University of KwaZulu-Natal

Bayesian Spatial Models With Application to HIV, TB and STI Modeling in Kenya

by

Ngesa Oscar Owino

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

July 2014

Declaration of Authorship

I, NGESA OSCAR OWINO, declare that this thesis titled, 'Bayesian Spatial Models With Application to HIV, TB and STI Modeling in Kenya' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- No part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

Signed: _____ Date: _____

Approved by Supervisors:

1. Prof. Dr. Henry Mwambi Signed: _____ Date: _____

2. Dr. Thomas Achia Signed: _____ Date: _____

List of Publications and Awards

1. O. Ngesa, T. Achia, and H. Mwambi, "A Flexible Random Effects Distribution in Disease Mapping Models", South African Statistical Journal, 2014, vol 48(1), 83-93.

2. O. Ngesa, T. Achia, and H. Mwambi, "Spatial Joint Disease Modeling and Mapping with Application to HIV and HSV-2", In Proceedings of the 55th Annual Conference of the South African Statistical Association, 2013, pp. 61-68.

3. O. Ngesa, T. Achia, and H. Mwambi, "Bayesian Hierarchical Generalized Gaussian Model With Sampling Weights In Complex Survey Data, accepted and presented during the International Conference on Mathematical Sciences and Statistics, 5-7 Feb 2013, Kuala-Lumpur, Malaysia.

4. O. Ngesa, T. Achia, and H. Mwambi, "Spatial Variation of HIV Infection in Kenya based on Complex Survey Data". *Submitted*.

5. O. Ngesa, T. Achia, and H. Mwambi, "A Double Random Effects Model for Spatially Correlated and Overdispersed Binary Data with Application to HSV-2 Variation in Kenya", *Submmitted*

6. First prize winner for oral presentation during the 2013 College of Agriculture, Engineering and Science, UKZN Postgraduate Research Day. Title of presentation: Joint Spatial Modeling of HIV and HSV-2".

7. Third prize winner for oral presentation during the 2012 College of Agriculture, Engineering and Science, UKZN Postgraduate Research Day. Title of presentation: Bayesian Spatial Modeling of HIV Variation Among Counties in Kenya: A Closer Look at Women Data".

"Untied but united."

Oscar Ngesa

University of KwaZulu-Natal

Abstract

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

Doctor of Philosophy

by Ngesa Oscar Owino

This dissertation is concerned with developing and extending statistical models in the area of spatial modeling with particular interest towards application to HIV, TB and HSV-2 data. Hierarchical spatial modeling is a common and useful approach for modeling complex spatially correlated data in many settings in epidemiological, public health and ecological studies. Chapter 1 of this thesis gives a chronological development of disease mapping models, from non-spatial to spatial and from single disease models to multiple disease models. In Chapter 2, a new model that relaxes the over-restrictive normal distribution assumption on the spatially unstructured random effect by using the generalised Gaussian distribution is introduced and investigated. The third chapter provides a framework for including sampling weights into the Bayesian hierarchical disease mapping model. In this model, design effect is used to re-scale the sample sizes. A new model for over dispersed spatially correlated binary data is developed in chapter 4 of this thesis; in this model, the over dispersion parameter is modeled by a beta random effect which is allowed to vary spatially also. In chapter 5, the common multiple spatial disease mapping models are reviewed and adopted for the binary data at hand since the original models were developed based on Poisson count data. The methodologies developed in this dissertation widen the toolbox for spatial analysis and disease mapping in applications in epidemiology and public health studies.

Acknowledgements

First and foremost I wish to thank the almighty God for bringing me this far and according me His mercies. Special thanks go to my supervisors Dr. Thomas Achia and Prof. Henry Mwambi for sharing their vast knowledge of statistics in general and spatial statistics in particular. Thank you for the many fruitful discussions and ideas and for lending me support throughout my study period. Many thanks goes to Prof. Samuel Manda who made several suggestions in this work.

I would like to thank the Ministry of Youth Affairs and Sports, in the Government of Kenya, for giving me a study leave that made this work possible. I thank all my colleagues at the Central Project Planning and Monitoring Unit (CPPMU) in this ministry for their support while I was away, in a particular manner, Mr. T.G Gakuu and Mr. Earnest Alela.

Thanks to my friends at Jomo Kenyatta University of Agriculture and Technology (JKUAT) for walking with me in this academic path and supporting my ambitions since my undergraduate; thanks to Dr. George Orwa, Prof. Dr. Romanus Odhiambo, Prof. Dr. John Kihoro, Dr. Waititu, Miss Jane Akinyi, Mr. Edward Mboya, Mrs Caro Mugo and the other teaching and non-teaching staff at the statistics department.

Thanks to my colleagues and friends at the University of KwaZulu-Natal for providing a pleasant and inspiring working environment. Thanks to Robert Mutwiri, Oyoo Yaye, Charles Otunga, Arts, Faustin, Nancy, Ijumba, Mushi, Kondwani, Stu, Dr. Ayoub, Mrs Swart, Dr. Chirove, Sbongakonke Masondo, Titilayo, Abdalla, Christel, Bev and Nzimande for their multidimensional support.

Last but not least, my family is thanked for their continuous support and for keeping me relevant in their lives even when I was away. I thank my girlfriend Alice Aduda for her unfailing love and my cousin Seth Owiti for his support, which started while we were together in JKUAT. In a special way I would like to thank Mr. George Amolo, Mr. Opiyo Ngesa and Dr. Alfred Owiti for seeing me through my early education.

Contents

List of Figures

List of Tables

Abbreviations

BYM	**B**esag, **Y**ork and **M**ollie
CAR	**C**onditional **A**uto**r**egressive
DHS	**D**emographic and **H**ealth **S**urveys
DIC	**D**eviance **I**information **C**riterion
DLTLD	**D**ivision of **L**eprosy, **T**uberculosis and **L**ung **D**isease
GGD	**G**eneralized **G**aussian **D**istribution
GLM	**G**eneralized **L**inear **M**odels
GLMM	**G**eneralized **L**inear **M**ixed **M**odels
HIV	**H**uman **I**mmunodeficiency **V**irus
HSV-1	**H**erpes **S**implex **V**irus-Type **1**
HSV-2	**H**erpes **S**implex **V**irus-Type **2**
ICAR	**I**ntrinsic **C**onditional **A**uto**r**egressive
KAIS	**K**enya **A**IDS **I**ndicator **S**urvey
MCAR	**M**ultivariate **C**onditional **A**uto**r**egressive
MSE	**M**ean **S**quared **E**rror
McMC	**M**arkov **c**hain **M**onte **C**arlo
STI	**S**exually **T**ransmitted **I**nfection
TB	**T**uberculosis
WHO	**W**orld **H**ealth **O**rganization

Dedicated to my parents, Joseph Ngesa, Gaudencia

Ngesa, Rose Ngesa and Peres Ngesa.

Dedicated to Anita Kerubo Ondieki.

Chapter 1

General Introduction

1.1 Overview

Of late, there has been a rising interest in the development and application of spatial statistical methods for analysis of geographically correlated data. This can be attributed to the increasing availability of geo-referenced data in many fields of study, for example public health and ecology. Most of the data collected by many African governments through surveys and sentinel surveillance are geo-referenced by districts, counties, provinces or other administrative units. The thesis is biased towards hierarchical models for count data, simply due to the availability of such data in this setting. We introduce this thesis by discussing key concepts in spatial modeling followed by the chronological development of spatial models, building up from single disease models to joint disease models.

1.2 Disease Mapping

Disease mapping refers to the estimation and presentation of summary measures of health outcomes[Rezaeian et al., 2007]. Some of the purposes of disease mapping include, to;

1. describe geographical variation of diseases.
2. generate hypotheses about a disease.
3. generate disease atlases.
4. detect clustering of a disease.

Mapping of disease incidence and prevalence is a common place in public health and epidemiology. Often the primary interest in disease mapping is to smooth and predict some response variables over a geographical domain of interest. The area-specific estimates of the diseases can be used by policy makers when making decisions on public health resources allocation. There are two fundamental characteristics of disease mapping, namely geographical distribution and disease.

1.3 Hierarchical Models for Disease Mapping

Complex disease mapping and spatial models stem from the generalized linear model. Consider a study domain D, partitioned into a set of regions, $i = 1, 2, \cdots, n$. Let y_i be the observed diseases counts in region i. These are typically the number

of cases prevalent or incident in region i. The counts are modeled as either Poisson or Binomial random variables in the generalized linear models (GLM) setting [Wakefield, 2007; Ghosh et al., 1999]. When the disease under discussion is rare, then the Poisson model can be used as an approximation to the binomial model. Covariates can be introduced into the models in the standard GLM fashion.

An extension of this model is the inclusion of an uncorrelated random effect, leading to the generalized linear mixed model (GLMM) with one random effect [Lawson et al., 2003]. The model is specified in hierarchy form, with two stages. In the first stage, the observed counts are conditionally independent given the values of the random effects. In the second stage, the distribution of the random effects is specified. This also allows for over-dispersion if the data model is Poisson.
Thus, so far, the model elicit a non-spatial correlation between the observations, and in fact the area-specific relative risk estimates are concessions between local data and global weighted averages obtained from the entire dataset. It is possible to introduce correlated random effects via a spatial covariance matrix. This can be achieved by considering the random effects to form a single vector following an appropriate distribution with a specified mean and a spatial variance-covariance matrix as opposed to the random effects being exchangeable. The most common assumption on the distribution of the random effects is a multivariate Gaussian distribution [Waller and Gotway, 2004; Gaetan and Guyon, 2010; Sherman, 2011].

The spatial variance-covariance matrix is made up of parametric functions defining the covariance structure based on location of any two units of study. For geostatistical data, the spatial covariance between two observations is dictated by the

distance between the two observations [Waller and Gotway, 2004; Cressie, 1993; Diggle et al., 1998]. In the case of lattice data, the neighbourhood can be specified based on the basis of sharing a border, the distance between the centroids of any pair of regions or a combination of these two.

Clayton and Kaldor [1987] introduced spatially structured prior distribution for the random effects. Estimation was done using empirical Bayes approach in which the relative risk estimate for a region was a compromise between local data and a weighted average of observations in the neighbourhood of that region. Besag et al. [1991] introduced the full Bayesian approach counterpart to the Clayton and Kaldor [1987] formulation. They implemented their model using Markov chain Monte Carlo (McMC) algorithms. Their specification gives an alternative to using the multivariate Gaussian models. It is called the conditional autoregressive model (CAR).

In CAR, the conditional distribution of a random effect in a region given all the others is simply the weighted average of all the other random effects. 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 is 1, otherwise it is 0. Best et al. [1999] discusses other weighting possibilities. All the weighting options discussed this far are assumed to be fixed when modeling. Lu et al. [2007] took another approach for weighting by estimating the weights from the data itself.

The CAR formulation has computational advantage over the multivariate Gaussian in the sense that the variance component in multivariate Gaussian requires

matrix inversion during its estimation, at each update when executing the algorithm leading to more computational burden; this is not required in CAR.

Besag et al. [1991] further extended this model by advocating for inclusion of both spatially unstructured random effects and spatially structured random effects, through the convolution model. This allows the model to borrow information both locally and globally. There is need to assign prior weight fairly to these two components so as to avoid either global over smoothing or local over smoothing. Bernardinelli et al. [1995] focussed their study on this and came up with a conclusion that the standard deviation of the conditional distribution of the spatially structured random effects should be 0.7 times the standard deviation of the spatially unstructured random effects. This conclusion is still open for debate. If the prior distributions of the precision parameters of the two random effects in the convolution model are taken to be non-informative, then only the sum of the two random effects will be identifiable and not the individual components.

Several authors have also proposed alternative formulations for the convolution model. Notably, Leroux et al. [1999] as opposed to Besag et al. [1991] formulation of a random intercept split into two components, the authors used only one random intercept and its variance covariance matrix was split into spatial and non-spatial components, with a parameter controlling the spatial dependency. For other authors with alternative proposals, see MacNab and Dean [2000] for a parametric bootstrap approach, and Green and Richardson [2002] for their hidden Markov field approach.

With all these alternative models, the Besag et al. [1991] formulation still enjoys

more application due to its close fit with common McMC implementations and also because of a wide variety of readily available software, e.g WinBUGS [Spiegelhalter et al., 2007] and R statistical software [RDevelopment, 2005] implementing it. Extensions of this model for zero inflated datasets has been considered [Lambert, 1992; Agarwal et al., 2002; Agarwal, 2006].

Hitherto, we have been discussing single disease modelling. There are two major approaches for handling multiple diseases, namely the shared component approach and the multivariate CAR approach [MacNab, 2010]. Bernadinelli et al. [1997] modelled two diseases by treating one disease to be a covariate while adjusting for sampling error on it. In a novel manner, Knorr-Held and Best [2001] extended BYM model to enable modeling two diseases jointly. Their formulation was called the shared component model, in which each of the random effects was shared by two diseases. Held et al. [2005] further extended this shared component model for the case of more than two diseases in a similar fashion. Wang and Wall [2003] also developed a general common spatial factor model, in which they hypothesised that a common spatially correlated latent factor causes the correlation between variables measured at the same location and correlations of each variable across locations.

Kim et al. [2001] proposed the two fold CAR model in which they allowed for sharing of information between neighbouring regions with respect to the same disease and also between the two diseases within the same region. Carlin and Banerjee [2003] and Gelfand and Vounatsou [2003] separately, at almost the same time, developed the multivariate conditional autoregressive (MCAR) model for modeling

multiple diseases under separability assumptions. The separability assumption dictates that the association structure decomposes into spatial and non-spatial components. The joint consideration comes in the sense that the spatial random effects were assumed to follow a joint distribution which allows for correlation of the components. They also assumed that there was a single parameter that controlled spatial dependency in all the diseases. Jin et al. [2005] further extended the model by introducing parameters that allowed for cross-covariance between diseases and regions. When there are only two diseases, this model reduces to the model proposed by Kim et al. [2001].

Reich et al. [2007] introduced a class of models that allow for two different classes of neighbourhood definitions. Another class of extension is the case of spatially varying coefficients. Instead of tying the spatial dependency on the random effects, the coefficients can be allowed to vary through the spatial domains, this allows for the relationship between responses and covariates to vary by region in the spatial domain [Hastie and Tibshirani, 1993; Hoover et al., 1998; Assuncao et al., 2002; Assunçao, 2003; Pavlov, 2003; Gamerman et al., 2003; Gelfand et al., 2003]. It is important to note that so far spatial dependency has been introduced into the model using random effects and spatially varying coefficients. It is possible to introduce the dependency via the observations themselves. Examples of these include autologistic model [Hoeting et al., 2000] and the autoPoison models [Besag, 1974; Griffith, 2002]. It is also worth noting that the overview provided above caters only for discrete variation of disease. This type of spatial data considered

is called areal data (lattice data). Two other basic types of spatial data are point-referenced data (geostatistical data) and point pattern data. For more discussions and exposure on geostatistical and point pattern spatial data types, we refer the readers to Diggle and Ribeiro [2007] and Cressie [1993] respectively.

1.4 Statement of the Problem

Several models have been developed to deal with disease models, both for single diseases and multiple diseases. Most of these models are based on the use of random effects, which is split into spatial and non-spatial components. Normality assumption is always used for the non-spatial component. These models have also been developed with knowledge that the data come from surveillance data or registry files. Very few models have been developed to deal with data from complex sample surveys.

There is need to consider models that allow flexibility in the normal random effects, this flexibility could be due to high peaked and low peakedness of the distributions. In this work we allow the random effects to have less than 3 or greater than 3, kurtosis values, allowing for low and high peakedness respectively.

There is also need to extend the models to accommodate data from complex surveys. This can be achieved by incorporating the sampling weights into the model. Overdispersion in count data is a common phenomenon. It is also possible for overdispersion and spatial autocorrelation to occur simultaneously. This thesis

also consider this dual problem and develops a model that caters for both overdispersion and spatial correlation while allowing the overdispersion parameter to vary spatially over space.

1.5 Main Objective

The main objective is to develop flexible models for disease mapping for both sentinel surveillance data and complex survey data.

1.6 Specific Objectives

The specific objectives for this thesis are:

- to review disease mapping models for single and multiple diseases.
- to incorporate survey weights in disease mapping with data from complex surveys.
- to develop models with less restrictive prior assumption on the unstructured heterogeneity random effect.
- to develop joint disease models for binary data
- to develop models that cater for overdispersion in spatially correlated binary data.

1.7 Dissertation Outline

This thesis is concerned with development of models and methods for the spatial analysis of diseases. The thesis is organised in form of chapters which represent full research papers that have been published in peer reviewed journals or submitted to the same. Each paper has been written as a stand-alone article that can be read separately from the rest of the thesis but draws separate conclusions that link to the overall research objectives. This dissertation is made up of six chapters, the section below describes the contents of each chapter.

Chapter 1: This chapter serves as an introduction to the study, giving recent developments in disease mapping for both single and multiple diseases, and the objectives of this study.

Chapter 2: This chapter extends the BYM model by introducing a flexible random effect distribution for the spatially unstructured random effect. The use of the generalised Gaussian distribution (GGD) is investigated using simulation studies and applied to tuberculosis data collected in Kenya in the year 2002 by the division of leprosy, TB and lung disease (DLTLD) under the Ministry of Health, Kenya. The GGD allows for the random effects to depart from the frequently assumed normal distribution.

Chapter 3: In this chapter a spatial model that incorporates sampling weights is developed. The weights are included in the model using design effects which adjusts the actual observed counts of diseases and the sample size in a binomial

setup. The developed model is applied to modeling HIV variation in Kenya for men and women.

Chapter 4: In this chapter, a model is developed to handle spatially correlated and overdispersed binary data. Overdispersion is a common phenomenon especially when the data follows Poisson distribution and binomial distribution. Overdispersion occurs when the usual mean-variance relationship is not adhered to in these distributions. In the proposed model, the overdispersion parameter is allowed to vary spatially in the regions under study. The model is used to model Herpes Simplex Virus-Type 2 (HSV-2) variation in Kenya.

Chapter 5: This chapter reviews the commonly encountered models for spatial joint modeling of diseases and adopted for bivariate spatial logistic models to suit the data at hand and the resulting models were used to jointly model HIV and HSV-2 in Kenya.

Chapter 6: Finally, this chapter gives a summary of the thesis in a nutshell. The findings are summarized and conclusions are derived from the preceding chapters. Topics for further study are highlighted in this section. A single reference list is given at the end of the dissertation.

Chapter 2

A Flexible Random Effects Distribution in Disease Mapping Models

Disease mapping has seen many applications in epidemiology and public health. The basic model used in disease mapping is the Besag, York and Mollie model, which incorporates two random effects, one which is spatially structured and the other random effect which is spatially unstructured. The normality assumption on the spatially unstructured random effect is very common. In this work, we investigate a more robust spatially unstructured random effect distribution by considering the symmetric generalized Gaussian distribution in the disease mapping problem. The distribution has the normal and Laplace distributions as special cases. The inference under this model are carried out under the Bayesian approach implemented in WinBUGS. The generalized Gaussian distribution is introduced in

WinBUGS using zero tricks. The usefulness of the proposed model is investigated with a simulation study and applied in real data; mapping tuberculosis in Kenya. In this paper we showed that the generalized Gaussian distribution can produce better results when the normality assumption is violated due to high peakedness or less peakedness in the data. For the case of data in which the random effects are truly normal, the generalized Gaussian distribution adjusts to a normal distribution as dictated by the data itself.

2.1 Introduction

Disease mapping refers to the estimation and presentation of summary measures of spatially observed health outcomes. The increased availability of georeferenced data and flexible computational softwares has seen rise in application of disease mapping in the areas of epidemiology and public health [Rezaeian et al., 2007; Everitt and Dunn, 2011]. Disease mapping can be used to describe geographical variation of diseases, identify clustering of diseases and generate atlas of diseases. A number of statistical reviews on disease mapping have been done [Wakefield, 2007; Clayton and Bernardinelli, 1992; Smans and Esteve, 1997; Wakefield et al., 2000; Manda et al., 2011].

The backbone model for univariate disease mapping is the Besag, York and Mollie (BYM) model proposed by Besag et al. [1991]. This model is a form of the generalized linear mixed effects model, with two random effects; a spatially unstructured random effect which is modelled using a normal prior and a spatially

structured random effect which is modelled using an intrinsic conditional autoregressive (ICAR) prior. The use of normal distribution to model the spatially unstructured random effects is mainly because of its computational simplicity. Assumption of normality on the uncorrelated random effect in models is common. Sometimes this assumption is incorrect because some random effects can in fact be platykurtic, leptokurtic or skewed; diverging from this general normality assumption [Box and Tiao, 1973]. When this normality assumption is violated, there is need to consider other models that would better suit the data at hand. The generalized Gaussian distribution can be used in cases where there is deviation from the normal kurtosis ($kurtosis = 3$) and when there is evidence of skewness in the data. It is a generalization of the common normal distribution to allow for these departures.

The general Gaussian distribution has two versions, both of which add a shape parameter to the normal distribution. The first version of the generalized Gaussian distribution includes normal and Laplace distributions. The continuous uniform distribution arises naturally as a limiting case for this distribution. All the distributions encompassed under this family are symmetric. In the second version, the shape parameter is used to incorporate skewness in the family of distributions. Positive values of the shape parameter produce distributions which are skewed to the left while negative values lead to right skewed distributions. In this work, we concentrate on the symmetric version of the generalized Gaussian distribution. The generalized Gaussian distribution, which we will denote by GGD, has three

parameters, the location parameter, μ, the scale parameter, σ^2 and the shape parameter, ϕ. The shape parameter dictates the amount of peakedness or kurtosis.

This work is structured as follows: in section 2, we review the BYM model, in section 3 we introduce the symmetric generalized Gaussian distribution and discuss its limiting distributions, in section 4, we carry out a simulation to study the effect of misspecifying the random effects, in section 5 we use the discussed models to analyze the tuberculosis (TB) data from Kenya and finally discussions and conclusions in section 6.

2.2 Review of the BYM model

The most commonly used model in single disease spatial analysis was proposed by Besag et al. [1991]. It was used to model disease prevalence in regions using the Poisson model. Let λ_i be the unknown relative risk for region i with respect to a standard population. Also let y_i denote the observed counts of disease in region i and e_i denote the expected count in the same region. The model assumed that the log of relative risk of disease can be broken down into a spatially structured component u_i and a spatially unstructured component v_i. This can be written mathematically as

$$y_i \sim Poisson\left(e_i\lambda_i\right), \tag{2.1}$$

with

$$log\left(\lambda_i\right) = u_i + v_i, \tag{2.2}$$

where u_i and v_i are random effects representing unobserved covariates, with u_i representing variables that if were observed would influence the spatial structure, while v_i represents the unobserved heterogeneity in region i. Besag et al. [1991] noted that in most cases, one of the random effects usually dominates the other. If u is stronger than v, then the estimated risk will show spatial structure and if v is stronger than u then the consequence will be to shrink the estimated means towards the overall mean. Besag et al. [1991] assumed that u and v were independent with the following priors:

$$p(v|\tau) \propto \tau^{\frac{-n}{2}} \exp\left\{-\frac{1}{2\tau}\sum_{i=1}^{n} v_i^2\right\}, \tag{2.3}$$

and

$$p(\boldsymbol{u}|k) \propto k^{\frac{-n}{2}} \exp\left\{-\frac{1}{2k}\sum_{i}\sum_{j\in N(i)} (u_i - u_j)^2\right\}. \tag{2.4}$$

Basically, equation (2.3) means that $\boldsymbol{v}$, the spatially unstructured component is a white noise Gaussian process with unknown variance τ, and equation (2.4) means that the spatially structured component $\boldsymbol{u}$, is a Gaussian Markov random field (GMRF) process with variance k, n being the number of regions under study and $N_{(i)}$ is the set of neighbours of region i. The neighbourhood can be defined in terms of Euclidean distance of the centroids of the regions, whether two regions share a border or a combination of these two. Besag et al. [1991] defined their neighbourhood based on shared border.

This implies that the conditional distributions of each u_i, given the rest, are given by

$$(u_i|\boldsymbol{u}_{-i}) \sim N\left(\frac{\sum_{j\in N_{(i)}} u_j}{d_i}, \frac{k}{d_i}\right), \tag{2.5}$$

with

$$E\left(u_i|\boldsymbol{u}_{-i}\right) = \frac{\sum_{j\in N_{(i)}} u_j}{d_i} \tag{2.6}$$

and

$$Var\left(u_i|\boldsymbol{u}_{-i}\right) = \frac{k}{d_i}, \tag{2.7}$$

where d_i is the number of neighbours of region i. This conditional distribution for u is called the intrinsic conditional autoregressive (ICAR) prior distribution. Besag et al. [1991] sampled the posterior distribution using the Gibbs sampler, an McMC algorithm.

2.3 The generalized Gaussian Distribution and its properties

Several authors, Wakefield [2007]; Besag et al. [1991]; Best et al. [1999], have mentioned that it is possible to replace the normality assumption of the spatially unstructured random effect with either the Laplace distribution or the Student's t distribution. In this work we explore the use of the generalized Gaussian distribution as a candidate for random effects. It has pleasant properties, allowing for the data to dictate the best fitting model for the random effects, whether normal or Laplace adaptively. The generalized Gaussian distribution can result in several

interesting distributions upon varying the shape parameter. Of great interest is the fact that the GGD reduces to a normal distribution when the shape parameter has a value of two and to the Laplace distribution when the shape parameter is one.

Definition 2.1. A random variable X is said to have a GGD if its probability density function is given by

$$f(x;\mu,\sigma,\phi) = \frac{1}{2\Gamma\left(1+\frac{1}{\phi}\right)\zeta(\phi,\sigma)} \exp\left(-\left|\frac{x-\mu}{\zeta(\phi,\sigma)}\right|^{\phi}\right) \tag{2.8}$$

where $x,\mu \in \mathbb{R}, \sigma > 0$ and $\zeta(\phi,\sigma) = \left[\frac{\sigma^2\Gamma\left(\frac{1}{\phi}\right)}{\Gamma\left(\frac{3}{\phi}\right)}\right]^{\frac{1}{2}}$. In this expression $\zeta(\phi,\sigma)$ is a scaling factor. See Nadarajah [2005] for further discussions on statistical properties of this distribution.

2.3.1 Special cases of the GGD

Property 1. *If $\phi = 1$ and $\sigma^2 = 2b$, then the pdf of the GGD becomes*

$$f(x;\mu,b,1) = \frac{1}{2b}\exp\left(-\left|\frac{(x-\mu)}{b}\right|\right). \tag{2.9}$$

Equation (2.9) is the Laplace probability density function with location parameter μ and scale parameter b.

Property 2. *If $\phi = 2$, (2.8) becomes*

$$f(x;\mu,\sigma,2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right). \tag{2.10}$$

Equation (2.10) *is the probability density function of a normal random variable with mean* μ *and variance* σ^2.

Property 3. *The uniform distribution is a limiting case of the GGD when* $\phi \to \infty$.

2.4 Simulation

In this section, we carry out a simulation study to determine the effect of wrongly specifying the distribution of the random effect in a BYM model. Three scenarios were considered in the simulation. In the first simulation, the datasets is generated through a random effect, v with a peaked kurtosis as follows: Assuming that there are 60 geographical regions and O_i is the number of disease counts observed in region i and E_i is the corresponding expected counts in that region. Without loss of generality, we further assume that no covariates are available for use.

1. Step 1: Generate 60 values of $v \sim GGD(0, 0.1, 1.02)$, which is platykurtic.
2. Step 2: Set $E = 40$ for all the regions.
3. Step 3: Calculate the relative risk $\theta = log(v)$
4. Step 4: Calculate $\lambda = E \times \theta$
5. Step 5: Generate the observed counts as $O \sim Poisson(60, mean = \lambda)$.

We fitted two Bayesian hierarchical models for the data set. The models were specified based on different assumptions on the random effects as follows:

$$O_i \sim Poisson(\mu_i) \tag{2.11}$$

and

$$\log(\mu_i) = \log(E_i) + v_i \tag{2.12}$$

with

- Model a: $v_i \sim GGD(0, \sigma_1^2, \phi)$,
- Model b: $v_i \sim N(0, \sigma_2^2)$.

The estimated relative risk $\hat{\theta}_i = log(v_i)$.

The simulation steps above were repeated $m = 1000$ times.

To compare the two models, we calculated the mean squared error (MSE), for each model, using the formula:

$$MSE = \frac{1}{1000} \sum_{j=1}^{1000} \frac{1}{60} \sum_{i=1}^{60} \left(\hat{\theta}_{ij} - \theta_{ij} \right)^2. \tag{2.13}$$

The model with a small MSE provides the best fit.

In the second scenario, the procedure above is repeated but changing the random effect generation mechanism in step 1, as $v \sim N(0, 0.1)$.

Similarly, in the third scenario, the procedure was repeated with the random effect being generated using $v \sim GGD(0, 0.1, 12)$, yielding platykurtic random effect.

From the simulation results, the generalized Gaussian random effect is seen to adapt well even in cases where the random effect strictly follows a normal distribution. In Table 2.1, the generalized Gaussian distribution produces lower mean squared error values as compared to the normal distribution in all the cases. When the random effects are platykurtic, the loss in efficiency incurred for using normal

TABLE 2.1: Model comparison under simulation

Generating dist.	parameters used	model used	MSE	% loss in efficiency
GGD(Leptokurtic)	$\mu = 0, \sigma^2 = 0.1, \phi = 1.02$	Normal	0.006497	17.8
		GGD	0.005514	0
Normal	$\mu = 0, \sigma^2 = 0.1$	Normal	0.003297	6.2
		GGD	0.003093	0
GGD(Platykurtic)	$\mu = 0, \sigma^2 = 0.1, \phi = 12$	Normal	0.002926	17.8
		GGD	0.002483	0

random effects is 17.841% and when the random effects are leptokurtic, the percentage loss in efficiency is 17.822%. When the random effects are generated using normal distribution, still the GGD has a lower mean squared error compared to the normal counterpart. The loss in efficiency for this case is low at 6.19%.

2.5 Application: Mapping of tuberculosis in Kenya

In this section we apply the model to TB data collected by the Ministry of Health, Kenya. The division of leprosy, TB and lung disease (DLTLD) is responsible for the data collection within this ministry. This central unit receives case finding reports from all counties on a quarterly basis and aggregates the values for the whole year. The Table 2.3 in Appendix 1a summarises the number of TB cases per county and the corresponding population estimates for the year 2002.

The following models, with increasing complexity, were fitted:

$$O_i \sim Poisson(\mu_i) \tag{2.14}$$

with

- Model 1: $\log(\mu_i) = \log(E_i) + \beta_0 + v_i; \quad v \sim N(0, \sigma_v^2)$
- Model 2: $\log(\mu_i) = \log(E_i) + \beta_0 + v_i; \quad v \sim GGD(0, \sigma_v^2, \phi)$
- Model 3: $\log(\mu_i) = \log(E_i) + \beta_0 + u_i; \quad u \sim ICAR$
- Model 4: $\log(\mu_i) = \log(E_i) + \beta_0 + v_i + u_i; \quad v \sim N(0, \sigma_v^2), u \sim ICAR$
- Model 5: $\log(\mu_i) = \log(E_i) + \beta_0 + v_i + u_i; \quad v \sim GGD(0, \sigma_v^2, \phi), u \sim ICAR$

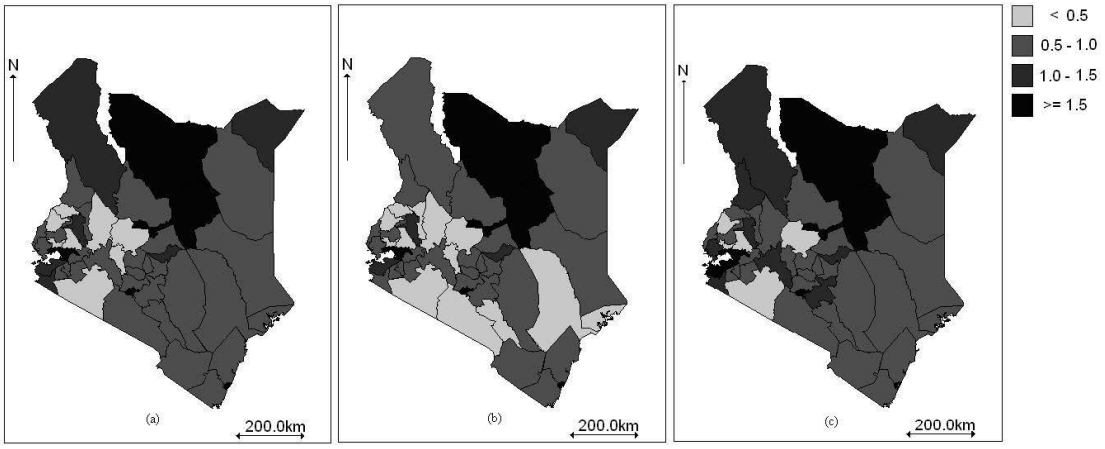
where O is the observed counts of cases of TB and E is the expected count of cases of TB. Model estimation was carried out using a Bayesian approach. All parameters in the models were assigned prior distributions. In this analysis, a non informative normal prior was assigned to the fixed effect coefficient β_0, the shape parameter ϕ was given a diffuse, uniform prior, and the variance parameters were assigned inverse gamma distributions. The models were implemented using WinBUGS version 1.4 [Spiegelhalter et al., 2007; Ntzoufras, 2011]. For each model, 50,000 Markov chain Monte Carlo (McMC) iterations were ran, with the initial 10,000 discarded to cater for the burn-in period and thereafter keeping every tenth sample value. The 4,000 iterations left were used for assessing convergence of the McMC and parameter estimation. We assessed McMC convergence of all models parameters by checking trace plots and autocorrelation plots of the McMC output, see Gelman et al. [2003]. The models were compared using the Deviance Information Criterion (DIC) as suggested by Spiegelhalter et al. [2002]. The best fitting model is one with the smallest DIC value. In this analysis, the unstructured

TABLE 2.2: Model comparison in mapping TB in Kenya

	Estimates					
Model	β_0	ϕ	σ_u^2	σ_v^2	pD	DIC
Model 1	-0.21(-0.31,-0.08)	-	-	0.07(0.03,0.19)	46.48	510.64
Model 2	-0.20(-0.38,-0.02)	2.27(1.07,6.27)	-	0.25(0.18,0.47)	45.76	507.81
Model 3	-0.22(-0.22,-0.19)	-	1.31(0.59,2.69)	-	49.76	515.12
Model 4	-0.22(-0.39,-0.11)	-	0.02(0.01,0.13)	0.06(0.03,0.11)	46.98	511.35
Model 5	-0.18(-0.30,-0.02)	3.92(0.82,7.83)	0.31(0.00,1.56)	0.16(0.00,0.34)	50.55	519.47

heterogeneity, modelled using the generalized Gaussian distribution was found to perform slightly better than the other models considered in this study. This can be seen in Table 2.2, based on the DIC values. Figure 2.1 shows the spatial distribution of TB in Kenya based on this best fitting model. This is a map of relative risk and its corresponding credible interval.

FIGURE 2.1: TB relative risk map(a) and the corresponding 95% lower(b) and upper(c) credible limits maps, respectively, produced by model 2.



2.6 Discussion

Routine framework for modelling correlated data is through the generalized linear mixed effects model in which a random effect is incorporated. The usual main assumption in a standard version of the setup is to model the between subject variations with random effects that are normally distributed; this choice to some extent has been driven, over the years, by computational ease or flexibility in generating the posterior distribution. The assumption of modelling random effects with a normal distribution has been both challenged and supported by several authors [McCulloch and Neuhaus, 2011; Litière et al., 2007, 2008]. A lot of work has been done in trying to find better fitting distributions in the recent past. Magder and Zeger [1996] proposed a smooth non-parametric maximum likelihood approach to modelling the random effects. Verbeke and Lesaffre [1997] proposed using a mixture of normal distributions for the random effects and they carried out their estimation using the expectation maximization (EM) algorithm. Zhang and Davidian [2001] proposed a semi-parametric linear mixed model in which they assumed that the random effects have a smooth density represented by semi-nonparametric truncated series expansion. Ho and Hu [2008] used a finite mixture of normal in a Bayesian setting with the number of components being estimated from the data automatically.

In disease mapping context, the same situation arises. The basic BYM model has two components, one which is spatially structured and the other component which is spatially unstructured. The spatially unstructured component is usually modelled using the normal distribution. In this work we propose the generalized

Gaussian distribution as a random effect distribution to replace the over-restrictive normal distribution for the unstructured heterogeneity. The special cases of the generalized Gaussian distribution, including the normal and Laplace distributions, are exposed. The generalized Gaussian distribution has an extra parameter to allow for high and low peakedness as dictated by the data.

The parameters in the models are estimated under Bayesian inference. The models were implemented in WinBUGS. The generalized Gaussian distribution is not a standard distribution in the WinBUGS software. We introduced this distribution in the software using zero tricks, see Appendix 1b. The models were compared using simulation studies and again with a real data set.

In the simulation study it was seen that the effect of misspecification of the random effects when the normal distribution is used in place of the generalized Gaussian distribution was high as compared to using the generalized Gaussian in place of the normal distribution. The generalized Gaussian distribution has all the nice properties of the normal distribution. In fact the normal distribution is a special case of the generalized Gaussian distribution. When the random effect distribution fails to adhere to the normality assumption due to peakedness, the generalized Gaussian distribution plays a big role in capturing this, something that the normal distribution cannot.

In the real data sets comparison, the generalized Gaussian distribution is seen to perform better than the normal distribution model. This model was used to produce county specific maps of relative risk of TB in Kenya. The maps are critical

in understanding disease epidemiology and also in helping policy makers to develop informed intervention programs and allocate scarce resources adequately.

One limitation of this model is that it only captures high and low peakedness departures from the normal distribution. It assumes that the random effects are symmetric. This assumption can at times also be wrong. More flexible random effects models, which can also capture skewness can be investigated.

Appendix 1a: Tuberculosis data

Table 2.3 gives the number of TB cases reported for each county and the corresponding population as at 2002 in Kenya.

TABLE 2.3: Number of TB cases reported for each county and the corresponding population size.

County	TB cases	Population	County	TB cases	Population
Baringo	572	461175	Mandera	993	286006
Bomet	729	437321	Marsabit	989	194960
Bungoma	1343	1134381	Meru	2380	1221068
Busia	1262	618068	Migori	1974	746904
Elgeyo Marakwet	395	326798	Mombasa	5889	755867
Embu	1145	497662	Muranga	1541	808488
Garissa	1000	480489	Nairobi	15979	2495170
Homa Bay	3159	829355	Nakuru	3413	1354899
Isiolo	611	107741	Nandi	720	659957
Kajiado	613	461174	Narok	610	604298
Kakamega	1979	1454722	Nyamira	763	548053
Kericho	1936	890544	Nyandarua	639	526742
Kiambu	2638	1523061	Nyeri	1536	722739
Kilifi	1521	923837	Samburu	340	163001
Kirinyaga	721	502243	Siaya	2034	790555
Kisii	2105	1052456	Taita Taveta	569	279951
Kisumu	4753	882705	Tana River	278	195965
Kitui	2166	908106	Tharaka-Nithi	1053	338616
Kwale	945	559901	Trans Nzoia	876	652005
Laikipia	344	365759	Turkana	1340	516833
Lamu	111	83985	Uasin Gishu	2384	707664
Machakos	2474	1005586	Vihiga	515	561538
Makueni	1119	856800	Wajir	743	377527
			West Pokot	915	349857

Appendix 1b: Specifying the GGD prior in WinBUGS using zero tricks

Since the GGD prior is not specified as a standard distribution in WinBUGS, we used the zero trick technique to specify this prior distribution. Suppose we wish to use a prior $GGD(\theta)$ which is not available, we first define a flat/non-informative prior for θ, say, $h(\theta)$. Next we define a dummy variable, say $Zeros$, with all its values set to zero. We also specify this dummy variable, $Zeros$, to be following a Poisson distribution with mean λ. Then we set λ to be equal to the negative log-likelihood of the prior distribution that we are interested in, the GGD, that is $\lambda = -log(GGD(\theta))$.

$$\begin{aligned} f(\theta|Zeros) &= f(Zeros = 0|\theta)h(\theta) \\ &= \frac{e^{-\lambda}\lambda^{Zeros}}{Zeros!}h(\theta) \\ &= \frac{e^{-\lambda}\lambda^{0}}{0!} \times 1 \\ &= e^{-\lambda} \\ &= e^{log(GGD(\theta))} \\ &= GGD(\theta) \end{aligned}$$

This is the prior distribution that that we wanted. The corresponding code in Winbugs software is given below.

```
#Generalized Gaussian prior distribution implementation using zero tricks

my_zeta<-pow(abs((sigmau*sigmau*exp(loggam(1/psi)))/(exp(loggam(3/psi)))),0.5)
```

```
normalizing<-1/((2*exp(loggam(1+1/psi)))*my_zeta)

logN<-log(normalizing)

for(i in 1: n)

{

v[i]~dunif(-10000,10000)

zeros[i]<-0

logGGD[i]<-logN-pow(abs(v[i]/my_zeta),psi)

zeros[i]~dpois(-logGGD[i])

}
```

Chapter 3

Spatial Variation of HIV Infection in Kenya based on Complex Survey Data

Human Immunodeficiency Virus (HIV) still remains a leading public health problem in Sub-sahara Africa and many parts of the world. Understanding geographical variation of HIV is key in formulating policies and interventions to combat it. Many governments, especially in developing countries rely on using Demographic and Health Surveys to monitor the progress of intervention programs put in place to fight HIV and other scourges.

The objective of this study is to demonstrate how to incorporate survey weights from complex survey data in spatial analysis, investigate the effect of weight inclusion on inference made and generate disease maps.

The weights are incorporated so as to take care of the nature of the complex survey design used to collect the data. The effect of survey weight incorporation is investigated by comparing variability between weight adjusted and unadjusted estimates. The method is applied to the Kenya Aids indicator survey dataset, collected by the government of Kenya, in 2007. It involves 19,840 individuals in the of age 15-64 years. Parameter estimation is carried out under Bayesian inference using Markov chain Monte Carlo methodology.

The weight adjustment was seen to reduce sampling variation and control for exaggeration of effects, this lead to quality inference. HIV prevalence for counties in Kenya is estimated for males and females and high prevalence counties identified. In this study, we have shown how to include survey weights into spatial modeling, so as to account for the unequal selection probabilities associated with complex surveys. The weighted analysis produces estimates with low variability as compared to unweighted analysis.

3.1 Introduction

Human Immunodeficiency Virus (HIV) still remains a leading public health problem in sub-Saharan Africa and many parts of the world. Governments in collaboration with private stakeholders have been tirelessly putting up new programs to prevent new HIV infections and to improve the quality of life of people affected and infected by HIV. Many advances like the introduction of free antiretroviral drugs, distribution of free condoms, free voluntary testing and counselling centres, and public awareness programmes are aimed at curbing this menace. National surveys

are important for assessing whether government and private sector led interventions and campaigns have an impact on the prevalence of the infection. There is need to monitor disease burden within administrative regions of a country.

Disease infection variation within a country's geographical regions is important in determining where more resources on prevention and treatment need to be focused [Waller and Gotway, 2004; Lawson et al., 1999]. Visualisation of disease distribution is an important procedure in understanding disease occurrence [Everitt and Dunn, 2011]. The use of maps in the context of disease distribution has developed rapidly in the public health sector [Lawson et al., 2003; Cliff, 1995; Kazembe et al., 2006; Manda et al., 2009]. Disease maps inform people about the geographical variation in disease burden [Wakefield, 2007].

Most countries carry out Demographic and Health Surveys (DHS) in order to understand people's comprehension of certain health issues and also determine prevalence and awareness about several diseases. The survey data are usually collected with an aim of being representative of the whole population. Complex survey designs are usually used to ensure proper representation of the population. In complex surveys, the individuals included in the sample usually have a sampling weight attached to them so as to downscale or upscale their representativeness of the population. This is used to acknowledge the fact that not all individuals in the population had a chance to be selected into this sample. If all the individuals had the same and equal chance of being selected into the population then this will imply a simple random sampling scheme hence each individual will have a sampling weight equal to one. This means that in complex survey designs individuals included in the study had unequal probabilities of selection, as opposed to simple

random sampling scheme.

Previous research [Ngigi, 2007; Montana et al., 2007], of geographical analysis of HIV has been done using standardised morbidity ratios, and ignoring the fact that the datasets used were collected using complex survey design where individuals included in the study had unequal probabilities of selection.

Chen et al. [2012], in their recent work on inclusion of sampling weights into analysis, for binary data, incorporated sampling weights by matching variances in a stratified random sampling survey. In this study, we incorporate the survey weights into the model using effective sample sizes through design effects [Kish, 1995]. Design effects measure the loss (or gain) in effectiveness by using a complex sampling scheme as opposed to simple random sampling. We then use a generalized linear mixed model [McCulloch and Neuhaus, 2005] with spatially unstructured random effects and spatially structured random effects being modelled by normal priors and intrinsic conditional autoregressive (ICAR) priors respectively [Besag et al., 1991]. The methods are applied to HIV data extracted from the Kenya aids indicator survey dataset of 2007, collected by the Government of Kenya. In this work, we also investigate if there is any difference in inference when we include the sampling weights in the model.

3.2 Methods

3.2.1 Data

The data for this study was extracted from the 2007 Kenya Aids Indicator Survey (KAIS), conducted by the Government of Kenya. The main objective of survey was to collect high quality data on the prevalence of HIV and sexually transmitted infections (STI) among adults, and to assess knowledge of HIV and STI in the populations. The survey collected a representative sample of households selected from the eight provinces in the country. It involved men and women in the age of 15-64 years. The primary sampling unit for the survey was a cluster, a collection of one or more enumeration areas, with an average of 100 households in each cluster. In all, 402 clusters were surveyed. Two questionnaires were used in the survey. The first one is a household questionnaire which collected information about the household head and the characteristics of the dwelling place. The second one, the individual questionnaire, collected information from men and women aged 15-64 years, about their demographic characteristics, and their knowledge on HIV and STI. All women and men aged 15-64 years in selected households who were either usual residents or visitors present the night before the survey were eligible to participate in the individual interview and blood draw, provided they gave informed consent. For minors aged 15-17 years, parental consent and minor assent were both required for participation. Participants could consent to the interview and blood draw or to the interview alone. The inclusion criteria may have captured non-Kenyans living as usual residents or visitors in a sampled

household. Design weights were used for households, individual interviews, and blood draws. The purpose of weighting was to correct for unequal probability of selection and to adjust for non-response to produce results that were representative of the larger population from which the sample was drawn. Base weights were adjusted for cluster non-response, household non-response, and individual non-response (both for the interview and the blood draw). Readers are referred to the final survey report for more details regarding survey methodologies used in the study [NASCOP, 2008]. Each individual was then asked for consent to provide a venous blood sample for HIV, HSV-2, syphilis testing and CD4 cell count. In total 17,940 individuals completed the individual questionnaires while 15,867 provided venous blood for testing.

3.2.2 Selection of variables

An initial univariate and multivariate exploratory data analysis carried out showed that the following variables were significantly associated with HIV infection among men: age at first sex, perceived risk of HIV, number of partners in the previous year, condom use, circumcision status, residential area, age, access to media and whether the person had STI. In a similar manner, for the women case, the following variables were identified to be associated with HIV infection: education level, age at first sex, perceived risk of HIV, number of partners in the previous year, residential area, age, frequency of away travels, marital status and whether the person had STI. The significant covariates identified were generally in agreement with existing literature [Montana et al., 2007, 2005; Cheluget et al., 2006; Johnson

and Way, 2006; Bailey et al., 2007; Weiss et al., 2000]. The identified individual level covariates were used to compute county covariates based on proportions.

3.2.3 Statistical models and analysis

This section introduces the models and their specifications. First the data are modeled without incorporating the survey weights and then followed by modeling with the survey weights adjustment. In the next subsection we fully describe the method used to incorporate the survey weights, followed by the hierarchical models considered.

3.2.3.1 Incorporating Sampling weights

Let y_{ij} be a binary response for the HIV status of individual j in area i ($i = 1, 2, \cdots, m \quad and \quad j = 1, 2, \cdots N_i$). Our initial interest is to estimate the small region specific proportion parameter p_i given by:

$$p_i = \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i}. \tag{3.1}$$

p_i can be estimated using the direct survey-weighted estimate given by

$$\hat{p}_i = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}}, \tag{3.2}$$

where n_i is the sample size of the small domain i and w_{ij} is the sampling weight associated with individual j in area i. In order to calculate the "effective sample

size", n_i^* for the i^{th} area, we first estimate the corresponding design effect, $deff_i$, see Kish [1995]. Design effect measures the loss or gain of effectiveness by using a complex sampling design instead of simple random sampling. Using this and following from You and Zhou [2011], $deff_i$ is estimated as

$$deff_i = \frac{s_i^2}{s_{ri}^2}, for \quad i = 1, 2, \cdots, m \tag{3.3}$$

where s_i^2 is the unbiased direct estimate of the variance of proportion based on the complex sampling design and s_{ri}^2 is the unbiased direct estimate of the variance of the proportion based on the simple random sampling design.

Using the estimated design effect, we calculate the effective sample size for each small area as follows;

$$n_i^* = \frac{n_i}{deff_i} \tag{3.4}$$

To find the effective number of HIV cases in each region, y_i^*, we multiplied the effective sample size and the area proportion $\hat{p}_i$ as calculated from equations (3.4) and (3.2) respectively as;

$$y_i^* = n_i^* \times \hat{p}_i \tag{3.5}$$

3.2.3.2 Hierarchical Model

The conditional distribution of y_i^* given p_i for this binary case is defined as

$$y_i^* | p_i \sim Bin\left(n_i^*, p_i\right), i = 1, 2, \cdots m. \tag{3.6}$$

where n_i^* and y_i^* are as defined in equations (3.4) and (3.5) respectively. Having described the distributional properties of the response variable y_i^*, we now present the statistical models that will be used. We consider random effects models, both spatially structured and spatially unstructured. The spatially unstructured random effect was modeled as using a normal prior, $v_i \sim N(0, \sigma_v^2)$. For the spatially structured random effect, u_i, we used the intrinsic conditional autoregressive(ICAR) prior, specified by Besag et al. [1991] as follows

$$u_i|\boldsymbol{u}_{-i}, j \neq i \sim N\left(\frac{\sum\limits_{j\in N_{(i)}} u_j}{d_i}, \frac{\sigma_u^2}{d_i}\right) \tag{3.7}$$

where $N_{(i)}$ denotes the set of neighbours of area i, d_i is the corresponding number of neighbours and $\boldsymbol{u}_{-i} = (u_1, \cdots, u_{i-1}, u_{i+1}, \cdots, u_m)$, is the vector of the random effects without the i^{th} component. Two regions are defined to be neighbours if they share a border. The variance component σ_u^2 accounts for spatial variation between the regions and is also used to capture the amount of variation explained by the spatial structure.

In this work, different models in increasing order of complexity arising from different assumptions on random effects were fitted. The models are (we present the systematic part of the models, the first stage for each model is the likelihood distribution given in (3.6)):

- Model 1: $logit(p_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$: Generalized linear model.
- Model 2: $logit(p_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + v_i$: Normal unstructured heterogeneity (UH) random effects model.

- Model 3: $logit(p_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i$: ICAR spatial random effects model.

- Model 4: $logit(p_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i + v_i$: Convolution model.

The modeling was done separately for men and women due to the difference in HIV prevalence between the two genders. In Kenya, HIV prevalence among women is 8.4% and 5.4% among men. Model estimation was carried out using a Bayesian approach [Gelman et al., 2003]. All parameters in the models were assigned prior distributions. In this analysis, non-informative normal priors were assigned to the fixed effect coefficients, β and the variance parameters were assigned non-informative inverse gamma distributions. The models were implemented using WinBUGS version 1.4, [Spiegelhalter et al., 2007; Ntzoufras, 2011]. For each model, 200,000 Markov chain Monte Carlo (McMC) iterations, [Mollie et al., 1996], were ran for each model, with the initial 20,000 discarded to cater for the burn-in period and there after keeping every tenth sample value. The 18,000 iterations left were used for assessing convergence of the McMC and parameter estimation. We assessed McMC convergence of all models parameters by checking trace plots and autocorrelation plots of the McMC output [Gelman et al., 2003]. The models were compared using the Deviance Information Criterion (DIC) [Spiegelhalter et al., 2002]. The best fitting model is one with the smallest DIC value. In cases where the difference in DIC between the models is not above 5, we selected the best model based on "simplicity" of the model, that is, model with both few parameters and random effects.

3.3 Results

In this analysis, laboratory HIV test results of a sample of 9,049 women and 6,818 men with full covariate information was used.

3.3.1 Effect of weight adjustment on parameter coefficients

In the men's data, the model with a spatially unstructured random effect only, was found to be the simplest model with a relatively lower DIC compared to the other models, see Table 3.1 and Table 3.2. Based on this best fitting model, the final conclusion on the effects of the covariates, that is on whether the variable is significant or not, was the same in both weight adjusted and weight unadjusted models. The coefficient estimates are also very close in both weight adjusted and weight unadjusted analyses. Proportion perceiving not to be at risk of HIV and proportion of men in the county who are not circumcised were found to be significantly associated with HIV infection.

TABLE 3.1: Men's results from weight unadjusted models: parameter estimates and corresponding 95% credible interval

County covariate	model 1	model 2	model 3	model 4
Constant	0.27(-3.39,3.70)	-0.86(-5.47,2.98)	0.49(-4.67,5.57)	-0.18(-4.35,5.54)
Prop. sex debut at 15-17	1.53(-0.51,3.56)	2.09(-0.51,4.84)	0.86(-1.93,3.49)	1.21(-1.47,3.68)
Prop. perceiving not at risk	-7.60(-12.7,-2.54)	-7.48(-14.4,-0.78)	-7.01(-14.31,0.52)	-7.01(-14.76,0.18)
Prop. who had one partner	-3.35(-5.29,-1.17)	-2.54(-5.20,0.50)	-3.46(-6.69,0.31)	-3.08(-6.51,-0.6)
Prop. not using condom	0.35(-1.54,2.29)	0.39(-2.45,3.06)	0.41(-2.06,3.03)	0.51(-2.27,3.12)
Prop. not circumcised	1.49(0.96,2.03)	1.77(0.96,2.65)	1.26(0.43,2.11)	1.44(0.64,2.26)
Prop. in rural area	-0.58(-1.11,-0.05)	-0.83(-1.85,0.13)	-1.00(-2.18,-0.02)	-0.97(-2.14,0.07)
Prop. without media access	-1.30(-2.68,-0.01)	-1.44(-3.22,0.32)	-1.47(-3.28,0.23)	-1.53(-3.38,0.31)
Prop. with STI	-4.23(-11.23,1.94)	-3.60(-12.2,4.94)	-5.53(-14.6,2.55)	-4.81(-13.7,3.09)
Average Age	-0.00(-0.06,0.055)	0.00(-0.07,0.08)	0.00(-0.10,0.09)	0.01(-0.10,0.10)
σ_v	-	0.18(0.00,0.50)	-	0.06(0.00,0.38)
σ_u	-	-	0.49(0.00,1.51)	0.30(0.00,1.12)
DIC	235.00	220.19	221.08	220.94

TABLE 3.2: Men's results from weight adjusted Models: parameter estimates and corresponding 95% credible interval

County covariate	model 1	model 2	model 3	model 4
Constant	-0.26(-3.12,3.19)	-0.00(-5.46,5.79)	-0.40(-5.39,4.48)	-0.36(-5.62,4.93)
Prop. sex debut at 15-17	0.41(-0.93,1.73)	1.33(-1.09,3.94)	0.28(-2.16,2.74)	0.86(-1.70,3.36)
Prop. perceiving not at risk	-8.71(-13.99,-3.7)	-7.24(-14.35,-0.12)	-5.67(-13.59,1.83)	-6.04(-14.45,2.28)
Prop. who had one partner	-2.82(-4.81,-1.0)	-2.53(-6.74,0.81)	-2.73(-5.96,0.52)	-2.26(-5.49,1.14)
Prop. not using condom	1.81(0.31,3.36)	0.69(-2.76,3.43)	1.58(-0.64,3.78)	1.22(-1.70,3.87)
Prop. not circumcised	1.48(0.98,1.96)	1.64(0.69,2.60)	1.25(0.37,2.11)	1.47(0.511,2.39)
Prop. in rural area	-0.757(-1.28,-0.23)	-0.83(-1.90,0.36)	-1.09(-2.34,0.04)	-0.97(-2.1,0.18)
Prop. without media access	-2.09(-2.90,-1.30)	-1.49(-3.16,0.28)	-1.86(-3.55,-0.30)	-1.66(-3.33,0.03)
Prop. with STI	-1.52(-6.60,3.256)	-2.34(-11.5,6.15)	-3.63(-11.50,3.55)	-2.62(-10.90,5.13)
Average Age	-0.00(-0.05,0.04)	-0.01(-0.12,0.08)	0.01(-0.08,0.09)	-0.01(-0.11,0.09)
σ_v	-	0.25(0.07,0.56)	-	0.13(0.00,0.44)
σ_u	-	-	0.66(0.18,1.58)	0.34(0.00,1.49)
DIC	261.67	226.74	227.64	227.13

In the women's data, the model with a spatially structured random effect only, was found to be the simple model with relatively lower DIC compared to the other models, for both weight adjusted models and weight unadjusted models, see Table 3.3 and Table 3.4. Based on this best fitting model, the final conclusion on whether the covariates were significant or not was the same in both weight adjusted and weight unadjusted models. Age at sex debut, proportion of people in the county who had one partner and the proportion of people in the county living in the rural area were found to be significantly associated with HIV infection.

TABLE 3.3: Women's results from weight unadjusted models: parameter estimates and corresponding 95% credible interval

County covariate	model 1	model 2	model 3	model 4
Constant	11.77(5.69,15.49)	14.07(9.33,19.7)	7.92(-0.11,15.96)	1.70(-2.64,12.89)
Prop. with no education	1.15(0.56,1.77)	0.03(-1.4,1.39)	1.08(-0.36,2.44)	1.15(-0.45,2.58)
Prop. sex debut at 15-17	4.32(3.15,5.49)	4.00(1.11,7.09)	3.44(1.39,5.78)	3.64(1.21,6.05)
Prop. perceiving not at risk	-1.12(-2.14,-0.19)	-0.87(-3.13,1.34)	-0.76(-3.03,1.58)	-0.96(-3.31,1.39)
Prop. who had one partner	-12.55(-16.05,-6.73)	-13.56(-17.05,-9.80)	-10.37(-17.14,-2.07)	-4.60(-14.21,1.59)
Prop. in rural area	-0.68(-1.16,-0.20)	-0.49(-1.54,0.61)	-1.12(-2.08,-0.10)	-1.31(-2.30,-0.27)
Prop. which did not stay away	-1.25(-2.01,-0.49)	-0.61(-2.35,1.27)	0.17(-1.38,1.86)	0.11(-1.61,2.05)
Prop. married with 1 partner	-1.54(-3.42,0.22)	-2.11(-6,1.88)	-1.32(-4.12,1.53)	-1.30(-4.55,1.74)
Prop. with STI	-7.53(-14.51,-0.84)	-6.37(-18.73,6.50)	-4.79(-17.34,7.88)	-3.67(-16.41,9.02)
Average Age	-0.05(-0.11,-0.00)	-0.08(-0.19,0.01)	-0.01(-0.10,0.06)	0.00(-0.11,0.10)
σ_v	-	0.28(0.11,0.59)	-	0.02(0.00,0.13)
σ_u	-	-	0.69(0.27,1.42)	0.69(0.21,1.45)
DIC	314.45	259.88	253.18	253.87

In both men's and women's data, even though the parameter estimates of the fixed

TABLE 3.4: Women's results from weight adjusted models: parameter estimates and corresponding 95% credible interval

County covariate	model 1	model 2	model 3	model 4
Constant	14.32(9.26,21.05)	24.9(20,31.51)	-2.61(-7.56,4.09)	-2.32(-7.33,2.61)
Prop. with no education	1.01(0.48,1.55)	-0.35(-1.77,1.00)	0.86(-0.74,2.82)	0.85(-0.96,2.47)
Prop. sex debut at 15-17	5.23(3.95,6.46)	3.19(0.27,6.21)	3.40(0.58,6.23)	3.63(0.38,6.53)
Prop. perceiving not at risk	-0.98(-2.17,0.10)	-0.80(-3.23,1.51)	-0.63(-3.41,2.10)	-0.75(-3.53,2.13)
Prop. who had one partner	-15.64(-22.35,-10.16)	-23.13(-29.67,-16.67)	-1.85(-3.49,-0.20)	0.03(-4.06,6.08)
Prop. in rural area	-1.22(-1.60,-0.85)	-0.12(-1.26,0.93)	-1.11(-1.97,-0.26)	-1.23(-2.44,-0.0)
Prop. which did not stay away	-0.81(-1.65,0.04)	-0.31(-2.41,1.59)	0.57(-1.50,2.82)	0.15(-1.99,2.64)
Prop. married with 1 partner	-1.29(-3.02,0.27)	-3.10(-6.94,0.49)	-2.51(-7.12,1.26)	-3.15(-6.79,0.37)
Prop. with STI	-10.1(-15.96,-4.43)	-7.67(-20.08,5.71)	-2.71(-16.85,12.22)	-3.45(-17.5,11.17)
Average Age	-0.04(-0.08,0.00)	-0.10(-0.19,-0.02)	0.05(-0.05,0.13)	0.01(-0.06,0.10)
σ_v	-	0.31(0.14,0.61)	-	0.04(0.00,0.25)
σ_u	-	-	0.97(0.42,1.91)	0.86(0.16,1.86)
DIC	328.31	259.84	252.71	252.42

effects coefficients were of the same order of magnitude for both weight adjusted and weight unadjusted models, the confidence interval for the weight adjusted models were generally narrower. In cases where the parameter estimates were a little bit different between the weight adjusted and weight unadjusted models, the weight adjusted model had slightly lower values.

Figure 3.1 and Figure 3.2 shows box plots of HIV prevalence in the 46 counties for men and women respectively, based on the best fitting weight adjusted models. The horizontal line gives the countrywide HIV prevalence for each gender, standing at 7.88% for women and 5.60% for men. Table

Table 3.5 gives the county HIV prevalence estimates based on the best fitting weight adjusted models for men and women.

TABLE 3.5: County HIV prevalence estimates and corresponding 95% credible interval based on best fitting weight adjusted models

		Men	Women
ID	County	Prevalence (95% CI)	Prevalence (95% CI)
1	Baringo	0.02(0.01,0.05)	0.02(0.01,0.05)
2	Bomet	0.11(0.08,0.13)	0.08(0.04,0.14)

3	Bungoma	0.02(0.00,0.04)	0.04(0.02,0.08)
4	Busia	0.05(0.02,0.07)	0.07(0.05,0.10)
5	Elgeyo Marakwet	0.05(0.02,0.09)	0.05(0.02,0.12)
6	Embu	0.02(0.01,0.04)	0.06(0.03,0.11)
7	Garissa	0.04(0.02,0.07)	0.01(0.01,0.02)
8	Homa Bay	0.26(0.19,0.34)	0.31(0.26,0.36)
9	Isiolo	0.03(0.00,0.09)	0.07(0.02,0.16)
10	Kajiado	0.03(0.01,0.08)	0.05(0.01,0.14)
11	Kakamega	0.05(0.04,0.08)	0.07(0.05,0.11)
12	Kericho	0.05(0.03,0.09)	0.12(0.07,0.18)
13	Kiambu	0.04(0.02,0.07)	0.04(0.02,0.07)
14	Kilifi	0.06(0.04,0.08)	0.08(0.05,0.12)
15	Kirinyaga	0.04(0.02,0.06)	0.06(0.03,0.09)
16	Kisii	0.04(0.02,0.07)	0.07(0.05,0.09)
17	Kisumu	0.13(0.08,0.20)	0.20(0.14,0.27)
18	Kitui	0.05(0.02,0.08)	0.07(0.04,0.12)
19	Kwale	0.04(0.02,0.06)	0.06(0.04,0.09)
20	Laikipia	0.04(0.02,0.08)	0.03(0.01,0.06)
21	Lamu	0.04(0.01,0.10)	0.01(0.00,0.04)
22	Machakos	0.03(0.02,0.06)	0.05(0.03,0.09)
23	Makueni	0.04(0.02,0.07)	0.09(0.06,0.11)
24	Mandera	0.01(0.00,0.04)	0.02(0.01,0.03)
25	Marsabit	0.01(0.00,0.03)	0.01(0.00,0.03)
26	Meru	0.02(0.01,0.04)	0.06(0.03,0.10)
27	Migori	0.17(0.12,0.23)	0.21(0.15,0.28)
28	Mombasa	0.08(0.05,0.12)	0.16(0.13,0.18)
29	Muranga	0.03(0.01,0.05)	0.04(0.03,0.06)
30	Nairobi	0.07(0.05,0.10)	0.11(0.08,0.14)
31	Nakuru	0.05(0.03,0.09)	0.11(0.06,0.17)
32	Nandi	0.08(0.04,0.15)	0.09(0.04,0.15)
33	Narok	0.04(0.01,0.09)	0.07(0.03,0.12)
34	Nyamira	0.05(0.02,0.09)	0.07(0.04,0.11)
35	Nyandarua	0.04(0.01,0.07)	0.04(0.02,0.08)
36	Nyeri	0.04(0.02,0.07)	0.03(0.02,0.05)
37	Siaya	0.14(0.09,0.20)	0.20(0.16,0.25)
38	Taita Taveta	0.06(0.04,0.09)	0.09(0.06,0.12)
39	Tana River	0.03(0.01,0.08)	0.03(0.01,0.05)
40	Tharaka-Nithi	0.04(0.02,0.07)	0.07(0.04,0.11)
41	Trans Nzoia	0.08(0.05,0.10)	0.09(0.04,0.18)
42	Turkana	0.10(0.06,0.15)	0.16(0.07,0.28)
43	Uasin Gishu	0.04(0.02,0.07)	0.07(0.05,0.09)
44	Vihiga	0.05(0.02,0.09)	0.06(0.03,0.09)
45	Wajir	0.01(0.00,0.03)	0.01(0.00,0.02)
46	West Pokot	0.03(0.01,0.06)	0.07(0.02,0.14)

3.3.2 Variation of HIV infection in Kenya

The best fitting weight adjusted models were used to produce smoothed maps to show the variation of HIV in Kenya, for men and women. Weight adjusted model 2 was used to produce the women's map while weight adjusted model 3, was used to produce the men's map.

From the choropleth map in Figure 3.3, the prevalence of HIV with respect to men among counties shows huge geographical discrepancy. There is high prevalence in the counties situated on the western side of the country, around Lake Victoria region and also in the Turkana county in the North West part of the country.

In the women's analysis, from the choropleth map in Figure 3.4, HIV prevalence was found to be higher around the Lake Victoria region counties, on the West part of the country. Turkana County, on the North West corner of the country is also found to have high prevalence of HIV, as well as the coastal counties situated in the South Eastern part of the country.

3.4 Discussion

In this study, we set out to develop a disease prevalence modeling and mapping approach that allows for acknowledgment of the complex nature of the method used to collect the survey data. Sampling weights are used to control two major forms of bias in survey sampling, namely non-response bias and non-coverage bias.

Introducing sampling weights during modelling in a generalized linear mixed modelling framework has been a challenge [Gelman, 2007; Heeringa et al., 2010]. Several attempts have been made on this topic in the recent past [DuMouchel and Duncan, 1983; Pfeffermann, 1993] but these procedures are not very flexible, and this is the main reason why the modeling approach is more popular for problems such as small area estimation [Fay and Herriot, 1979; Rao, 2005]. Another problem with weight adjusted estimates is that standard errors are hard to calculate. In as much as fitting regression models is simple for ordinary data, and of late fitting hierarchical models has followed suite, due to new state of the art softwares like WinBUGS, STATA, R, SAS etc, applying these models to survey data with acknowledgment of survey weights is still a challenge.

In this work, we model the prevalence of HIV based on data collected from a complex survey. We introduced the weights into the model using design effects, which is a design based approach. Our main interest is to compare the variation in inference when we model the data with these survey weights and when we ignore the survey weights (assuming simple random sampling was used in collecting the data), applied to Kenya HIV prevalence data.

To capture both regional variation and region specific characteristics, both spatially structured and spatially unstructured random effects were introduced in the models. The spatially structured random effects were modelled using a Markov random field while the spatially unstructured random effects were modelled using a zero mean Gaussian prior. Bayesian inference was used in the estimation due to the complex nature of the data and models.

Weight adjusted estimates were found to have low variability, their confidence intervals were narrow and were generally lower in value than the weight unadjusted analyses. In other words weight adjustment was able to control for exaggeration of effects and this led to quality inference. This was in line with Chen et al. [2012] findings about sampling weight adjustment analysis, carried out for small area estimation.

In terms of whether the county covariates of interest in the model were significant or not, both weight adjusted and weight unadjusted analyses gave the same results. The order in magnitude of the coefficients was also the same for both weight adjusted and weight unadjusted models. This result seems to support an argument by Fienberg [2009], which highlights that including survey weights in statistical models is a futile venture in Bayesian analysis.

Disease prevalence maps provide a quick visualisation of the geographic variation of disease burden in a country. For a given county, this is critical especially in the dispensation of the new constitution with counties being the new administrative units.

There was significant variation of HIV prevalence among the counties, which to the best of our knowledge could be explained by the socio-economic and cultural practices of communities residing in the counties. As noted in other studies, [Ngigi, 2007; Montana et al., 2007; Cheluget et al., 2006; Oluoch et al., 2011], the prevalence of HIV for both men and women was found to be mainly higher in the Lake Victoria region counties, that is the Western part of the country.

One limitation in this study is the restricted number of variables available and their nature from this national survey data. For example, condom use covariate was not well captured since the survey only asks if someone used a condom with the last partner, it does not capture the consistency in condom use.

3.5 Conclusions

We have shown a statistical method of incorporating sampling weights and the complex structure of survey data in disease mapping analyses. This method can be applied to any survey data provided the information of every individual's residence region is available in the data. Calculations of the effective sample sizes and effective observations for the regions using design effects is straight forward and can be applied with ease.

The method produces maps which can show geographical variation of the disease burden at a glance, based on survey data. These maps are important to policy makers in both government and private sectors. Policy makers can use these maps in formulating policies and programs suited for each county. In particular such easy to use tools can be very useful in optimal allocation of resources aimed at controlling the disease. The separation of men and women analyses can further help health policy makers to come up with gender specific policies.

3.6 Acknowledgements

The research is supported by the UKZN's College of Agriculture, Engineering and Science postgraduate bursary. The authors wish to thank the Kenya National Bureau of Statistics for providing the KAIS (2007) dataset used in this work.

FIGURE 3.1: County HIV prevalence box-plot for men using weight adjusted data and best fitting model 2.

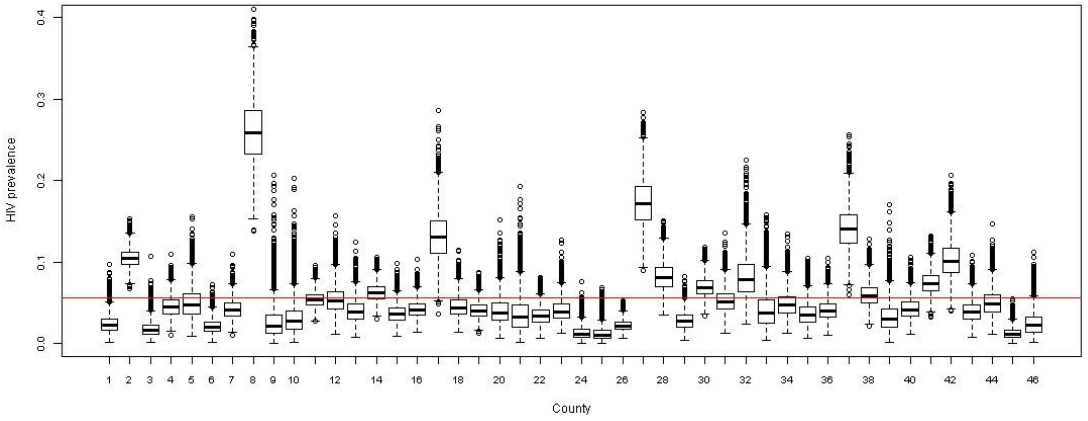


FIGURE 3.2: County HIV prevalence box-plot for women using weight adjusted data and best fitting model 3

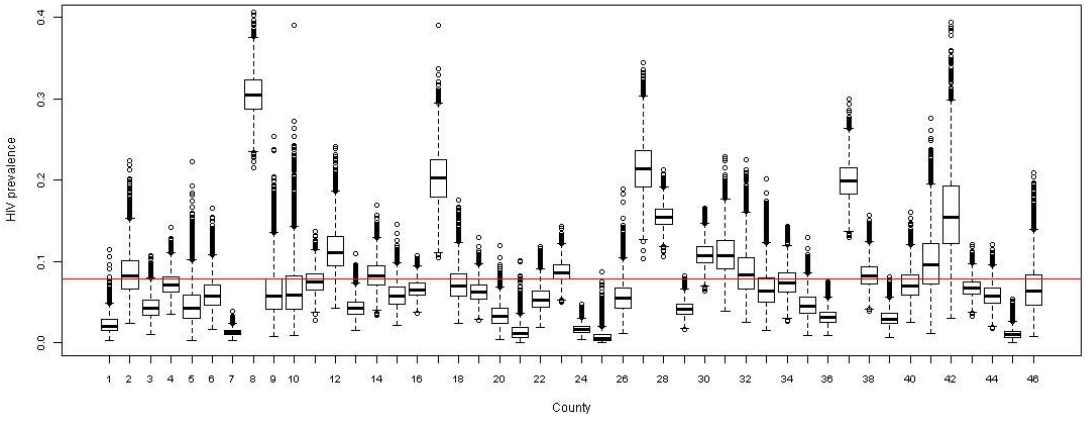


FIGURE 3.3: HIV prevalence map for men based on weight adjusted data and model 2

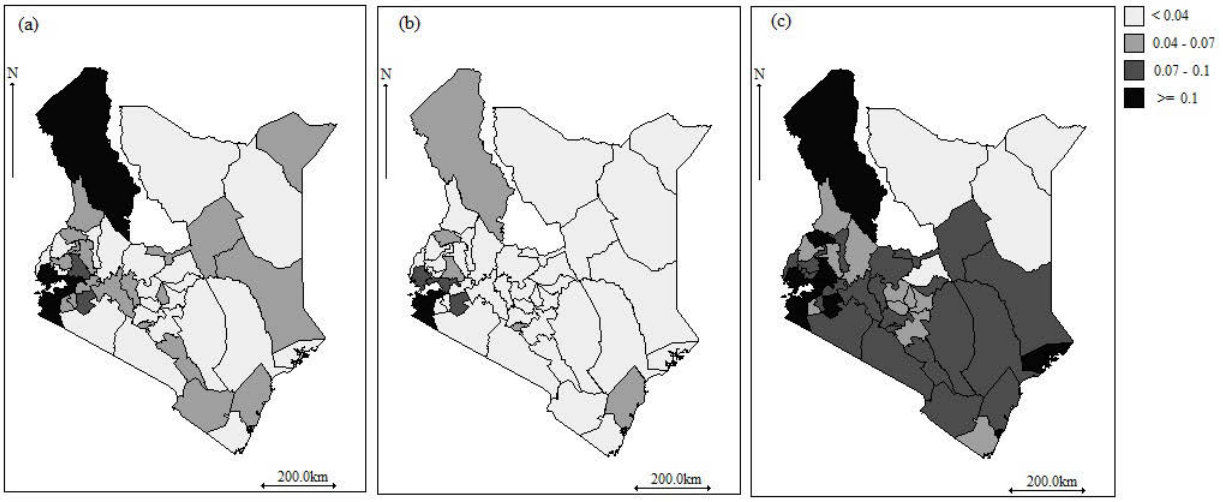
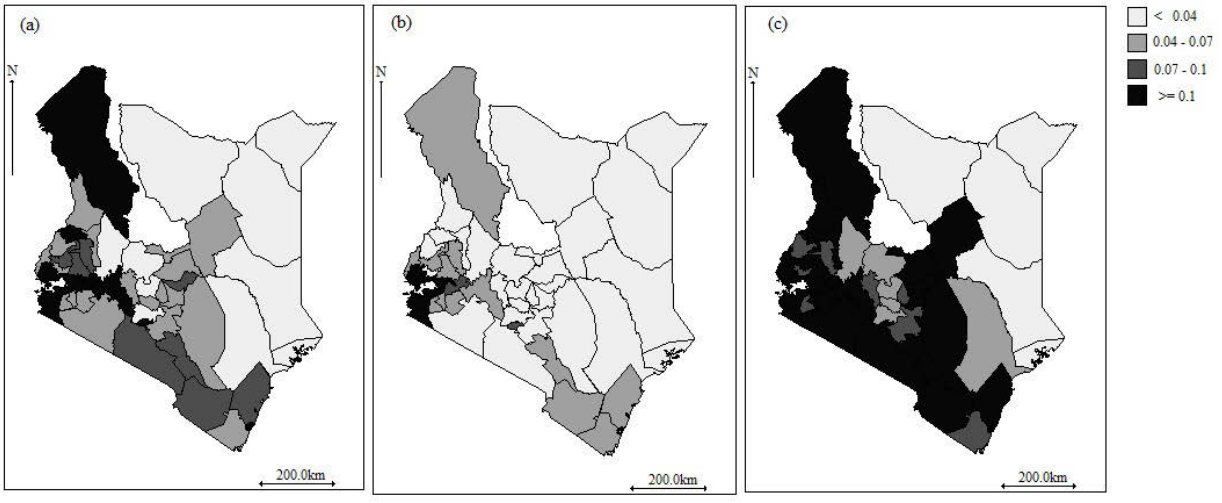


FIGURE 3.4: HIV prevalence map for women based on weight adjusted data and model 3



Chapter 4

A Double Random Effects Model for Spatially Correlated and Overdispersed Binary Data with Application to HSV-2 Variation in Kenya

Herpes Simplex Virus-Type 2 (HSV-2) still remains a neglected disease in many parts of the world, especially in Africa. This is evident by the low publicity and few testing centers that are accorded to the disease. Overdispersion is a common phenomenon in count data. Several models have been developed to accommodate for overdispersion in modeling count data. The models developed usually either ignore spatial correlation or allows for spatial correlation but assumes that the

overdispersion parameter is not influenced spatially. In this work, a spatial model for overdispersed binary data which allows for the overdispersion parameter to vary over the region under study is developed. This model was compared to existing models and was found to provide a better fit to the data at hand. The proposed model was used to develop smoothed HSV-2 prevalence maps for Kenya. These maps can be very informative to policy makers especially if there are limited resources available in setting up testing centers and a decision has to be made on where the centers are to be located across the country.

4.1 Introduction

Herpes simplex virus type 2 (HSV-2) is the most frequent cause of genital herpes. It is one of the most prevalent sexually transmitted infections (STIs) worldwide [Cusini and Ghislanzoni, 2001]. HSV-2 causes significant neurological morbidity in Africa that has plagued human health.

The disease is focal in transmission with the primary infection affecting skin or mucosal surfaces, consequently the ganglia becomes susceptible to be invaded by viral deoxyribonucleic acid (DNA), being transported to these site with retrogade axonal transport . During these stage viral genome undergoes circulization with sensory neurons being the primary target hence a state of latency being observed [Aumakhan et al., 2010]. Antigenic variation mechanism has also been associated with the disease with strong mounting of T-cell and humoral immunity. Stress triggers up to two fold in terms of transcription and replication of the virus with

the overall clinical inflammatory complications of mucosal lesions and skin being shed [Horbul et al., 2011].

HSV-2 mainly affects the genital area and is primarily transmitted by sexual contact with an infected person though cross-infection may be possible from orogenital sex. Even though HSV-2 might not be generally dangerous, it is a nuisance, can be excruciating and may cause emotionally trauma to the infected person. As per now, there is no cure for HSV-2; once you have the virus in your body system, you can only be put on a treatment therapy to suppress it but it will still remain in the body [Beauman, 2005]. Many individuals have none or mild symptoms of the herpes virus and are therefore unaware of the infection. Persons infected with HSV-2 do not necessarily develop a clinical disease, but most, from time to time produce virus from the genital tract [Lafferty et al., 1987]. Several studies have indicated that genital herpes is associated with high risk of HIV transmission and acquisition [Celum et al., 2008; Watson-Jones et al., 2008]. Ngesa et al. [2013] carried out a joint analysis to investigate the spatial correlation between HIV and HSV-2 among men at county level in Kenya; a significant positive correlation was established between these two infections. In Africa, the diseases is not well reported/studied and very few testing centres exist for this disease.

Spatial analysis of HSV-2 prevalence can give insight for understanding the geographical distribution of the disease and create strategies for setting up testing programmes in areas with high prevalence of this disease. In modelling binary data, especially disease presence or absence in individuals within a region, binomial distribution is a common assumption on the data generating mechanism.

The main objective of the study is to develop an appropriate statistical model to assess the prevalence of HSV-2 by county in Kenya, allowing for spatial variations in the regions while accounting for overdispersion in the data. Overdispersion results when the data appear more dispersed than is expected under some reference model. It occurs in data distributions in which the observed counts have variances that are functions which depend on the value of the mean. That is, the variance of Y depends on the expected value of Y, which is estimated from the data. In binomial and Poisson distributions, theoretically, the dispersion parameter is usually assumed to have a given a value of 1, any departure from this value leads to overdispersion (> 1) or underdispersion (< 1).

Several over dispersion modelling strategies exist for binomial data. Two major approaches for handling this is by assuming a distribution for the probability of success or using random effects. In beta binomial models, the probability of success is assumed to follow a beta distribution while the data generating mechanism follows a binomial distribution. Hinde and Demétrio [1998] provide a unified approach for modelling over-dispersion and clustering. Molenberghs et al. [2007] and Molenberghs et al. [2010] introduced the combined modeling framework for handling over-dispersion and repeated measures. Molenberghs et al. [2012] and Kassahun et al. [2012] focused on combined modelling framework for binomial data. In Kassahun et al. [2012] and Molenberghs et al. [2012], non-spatial binomial data was considered and the non-spatial random effects were introduced via the mean function.

In this study, we develop a model which allows for the dispersion parameter in

the binomial model to be modelled separately and allowed also to vary spatially among the regions under study. None of the above studies have allowed for this in their analysis. This new model is compared with existing over-dispersion modeling strategies using the deviance information criterion (DIC) [Spiegelhalter et al., 2002]. The application of the model yields new and important insights in geographical variation of HSV-2 in Kenya. We employ a fully Bayesian approach in the estimation of the parameters. The models are implemented in WinBUGS software [Spiegelhalter et al., 2007]. The programs used for the analyses are available upon request from the first author.

4.2 Data

The data for this study was extracted from the 2007 Kenya Aids Indicator Survey (KAIS), conducted by the Government of Kenya. The main objective of survey was to collect high quality data on the prevalence of HIV and sexually transmitted infections (STI) among adults, and to assess knowledge of HIV and STI in the populations. The survey collected a representative sample of households selected from the eight provinces in the country. It involved all men and women in the age of 15-64 years. Two questionnaires were used in the survey. The first one is a household questionnaire which collected information about the household head and the characteristics of the dwelling place. The second one, the individual questionnaire, collected information from men and women aged 15-64 years, about their demographic characteristics, and their knowledge on HIV and STI. Each individual was then asked for consent to provide a venous blood sample for HIV,

HSV-2, syphilis testing and CD4 cell count. In total, 6,606 men, aged 15-64 years who provided venous blood for HSV-2 testing and also had full covariate information was extracted from the KAIS data and were used in the analysis. The following covariates were used in the analysis: education level, circumcision status, place of residence (urban/rural) and age of respondent.

4.3 Disease mapping models

In this section, we review some of the frequently encountered models for disease mapping when the data is binary and their overdispersion counterparts. We review the binomial model as an exponential family member, beta-binomial model, logistic-normal model and the convolution model. The proposed model is also outlined in this section.

4.3.1 Binomial model and exponential family

The basic model for modeling binary count data is the binomial model. The binomial distribution has the structure;

$$f(y|p) = \binom{n}{x} p^y (1-p)^{n-y}, y = 0, 1, \cdots, n \tag{4.1}$$

The binomial distribution is a member of the exponential family [McCulloch and Neuhaus, 2005]. The exponential family distributions take the form,

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}, \tag{4.2}$$

where θ and ϕ are called the natural parameter and the scale parameter respectively. For the binomial data, this can be shown as

$$f(y|p) = \binom{n}{x} p^y (1-p)^{n-y} \tag{4.3}$$

$$= \exp\left[\frac{yln\left(\frac{p}{1-p}\right) + nln\,(1-p)}{1} + ln\left(\binom{n}{y}\right)\right] \tag{4.4}$$

Comparing this with the general exponential family structure, the dispersion parameter, $\phi = 1$. This model is not adequate for data which has overdispersion.

4.3.2 Beta-binomial

A beta-binomial model can be used to model binary count data in cases where there is overdispersion or underdispersion, i.e $\phi \neq 1$. The model is formulated as follows and has two levels of hierarchy.

$$y_i \sim Binomial(n_i, p_i)$$

$$p_i \sim Beta(\alpha, \beta)$$

In this model, the prior means and variances of p_i are

$$E_{prior}(p_i) = \frac{\alpha}{\alpha+\beta}, \tag{4.5}$$

$$Var_{prior}(p_i) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4.6}$$

This prior, with fixed α and β leads to a closed form posterior distribution, a beta distribution, i.e

$$Posterior = \frac{Likelihood \times Prior}{Constant}$$

.

$$P(\boldsymbol{p}|\boldsymbol{y},\boldsymbol{n},\alpha,\beta) = \prod_{i=1}^{h}\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}\frac{p_i^{\alpha-1}(1-p_i)^{\beta-1}}{[B(\alpha,\beta)]^h} \tag{4.7}$$

$$= \prod_{i=1}^{h}\binom{n_i}{y_i}\frac{p_i^{y_i+\alpha-1}(1-p_i)^{n_i-y_i+\beta-1}}{[B(\alpha,\beta)]^h} \tag{4.8}$$

This is a joint distribution of h independent beta distributions with parameters $y_i+\alpha$ and $n_i-y_i+\beta$ i.e $Beta\,(y_i+\alpha, n_i-y_i+\beta)$.

Therefore the posterior distribution of p_i has means and variances given by

$$E_{posterior}(p_i) = \frac{y_i+\alpha}{n_i+\alpha+\beta}, \tag{4.9}$$

$$Var_{posterior}(p_i) = \frac{(y_i+\alpha)(n_i-y_i+\beta)}{(n_i+\alpha+\beta)^2(n_i+\alpha+\beta+1)} \tag{4.10}$$

This model is preferred due to the beta-beta conjugacy in prior and posterior distributions. The model captures overdispersion in the data but is inadequate if there exists spatial correlation in the data.

4.3.3 Logistic-normal model

This is an alternative model to the beta-binomial model and it allows for introduction of covariates in the modeling framework. In this model, it is assumed that a direct linkage exists between a linear predictor, η and the parameter of interest p_i. The commonly used link function for binary data is the logit link i.e

$$y_i \sim Binomial(n_i, p_i)$$

$$logit(p_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}$$

This falls under the generalized linear modeling (GLM) framework [McCullagh and Nelder, 1989]. A set of random effects can be introduced into the model to capture extra variation. This random effect is usually called uncorrelated heterogeneity (UH), i.e

$$logit(p_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} + v_i,$$

where $v_i \sim N(0, \sigma_v^2)$. This is the logistic-normal model and it falls in the broad framework of models called generalized linear mixed models (GLMM). The weakness with the above model is that it only allows spatially unstructured variability. If there are spatially structured effects then such a model cannot suffice.

4.3.4 Convolution model

It is possible to include two random effects, u_i and v_i in which u_i is spatially structured and v_i is spatially unstructured. This type of model (though in a

different form) was formulated by Besag et al. [1991], and it is described as follows.

$$y_i \sim Binomial(n_i, p_i)$$

$$logit(p_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} + u_i + v_i,$$

where u_i and v_i are variables representing unobserved covariates. In this formulation, u_i represents variables that if were observed would bring in influential spatial structure while v_i represents the unobserved spatially unstructured variables. In most cases, one of them usually dominates the other [Besag et al., 1991]. If u is stronger than v then the estimated risk will show spatial structure and if v is stronger than u then the consequence will be to shrink the estimated means towards the overall mean. Besag et al. [1991] assumed that u and v were independent with the following priors;

$$p(\boldsymbol{v}|\tau) \propto \tau^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\tau}\sum_{i=1} n v_i^2\right\}, \tag{4.11}$$

and

$$p(\boldsymbol{u}|k) \propto k^{-\frac{n}{2}} \exp\left\{-\frac{1}{2k}\sum_{j\in N(i)} n(u_i - u_j)\right\}. \tag{4.12}$$

In this equation, $N_{(i)}$ is the set of neighbours of region i and n is the number of regions in the study. Basically, equation (4.11) means that $\boldsymbol{v}$, the spatially unstructured random effect is a white noise Gaussian process with unknown variance τ and equation (4.12) means that the spatially structured random effect, $\boldsymbol{u}$, is a Gaussian Markov random field process. This distribution for spatially structured

random effect is also known as the conditional autoregressive model, conditionally it is normal distributed and given by

$$\left(u_i|\boldsymbol{u}_{N(i)}\right) \sim N\left(\frac{\sum_{j\in N_{(i)}} u_j}{d_i}, \frac{k}{d_i}\right), \tag{4.13}$$

$N_{(i)}$ and d_i denotes the set of neighbours of area i and the number of its neighbours respectively. In this model, the spatially unstructured component can also be viewed to capture overdispersion while the spatial structured component allows for spatial correlation.

4.3.5 Combined model

Molenberghs et al. [2012] and Kassahun et al. [2012] introduced the combined modelling strategy for binary random variables. In their models, they introduced two independent random effects, one for overdispersion and the other one to cater for clustering. These two random effects were introduced in the generalised linear mixed modelling framework. The models took the form:

$$y_{ij} \sim Bernoulli(p_{ij}),$$

$$p_{ij} = \theta_{ij}\omega_{ij},$$

$$logit(\omega_{ij}) = X_{ij}^T\beta + v_i$$

$$\theta_{ij} \sim Beta(\alpha, \beta),$$

$$v_i \sim N(0, \sigma_v^2).$$

4.3.6 Proposed spatial beta-binomial model

In this work, we propose a spatial beta-binomial model which extends on the work of Molenberghs et al. [2012] and Kassahun et al. [2012]. In their models, the additional random effect to capture overdispersion was assumed not to be spatially influenced and independent of the neighbourhood structure. The proposed model allows for the overdispersion parameter to also vary spatially over the area under study.

It is worth noting that, in binary data, overdispersion does not occur when the data is assumed to be independent and identically distributed Bernoulli. When there are hierarchies in the Bernoulli data or when the binary data accumulate to binomial data, then overdispersion can occur. Since these Bernoulli data are now correlated in the clusters, then there is a possibility for overdispersion.

The proposed model is developed for the analysis HSV-2 prevalence data which is clustered in counties. The model is formulated as follows. Let y_{ij} be the HSV-2 status of individual j in county i, where $i = 1, 2, \cdots N$ and $j = 1, 2, \cdots, n_i$. Then

$$y_{ij} \sim Bernoulli(p_{ij}),$$

$$p_{ij} = \theta_{ij}\omega_{ij},$$

$$logit(\omega_{ij}) = X_{ij}^T\beta + u_{1i} + v_{1i}$$

$$\theta_{ij} \sim Beta(a_i, b_i),$$

$$a_i = \mu_i^* \psi_i$$

$$b_i = \psi_i(1 - \mu_i^*)$$

$$logit(\mu_i^*) = u_{2i} + v_{2i}$$

$$\psi_i = exp(u_{3i} + v_{3i})$$

$$\left(u_{li} | \boldsymbol{u}_{lN(i)}\right) \sim N\left(\frac{\sum_{j \in N_{(i)}} u_{lj}}{d_{li}}, \frac{k_l}{d_{li}}\right), l = 1, 2, 3.$$

$$v_{li} \sim N(0, \sigma_{vl}^2), l = 1, 2, 3$$

4.4 Estimation of parameters

We chose to use Bayesian inference in estimating the parameters in the model with Markov chain Monte Carlo(McMC) technique. In Bayesian inference, parameters are treated as random variables and are given the so called prior distributions. These prior distributions are updated with the observed data to give the posterior distributions of the parameters of interest.

In our models, the following prior distributions were used. The spatially structured component were assigned a conditional autoregressive prior and their corresponding precision parameters $\frac{1}{k_l}$ were given a non-informative gamma distributed

priors, i.e $\frac{1}{k_l} \sim gamma(0.1, 0.0001)$. All the fixed effect parameters $\boldsymbol{\beta}$ were given non-informative normally distributed priors, i.e $\beta \sim N(0, 0.0001)$.

The models were implemented using WinBUGS version 1.4 [Spiegelhalter et al., 2007]. For each model, 85,000 McMC [Mollie et al., 1996] iterations were ran, with the initial 15,000 discarded to cater for the burn-in period and there after keeping every tenth sample value. The 7,000 iterations left were used for assessing convergence of the McMC and parameter estimation.

We assessed McMC convergence of all models parameters by checking trace plots and autocorrelation plots of the McMC output [Gelman et al., 2003]. The models were compared using the Deviance Information Criterion (DIC) [Spiegelhalter et al., 2002]. The best fitting model is one with the smallest DIC value.

4.5 Results

This results section has been divided into two subsections; the first subsection compares the performance of the proposed spatial beta-binomial model against all the other models that were discussed and the second section uses the best fitting model to identify determinants of HSV-2 infection in Kenya among men.

4.5.1 Models comparison

The HSV-2 data for men in Kenya was analyzed using the following models: simple logistic model (M1), logistic beta-binomial model (M2), logistic normal beta-binomial model (M3), logistic convolution model (M4), spatial combined model (M5), and finally the proposed spatial beta-binomial model (M6). Comparing the models based on the DIC, from Table 4.1, the spatial combined model ($DIC = 6202.79$) performed better than simple logistic model ($DIC = 6419.36$),Logistic beta-binomial model ($DIC = 6241.59$),logistic convolution model ($DIC = 6221.72$) and logistic normal beta-binomial model ($DIC = 6215.27$). The proposed spatial beta-binomial model ($DIC = 6164.90$) outperformed all the models.

TABLE 4.1: Parameter Estimates and Corresponding Credible Interval

Effect		Simple Logistic	Logistic Beta Binomial
Fixed Effects			
Intercept		-2.27(-2.61,-1.89)	-1.65(-1.95,-1.15)
Urban	Yes	0.33(0.18,0.48)	0.25(-0.04,0.49)
Circumcised	Yes	-1.45(-1.61,-1.26)	-1.34(-1.58,-1.08)
Education	Primary	0.25(0.09,0.41)	0.14(-0.07,0.34)
	Secondary	0.02(-0.16,0.18)	-0.11(-0.33,0.10)
	Higher	-0.37(-0.59,-0.16)	-0.55(-0.81,-0.29)
Age	20-24	1.02(0.64,1.37)	0.81(0.22,1.18)
	25-29	1.93(1.51,2.32)	1.83(1.21,2.16)
	30-34	2.45(2.04,2.82)	2.49(1.92,2.85)
	35-39	2.89(2.52,3.24)	2.97(2.43,3.30)
	40-44	3.26(2.89,3.71)	3.49(2.77,3.88)
	45-49	3.32(2.93,3.74)	3.68(2.95,4.07)
	50-54	3.02(2.65,3.42)	3.17(2.53,3.71)
	55-59	3.13(2.76,3.54)	3.35(2.80,3.82)
	60-64	2.88(2.54,3.39)	3.02(2.30,3.50)
Random effects			
a		-	3.22(3.01,3.80)
b		-	1.39(1.17,1.50)
σ_v		-	-
σ_u		-	-
DIC		6419.36	6241.59
		Logistic normal beta-binomial	Logistic convolution

Fixed Effects			
Intercept		-2.34(-2.75,-1.78)	
Urban	Yes	0.39(0.10,0.70)	0.33(0.10,0.58)
Circumcised	Yes	-1.05(-1.34,-0.84)	-0.92(-1.20,-0.66)
Education	Primary	0.16(-0.01,0.33)	0.14(-0.03,0.28)
	Secondary	-0.04(-0.25,0.16)	-0.10(-0.29,0.08)
	Higher	-0.40(-0.63,-0.16)	-0.42(-0.62,-0.21)
Age	20-24	0.83(0.22,1.24)	1.01(0.70,1.32)
	25-29	1.83(1.21,2.18)	1.95(1.64,2.30)
	30-34	2.48(1.80,2.91)	2.54(2.20,2.84)
	35-39	2.93(2.23,3.34)	2.92(2.59,3.21)
	40-44	3.39(2.67,3.82)	3.31(3.00,3.60)
	45-49	3.47(2.82,3.93)	3.39(3.08,3.67)
	50-54	3.15(2.40,3.61)	3.08(2.77,3.40)
	55-59	3.20(2.49,3.70)	3.14(2.79,3.48)
	60-64	3.01(2.28,3.50)	3.00(2.62,3.29)
Random effects			
a		4.73(4.15,4.99)	-
b		1.19(1.10,1.40)	-
σ_v		2.25(1.25,3.87)	0.04(0.00,0.14)
σ_u		-	0.40(0.15,0.74)
DIC		6215.27	6221.72
		Spatial combined model	Spatial beta-binomial model
Fixed Effects			
Intercept		-2.35(-2.75,-1.88)	-1.97(-2.49,-1.48)
Urban	Yes	0.52(0.26,0.81)	0.52(0.17,0.89)
Circumcised	Yes	-1.04(-1.33,-0.75)	-1.06(-1.39,-0.73)
Education	Primary	0.14(-0.08,0.35)	0.11(-0.16,0.35)
	Secondary	-0.09(-0.31,0.12)	-0.21(-0.55,0.084)
	Higher	-0.47(-0.74,-0.24)	-0.67(-1.10,-0.32)
Age	20-24	0.94(0.68,1.26)	1.13(0.73,1.54)
	25-29	1.98(1.67,2.31)	2.23(1.82,2.68)
	30-34	2.63(2.36,2.90)	3.03(2.56,3.59)
	35-39	3.14(2.77,3.53)	3.64(3.09,4.40)
	40-44	3.65(3.28,3.98)	4.69(3.67,10.29)
	45-49	3.75(3.40,4.15)	4.85(3.81,11.17)
	50-54	3.34(2.96,3.73)	3.98(3.30,5.08)
	55-59	3.45(2.97,3.93)	4.14(3.42,5.63)
	60-64	3.23(2.91,3.49)	3.77(3.11,4.76)
Random effects			
a		3.77(3.03,4.87)	-
b		1.33(1.12,1.49)	-
σ_{u1}			0.28(0.01,0.85)
σ_{v1}		0.57(0.17,1.30)	0.26(0.00,1.36)
σ_{u2}		-	0.02(0.00,0.22)
σ_{v2}		-	0.03(0.00,0.29)

σ_{u3}	-	0.33(0.23,0.43)
σ_{v3}	-	0.02(0.00,0.19)
DIC	6202.79	6164.90

4.5.2 Spatial distribution of HSV-2 infection in Kenya and its determinants

The following discussions on the effect of the fixed effects is based on this best fitting spatial beta-binomial model. The risk of getting HSV-2 among men was higher for those in urban as compared to those in rural, $0.52(0.17, 0.89)$. The estimated model regression coefficients show that circumcision in males reduces the chance of being infected by HSV-2, $-1.06(-1.39, -0.73)$. Having higher education reduces the risk of being infected with HSV-2, $-0.67(-1.10, -0.32)$. Males in the age groups $40 - 49$ are at higher risk of being infected with HSV-2 compared to the other age groups. The spatial beta-binomial model was used to produce the smoothed maps of HSV-2 prevalence in Kenya by county as show in Figure 4.1. HSV-2 is more prevalent in the Western part of the country around Lake Victoria region. It is also more prevalent in the Southern and South West region of the country.

FIGURE 4.1: HSV-2 prevalence map(a) and the corresponding 95% lower(b) and upper(c) credible limits maps, respectively, based on the proposed spatial beta-binomial model

4.6 Discussion

Analysis of the HSV-2 prevalence in Men's data from Kenya showed that in the presence of overdispersion and spatial autocorelation, the proposed model produces a better fit than all the previous models that were discussed. In this work, we model spatially overdispersed binary data by introducing two sets of random effects; beta random effects and spatially structured random effects. The spatially structured random effects are modeled using Conditional Autoregressive (CAR) model [Besag et al., 1991]. In this model, the overdispersion parameter is also allowed to vary spatially over the region under study. The proposed model is called spatial beta-binomial model. This model can be viewed as a substitute to the commonly encountered convolution model since the beta random effects introduced capture overdispersion while the normal CAR random effects capture the spatial correlation inherent in the data. The models were implemented in WinBUGS software where non-informative priors were assigned to parameters and hyperparameters in the models.

Several disease mapping models ranging from standard models to models which cater for overdispersion and spatial correlation were compared. We compared the logistic model, logistic-normal model, logistic beta-binomial model, logistic beta-binomial normal model, logistic convolution model and the proposed spatial beta-binomial model.

The spatial beta-binomial model was found to perform better in terms of deviance information criterion. From the data analysis carried out, based on this model, the chance of getting infected by HSV-2 among men was found to be higher for those in urban as compared to those in rural. Males who are circumcised had a lower chance of being infected by HSV-2. Males who had attained higher education level were found to have reduced the chance of being infected with HSV-2. Males in the age groups $40-49$ were found to have a higher chance of being infected with HSV-2 compared to the other age groups.

The proposed model was used to produce county specific HSV-2 prevalence smoothed maps for Kenya. These maps are important to policy makers in both government and private sectors. Policy makers can use these maps in formulating policies and intervention programs suited for each county.

From previous research findings that HSV-2 is highly correlated with HIV and that HSV-2 accelerates the transmission and acquisition of HIV [Celum et al., 2008; Watson-Jones et al., 2008]; the HSV-2 high prevalence areas can be targeted with tailor made intervention programs with the hope of curbing the acquisition and transmission of HIV.

This study is based on a cross-sectional survey hence it is only possible to make ecological association and it will be wrong to imply any causal association. Based on this, we therefore caution readers to interpret the findings with great caution. Causality can only be established by studies which have been carefully designed to investigate such.

As part of future research, the proposed model can be investigated for further extension to multivariate setting where several diseases have been observed in each spatial unit.

Chapter 5

Spatial Joint Disease Modeling and Mapping with Application to HIV and HSV-2

The joint modeling of epidemiological and public health outcomes within a spatial statistical context opens numerous opportunities for understanding disease aetiology. This paper reviews statistical properties of common joint modeling approaches and develops a model for the joint variation of the human immunodeficiency virus (HIV) and the herpes simplex virus-type 2 (HSV-2). The dataset used in this study consisted of men age 15-49 years from the 2007 Kenya Aids indicator survey data. Bivariate spatial logistic models are developed at the individual level and used to identify comorbidity of HIV and HSV-2. The joint spatial modelling strategy helps in stabilizing parameter estimates by borrowing strength between different diseases and also between neighbouring regions. A Bayesian approach

was used and the models were implemented in WinBUGS software. Both diseases showed significant spatial variation with highest disease burdens occurring around the Lake Victoria region. There was a significant positive correlation between HIV and HSV-2, a result in line with other studies. The correlation between the two diseases could mean that occurrence of one disease could be accelerating transmission and/or acquisition of the other disease. HSV-2 is not widely publicized and there are very few centres that test for it. There is need to put more effort and investment in controlling HSV-2 with the hope of reducing acquisition and transmission of HIV.

5.1 Introduction

Human immunodeficiency virus still remains a major concern in the global community. It is estimated that 34 million people were living with the virus at the end of 2011 [WHO and UNICEF, 2012]. Sub-Saharan Africa is the worst hit and it is estimated that 23.5 million people are living with HIV in this region [WHO and UNICEF, 2012]. In Kenya, the nationwide prevalence is estimated at 7.1% among adults aged 15-64 years. The prevalence of HIV is 8.4% among women and 5.4% among men [NASCOP, 2008]. However, the HIV prevalence varies considerably between sub-regions in the country. Counties like Bungoma, Embu, Isiolo, Kajiado, Machakos, Meru, Wajir and West Pokot having a prevalence of less than 3% while counties close to the lake Victoria region, in the west having prevalence ranging between 15% and 25% [NASCOP, 2008].

HIV is suspected to be associated with herpes simplex virus [Celum et al., 2008; Watson-Jones et al., 2008]. Herpes simplex virus is the causative agent of genital herpes. Two serotypes of Herpes simplex virus have been identified, namely herpes simplex virus-type 1 (HSV-1) and herpes simplex virus-type 2 (HSV-2). There is no cure for herpes; once you have the virus in your body system, you can only be put on a treatment therapy to suppress it but it will still remain in the body [Beauman, 2005]. Many individuals have non or mild symptoms of the herpes virus and are therefore unaware of the infection. Several studies have indicated that genital herpes is associated with high risk of HIV transmission and acquisition [Celum et al., 2008; Watson-Jones et al., 2008].

A lot of work has been done to estimate prevalence at regions below the national level. This falls in small area estimation techniques [Rao, 2005]. The estimates in these sub-regions can be highly unreliable if the sample sizes in those sub-regions are small. In single disease modeling, spatial smoothing techniques can be used to borrow information from other neighbouring regions so as to get robust estimates for each region. In the same spirit, modeling two diseases which have some association can help in stabilizing the sub-region estimates by borrowing information from the other disease and also borrowing information from its neighbours.

Several approaches exist for spatial joint modeling of diseases [Manda et al., 2011; Kazembe and Namangale, 2007; Manda et al., 2012]. The two main encountered models are the shared component model by Held et al. [2005] and the multivariate conditional autoregressive model by Carlin and Banerjee [2003]. In this study, we jointly model HIV and HSV-2 at individual level using these two methods and

compare and contrast the results vis a vis separate modeling of the diseases. The models are applied to the men's data extracted from the Kenya AIDS indicator survey, collected by the government of Kenya in 2007.

5.2 Data

The data for this study was extracted from the 2007 Kenya AIDS indicator survey (KAIS), conducted by the Government of Kenya. The main objective of survey was to collect high quality data on the prevalence of HIV and Sexually Transmitted Infections (STI) among adults, and to assess knowledge of HIV and STI in the populations. The survey collected a representative sample of households selected from the eight provinces in the country. It involved all men and women in the age of 15-64 years. Two questionnaires were used in the survey. A household questionnaire which collected information about the household head and the characteristics of the dwelling place. The second one, the individual questionnaire which collected information from men and women aged 15-64 years, about their demographic characteristics, and their knowledge on HIV and STI. Each individual was then asked for consent to provide a venous blood sample for HIV and HSV-2 testing. In this study we use the men's data from this survey. In total, 6,606 men who provided venous blood for testing and also had full covariate information were used in the analysis. The following covariates were used in the analysis: education level, circumcision status, place of residence(urban/rural) and age of respondent.

5.3 Review of Models

In this section we review the conditional autoregressive (CAR) distribution and the two proposed spatial joint modeling strategies to be used on this data, namely the multivariate conditional autoregressive (MCAR) and the shared component models.

5.3.1 Conditional autoregressive distribution

Consider a vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_p)^T$ of p components that follow a multivariate Gaussian distribution with mean zero and variance-covariance matrix $\boldsymbol{B}^{-1}$, where $\boldsymbol{B}$ is a $p \times p$ symmetric and positive definite matrix. It follows that the joint pdf of $\boldsymbol{\phi}$ is given by

$$p(\boldsymbol{\phi}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{B}|^{\frac{1}{2}} \exp\left\{\frac{1}{2}\boldsymbol{\phi}^T \boldsymbol{B}\boldsymbol{\phi}\right\} \tag{5.1}$$

The conditional distribution of one of the components given the remaining ones, in terms of the elements of matrix $\boldsymbol{B}$, is given by

$$p(\phi_i|\boldsymbol{\phi}_{-i}) \propto \exp\left\{\frac{-b_{ii}}{2}\left(\phi_i - \sum_{j\neq i}\frac{-b_{ij}}{b_{ii}}\phi_j\right)^2\right\} \tag{5.2}$$

This can be written in short form as $\phi_i|\boldsymbol{\phi}_{-i} \sim N(-\sum_{j\neq i}\frac{b_{ij}}{b_{ii}}\phi_j, \frac{1}{b_{ii}})$.

Besag [1974], using Hammersley-Clifford's theorem and Brook's lemma, showed the conditions under which the full conditional distributions specified above uniquely define a full joint distribution. If we simplify the notations in the conditional distribution by letting $\frac{-b_{ij}}{b_{ii}} = c_{ij}$, $\quad c_{ii} = 0$ and $\frac{1}{b_{ii}} = \sigma_i^2$, then proceed to form two

matrices $\boldsymbol{C}$ and $\boldsymbol{M}$, with $\boldsymbol{C}$ having elements c_{ij} and c_{ii}, and M being a $diag(\sigma_i^2)$ matrix. The inverse of the dispersion matrix, $\boldsymbol{B}$ is then related to $\boldsymbol{C}$ and $\boldsymbol{M}$ as follows;

$$\boldsymbol{B} = \boldsymbol{M}^{-1}(\boldsymbol{I} - \boldsymbol{C}) \tag{5.3}$$

Then the joint distribution of $\boldsymbol{\phi}$ is $MVN(0, \boldsymbol{M}^{-1}(\boldsymbol{I} - \boldsymbol{C}))$. The trick is that $\boldsymbol{C}$ and $\boldsymbol{M}$ must be modeled properly so as to ensure symmetry in $\boldsymbol{B}$. The condition $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$ usually guarantees the required symmetry. The $\boldsymbol{C}$ matrix is also specified to show relationship between neighbours.

A commonly used adjacency matrix for lattice data is one due to Besag [1974]. He defined the the elements of the matrix C matrix as $c_{ii} = 0$ and $c_{ij} = \frac{1}{m_i}$ if j is adjacent to i and zero otherwise. Here m_i is the number of neighbours of region i. Define another matrix $\boldsymbol{W}$ to hold the adjacency structure, where $w_{ii} = 0$, $w_{ij} = 1$ if region i and region j are neighbours and zero otherwise. It then follows that $\boldsymbol{C} = \boldsymbol{W}_s$ where $\boldsymbol{W}_s = diag(\frac{1}{m_i})$. This notation implies the following specification to the matrix $\boldsymbol{B}$; $b_{ii} = \lambda m_i$ and, $b_{ij} = -\lambda$ if region j is adjacent to region i and zero otherwise. This further implies that

$$B = \lambda(diag(m_i) - C). \tag{5.4}$$

From equation (5.3), $\boldsymbol{M}^{-1}(\boldsymbol{I} - \boldsymbol{C})$ has to be positive definite for the conditional distributions to give rise to a valid joint pdf. Besag [1974] definition of the adjacency matrix leads to an improper joint pdf. This can be overcome by introducing

a parameter α into the precision matrix $\boldsymbol{B}$, to yield

$$\boldsymbol{B} = \boldsymbol{M}^{-1}(\boldsymbol{I} - \alpha \boldsymbol{C}) \tag{5.5}$$

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

5.3.2 Multivariate Conditional Autoregressive 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 which the conditional multivariate distributions uniquely determine the corresponding multivariate joint pdf. Using these results, Carlin and Banerjee [2003] developed the MCAR as follows. Let $\boldsymbol{\Phi}^T = (\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T, \cdots, \boldsymbol{\phi}_p^T)$, where each $\boldsymbol{\phi}_i$ is an $n \times 1$ vector. Then $\boldsymbol{\Phi}$ is an $np \times 1$ vector. Also let $\boldsymbol{\Phi}$ have a multivariate Gaussian distribution with mean $\mathbf{0}$ and dispersion matrix $\boldsymbol{B}$, written as

$$P(\boldsymbol{\Phi}) = (2\pi)^{\frac{-np}{2}} |\boldsymbol{B}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\Phi}^T \boldsymbol{B} \boldsymbol{\Phi}\right\} \tag{5.6}$$

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

are given by

$$P(\boldsymbol{\phi}_i|\boldsymbol{\phi}_{-i}) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\phi}_i - \boldsymbol{B}_{ii}^{-1}\sum_{j\neq i}(-\boldsymbol{B}_{ij})\,\boldsymbol{\phi}_j\right)^T \boldsymbol{B}_{ii}\left(\boldsymbol{\phi}_i - \boldsymbol{B}_{ii}^{-1}\sum_{j\neq i}(-\boldsymbol{B}_{ij})\,\boldsymbol{\phi}_j\right)\right] \tag{5.7}$$

This implies that $\boldsymbol{\phi}_i|\boldsymbol{\phi}_{-i} \sim N_n\left(\boldsymbol{B}_{ii}^{-1}\sum_{j\neq i}(-\boldsymbol{B}_{ij})\,\boldsymbol{\phi}_j, \boldsymbol{B}_{ii}^{-1}\right)$. The full conditional probability density functions are

$$P\left(\boldsymbol{\phi}_i|\boldsymbol{\phi}_{-i}\right) = N_n\left(\sum_{j\neq i} C_{ij}\boldsymbol{\phi}_j, \Sigma_i\right), i = 1, 2, \cdots, p, \tag{5.8}$$

where $\boldsymbol{\Sigma}_i$ and $\boldsymbol{C}_{ij}$ are $n\times n$ matrices. $\boldsymbol{\Sigma}_i$ is also symmetric and positive definite. We now write Σ_i and C_{ij} in terms of $\boldsymbol{B}$, the precision matrix of the joint distribution as $C_{ij} = -\boldsymbol{B}_{ii}^{-1}\boldsymbol{B}_{ij}$ and $\boldsymbol{\Sigma}_i = \boldsymbol{B}_{ii}^{-1}$. If we set $\boldsymbol{\Sigma}$ to be a block diagonal matrix with $\boldsymbol{\Sigma}_i$ blocks and $\boldsymbol{C}$ as a partitioned matrix with blocks $\boldsymbol{C}_{ij}$ and $\boldsymbol{C}_{ii} = \boldsymbol{0}_{n\times n}$, then

$$\boldsymbol{B} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{I} - \boldsymbol{C}) \tag{5.9}$$

A propriety parameter α can be added into the precision matrix in equation (5.8) to yield

$$\boldsymbol{B} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{I} - \alpha\boldsymbol{C}) \tag{5.10}$$

For $\boldsymbol{B}$ to be symmetric then a condition to satisfy this is that $\boldsymbol{C}_{ij}\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_i\boldsymbol{C}_{ji}^T$. Carlin and Banerjee [2003] denoted this distribution by $MCAR(\boldsymbol{C}, \boldsymbol{\Sigma})$. During implementation, α and $\boldsymbol{\Sigma}$ are given appropriate priors, most often, uniform distribution for α and $Wishart(\rho, \boldsymbol{\Sigma}_0)$ for $\boldsymbol{\Sigma}$.

5.3.3 Shared Component Model

Held et al. [2005] introduced a new joint modeling paradigm known as the shared component model. The idea was borrowed from Knorr-Held and Best [2001], in their work on joint disease clusters detection. For the case of two diseases, two components were introduced into the model; one component which is relevant to the two diseases and another one which is specific to one of the diseases. The two components represent unobserved spatial variables that affect the risk of the disease(s). Let y_{i1} be the observed cases of disease 1 in region i and e_{1i} be the corresponding expected number of cases for the same disease. Similarly y_{i2} and e_{2i} are the observed and expected number of cases for the second disease in region i. In their model, they assumed that;

$$y_{i1} \sim Poisson\left(e_{i1} \exp(\eta_{i1})\right),$$

$$y_{i2} \sim Poisson\left(e_{i2} \exp(\eta_{i2})\right),$$

and the log relative risks were modeled using normal random variables with,

$$\eta_{i1} \sim N\left(\alpha_1 + u_{1i}\delta + u_{2i}, \tau_1\right),$$

$$\eta_{i2} \sim N\left(\alpha_2 + \frac{u_{1i}}{\delta}, \tau_2\right).$$

$\boldsymbol{u}_1$ is the shared component while $\boldsymbol{u}_2$ is the component specific to the first disease only. They were assumed to follow Gaussian Markov random fields (GMRF) with precision parameters τ_1 and τ_2 respectively. The non-negative parameter δ was

included in the model to allow the two diseases to have different risk gradients in the shared component. They assumed that $\log \delta \sim N(0, \sigma^2)$, with the value of σ^2 pre-set to 0.17. The parameters α_1 and α_2 are the intercepts for disease 1 and disease 2 respectively and were assumed to be having uniform priors.

5.4 Models Specification

In this section, the reviewed models are adopted to suit the Bernoulli data at hand. Let y_{ijk} be the disease status(0/1) of disease k, $k = 1$ for HIV, $k = 2$ for HSV-2, for individual j in county $i, i = 1, 2, \cdots 46$. We further assume that the observed outcomes arise from a bivariate Bernoulli distribution, with p_{ijk} as the probability of disease k occurring in individual j in area i. The data generating model is defined as

$$y_{ijk} \sim Bernoulli(p_{ijk}), \tag{5.11}$$

and for each model, the covariates are introduced as discussed below. In the shared component model, the covariates and random effects are introduced as follows

$$logit(p_{ij1}) = \alpha_1 + \boldsymbol{X}^T\boldsymbol{\beta}_1 + u_{1i}\delta + u_{2i},$$

$$logit(p_{ij2}) = \alpha_2 + \boldsymbol{X}^T\boldsymbol{\beta}_2 + \frac{u_{1i}}{\delta}.$$

$\boldsymbol{u}_1$ is the shared component while $\boldsymbol{u}_2$ is the component specific to the first disease only. These two components are modeled using conditional autoregressive priors with precision parameters τ_1 and τ_2 respectively.

In the multivariate CAR model, the covariates and random effects are introduced as follows

$$logit(p_{ij1}) = \alpha_1 + \boldsymbol{X}^T\boldsymbol{\beta}_1 + u_{1i},$$

$$logit(p_{ij2}) = \alpha_2 + \boldsymbol{X}^T\boldsymbol{\beta}_2 + u_{2i}.$$

where $\boldsymbol{u} = (u_1, u_2)^T$ is modeled using a multivariate condition autoregressive prior that is $\boldsymbol{u} \sim MCAR(1, \Sigma)$, where Σ is the covariance matrix inducing correlation.

In the separate analyses, the covariates and random effects are introduced as follows

$$logit(p_{ij1}) = \alpha_1 + \boldsymbol{X}^T\boldsymbol{\beta}_1 + u_{1i} + v_{1i},$$

$$logit(p_{ij2}) = \alpha_2 + \boldsymbol{X}^T\boldsymbol{\beta}_2 + u_{2i} + v_{2i}.$$

where, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are modeled using independent conditional autoregressive priors while $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are modeled using independent normal distributions.

Also, α_k is the intercept for disease k while the terms $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$ are vectors of regression parameters corresponding to the set of covariates (fixed effects). Model estimation was carried out using the Bayesian approach and appropriate prior distributions were specified for all parameters of the models. In addition to the priors given to the random effects discussed in the models above, non-informative priors were assigned to the regression coefficients. For the intercepts, diffuse priors were assumed, that is, $p(\alpha_k) \propto 1$, while for the covariate coefficients highly dispersed normal distribution priors were chosen, that is, $p(\boldsymbol{\beta}) \sim N(0, 10000)$. In the shared component model, the extra parameter δ was given a prior as

$\log \delta \sim N(0, \sigma^2)$ while in the multivariate setting the covariance matrix was given an inverse Wishart prior as $\Sigma \sim IW(r, R)$, with R being assumed to be an identity matrix.

5.5 Results

The estimated covariate effects for the different models fitted are presented in an odds ratio scale in Table 5.1. The table summarizes the results of the models including parameter estimates (odds ratio) and their confidence intervals. The multivariate CAR model was found to have the lowest DIC hence the best fitting model for the data. In the subsequent discussion we only report results of this best fitting model. The correlation between HIV and HSV-2 was found to be positive and significant, $0.37(0.001, 0.69)$. The odds of getting HIV and HSV-2 were higher for those in urban as compared to those in rural, $1.41(1.02, 1.94)$ and $1.36(1.11, 1.65)$ respectively. The odds of being infected by HIV and HSV-2 was lower for the circumcised males as compared to the uncircumcised males, $0.25(0.17, 0.32)$ and $0.42(0.34, 0.50)$, respectively). The odds of getting infected by HIV and HSV-2 were lower for those with post primary education but higher for those with primary education as compared to those with no education at all. The odds of getting HIV and HSV-2 were high for those in the age groups above $15 - 19$ years. The age groups in the category $25 - 39$ years had the highest risk of both HIV and HSV-2 infection. The choropleth maps in Figure 5.1 give the county-specific HIV and HSV-2 smoothed prevalence estimates from the joint model. High rates of HIV and HSV-2 were found to be concentrated in the western

and southern parts of the country. The highest burden of the two diseases occur at the Lake Victoria region.

TABLE 5.1: Posterior means(95%CI) estimates for HIV and HSV-2 smoothed prevalence parameters

		Separate models				Joint models			
		Fixed Effects		Mixed Effects		Multivariate CAR		Shared Component	
Covariates		HIV	HSV-2	HIV	HSV-2	HIV	HSV-2	HIV	HSV-2
Rural	Yes	1	1	1	1	1	1	1	1
	No	1.36 (1.02,1.72)	1.4 (1.21,1.59)	1.44 (1.00,1.92)	1.34 (1.10,1.64)	1.41 (1.02,1.94)	1.36 (1.11,1.65)	1.32 (0.87,1.77)	1.32 (1.03,1.63)
Circum.	No	1	1	1	1	1	1	1	1
	Yes	1.29 (0.88,1.70)	0.96 (0.69,1.39)	0.22 (0.15,0.31)	0.40 (0.32,0.51)	0.25 (0.17,0.32)	0.42 (0.34,0.50)	0.25 (0.17,0.39)	0.48 (0.39,0.61)
Educ.	None	1	1	1	1	1	1	1	1
	Prim.	1.29 (0.88,1.70)	1.26 (1.05,1.47)	1.25 (0.85,1.62)	1.14 (0.96,1.31)	1.20 (0.86,1.63)	1.14 (0.97,1.33)	1.09 (0.77,1.42)	1.14(0.95,1.33)
	Sec.	0.95 (0.69,1.39)	1.03 (0.86,1.20)	0.92 (0.64,1.34)	0.93 (0.78,1.15)	0.90 (0.63,1.28)	0.93 (0.77,1.13)	0.82 (0.57,1.12)	0.92 (0.77,1.09)
	Higher	0.53 (0.31,0.87)	0.70 (0.54,0.86)	0.56 (0.36,0.84)	0.69 (0.55,0.88)	0.63 (0.40,0.97)	0.67 (0.51,0.85)	0.52 (0.31,0.82)	0.67 (0.53,0.86)
Age	15-19	1	1	1	1	1	1	1	1
	20-24	1.52(0.33,2.90)	2.26(0.52,3.19)	1.73(0.38,3.23)	2.35(0.52,3.56)	1.52(0.32,2.86)	2.33(0.45,3.28)	1.46(0.54,2.77)	2.51(1.48,3.76)
	25-29	5.36(0.98,9.59)	5.40(1.11,7.80)	6.25(0.82,11.06)	6.07(1.00,9.44)	5.03(1.01,9.25)	6.08(1.18,8.37)	5.09(2.32,8.52)	6.58(4.29,9.28)
	30-34	6.65(1.11,12.09)	9.20(1.87,13.51)	8.21(1.03,14.51)	10.76(1.79,16.42)	6.69(1.19,12.86)	10.77(1.83,14.58)	7.00(3.27,11.7)	11.77(7.25,16.87)
	35-39	9.00(1.20,17.22)	13.84(2.65,19.62)	10.51(1.14,18.49)	15.97(2.61,24.1)	8.25(1.08,15.92)	15.90(2.63,22.23)	8.50(4.11,13.9)	17.29(11.02,24.94)
	40-44	8.77(1.52,14.88)	20.33(3.606,28.68)	10.57(0.98,18.56)	23.91(3.80,36.44)	8.45(1.24,16.00)	23.80(4.18,33.2)	8.34(3.98,13.48)	25.68(16.57,38.68)
	45-49	6.17(0.72,10.75)	21.74(3.81,31.72)	7.55(1.04,13.58)	24.74(4.13,38.58)	5.76(0.74,10.73)	25.29(3.91,35.19)	5.72(2.18,9.52)	28.54(17.85,42.29)
	50-54	6.52(0.97,12.17)	16.28(3.03,23.37)	7.64(0.80,13.89)	18.86(2.62,28.45)	5.95(0.77,12.87)	18.85(2.39,27.36)	5.82(2.36,9.64)	21.33(12.42,31.48)
	55-59	2.35(0.63,4.17)	18.27(2.39,27.75)	2.42(0.45,4.84)	20.08(2.87,32.77)	2.46(0.49,5.12)	20.41(3.03,28.31)	2.09(0.69,4.43)	22.95(14.78,34.85)
	60-64	2.40(0.45,4.57)	13.78(2.90,20.92)	3.03(0.37,6.41)	17.31(1.95,26.58)	1.72(0.42,4.20)	17.07(2.77,25.53)	2.26(1.05,4.96)	18.21(10.27,27.58)
Random effects									
Unstr. std. dev.		-	-	0.19(0.04,0.35)	0.003(0.00,0.01)	-	-		
Str. std. dev.		-	-	0.003(0.00,0.01)	0.54(0.06,1.02)	0.84(0.52,1.15)	0.77(0.51,1.03)	2.71(0.79,5.05)	
Corr.		-	-	-	-	0.37(0.001,0.69)			
Ind. DIC		2473.45	6419.64	2428.35	6217.02	2425.62	6214.6	2426.63	6336.05
Full DIC		8893.09		8645.37		8640.22		8762.68	

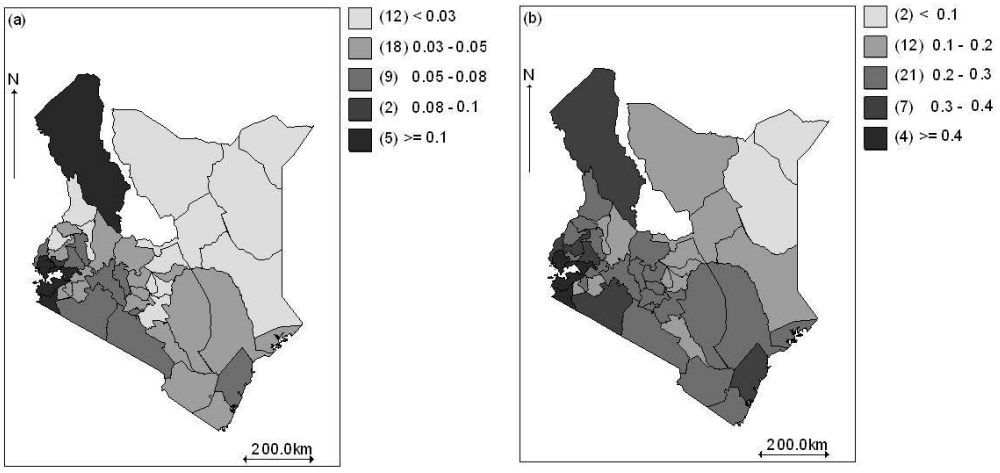


FIGURE 5.1: (a) HIV prevalence among men by county and (b) HSV-2 prevalence among men by county from the best fitting joint model.

5.6 Discussion

In this work, we have reviewed models for spatial joint modeling of diseases. The models were then adopted for bivariate spatial logistic models to suit the data at hand and the resulting models were used to jointly model HIV and HSV-2 in Kenya.

Joint modeling has several advantages, first, statistically, joint modeling helps to stabilize parameter estimates especially in small area estimation where sample sizes at sub-regions with respect to each disease are small and secondly, in epidemiological perspective, joint modeling helps in determining divergent and similar patterns of disease and understanding diseases association.

The models were compared using the deviance information criterion. The multivariate conditional autoregressive model was the best fitting model. Under this model, the spatial correlation between HIV and HSV-2 was found to be significant.

This finding is in line with other studies [Celum et al., 2008; Watson-Jones et al., 2008].

Joint modeling of diseases has recently seen growth in application with the main aim of understanding aetiology of several diseases. Most recent applications with similar modeling strategy include Manda et al. [2011], Manda et al. [2012] and Kazembe and Namangale [2007]. In these works, the authors fitted only one type of model, either multivariate CAR or shared component model. In our work we fit separate models, shared component model and multivariate conditional autoregressive model and compare the performance of each on this data.

The MCAR model was used to produce smoothed prevalence maps which can show geographical variation of the diseases burden at a glance, based on survey data. These maps are important to policy makers in both government and private sectors. Policy makers can use these maps in formulating policies and programs suited for each county. In particular such easy to use tools can be very useful in optimal allocation of resources aimed at controlling the diseases.

Chapter 6

Conclusion and Future Research

This dissertation has been concerned with developing and extending statistical models in the area of spatial modelling with inclination towards application to HIV, TB and HSV-2 data. The models under consideration cater for areal(lattice) data only; geostatistical data and point pattern data were not delved into in this work. No study goes without limitations. A major limitation of our study is that the data used for county estimation was collected when the country was still based on the old administrative units (provinces), the data was not powered to carry out estimation at these new administrative units. The study rides on the advantage that these new administrative units called 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. Also, in the KAIS (2007) data, the way some variables were captured was not useful; ever using a condom should be replaced with consistent use of condom. Despite these limitations, the models developed in this thesis will find wide application in

spatial analysis.

Chapter 2 introduces a new model that relaxes the over-restrictive normal distribution assumption on the spatially unstructured random effect by using the generalised Gaussian distribution. In chapter 3, a framework for including sampling weights into the Bayesian hierarchical disease mapping model is given; in this model, design effect is used to re-scale the sample sizes. A new model is developed in chapter 4 to cater for overdispersed spatially correlated binary data; in this model, the overdispersion parameter is modelled by a beta random effect which is allowed to vary spatially also . In chapter 5, the common multiple spatial disease mapping models are reviewed and adopted for the binary counterpart since the original models were meant for Poisson data.

Future work will consider extensions of the models presented in Chapters 2, 4 and 5. In chapter 2, a topic of interest could be to consider a more generalised Gaussian distribution for the random effects that allows for skewness.

Further studies can be directed into using more flexible distributions for the unstructured random effects in a multiple spatial disease modelling in lieu of the multivariate Gaussian distribution. A generalised multivariate Gaussian distribution, is a possible candidate in this framework.

A major extension of all the models discussed in this thesis is incorporation of the temporal component. Since the KAIS survey will be carried out severally in future, these models can be extended to accommodate for this time portion. Several authors have considered temporal extensions to hierarchical spatial models based on a parametric description of time trends, on independent risk estimates for each

time period, or on the definition of the joint covariance matrix for all the periods as a Kronecker product of matrices [Cressie and Wikle, 2011]. Waller et al. [1997] introduced a spatio-temporal model where the spatial effects are nested within time. MacNab and Dean [2002] proposed a generalized additive mixed model (GAMM), where B-spline smoothing over the temporal dimension provides a flexible means of accommodating overall time effects as well as region-specific time effects. Martínez-Beneito et al. [2008] came up with an autoregressive approach to spatio-temporal disease mapping by fusing ideas from autoregressive time series in order to link information in time and by spatial modelling to link information in space. Schrödle and Held [2011] also carried out a spatio-temporal disease mapping, utilising the Integrated Nested Laplace approximation method. A multivariate spatio-temporal analysis may be explored and could lead to improved precision for the estimation of the underlying disease risks, by borrowing strength from other diseases as well as from neighboring areas and/or time points. All the developments in overdispersed spatial models, including the development of chapter 4, have been based on single disease modelling. Further research can be focussed on extending these models for multiple diseases.

Appendix A

WinBUGS Codes for chapter Two Models

```
model}
# Likelihood
for (i in 1 : N) {
O[i]  ~ dpois(mu[i])
log(mu[i]) <- log(E[i]) + alpha0 + v[i]+u[i]
RR[i] <- exp(alpha0 + v[i]+u[i]) # Area-specific relative risk (for maps)
u[i]~dnorm(0,precu)
}
# CAR prior distribution for random effects:
v[1:N] ~ car.normal(adj[], weights[], num[], precv)
for(k in 1:sumNumNeigh) {
weights[k] <- 1
}
# Other priors:
alpha0  ~ dflat()
precv  ~ dgamma(0.5, 0.0005) # prior on precision
precu  ~ dgamma(0.01, 0.01)
sigmav <- sqrt(1 /  precv) # standard deviation of v
  sigmau <- sqrt(1 /  precu) # standard deviation of u
}
Data
Initials
list(precv = 1,precu=1, alpha0 = 0,
v=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

```
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
u=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
list(precv = 1,precu=1, alpha0 = 0,
v=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
u=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
#GGD MODEL
model { # Likelihood
for (i in 1 : N) {
O[i]  ~ dpois(mu[i])
log(mu[i]) <- log(E[i]) + alpha0 + v[i]+u[i]
RR[i] <- exp(alpha0 + v[i]+u[i]) # Area-specific relative risk (for maps)}
#Generalized Gaussian Distribution implementation using zero tricks
my_zeta<-pow(abs((sigmau*sigmau*exp(loggam(1/psi)))/(exp(loggam(3/psi)))),0.5)
normalizing<-1/((2*exp(loggam(1+1/psi)))*my_zeta)
logN<-log(normalizing)
for(i in 1: N)
{
u[i]~dunif(-10000,10000)
zeros[i]<-0
logGGD[i]<-logN-pow(abs(u[i]/my_zeta),psi)
zeros[i]~dpois(logGGD[i])
}
#priors
psi~dunif(0,3)
sigmau~dgamma(1,1)
# CAR prior distribution for random effects:
v[1:N] ~ car.normal(adj[], weights[], num[], precv)
for(k in 1:sumNumNeigh)  { weights[k] <- 1 }
# Other priors:
alpha0  ~ dflat()
precv  ~ dgamma(0.5, 0.0005)  # prior on precision
# precu  ~ dgamma(0.01, 0.01)
sigmav <- sqrt(1 /  precv) # standard deviation of v
#  sigmau <- sqrt(1 /  precu) # standard deviation of u
}
#Data
#Initials
list(precv = 1, alpha0 = 0, psi=1,sigmau=0.5,
v=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
u=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
list(precv = 1,precu=1, alpha0 = 0,
v=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

```
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
u=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
```

Appendix B

WinBUGS Codes for Chapter Three Models

```
#####################################################################
###                          WEIGHTED                             ###
###                          ANALYSIS  WOMEN                      ###
#####################################################################
###############BEST FITTING NON SPATIAL MODEL#################
####Variables#####
#x1=firstsex_15-17, #x2=Percept_norisk, #x3=One_partner
#x4=prop_not_using_condom, #x5=prop_No_condom
#x6=prop_Not_Circumcised, #x7=prop_rural, #x8=No_media_access
#x9=propwith_STI, #x10=age
model
{ for(i in 1:N)
{ n_adj[i]<-n[i]/deff[i]
y_adj[i]<-p_est[i]*n_adj[i]
y_adj[i]~dbin(p[i],n_adj[i])
logit(p[i])<-beta[1] +beta[2]*x1[i]+beta[3]*x2[i]+beta[4]*x3[i]
}
#prior distribution for the current model
for(j in 1:ncov)
{beta[j]~dnorm(0.0,0.0001)}
}
#initialisation
list(beta=c(0,0,0,0))
# data
```

```
#################################################################
##############UNSTRUCTURED HETEROGENEITY MODEL###########
####Variables#####: #x1= mean county age, #x2= sex debut average age,
 #x3= PropAway
model
{ for(i in 1:N)
{ n_adj[i]<-n[i]/deff[i]
y_adj[i]<-p_est[i]*n_adj[i]
y_adj[i]~dbin(p[i],n_adj[i])
logit(p[i])<-beta[1] +beta[2]*x1[i]+beta[3]*x2[i]+beta[4]*x3[i]+v[i]
v[i]~dnorm(0,tau.v) }
sigma.v<-(1/tau.v)
#prior distribution for the current model
for(j in 1:ncov){beta[j]~dnorm(0.0,0.0001) }
tau.v~dgamma(0.1,0.0001)
}
#initialisation
list(beta=c(0,0,0,0),tau.v=0.01,v=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0 ,0))
# data
#################################################################
########################CAR MODEL#############################
####Variables#####: #x1= mean county age #x2= sex debut average age
 #x3= PropAway
model
{ for(i in 1:N)
{ n_adj[i]<-n[i]/deff[i]
y_adj[i]<-p_est[i]*n_adj[i]
y_adj[i]~dbin(p[i],n_adj[i])
logit(p[i])<-beta[1] +beta[2]*x1[i]+beta[3]*x2[i]+beta[4]*x3[i]+u[i]}
# CAR
u[1:N]~car.normal(adj[],weights[],num[],tau.u)
for(k in 1:sumNumNeigh){ weights[k]<-1}
for(j in 1:ncov){beta[j]~dnorm(0.0,0.0001) }
tau.u~dgamma(0.1,0.0001)
sigma.u<-(1/tau.u)
}
#initialisation
list(beta=c(0,0,0,0),tau.u=0.01,u=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,  0, 0, 0,
  0, 0, 0, 0, 0, 0 ,0))
# data
#################################################################
####################CONVOLUTION MODEL########################
####Variables##### #x1= mean county age #x2= sex debut average age
#x3= PropAway
```

```
model
{ for(i in 1:N)
{ n_adj[i]<-n[i]/deff[i]
y_adj[i]<-p_est[i]*n_adj[i]
y_adj[i]~dbin(p[i],n_adj[i])
logit(p[i])<-beta[1] +beta[2]*x1[i]+beta[3]*x2[i]+beta[4]*x3[i]+u[i]+v[i]
v[i]~dnorm(0,tau.v) }
sigma.v<-(1/tau.v)
grandp<-sum(y_adj[])/sum(n_adj[])
for(i in 1:N) { RR[i]<-p[i]/grandp }
# CAR
u[1:N]~car.normal(adj[],weights[],num[],tau.u)
for(k in 1:sumNumNeigh){ weights[k]<-1 }
#prior distribution for the current model
for(j in 1:ncov){beta[j]~dnorm(0.0,0.0001)}
tau.u~dgamma(0.1,0.0001)
tau.v~dgamma(0.1,0.0001)
sigma.u<-(1/tau.u)
}
#initialisation
list(beta=c(0,0,0,0),tau.u=0.01,tau.v=0.01,u=c(0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0,  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ,0),v=c(0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0,   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0,    0, 0, 0,  0, 0, 0, 0, 0, 0, 0 ,0))
# data
#################################################################
#################################################################
#################################################################
###                          WEIGHTED                         ###
###                          ANALYSIS   MEN                   ###
#################################################################
###############BEST FITTING NON SPATIAL MODEL#################
####Variables#####: #x1=firstsex_15-17#x2=Percept_norisk
#x3=One_partner #x4=prop_not_using_condom#x5=prop_No_condom
#x6=prop_Not_Circumcised
 #x7=prop_rural #x8=No_media_access #x9=propwith_STI#x10=age
model
{ for(i in 1:N)
{ y[i]~dbin(p[i],n[i])
logit(p[i])<-beta[1] +beta[2]*x1[i]+beta[3]*x2[i]+beta[4]*x3[i]+
beta[5]*x4[i]+beta[6]*x5[i]+beta[7]*x6[i]+beta[8]*x7[i]+beta[9]*x8[i]+
beta[10]*x9[i]+beta[11]*x10[i] }
#prior distribution for the current model
for(j in 1:ncov){beta[j]~dnorm(0.0,0.0001)}
}
#initialisation
```

```
list(beta=c(0,0,0,0,0,0,0,0,0,0,0))
# data
#N=46,ncov=11
#####################################################################
##############UNSTRUCTURED HETEROGENEITY MODEL###########
####Variables#### #x1= mean county age #x2= proportion with no education
#x3= Proportion perceiving not at risk of HIV #x4=propotion not circumcised
model
{ for(i in 1:N)
{n_adj[i]<-n[i]/deff[i]
y_adj[i]<-p_est[i]*n_adj[i]
y_adj[i]~dbin(p[i],n_adj[i])
logit(p[i])<-beta[1] +beta[2]*x1[i]+beta[3]*x2[i]+beta[4]*x3[i]+
beta[5]*x4[i]+v[i]
v[i]~dnorm(0,tau.v) }
sigma.v<-(1/tau.v)
#prior distribution for the current model
for(j in 1:ncov){beta[j]~dnorm(0.0,0.0001)}
tau.v~dgamma(0.1,0.0001)
}
#initialisation
list(beta=c(0,0,0,0,0),tau.v=0.01,v=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0 ,0))
# data
#####################################################################
########################CAR MODEL############################
####Variables##### #x1= mean county age #x2= proportion with no education
#x3= Proportion perceiving not at risk of HIV #x4=propotion not circumcised
model
{ for(i in 1:N)
{ n_adj[i]<-n[i]/deff[i]
y_adj[i]<-p_est[i]*n_adj[i]
y_adj[i]~dbin(p[i],n_adj[i])
logit(p[i])<-beta[1] +beta[2]*x1[i]+beta[3]*x2[i]+beta[4]*x3[i]+
beta[5]*x4[i]+u[i] }
# CAR
u[1:N]~car.normal(adj[],weights[],num[],tau.u)
for(k in 1:sumNumNeigh) { weights[k]<-1}
for(j in 1:ncov) { beta[j]~dnorm(0.0,0.0001)}
tau.u~dgamma(0.1,0.0001)
sigma.u<-(1/tau.u)
}
#initialisation
list(beta=c(0,0,0,0,0),tau.u=0.01,u=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0,  0, 0, 0, 0, 0, 0, 0, 0, 0 ,0))
```

```
# data
#################################################################
###################CONVOLUTION MODEL#########################
####Variables#####
#x1= mean county age #x2= proportion with no education
#x3= Proportion perceiving not at risk of HIV #x4=propotion not circumcised
model
{ for(i in 1:N)
{n_adj[i]<-n[i]/deff[i]
y_adj[i]<-p_est[i]*n_adj[i]
y_adj[i]~dbin(p[i],n_adj[i])
logit(p[i])<-beta[1] +beta[2]*x1[i]+beta[3]*x2[i]+beta[4]*x3[i]+
beta[5]*x4[i]+u[i]+v[i]
v[i]~dnorm(0,tau.v) }
sigma.v<-(1/tau.v)
grandp<-sum(y_adj[])/sum(n_adj[])
for(i in 1:N) {RR[i]<-p[i]/grandp }
# CAR
u[1:N]~car.normal(adj[],weights[],num[],tau.u)
for(k in 1:sumNumNeigh) {weights[k]<-1}
#prior distribution for the current model
for(j in 1:ncov){beta[j]~dnorm(0.0,0.0001) }
tau.u~dgamma(0.1,0.0001)
tau.v~dgamma(0.1,0.0001)
sigma.u<-(1/tau.u)
}
#initialisation
list(beta=c(0,0,0,0,0),tau.u=0.01,tau.v=0.01,u=c(0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,  0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ,0),v=c(0, 0, 0, 0, 0, 0,   0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,   0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ,0))
# data
#################################################################
```

Appendix C

WinBUGS Codes for chapter Four Models

```
##########Simple logistic )##############################
model
{
#likelihood
for(i in 1: N)
{###Definition of Variables##
###circumcised=1, notcircumcised=1##
D.Circumcised[i]<-equals(Circumcised[i],2)
###urban=2, rural=1##
D.Urban[i]<-equals(Urban[i],2)
###None=1,primary=2,secondary=3,higher=4##
D.education2[i]<-equals(education[i],2); D.education3[i]<-equals(education[i],3)
D.education4[i]<-equals(education[i],4)
###Age=15-19,20-24,25-29,30-34,....60-64##
D.Age2[i]<-equals(Age[i],2); D.Age3[i]<-equals(Age[i],3)
D.Age4[i]<-equals(Age[i],4); D.Age5[i]<-equals(Age[i],5)
D.Age6[i]<-equals(Age[i],6); D.Age7[i]<-equals(Age[i],7)
D.Age8[i]<-equals(Age[i],8); D.Age9[i]<-equals(Age[i],9)
D.Age10[i]<-equals(Age[i],10)
herpes[i]~dbern(p1[i])
p1[i]<-min(1,max(0,pb[i]))
logit(pb[i])<-beta1+edu1[2]*D.education2[i]+
edu1[3]*D.education3[i]+edu1[4]*D.education4[i]+
 Circum1*D.Circumcised[i]+Urb1*D.Urban[i]+Age1[2]*D.Age2[i]+
```

```
 Age1[3]*D.Age3[i]+Age1[4]*D.Age4[i]+Age1[5]*D.Age5[i]+
Age1[6]*D.Age6[i]+Age1[7]*D.Age7[i]+Age1[8]*D.Age8[i]+
Age1[9]*D.Age9[i]+Age1[10]*D.Age10[i]
}
#nuissance parameters set to zero..to cater for reference levels
Age1[1]<-0; Age2[1]<-0; edu1[1]<-0; edu2[1]<-0
#prior for gamma random effects
#priors
beta1~dnorm(0,0.0001); circum1~dnorm(0,0.0001); Urb1~dnorm(0,0.0001)
#Age coefficients
for(k in 2:10) {Age1[k]~dnorm(0,0.0001) }
#Education coefficients
for(j in 2: 4) { edu1[j]~dnorm(0,0.0001) }
#ODDS ratios
#Education coefficients
for(j in 1: 4){ OREdu1[j]<-exp(edu1[j]) }
#Age coefficients
for(k in 1:10){ ORAge1[k]<-exp(Age1[k]) }
ORCircum1<-exp(Circum1); ORUrb1<-exp(Urb1)
for(i in 1: N)
{ for(j in 1: Nareas) {PH[j,i]<-(p1[i])*(equals(county[i],j)) }
}
for(j in 1: Nareas)
{ for(i in 1: N){count[j,i]<-equals(county[i],j) }
number[j]<-sum(count[j,])
PHIVC[j]<-sum(PH[j,])/number[j] }
}
#DATA
#INITIALS
list(beta1=0,Circum1=0,Urb1=0,Age1=c(NA,0,0,0,0,0,0,0,0,0),edu1=c(NA,0,0,0))
##########logistic convolution )##############################
model
{
#likelihood
for(i in 1: N)
{
###Definition of Variables##
###circumcised=1, notcircumcised=1##
D.Circumcised[i]<-equals(Circumcised[i],2)
###urban=2, rural=1##
D.Urban[i]<-equals(Urban[i],2)
###None=1,primary=2,secondary=3,higher=4##
D.education2[i]<-equals(education[i],2);D.education3[i]<-equals(education[i],3)
D.education4[i]<-equals(education[i],4)
###Age=15-19,20-24,25-29,30-34,....60-64##
D.Age2[i]<-equals(Age[i],2); D.Age3[i]<-equals(Age[i],3)
D.Age4[i]<-equals(Age[i],4); D.Age5[i]<-equals(Age[i],5)
```

```
D.Age6[i]<-equals(Age[i],6); D.Age7[i]<-equals(Age[i],7)
D.Age8[i]<-equals(Age[i],8); D.Age9[i]<-equals(Age[i],9)
D.Age10[i]<-equals(Age[i],10)
herpes[i]~dbern(p1[i])
p1[i]<-min(1,max(0,pb[i]))
logit(pb[i])<-beta1+edu1[2]*D.education2[i]+
edu1[3]*D.education3[i]+edu1[4]*D.education4[i]+
Circum1*D.Circumcised[i]+Urb1*D.Urban[i]+Age1[2]*D.Age2[i]+
Age1[3]*D.Age3[i]+Age1[4]*D.Age4[i]+Age1[5]*D.Age5[i]+
Age1[6]*D.Age6[i]+Age1[7]*D.Age7[i]+Age1[8]*D.Age8[i]+
Age1[9]*D.Age9[i]+Age1[10]*D.Age10[i]+U1[county[i]]+V1[county[i]]
}
#nuissance parameters set to zero..to cater for reference levels
Age1[1]<-0; Age2[1]<-0; edu1[1]<-0; edu2[1]<-0
#prior for gamma random effects
#priors
beta1~dnorm(0,0.0001); Circum1~dnorm(0,0.0001)
Urb1~dnorm(0,0.0001);
#Age coefficients
for(k in 2:10)
{ Age1[k]~dnorm(0,0.0001) }
#Education coefficients
for(j in 2: 4) { edu1[j]~dnorm(0,0.0001) }
#ODDS ratios
#Education coefficients
for(j in 1: 4) {ORedu1[j]<-exp(edu1[j])}
#Age coefficients
for(k in 1:10) { ORAge1[k]<-exp(Age1[k]) }
ORCircum1<-exp(Circum1); ORUrb1<-exp(Urb1)
omega.v1 ~ dgamma(0.1, 0.0001) ; omega.spatial1 ~ dgamma(0.1, 0.0001)
omega.v1sq<-1/omega.v1 ; omega.spatial1sq<-1/omega.spatial1
for(j in 1: Nareas) {  V1[j] ~ dnorm(0, omega.v1) }
U1[1:Nareas] ~ car.normal(adj[],weights1[],num[],omega.spatial1)
 for (k in 1:sumNumNeigh) { weights1[k] <- 1  }
for(i in 1: N)
{ for(j in 1: Nareas) {PH[j,i]<-(p1[i])*(equals(county[i],j)) }
}
for(j in 1: Nareas)
{ for(i in 1: N){count[j,i]<-equals(county[i],j)}
number[j]<-sum(count[j,])
PHIVC[j]<-sum(PH[j,])/number[j]
}
}
#DATA
#INITIALS
list(beta1=0,Circum1=0,Urb1=0,Age1=c(NA,0,0,0,0,0,0,0,0,0),
edu1=c(NA,0,0,0),omega.v1=0.01,omega.spatial1=0.01,
```

```
U1=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),V1=c(0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0))
########## spatial beta binomial )#############################
model
{
#likelihood
for(i in 1: N)
{
###Definition of Variables##
###circumcised=1, notcircumcised=1##
D.Circumcised[i]<-equals(Circumcised[i],2)
###urban=2, rural=1##
D.Urban[i]<-equals(Urban[i],2)
###None=1,primary=2,secondary=3,higher=4##
D.education2[i]<-equals(education[i],2); D.education3[i]<-equals(education[i],3)
D.education4[i]<-equals(education[i],4)
###Age=15-19,20-24,25-29,30-34,....60-64##
D.Age2[i]<-equals(Age[i],2); D.Age3[i]<-equals(Age[i],3)
D.Age4[i]<-equals(Age[i],4); D.Age5[i]<-equals(Age[i],5)
D.Age6[i]<-equals(Age[i],6); D.Age7[i]<-equals(Age[i],7)
D.Age8[i]<-equals(Age[i],8); D.Age9[i]<-equals(Age[i],9)
D.Age10[i]<-equals(Age[i],10)
herpes[i]~dbern(p1[i])
p1[i]<-min(1,max(0,pb[i]))
pb[i]<-theta[county[i]]*omega[i]
logit(omega[i])<-beta1+edu1[2]*D.education2[i]+
edu1[3]*D.education3[i]+edu1[4]*D.education4[i]+
 Circum1*D.Circumcised[i]+Urb1*D.Urban[i]+Age1[2]*D.Age2[i]+
 Age1[3]*D.Age3[i]+Age1[4]*D.Age4[i]+Age1[5]*D.Age5[i]+
Age1[6]*D.Age6[i]+Age1[7]*D.Age7[i]+Age1[8]*D.Age8[i]+
Age1[9]*D.Age9[i]+Age1[10]*D.Age10[i]+U1[county[i]]+V1[county[i]]
}
#nuissance parameters set to zero..to cater for reference levels
Age1[1]<-0; Age2[1]<-0; edu1[1]<-0; edu2[1]<-0
#prior for gamma random effects
a~dunif(3,5); b~dunif(1.1,1.5)
#priors
beta1~dnorm(0,0.0001);  Circum1~dnorm(0,0.0001)
Urb1~dnorm(0,0.0001)
#Age coefficients
for(k in 2:10) { Age1[k]~dnorm(0,0.0001) }
#Education coefficients
for(j in 2: 4) { edu1[j]~dnorm(0,0.0001) }
#ODDS ratios
#Education coefficients
```

```
for(j in 1: 4) { ORedu1[j]<-exp(edu1[j])}
#Age coefficients
for(k in 1:10) { ORAge1[k]<-exp(Age1[k]) }
ORCircum1<-exp(Circum1); ORUrb1<-exp(Urb1)
omega.v1 ~ dgamma(0.1, 0.0001); omega.spatial1 ~ dgamma(0.1, 0.0001)
omega.v1sq<-1/omega.v1; omega.spatial1sq<-1/omega.spatial1
for(j in 1: Nareas) {
 V1[j] ~ dnorm(0, omega.v1) ;theta[j]~dbeta(a,b) }
U1[1:Nareas] ~ car.normal(adj[],weights1[], num[],omega.spatial1)
 for (k in 1:sumNumNeigh) {   weights1[k] <- 1  }
for(i in 1: N)
{ for(j in 1: Nareas) {PH[j,i]<-(p1[i])*(equals(county[i],j)) }
}
for(j in 1: Nareas)
{ for(i in 1: N) { count[j,i]<-equals(county[i],j) }
number[j]<-sum(count[j,]); PHIVC[j]<-sum(PH[j,])/number[j]
}
}
#DATA
#INITIALS
list(a=4,b=1.2,beta1=0,Circum1=0,Urb1=0, Age1=c(NA,0,0,0,0,0,0,0,0,0),
edu1=c(NA,0,0,0),omega.spatial1=0.01,omega.v1=0.01,
U1=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
V1=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
###########logistic beta binomialrandom effects)############
model
{
#likelihood
for(i in 1: N)
{
###Definition of Variables##
###circumcised=1, notcircumcised=1##
D.Circumcised[i]<-equals(Circumcised[i],2)
###urban=2, rural=1##
D.Urban[i]<-equals(Urban[i],2)
###None=1,primary=2,secondary=3,higher=4##
D.education2[i]<-equals(education[i],2);D.education3[i]<-equals(education[i],3)
D.education4[i]<-equals(education[i],4)
###Age=15-19,20-24,25-29,30-34,....60-64##
D.Age2[i]<-equals(Age[i],2);D.Age3[i]<-equals(Age[i],3)
D.Age4[i]<-equals(Age[i],4);D.Age5[i]<-equals(Age[i],5)
D.Age6[i]<-equals(Age[i],6);D.Age7[i]<-equals(Age[i],7)
D.Age8[i]<-equals(Age[i],8);D.Age9[i]<-equals(Age[i],9)
D.Age10[i]<-equals(Age[i],10)
herpes[i]~dbern(p1[i])
```

```
p1[i]<-min(1,max(0,pb[i]))
pb[i]<-theta[i]*omega[i]
theta[i]~dbeta(a,b)
logit(omega[i])<-beta1+edu1[2]*D.education2[i]+
edu1[3]*D.education3[i]+edu1[4]*D.education4[i]+
Circum1*D.Circumcised[i]+Urb1*D.Urban[i]+Age1[2]*D.Age2[i]+
Age1[3]*D.Age3[i]+Age1[4]*D.Age4[i]+Age1[5]*D.Age5[i]+
Age1[6]*D.Age6[i]+Age1[7]*D.Age7[i]+Age1[8]*D.Age8[i]+
Age1[9]*D.Age9[i]+Age1[10]*D.Age10[i]+V1[county[i]]
}
#nuissance parameters set to zero..to cater for reference levels
Age1[1]<-0; Age2[1]<-0; edu1[1]<-0; edu2[1]<-0
#prior for gamma random effects
a~dunif(3,5); b~dunif(1.1,1.5)
#priors
beta1~dnorm(0,0.0001); Circum1~dnorm(0,0.0001)
Urb1~dnorm(0,0.0001);
#Age coefficients
for(k in 2:10) { Age1[k]~dnorm(0,0.0001) }
#Education coefficients
for(j in 2: 4) { edu1[j]~dnorm(0,0.0001) }
#ODDS ratios
#Education coefficients
for(j in 1: 4) { ORedu1[j]<-exp(edu1[j]) }
#Age coefficients
for(k in 1:10) { ORAge1[k]<-exp(Age1[k])}
ORCircum1<-exp(Circum1); ORUrb1<-exp(Urb1)
omega.v1 ~ dgamma(0.1, 0.0001)
# omega.spatial1 ~ dgamma(0.1, 0.0001)
omega.v1sq<-1/omega.v1
# omega.spatial1sq<-1/omega.spatial1
for(j in 1: Nareas) { V1[j] ~ dnorm(0, omega.v1)  }
#U1[1 : Nareas] ~ car.normal(adj[], weights1[], num[], omega.spatial1)
 #for (k in 1:sumNumNeigh) {
       #        weights1[k] <- 1
    #        }
for(i in 1: N)
{ for(j in 1: Nareas)
{ PH[j,i]<-(p1[i])*(equals(county[i],j)) }
}
for(j in 1: Nareas)
{ for(i in 1: N){ count[j,i]<-equals(county[i],j) }
number[j]<-sum(count[j,])
PHIVC[j]<-sum(PH[j,])/number[j]
}
}
#DATA
```

```
#INITIALS
list(a=4,b=1.2,beta1=0,Circum1=0,Urb1=0,
Age1=c(NA,0,0,0,0,0,0,0,0,0),edu1=c(NA,0,0,0),
omega.v1=0.01,V1=c(0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0))
##########Combine spatial beta binomial )##############
model
{
#likelihood
for(i in 1: N)
{
###Definition of Variables##
###circumcised=1, notcircumcised=1##
D.Circumcised[i]<-equals(Circumcised[i],2)
###urban=2, rural=1##
D.Urban[i]<-equals(Urban[i],2)
###None=1,primary=2,secondary=3,higher=4##
D.education2[i]<-equals(education[i],2);D.education3[i]<-equals(education[i],3)
D.education4[i]<-equals(education[i],4)
###Age=15-19,20-24,25-29,30-34,....60-64##
D.Age2[i]<-equals(Age[i],2); D.Age3[i]<-equals(Age[i],3)
D.Age4[i]<-equals(Age[i],4); D.Age5[i]<-equals(Age[i],5)
D.Age6[i]<-equals(Age[i],6); D.Age7[i]<-equals(Age[i],7)
D.Age8[i]<-equals(Age[i],8); D.Age9[i]<-equals(Age[i],9)
D.Age10[i]<-equals(Age[i],10)
herpes[i]~dbern(p1[i])
p1[i]<-min(1,max(0,pb[i]))
pb[i]<-theta[county[i]]*omega[i]
logit(omega[i])<-beta1+edu1[2]*D.education2[i]+
edu1[3]*D.education3[i]+edu1[4]*D.education4[i]+
Circum1*D.Circumcised[i]+Urb1*D.Urban[i]+Age1[2]*D.Age2[i]+
Age1[3]*D.Age3[i]+Age1[4]*D.Age4[i]+Age1[5]*D.Age5[i]+
Age1[6]*D.Age6[i]+Age1[7]*D.Age7[i]+Age1[8]*D.Age8[i]+
Age1[9]*D.Age9[i]+Age1[10]*D.Age10[i]+U1[county[i]]+V1[county[i]]
}
#nuissance parameters set to zero..to cater for reference levels
Age1[1]<-0; Age2[1]<-0; edu1[1]<-0; edu2[1]<-0
#prior for gamma random effects
a~dunif(3,5); b~dunif(1.1,1.5)
#priors
beta1~dnorm(0,0.0001);  Circum1~dnorm(0,0.0001)
Urb1~dnorm(0,0.0001)
#Age coefficients
for(k in 2:10) { Age1[k]~dnorm(0,0.0001) }
#Education coefficients
for(j in 2: 4) { edu1[j]~dnorm(0,0.0001)}
```

```
#ODDS ratios
#Education coefficients
for(j in 1: 4) { ORedu1[j]<-exp(edu1[j]) }
#Age coefficients
for(k in 1:10) { ORAge1[k]<-exp(Age1[k]) }
ORCircum1<-exp(Circum1); ORUrb1<-exp(Urb1)
omega.v1 ~ dgamma(0.1, 0.0001); omega.spatial1 ~ dgamma(0.1, 0.0001)
omega.v1sq<-1/omega.v1;  omega.spatial1sq<-1/omega.spatial1
for(j in 1: Nareas) { V1[j] ~ dnorm(0, omega.v1)
theta[j]~dbeta(a,b) }
U1[1 :Nareas] ~ car.normal(adj[],weights1[],num[],omega.spatial1)
 for (k in 1:sumNumNeigh) {   weights1[k] <- 1   }
for(i in 1: N)
{ for(j in 1: Nareas) {PH[j,i]<-(p1[i])*(equals(county[i],j)) }
}
for(j in 1: Nareas)
{ for(i in 1: N){ count[j,i]<-equals(county[i],j) }
number[j]<-sum(count[j,])
PHIVC[j]<-sum(PH[j,])/number[j]
}
}
#DATA
#INITIALS
list(a=4,b=1.2,beta1=0,Circum1=0,Urb1=0,
Age1=c(NA,0,0,0,0,0,0,0,0,0),edu1=c(NA,0,0,0),
omega.spatial1=0.01,omega.v1=0.01,
U1=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
V1=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
##### spatial beta binomial effects version random beta)#####
model
{
#likelihood
for(i in 1: N)
{
###Definition of Variables##
###circumcised=1, notcircumcised=1##
D.Circumcised[i]<-equals(Circumcised[i],2)
###urban=2, rural=1##
D.Urban[i]<-equals(Urban[i],2)
###None=1,primary=2,secondary=3,higher=4##
D.education2[i]<-equals(education[i],2)
D.education3[i]<-equals(education[i],3)
D.education4[i]<-equals(education[i],4)
###Age=15-19,20-24,25-29,30-34,....60-64##
D.Age2[i]<-equals(Age[i],2); D.Age3[i]<-equals(Age[i],3)
```

```
D.Age4[i]<-equals(Age[i],4); D.Age5[i]<-equals(Age[i],5)
D.Age6[i]<-equals(Age[i],6); D.Age7[i]<-equals(Age[i],7)
D.Age8[i]<-equals(Age[i],8); D.Age9[i]<-equals(Age[i],9)
D.Age10[i]<-equals(Age[i],10)
herpes[i]~dbern(p1[i])
p1[i]<-min(1,max(0,pb[i]))
pb[i]<-theta[county[i]]*omega[i]
logit(omega[i])<-beta1+edu1[2]*D.education2[i]+
edu1[3]*D.education3[i]+edu1[4]*D.education4[i]+
Circum1*D.Circumcised[i]+Urb1*D.Urban[i]+Age1[2]*D.Age2[i]+
Age1[3]*D.Age3[i]+Age1[4]*D.Age4[i]+Age1[5]*D.Age5[i]+
Age1[6]*D.Age6[i]+Age1[7]*D.Age7[i]+Age1[8]*D.Age8[i]+
Age1[9]*D.Age9[i]+Age1[10]*D.Age10[i]+U1[county[i]]+V1[county[i]]
}
#nuissance parameters set to zero..to cater for reference levels
Age1[1]<-0; Age2[1]<-0; edu1[1]<-0; edu2[1]<-0
#prior for gamma random effects
#a~dunif(3,5)
#b~dunif(1.1,1.5)
#priors
beta1~dnorm(0,0.0001);   Circum1~dnorm(0,0.0001)
Urb1~dnorm(0,0.0001);
#Age coefficients
for(k in 2:10) { Age1[k]~dnorm(0,0.0001) }
#Education coefficients
for(j in 2: 4) { edu1[j]~dnorm(0,0.0001) }
#ODDS ratios
#Education coefficients
for(j in 1: 4) { ORedu1[j]<-exp(edu1[j]) }
#Age coefficients
for(k in 1:10) { ORAge1[k]<-exp(Age1[k]) }
ORCircum1<-exp(Circum1); ORUrb1<-exp(Urb1)
omega.v1 ~ dgamma(0.1, 0.0001); omega.spatial1 ~ dgamma(0.1, 0.0001)
omega.v2 ~ dgamma(0.1, 0.0001); omega.spatial2 ~ dgamma(0.1, 0.0001)
omega.v3 ~ dgamma(0.1, 0.0001) ; omega.spatial3 ~ dgamma(0.1, 0.0001)
omega.v1sq<-1/omega.v1 ; omega.spatial1sq<-1/omega.spatial1
omega.v2sq<-1/omega.v2 ; omega.spatial2sq<-1/omega.spatial2
omega.v3sq<-1/omega.v3; omega.spatial3sq<-1/omega.spatial3
for(j in 1: Nareas)
{
V1[j] ~ dnorm(0, omega.v1)
                    V2[j] ~ dnorm(0, omega.v1)
                    V3[j] ~ dnorm(0, omega.v1)
theta[j]~dbeta(a1[j],b1[j])
a1[j]<-max(1.01,a[j])
b1[j]<-max(1.01,b[j])
a[j]<-mu[j]*psi[j]
```

```
b[j]<-psi[j] - psi[j]*mu[j]
logit(mu[j])<-U2[j]+V2[j]
psi[j]<-exp(U3[j]+V3[j])
}
U1[1:Nareas] ~ car.normal(adj[], weights1[], num[], omega.spatial1)
U2[1:Nareas] ~ car.normal(adj[], weights2[], num[], omega.spatial2)
U3[1:Nareas] ~ car.normal(adj[], weights3[], num[], omega.spatial3)
 for (k in 1:sumNumNeigh) {
               weights1[k] <- 1 ;weights2[k] <- 1; weights3[k] <- 1
            }
for(i in 1: N)
{
for(j in 1: Nareas) {PH[j,i]<-(p1[i])*(equals(county[i],j)) }
}
for(j in 1: Nareas)
{
for(i in 1: N) { count[j,i]<-equals(county[i],j) }
number[j]<-sum(count[j,])
PHIVC[j]<-sum(PH[j,])/number[j]
}
}
#DATA
#INITIALS
list(beta1=0,Circum1=0,Urb1=0,Age1=c(NA,0,0,0,0,0,0,0,0,0),
edu1=c(NA,0,0,0),omega.spatial1=0.01,omega.spatial2=0.01,
omega.spatial3=0.01,omega.v3=0.01,omega.v2=0.01,
omega.v1=0.01,U1=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0),U2=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
U3=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),V1=c(0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0),V2=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
V3=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
```

Appendix D

WinBUGS Codes for chapter Five Models

```
################################################
########## Separate Analyses  #################
################################################
model
{
#likelihood
for(i in 1: N)
{#N=3759
#for HIV
hiv[i]~dbern(p1[i])
p1[i]<-min(1,max(0,PHIV[i]))
logit(PHIV[i])<-beta1[1]+beta1[2]*education[i]+
 beta1[3]*Circumcised[i]
+beta1[5]*Urban[i]+beta1[5]*Age[i]
#for herpes
herpes[i]~dbern(p2[i])
p2[i]<-min(1,max(0,PHRP[i]))
logit(PHRP[i])<-beta2[1]+beta2[2]*education[i]+
 beta2[3]*Circumcised[i]+
beta2[4]*Urban[i]+beta2[5]*Age[i]
y11[i]<-hiv[i]
y11[i]~dbern(p11[i])
p1[i]<-min(1,max(0,PHIV[i]))
logit(PHIV[i])<-beta1[1]+beta1[2]*education[i]+
```

```
 beta1[3]*Circumcised[i]+
beta1[5]*Urban[i]+beta1[5]*Age[i]
#for herpes
herpes[i]~dbern(p2[i])
p2[i]<-min(1,max(0,PHRP[i]))
logit(PHRP[i])<-beta2[1]+beta2[2]*education[i]+
 beta2[3]*Circumcised[i]+
beta2[4]*Urban[i]+beta2[5]*Age[i]
}
#Getting Odds ratio from logOdds, by taking exponent of the coefficients
for(i in 2:8){ Oddsbeta1[i]<-exp(beta1[i]); Oddsbeta2[i]<-exp(beta2[i])}
#prior
for(j in 1: 5)
{ beta1[j]~dnorm(0,0.0001); beta2[j]~dnorm(0,0.0001) }
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]
}
}
#DATA
#INITIALS
list(beta1=c(0,0,0,0,0),beta2=c(0,0,0,0,0))
########################################
#########Multivariate CAR model##########
########################################
model
{
#likelihood
for(i in 1: N)
{#N=3759
#for HIV
HIV[i]~dbern(p1[i])
p1[i]<-min(1,max(0,PHIV[i]))
logit(PHIV[i])<-beta1[1]+beta1[2]*perceived_Risk[i]+
beta1[3]*Ever_used_condom[i]+beta1[4]*Circumcised[i]+
beta1[5]*Urban[i]+beta1[6]*Age[i]+beta1[7]*away[i]+
```

```
beta1[8]*Sex_resp2[i]+S[1,county[i]]
#for herpes
herpes[i]~dbern(p2[i])
p2[i]<-min(1,max(0,PHRP[i]))
logit(PHRP[i])<-beta2[1]+beta2[2]*perceived_Risk[i]+
beta2[3]*Ever_used_condom[i]+beta2[4]*Circumcised[i]+
beta2[5]*Urban[i]+beta2[6]*Age[i]+beta2[7]*away[i]+
beta2[8]*Sex_resp2[i]+S[2,county[i]]
}
#Getting Odds ratio from logOdds, by taking exponent of the coefficients
for(i in 2:8){ Oddsbeta1[i]<-exp(beta1[i]); Oddsbeta2[i]<-exp(beta2[i])}
  # 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, ] (oral cancer)
            sigma[1] <- sqrt(sigma2[1, 1])
         # conditional SD of S[2,] (lung cancer)
         sigma[2] <- sqrt(sigma2[2, 2])
         # within-area conditional correlation
            corr <- sigma2[1, 2] / (sigma[1] * sigma[2])
         # between oral and lung cancers.
              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(j in 1: 8){ beta1[j]~dnorm(0,0.0001); beta2[j]~dnorm(0,0.0001)}
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]
}
}
#DATA
#INITIALS
list(beta1=c(0,0,0,0,0,0,0,0),beta2=c(0,0,0,0,0,0,0,0),
```

```
omega=structure(.Data=c(1,1,1,1), .Dim=c(2,2)))
##############################################
###      Shared Component model ##############
##############################################
model
{
#likelihood
for(i in 1: N)
{
#N=3759
#for HIV
hiv[i]~dbern(p1[i])
p1[i]<-min(1,max(0,PHIV[i]))
logit(PHIV[i])<-beta1[1]+beta1[2]*perceived_Risk[i]+
 beta1[3]*Ever_used_condom[i]+
beta1[4]*Circumcised[i]+beta1[5]*Urban[i]+
beta1[6]*Age[i]+beta1[7]*away[i]+
beta1[8]*Sex_resp2[i]+S[2,county[i]]
#for herpes
herpes[i]~dbern(p2[i])
p2[i]<-min(1,max(0,PHRP[i]))
logit(PHRP[i])<-beta2[1]+beta2[2]*perceived_Risk[i]+
beta2[3]*Ever_used_condom[i]+beta2[4]*Circumcised[i]+
beta2[5]*Urban[i]+beta2[6]*Age[i]+beta2[7]*away[i]+
beta2[8]*Sex_resp2[i]+S[2,county[i]]
}

#Getting Odds ratio from logOdds, by taking exponent of the coefficients
for(i in 2:8)
{
Oddsbeta1[i]<-exp(beta1[i]); Oddsbeta2[i]<-exp(beta2[i])
}
         for(i in 1:Nareas)
{
# Define log relative risk in terms of disease-specific
#(psi) and shared (phi)
 # random effects
 # changed order of k and i index for psi
 #(needed because car.normal assumes
 # right hand index is areas)
                S[1, i] <- phi[i] * delta  + psi[1, i]
                S[2, i] <- phi[i] / delta  + psi[2, i]
        }
# Spatial priors (BYM) for the disease-specific random effects
              for (k in 1 : Ndiseases) {
               for (i in 1 : Nareas) {
# convolution prior = sum of unstructured and spatial effects
```

```
                    psi[k, i] <- U.sp[k, i] + S.sp[k, i]
# unstructured disease-specific random effects
                    U.sp[k, i] ~ dnorm(0, tau.unstr[k])
                  }
 # spatial disease-specific effects
S.sp[k,1:Nareas] ~ car.normal(adj[],weights[], num[],tau.spatial[k])
               }
  # Spatial priors (BYM) for the shared random effects
              for (i in 1:Nareas) {
  # convolution prior = sum of unstructured and spatial effects
                phi[i] <- U.sh[i] + S.sh[i]
  # unstructured shared random effects
                U.sh[i] ~ dnorm(0, omega.unstr)
             }
  # spatial shared random effects
S.sh[1:Nareas] ~ car.normal(adj[], weights[], num[], omega.spatial)
            for (k in 1:sumNumNeigh) {   weights[k] <- 1    }
#prior
for(j in 1: 8){ beta1[j]~dnorm(0,0.0001); beta2[j]~dnorm(0,0.0001)}
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]
}
# Other priors
              for (k in 1:Ndiseases) {
                    tau.unstr[k] ~ dgamma(0.1, 0.0001)
                    tau.spatial[k] ~ dgamma(0.1, 0.0001)
                  }
                  omega.unstr ~ dgamma(0.1, 0.0001)
                  omega.spatial ~ dgamma(0.1, 0.0001)
 # scaling factor for relative strength of shared component
 #for each disease
            logdelta ~ dnorm(0, 5.9)
```

```
# (prior assumes 95% probability that delta^2 is between 1/5 and 5;
               delta <- exp(logdelta)
# lognormal assumption is invariant to which disease is labelled 1
         # and which is labelled 2)
 # ratio (relative risk of disease 1 associated with shared component)
 # to (relative risk of disease 2 associated with shared component)
# # - see Knorr-Held and Best (2001) for further details
# # RR.ratio <- pow(delta, 2)
#Mapping issues
# Relative risks and other summary quantities
# The GeoBUGS map tool can only map vectors, so need to create
#separate vector  of quantities to be mapped, rather than an
#array (i.e. totalRR[i,k] won't work!)
}
#DATA
#INITIALS
list(beta1=c(0,0,0,0,0,0,0,0),beta2=c(0,0,0,0,0,0,0,0),
omega.unstr=0.01,omega.spatial=0.01,logdelta=2,
tau.unstr=c(0.01,0.01),tau.spatial=c(0.01,0.01))
```

Bibliography

M. Rezaeian, G. Dunn, S. St Leger, and L. Appleby. Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. *J Epidemiol Commun H*, 61(2):98–102, 2007.

J. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183, 2007.

M Ghosh, K Natarajan, L.A Waller, and D Kim. Hierarchical bayes glms for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference*, 75(2):305–318, 1999.

A.B. Lawson, W.J. Browne, C.L.V. Rodeiro, and J. Wiley. *Disease mapping with WinBUGS and MLwiN*. Wiley Online Library, 2003.

L.A. Waller and C.A. Gotway. *Applied spatial statistics for public health data*, volume 368. Wiley-Interscience, 2004.

C. Gaetan and X. Guyon. *Spatial Statistics and Modeling, Series in Statistics*. Springer, 2010.

M. Sherman. *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Wiley, 2011.

N.A.C. Cressie. *Statistics for Spatial Data*. Wiley-Interscience, 1993.

P.J Diggle, J.A Tawn, and R.A Moyeed. Model-based geostatistics. *J Roy Stat Soc C-app*, 47(3):299–350, 1998.

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

J. Besag, J. York, and A. Mollie. Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann Inst Stat Math.*, 43:1–59, 1991.

N.G Best, A Thomas, L.A Waller, E.M Conlon, and R.A Arnold. Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, volume 6, pages 131–156. Oxford University Press, USA, 1999.

H. Lu, C.S Reilly, S. Banerjee, and B.P Carlin. Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics*, 14(4):433–452, 2007.

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

B.G Leroux, X. Lei, and N. Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. *Statistical Models in Epidemiology, the Environment and Clinical Trials*, 116:179–192, 1999.

Y.C MacNab and C.B Dean. Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models. *Statistics in medicine*, 19(17-18): 2421–2435, 2000.

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

D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. Winbugs user manual version 1.4 january 2003. upgraded to version 1.4.3, 2007.

Core RDevelopment. R: A language and environment for statistical computing, 2005.

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

D.K. Agarwal, A.E Gelfand, and S. Citron-Pousty. Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9(4): 341–355, 2002.

D.K. Agarwal. Two-fold spatial zero-inflated models for analysing isopod settlement patterns. *Bayesian Statistics and its Applications*, 2006.

Y.C MacNab. On bayesian shared component disease mapping and ecological regression with errors in covariates. *Statistics in medicine*, 29(11):1239–1249, 2010.

L Bernadinelli, Cristian Pascutto, NG Best, and WR Gilks. Disease mapping with errors in covariates. *Statistics in Medicine*, 16(7):741–752, 1997.

L. Knorr-Held and N.G. Best. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):73–85, 2001.

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

F. Wang and M. Wall. Generalized common spatial factor model. *Biostatistics*, 4 (4):569–582, 2003.

H. Kim, D. Sun, and R.K. Tsutakawa. A bivariate bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*, 96(456):1506–1521, 2001.

B.P. Carlin and S. Banerjee. Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian statistics*, 7:45–64, 2003.

A.E. Gelfand and P. Vounatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15, 2003.

X. Jin, B.P Carlin, and S. Banerjee. Generalized hierarchical multivariate car models for areal data. *Biometrics*, 61(4):950–961, 2005.

B.J Reich, J.S. Hodges, and B.P. Carlin. Spatial analyses of periodontal data using conditionally autoregressive priors having two classes of neighbor relations. *Journal of the American Statistical Association*, 102(477):44–55, 2007.

T Hastie and R Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993.

D.R. Hoover, J.A. Rice, C.O Wu, and L. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85 (4):809–822, 1998.

R.M. Assuncao, J.E. Potter, and S.M. Cavenaghi. A bayesian space varying parameter model applied to estimating fertility schedules. *Statistics in Medicine*, 21(14):2057–2075, 2002.

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

A.D Pavlov. Space-varying regression coefficients: a semi-parametric approach applied to real estate markets. *Real Estate Economics*, 28(2):249–283, 2003.

D. Gamerman, A.R.B Moreira, and H. Rue. Space-varying regression models: specifications and simulation. *Computational Statistics & Data Analysis*, 42(3): 513–533, 2003.

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

J.A. Hoeting, M. Leecaster, and D. Bowden. An improved model for spatially correlated binary responses. *Journal of agricultural, biological, and environmental statistics*, 5(1):102–114, 2000.

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

D.A Griffith. A spatial filtering specification for the auto-poisson model. *Statistics & Probability Letters*, 58(3):245–251, 2002.

Peter Diggle and Paulo Justiniano Ribeiro. *Model-based geostatistics*. Springer, 2007.

B.S Everitt and G. Dunn. *Applied Multivariate Analysis. 2001*. Arnold, London, 2011.

D. Clayton and L. Bernardinelli. *Bayesian methods for mapping disease risk: Geographical and environmental epidemiology, methods for small-area studies*. Oxford University Press, Oxford, 1992.

M. Smans and J. Esteve. *Pratical approaches to disease mapping: Geographical and Environmental Epidemiology, Methods for Small area studies*. Oxford University Press, 1997.

J.C Wakefield, N.G Best, and L Waller. *Bayesian approaches to disease mapping, Spatial epidemiology: methods and applications*. Oxford: Oxford University Press, 2000.

S.M Manda, R.G Feltbower, and M.S Gilthorpe. Review and empirical comparison of joint mapping of multiple diseases. *Southern African Journal of Epidemiology and Infection*, 27(4):169–182, 2011.

G.E. Box and G.C. Tiao. *Bayesian inference in statistical analysis*. Addison-Wesley Pub. Co., 1973.

S. Nadarajah. A generalized normal distribution. *Journal of Applied Statistics*, 32 (7):685–694, 2005.

I. Ntzoufras. *Bayesian modeling using WinBUGS*, volume 698. Wiley, 2011.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, 2003.

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

C.E McCulloch and J.M Neuhaus. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science*, 26(3): 388–402, 2011.

S. Litière, A. Alonso, and G. Molenberghs. Type i and type ii error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63(4): 1038–1044, 2007.

S. Litière, A. Alonso, and G. Molenberghs. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in medicine*, 27(16):3125–3144, 2008.

L.S Magder and S.L Zeger. A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association*, 91(435):1141–1151, 1996.

G. Verbeke and E. Lesaffre. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23(4):541–556, 1997.

D. Zhang and M. Davidian. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57(3):795–802, 2001.

R.K.W Ho and I. Hu. Flexible modelling of random effects in linear mixed modelsa bayesian approach. *Computational Statistics & Data Analysis*, 52(3):1347–1361, 2008.

A. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J.F. Viel, R. Bertollini, et al. *Disease mapping and risk assessment for public health.* John Wiley & Sons, 1999.

A.D Cliff. Analysing geographically-related disease data. *Statistical methods in medical research*, 4(2):93–101, 1995.

L.N. Kazembe, I. Kleinschmidt, T.H. Holtz, and B.L. Sharp. Spatial analysis and mapping of malaria risk in malawi using point-referenced prevalence of infection data. *International Journal of Health Geographics*, 5(1):41, 2006.

S.O.M. Manda, R.G. Feltbower, and M.S. Gilthorpe. Investigating spatio-temporal similarities in the epidemiology of childhood leukaemia and diabetes. *Eur J Epidemiol*, 24(12):743–752, 2009.

M.M. Ngigi. *A Geographical study on the HIV/AIDS pandemic in Kenya.* PhD thesis, University of Tsukuba, 2007.

L. Montana, M. Neuman, and V. Mishra. Spatial modeling of hiv prevalence in kenya. In *Demographic and Health Research*, 2007.

C.X. Chen, T. Lumley, and J. Wakefield. The use of sampling weights in bayesian hierarchical models for small area estimation. Technical report, University of Washington, 2012.

L. Kish. Methods for design effects. *Journal of Official Statistics*, 11:55–55, 1995.

C.E. McCulloch and J.M. Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2005.

Republic of Kenya NASCOP. Kenya aids indicator survey 2007 report, 2008.

L. Montana, M. Neuman, V. Mishra, and R. Hong. Spatial modeling of hiv prevalence in cameroon, kenya, and tanzania. In *Population Association of America Annual Conference*, 2005.

B. Cheluget, G. Baltazar, P. Orege, M. Ibrahim, LH Marum, and J. Stover. Evidence for population level declines in adult hiv prevalence in kenya. *Sex Transm Infect*, 82(suppl 1):i21–i26, 2006.

K. Johnson and A. Way. 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, 2006.

R.C. Bailey, S. Moses, C.B. Parker, K. Agot, I. Maclean, J.N. Krieger, C.F.M. Williams, R.T. Campbell, and J.O. Ndinya-Achola. Male circumcision for hiv

prevention in young men in kisumu, kenya: a randomised controlled trial. *The Lancet*, 369(9562):643–656, 2007.

H.A. Weiss, M.A. Quigley, and R.J. Hayes. Male circumcision and risk of hiv infection in sub-saharan africa: a systematic review and meta-analysis. *Aids*, 14 (15):2361, 2000.

Y. You and Q.M. Zhou. Hierarchical bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, 37:25–37, 2011.

A. Mollie, W.R Gilks, S. Richardson, and D.J Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman Hall Nueva York, 1996.

A. Gelman. Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164, 2007.

S.G Heeringa, B.T West, and P.A Berglund. *Applied survey data analysis*. Chapman & Hall/CRC, 2010.

W.H DuMouchel and G.J Duncan. Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78(383):535–543, 1983.

D. Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2):317–337, 1993.

R.E Fay and R.A Herriot. Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

J.N.K Rao. *Small area estimation*, volume 331. John Wiley and Sons, 2005.

S.E Fienberg. The relevance or irrelevance of weights for confidentiality and statistical analyses. *Journal of Privacy and Confidentiality*, 1(2):183–195, 2009.

T. Oluoch, I. Mohammed, R. Bunnell, R. Kaiser, A.A. Kim, A. Gichangi, M. Mwangi, S. Dadabhai, L. Marum, A. Orago, et al. Correlates of hiv infection among sexually active adults in kenya: A national population-based survey. *The open AIDS journal*, 5:125, 2011.

M. Cusini and M. Ghislanzoni. The importance of diagnosing genital herpes. *Journal of Antimicrobial Chemotherapy*, 47(suppl 1):9–16, 2001.

B. Aumakhan, C.A. Gaydos, T.C. Quinn, C. Beyrer, L. Benning, H. Minkoff, D.J. Merenstein, M. Cohen, R. Greenblatt, M. Nowicki, et al. Clinical reactivations of herpes simplex virus type 2 infection and human immunodeficiency virus disease progression markers. *PloS one*, 5(4):e9973, 2010.

J.E. Horbul, S.C. Schmechel, B.R.L Miller, S.A. Rice, and P.J. Southern. Herpes simplex virus-induced epithelial damage and susceptibility to human immunodeficiency virus type 1 infection in human cervical organ culture. *PloS one*, 6(7): e22638, 2011.

J.G Beauman. Genital herpes: a review. *American family physician*, 72(8):1527, 2005.

W. E. Lafferty, R. W. Coombs, J. Benedetti, C. Critchlow, and L. Corey. Recurrences after oral and genital herpes simplex virus infection. *New England Journal of Medicine*, 316(23):1444–1449, 1987.

C. Celum, A. Wald, J. Hughes, J. Sanchez, S. Reid, S. Delany-Moretlwe, F. Cowan, M. Casapia, A. Ortiz, J. Fuchs, et al. Effect of aciclovir on hiv-1 acquisition in herpes simplex virus 2 seropositive women and men who have sex with men: a randomised, double-blind, placebo-controlled trial. *Lancet*, 371(9630):2109–2119, 2008.

D. Watson-Jones, H.A Weiss, M. Rusizoka, J. Changalucha, K. Baisley, K. Mugeye, C. Tanton, D. Ross, D. Everett, T. Clayton, et al. Effect of herpes simplex suppression on incidence of hiv among women in tanzania. *New England Journal of Medicine*, 358(15):1560–1571, 2008.

O. Ngesa, T. Achia, and H. Mwambi. Spatial joint disease modeling and mapping with application to hiv and hsv-2. In *Proceedings of the 55th Annual Conference of the South African Statistical Association*, pages 61–68. South African Statistical Association, 2013.

J. Hinde and C.G.B Demétrio. Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170, 1998.

G. Molenberghs, G. Verbeke, and C.G.B Demetrio. An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis*, 13(4):513–531, 2007.

G. Molenberghs, G. Verbeke, C.G.B Demétrio, and A. Vieira. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, 25(3):325–347, 2010.

G. Molenberghs, G. Verbeke, S. Iddi, and C.G.B Demétrio. A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis*, 111:94–109, 2012.

W. Kassahun, T. Neyens, G. Molenberghs, C. Faes, and G. Verbeke. Modeling overdispersed longitudinal binary data using a combined beta and normal random-effects model. *Archives of Public Health*, 70(1):1–13, 2012.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

WHO and UNICEF. Unaids. global hiv/aids responseepidemic update and health sector progress towards universal accessprogress report 2011, 2012.

L.N. Kazembe and J.J. Namangale. A bayesian multinomial model to analyse spatial patterns of childhood co-morbidity in malawi. *Eur J Epidemiol*, 22(8): 545–556, 2007.

S.O.M. Manda, C.J. Lombard, and T. Mosala. Divergent spatial patterns in the prevalence of the human immunodeficiency virus (hiv) and syphilis in south african pregnant women. *Geospatial Health*, 6(2):221–231, 2012.

DA Harville. *Matrix Algebra from a Statistician's Perspective*. Springer, New York, 1997.

K.V Mardia. Multi-dimensional multivariate gaussian markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284, 1988.

N. Cressie and C.K Wikle. *Statistics for spatio-temporal data*. Wiley. com, 2011.

Lance A Waller, Bradley P Carlin, Hong Xia, and Alan E Gelfand. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92(438):607–617, 1997.

Y.C MacNab and C.B Dean. Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in Medicine*, 21(3):347–358, 2002.

MA Martínez-Beneito, A López-Quilez, and P Botella-Rocamora. An autoregressive approach to spatio-temporal disease mapping. *Statistics in medicine*, 27 (15):2874–2889, 2008.

Birgit Schrödle and Leonhard Held. Spatio-temporal disease mapping using inla. *Environmetrics*, 22(6):725–734, 2011.