbeta0<-diag(1000,)

Sbeta0

tinv<-diag(1,1)

prior<-list(a0=2,b0=0.1,nu0=4,tinv=tinv,mub=rep(0,1),Sb=diag(1000,1),

beta0=beta0,Sbeta0=Sbeta0)
```

```
# Initial state

state <- NULL


# MCMC parameters


nburn <- 5000

nsave <- 5000

nskip <- 0

ndisplay <- 1000

mcmc <- list(nburn=nburn,nsave=nsave,nskip=nskip,ndisplay=ndisplay)


# Fit the Probit model

fit1 <- DPglmm(fixed=HIV~educationlevel+age_first_sex+perceived_Risk+

partners_last_1yr+Urban+Freq_of_travel_away+MaritalStatusX+

STI,random=~1|countyX,

family=binomial(logit),prior=prior,mcmc=mcmc,state=state,status=TRUE)

summary(fit1)


############################POLYA TREES####################


prior <- list(alpha=1,

M=4,

frstlprob=FALSE,

nu0=4,
```

```
tinv=diag(1,1),

mub=rep(0,1),

Sb=diag(1000,1),

beta0=rep(0,8),

Sbeta0=diag(10000,8))

# Initial state

state <- NULL

# MCMC parameters

nburn <- 5000

nsave <- 25000

nskip <- 20

ndisplay <- 1000

mcmc <- list(nburn=nburn,

nsave=nsave,

nskip=nskip,

ndisplay=ndisplay,

tune1=0.5,tune2=0.5,

samplef=1)

# Fitting the Logit model

fit1 <- PTglmm(fixed=HIV~educationlevel+age_first_sex+perceived_Risk+

partners_last_1yr+Urban+Freq_of_travel_away+MaritalStatusX+STI,random=~1|countyX

family=binomial(logit),prior=prior,mcmc=mcmc,

state=state,status=TRUE)

summary(fit1)
```

```
########################Multivariate CAR model#########################


model

{

#spline

for(i in 1: N)

{

for(l in 1:degree+1)

{

X[i,l]<-pow(Age[i],l-1)

}

}

for(i in 1: N)

{

for(k in 1:20)

{

u[i,k]<-(Age[i]-knot[k])*step(Age[i]-knot[k])

Z[i,k]<-pow(u[i,k],degree)

}

}


#likelihood
```

```
for(i in 1: N)
{
#for HIV
hiv[i]~dbern(p1[i])
p1[i]<-min(1,max(0,PHIV[i]))

logit(PHIV[i])<-beta1+edu1[county[i]]*education[i]+Marrd1[county[i]]*Married[i]+
Perc1[county[i]]*Perceived[i]+AgeF1[county[i]]*AgeatF[i]+STI1[county[i]]*STI[i]+
Stay1[county[i]]*Stayaway[i]+Urb1[county[i]]*Urban[i]+
Partn1[county[i]]*Partner[i]+U[county[i],1]+spline1[i]
spline1[i]<-inprod(b1[ ], Z[i, ])+inprod(betaS1[ ], X[i, ])

#for herpes
herpes[i]~dbern(p2[i])
p2[i]<-min(1,max(0,PHRP[i]))
logit(PHRP[i])<-beta2+edu2[county[i]]*education[i]+Marrd2[county[i]]*Married[i]+
Perc2[county[i]]*Perceived[i]+AgeF2[county[i]]*AgeatF[i]+STI2[county[i]]*STI[i]+
Stay2[county[i]]*Stayaway[i]+Urb2[county[i]]*Urban[i]+
Partn2[county[i]]*Partner[i]+U[county[i],2]+spline2[i]

spline2[i]<-inprod(b2[ ], Z[i, ])+inprod(betaS2[ ], X[i, ])
}
for (k in 1:sumNumNeigh) {
weights1[k] <- 1
```

```
}

omega.spatial1 ~ dgamma(0.5, 0.0005)

omega.spatial1sq<-1/omega.spatial1


#Education coefficients

edu1[1 : 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)

edu2[1: 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)


#Married coefficients

 Marrd1[1 : 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)

 Marrd2[1 : 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)


#perceived risk

 Perc1[1 : 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)

 Perc2[1 : 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)


#age at first sex coefficients

 AgeF1[1 : 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)

 AgeF2[1: 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)


#stay away coeff

 Stay1[1: 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)

 Stay2[1 : 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)
```

```
#Partners

 Partn1[1: 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)

 Partn2[1 : 46] ~ car.normal(adj[], weights1[], num[], omega.spatial1)


#STI

 STI1[1:46]~ car.normal(adj[], weights1[], num[], omega.spatial1)

 STI2[1:46]~ car.normal(adj[], weights1[], num[], omega.spatial1)


#Urb

 Urb1[1:46]~ car.normal(adj[], weights1[], num[], omega.spatial1)

 Urb2[1:46]~ car.normal(adj[], weights1[], num[], omega.spatial1)


#prior for intercept

beta1~dnorm(0.01,0.01)

beta2~dnorm(0.01,0.01)

for(l in 1:degree+1)

{

betaS1[l]~dnorm(0,0.0001 )


}

#priorsplines


for(k in 1:20)

{
```

```
b1[k]~dnorm(0,taub1 )

}

taub1~dgamma(1000,0.001)

for(l in 1:degree+1)

{

betaS2[l]~dnorm(0,0.0001 )

}

#priorsplines

for(k in 1:20)

{

b2[k]~dnorm(0,taub2 )



}

taub2~dgamma(1000,0.001)

#prior

for(i in 1: N)

{

for(j in 1: 46)

{

PH[j,i]<-(PHIV[i])*(equals(county[i],j))

PHPS[j,i]<-(PHRP[i])*(equals(county[i],j))

}

}

for(j in 1: 46)
```

```
{
for(i in 1: N)
{
count[j,i]<-equals(county[i],j)
}
number[j]<-sum(count[j,])
PCHV[j]<-sum(PH[j,])/number[j]
PCHPS[j]<-sum(PHPS[j,])/number[j]


}


#unstructured prior
for(i in 1:Nareas)
{
U[i, 1:Ndiseases] ~ dmnorm(zero[], tau[ , ])
}
# Precision matrix of MV Normal
  tau[1:Ndiseases, 1:Ndiseases] ~ dwish(Q[ , ], Ndiseases)
# Covariance matrix of MV Normal
  sigma2.U[1:2, 1:2] <- inverse(tau[ , ])


  sigma.U[1] <- sqrt(sigma2.U[1, 1])
  sigma.U[2] <- sqrt(sigma2.U[2, 2])
# within-area correlation between unstructured component of variation in HIVand
```

```
# HSV-2

  corr.U <- sigma2.U[1, 2] / (sigma.U[1] * sigma.U[2]

}

#Data

#INITIALS
```

# Bibliography

Adrienne, L. and Mbiti, I. (2012). Access, sorting, and achievement: the short-run effects of free primary education in Kenya. *American Economic Journal: Applied Economics*, 4(4):226–253.

AMIS (2007). Angola malaria indicator survey 2006-2007.

AMIS (2012). Angola malaria indicator survey 2011.

Amornkul, N., Vandenhoudt, H., Nasokho, P., Odhiambo, F., and Mwaengo, D. (2009). HIV prevalence and associated risk factors among individuals aged 13-34 years in Rural Western Kenya. *PloS one*, 4(7).

Assunçao, R., Assunçao, J., and Lemos, M. (1998). Induced technical change: a Bayesian spatial varying parameter model. In *Proceedings of XVI Latin American Meeting of the Econometric Society. Catholic University of Peru: Lima.*

Assunçao, R. M. (2003). Space varying coefficient models for small area data. *Environmetrics*, 14(5):453–473.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, pages 171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46(2):199–208.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, pages 159–188.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Bernardinelli, L., Clayton, D., and Montomoli, C. (1995). Bayesian estimates of disease maps: how important are priors? *Statistics in medicine*, 14(21-22):2411–2431.

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

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

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

Best, N., Cockings, S., Bennett, J., Wakefield, J., and Elliott, P. (2001). Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):155–174.

Browne, W. J., Goldstein, H., and Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, 1(2):103–124.

Burgoyne, A. and Drummond, P. (2008). Knowledge of HIV and AIDS in women in Sub-Saharan Africa. *African Journal of Reproductive Health*, 12:14–31.

Burgoyne, A. D. and Drummond, P. D. (2009). Knowledge of HIV and AIDS in women in Sub-Saharan Africa. *African Journal of Reproductive Health*, 12(2):14–31.

Carlin, B. P. and Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data. *Bayesian statistics*, 7:45–63.

Carroll, M. and Ruppert, D. (2003). *Semi-parametric Regression*. Cambridge university Press, Camridge.

CBS (2004). Central Bureau of Statistics (CBS) [Kenya], Ministry of Health (MOH) [Kenya], and ORC Macro. Report.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., et al. (2006). Deviance information criteria for missing data models. *Bayesian analysis*, 1(4):651–673.

Clark, S. (2004). Early marriage and HIV risks in Sub Saharan Africa. *Studies in family planning*, 35(3):149–160.

Clayton, D. and Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. *Geographical and environmental epidemiology: methods for small-area studies*, pages 205–220.

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

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. *Statistical models in S*, pages 309–376.

Cohen, M. S. (1998). Sexually transmitted diseases enhance HIV transmission: no longer a hypothesis. *The Lancet*, 351:S5–S7.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, 4(5):613–617.

Currie, I. D., Durban, M., and Eilers, P. H. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):259–280.

Dabney, A. R. and Wakefield, J. C. (2005). Issues in the mapping of two diseases. *Statistical methods in medical research*, 14(1):83–112.

Diggle, P. J., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

DiMatteo, I., Genovese, C., and Kass, R. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071.

Dominguez-Molina, J., González-Farías, G., and Gupta, A. (2003). The multivariate closed skew normal distribution. Report, Technical report.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., and Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing

properties of Conditional Autoregressive (CAR) models. *International journal of health geographics*, 6(1):1.

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

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

Fahrmeir, L. and Knorr-Held, L. (1997). Dynamic and semiparametric models.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer, 2 edition.

Fahrmeir, L. and Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association*, 91(436):1584–1594.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Ferreira, A. and Garcia, N. L. (2001). Simulation study for misspecifications on a frailty model. *Brazilian Journal of Probability and Statistics*, pages 121–134.

Fotheringham, S., Chris, B., and Martin, C. (2003). *Geographically weighted Regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Gaetan, C., Guyon, X., and Bleakley, K. (2010). *Spatial statistics and modeling*, volume 81. Springer.

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

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Ghebremichael, M., Larsen, U., and Painstil, E. (2009). Association of Age at first sex with HIV-1, HSV-2, and other sexual transmitted infections among women in Northern Tanzania. *National Center for Biotechnology Information*, 36(9):570–576.

Green, P. and Silverman, B. (1993). *Nonparametric regression and generalized linear models: A roughness penalty approach*. CRC Press.

Green, P. J. and Richardson, S. (2002). Hidden markov models and disease mapping. *Journal of the American statistical association*, 97(460):1055–1070.

Griffith, D. A. (2001). A spatial filtering specication for the auto-poisson model.

Hanson, T. E. (2012). Inference for mixtures of finite polya tree models. *Journal of the American Statistical Association*.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*, volume 1. Springer.

Hastie, T. and Tibshirani, R. (1995). Genaralized additive models for medical research. *Statistical Methods in Research*, 4:187.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC Press.

Hay, S. I., Guerra, C. A., Tatem, A. J., Atkinson, P. M., and Snow, R. W. (2005). Tropical infectious diseases: Urbanization, malaria transmission and disease burden in africa. *Nature Reviews Microbiology*, 3(1):81–90.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320.

Hinton, G. E. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13. ACM.

Hoeting, J. A., Leecaster, M., and Bowden, D. (2000). An improved model for spatially correlated binary responses. *Journal of agricultural, biological, and environmental statistics*, pages 102–114.

Hosseini, F., Eidsvik, J., and Mohammadzadeh, M. (2011). Approximate bayesian inference in spatial glmm with skew normal latent variables. *Computational Statistics & Data Analysis*, 55(4):1791–1806.

Jara, A., Hanson, T. E., and Lesaffre, E. (2012). Robustifying generalized linear mixed models using a new class of mixtures of multivariate polya trees. *Journal of Computational and Graphical Statistics*.

Johnson, K. and Way, A. (2006). Risk factors for HIV infection in a national adult population: evidence from the 2003 Kenya Demographic and Health Survey. *Journal of Acquired Immune Deficiency Syndromes*, 42(5):627–636.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (2002). *Continuous multivariate distributions, volume 1, models and applications*, volume 59. New York: John Wiley & Sons.

Johnson, W. and Christensen, R. (1989). Nonparametric bayesian analysis of the accelerated failure time model. *Statistics & Probability Letters*, 8(2):179–184.

Kazembe, L., Chirwa, T., Simbeye, J., and Namangale, J. (2008). Applications of bayesian approach in modelling risk of malaria-related hospital mortality. *BMC medical research methodology*, 8(1):6.

Kenya (1997). Ministry of Health, Sessional Paper No.4 of 1997 on AIDS in Kenya.

Kleinman, K. P. and Ibrahim, J. G. (1998). A semiparametric bayesian approach to the random effects model. *Biometrics*, pages 921–938.

Kleinschmidt, I., MacPhail, C., Natashia, M., Pettifor, A., and Rees, H. (2007). Geographic distribution of human immunodeficiency virus in South Africa. *The American journal of tropical medicine and hygiene*, 77(6):1163–1169.

Knorr-Held, L. (1999). Bayesian modelling of inseparable space-time variation in disease risk.

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

Komárek, A. and Lesaffre, E. (2008). Generalized linear mixed model with a penalized gaussian mixture as a random effects distribution. *Computational Statistics & Data Analysis*, 52(7):3441–3458.

Kuss, M. and Rasmussen, C. E. (2005). Assessing approximate inference for binary gaussian process classification. *Journal of Machine Learning Research*, 6(Oct):1679–1704.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Lang, S., Fronk, E., and Fahrmeir, L. (2002). Function estimation with locally adaptive dynamic models. *Computational Statistics*, 17(4):479–500.

Langford, H., Leyland, A., Rasbash, J., and Goldstein, H. (1999). Multilevel modelling of the geographical distributions of diseases. *Applied Statistics*, 48:253–268.

Lavine, M. (1992). Some aspects of polya tree distributions for statistical modelling. *The annals of statistics*, pages 1222–1235.

Lawson, A. B., Browne, W. J., and Rodeiro, C. L. V. (2003). *Disease mapping with WinBUGS and MLwiN*, volume 11. John Wiley & Sons.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer.

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

Looker, K. J., Garnett, G. P., and Schmid, G. P. (2008). An estimate of the global prevalence and incidence of herpes simplex virus type 2 infection. *Bulletin of the World Health Organization*, 86(10):805–812A.

López-Quilez, A. and Munoz, F. (2009). Review of spatio-temporal models for disease mapping.

Lu, H., Reilly, C. S., Banerjee, S., and Carlin, B. P. (2007). Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics*, 14(4):433–452.

Mabaso, M. L., Vounatsou, P., Midzi, S., Da Silva, J., and Smith, T. (2006). Spatio-temporal analysis of the role of climate in inter-annual variation of malaria incidence in Zimbabwe. *International Journal of Health Geographics*, 5(1):1.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741.

MacNab, Y. and Dean, C. (2000). Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models. *Statistics in Medicine*, 19(17-18):2421–2435.

MacNab, Y. C. (2010). On bayesian shared component disease mapping and ecological regression with errors in covariates. *Statistics in medicine*, 29(11):1239–1249.

MacNab, Y. C. and Dean, C. (2002). Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in medicine*, 21(3):347–358.

Manda, O. and Leyland, H. (2007). An empirical comparison of maximum likelihood and Bayesian estimation methods for multivariate disease mapping. *South African Statistical Journal*, 41(4):1–21.

Manda, S. M., Feltbower, R. G., and Gilthorpe, M. S. (2012). Review and empirical comparison of joint mapping of multiple diseases: review. *Southern African Journal of Epidemiology and Infection*, 27(4):169–182.

Manda, S. O. (2011). A nonparametric frailty model for clustered survival data. *Communications in Statistics—Theory and Methods*, 40(5):863–875.

Mardia, K. (1988). Multi-dimensional multivariate Gaussian markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284.

Milman, V. D. and Schechtman, G. (2009). *Asymptotic theory of finite dimensional normed spaces: Isoperimetric inequalities in riemannian manifolds*, volume 1200. Springer.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology.

Mishra, V., Montana, L., and Neuman, M. (2007). Spatial modeling of HIV prevalence in kenya. In Demographic and Health Research.

NASCOP (2007). Ministry of Health, Kenya: Kenya AIDS Indicator Survey report.

NASCOP (2012). Ministry of Health, Kenya: Kenya AIDS Indicator Survey report. Report.

Ngesa, O., Achia, T., and Mwambi, H. (2014a). A flexible random effects distribution in disease mapping models. *South African Statistical Journal*, 48(1):83–93.

Ngesa, O., Mwambi, H., and Achia, T. (2014b). Bayesian Spatial Semi-parametric Modeling of HIV variation in Kenya. *PloS one*, 9(7).

Njau, J., Goodman, C., Kachur, S., Palmer, N., Khatib, R., Abdulla, S., Mills, A., and Bloland, P. (2006). Fever treatment and household wealth: the challenge posed for rolling out combination therapy for malaria. *Tropical Medicine & International Health*, 11(3):299–313.

Nyarko, S. H. and Cobblah, A. (2014). Sociodemographic determinants of Malaria among under-five children in Ghana. *Malaria research and treatment*, 2014.

O'Hagan, A. and Kingman, J. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–42.

Pates, H. and Curtis, C. (2005). Mosquito behavior and vector control. *Annu. Rev. Entomol.*, 50:53–70.

Pérez-Escamilla, R., Dessalines, M., Finnigan, M., Pachón, H., Hromi-Fiedler, A., and Gupta, N. (2009). Household food insecurity is associated with childhood malaria in rural Haiti. *The Journal of nutrition*, 139(11):2132–2138.

Rezaeian, M., Dunn, G., St Leger, S., and Appleby, L. (2007). Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. *Journal of epidemiology and community health*, 61(2):98–102.

Røttingen, J.-A., Cameron, D. W., and Garnett, G. P. (2001). A systematic review of the epidemiologic interactions between classic sexually transmitted diseases and hiv: how much really is known? *Sexually transmitted diseases*, 28(10):579–597.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 325–338.

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

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

Ruebush, T., Burkot, T., de Oliveira, A., da Silva, J., Renshaw, M., et al. (2005). Presidents Initiative on Malaria needs assessment Angola 9-18 August 2005.

Seeger, M. W. (2008). Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9(Apr):759–813.

Sherman, M. (2011). *Spatial statistics and spatio-temporal data: covariance functions and directional properties*. John Wiley & Sons.

Speckman, P. L. and Sun, D. (2003). Fully bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90(2):289–302.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493.

Spiegelhalter, T., Best, N., and Lunn, D. (2007). Winbugs user manual version 1.4.3.

Sun, D., Tsutakawa, R. K., Kim, H., He, Z., et al. (2000). Spatio-temporal interaction with disease mapping. *Statistics in Medicine*, 19(15):2015–2035.

TACAIDS (2013). Tanzania HIV/AIDS and Malaria indicator survey 2011-2012. Report.

UNAIDS (2013). Report on global AIDS epidemic. Geneva, Switzeland. Report, UNAIDS/WHO.

UNICEF (2013). Massive scale-up targets malnutrition in Angola. Technical report.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 364–372.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.

Walker, S. G. and Mallick, B. K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):845–860.

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.

Waller, L. A. and Gotway, C. A. (2004). *Applied spatial statistics for public health data*, volume 368. John Wiley & Sons.

Wand, H., Whitaker, C., and Ramjee, G. (2001). Geoadditive models to assess spatial variation of HIV infections among women in Local communities of Durban, South Africa. *Intenational Journal of Health Geographics*, 10:1–9.

Weiss, H. (2004). Epidemiology of herpes simplex virus type 2 infection in the developing world. *National Center for Biotechnology Information*, 11:24A–35A.

Wheeler, D. and Waller, L. (2009). Comparing spatially varying coefficient models: A case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical systems*, 11(1):1–22.

White, G. (1974). Anopheles gambiae complex and disease transmission in Africa. *Transactions of the Royal Society of Tropical Medicine and hygiene*, 68(4):278–298.

WHO (2013). Report on sexually transmitted infections (stis).

WHO (2014). Report on global health observatory data.

WHO (2016). World health organization malaria fact sheet.

Yue, Y. R., Speckman, P. L., and Sun, D. (2012). Priors for Bayesian adaptive spline smoothing. *Annals of the Institute of Statistical Mathematics*, 64(3):577–613.

Yusuf, O. B., Adeoye, B. W., Oladepo, O. O., Peters, D. H., and Bishai, D. (2010). Poverty and fever vulnerability in Nigeria: a multilevel analysis. *Malaria journal*, 9(1):1.