

Flexible Bayesian Hierarchical Spatial Modeling in Disease Mapping

Kassahun Abere Ayalew

August 2022

Flexible Bayesian Hierarchical Spatial Modeling in Disease Mapping

DISSERTATION

submitted to the Department of Statistics, School of Mathematics, Statistics and
Computer Science in the College of Agriculture, Engineering, and Science in
fulfillment for the degree of Doctor of Philosophy in Statistics.

by

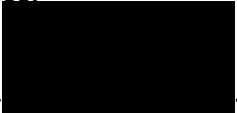
KASSAHUN ABERE AYALEW

DECLARATION

I hereby declare that this thesis is an original work and has not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person. Each contribution to, and quotation in this dissertation from the work(s) of other people has been duly acknowledge.

Kassahun Abere Ayalew :.....

Approved by Supervisor:

Prof. Samuel Manda : ......

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the almighty God for giving me the strength, wisdom and courage. I would like to express my deepest gratitude to my supervisor professor Samuel Manda for his guidance, encouragement, helpful feedback, kindness, sharing valuable resources including his vast knowledge about statistics during my PhD time.

A special thanks you to professor Bo Cai who also provided valuable guidance, feedback and advise to my PhD work; and it was unfortunate that you had to be excluded from the supervision team at last minute due to some issues.

Also I am indebted to the College of Agriculture, Engineering and Science School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal for waiving the tuition fee for doing my PhD. Thanks to the National Institute for Communicable Diseases and Human Science Research Counsel for allowing me to use their data in this dissertation.

Thank you to friends and colleagues for the support and encouragement. Special thanks to my family members especially my mother for her prayer, love, and all the sacrifices she made in raising me.

ABSTRACT

The Gaussian Intrinsic Conditional Autoregressive (ICAR) spatial model, which usually has two components, namely an ICAR for spatial smoothing and standard random effects for non-spatial heterogeneity, is used to estimate spatial distributions of disease risks. The normality assumption in this model may not always be correct and misspecification of the distribution of random effects could result in biased estimation of the spatial distribution of disease risk, which could lead to misleading conclusions and policy recommendations. Limited research studies have been done where the estimation of the spatial distributions of diseases under the ICAR-normal model were compared to those obtained from fitting ICAR-nonnormal model. The results from these studies indicated that the ICAR-nonnormal models performed better than the ICAR-normal in terms of accuracy, efficiency and predictive capacity. However, these efforts have not fully addressed the effect on the estimation of spatial distributions under flexible specification of ICAR models in disease mapping.

The overall aim of this PhD thesis was to develop approaches that relax the normality assumption that is often used in modeling and fitting of ICAR models in the estimation of spatial patterns of diseases. In particular, the thesis considered the skew-normal and skew-Laplace distributions under the univariate, and skew-normal for the multivariate specifications to estimate the spatial distributions of either univariable or multivariable areal data. The thesis also considered non-parametric specification of the multivariate spatial effects in the ICAR model, which is a novel extension of an earlier work. The estimation of the models was done using Bayesian statistical approaches.

The performances of our suggested alternatives to the ICAR-normal model were evaluated by simulating studies as well as with practical application to the estimation of district-level distribution of HIV prevalence and treatment coverage using health survey data in South Africa. Results from the simulation studies and analysis of real data demonstrated that our approaches performed better in the prediction of spatial distributions for univariable and multivariable areal data in disease mapping approaches.

This PhD has shown the limitations of relying on the ICAR-normal model for the estimations of spatial distributions for all spatial analyses, even when the data could be asymmetric and non-normal. In such scenarios, skewed-ICAR and nonparametric ICAR approaches could provide better and unbiased estimation of the spatial pattern of diseases.

LIST OF JOURNAL PUBLICATION AND CONFERENCE PAPERS

1. Ayalew KA, Manda S, Cai B. A Comparison of Bayesian Spatial Models for HIV Mapping in South Africa. *Int J Environ Res Public Health*. 2021 Oct 26;18(21).
2. Kassahun Ayalew, Samuel Manda, Din Chen. Skewed random effects distribution in conditionally autoregressive spatial models for estimating HIV prevalence at local level in South Africa. Presented at Joint Conference of the Sub-Saharan Africa Network (SUSAN) of the International Biometrics Society (IBS) and DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB), Cape Town, 8 - 11 September 2019
3. Kassahun Ayalew, Samuel Manda, Bo Cai. Non-Normal Statistical Estimation of Spatial Distribution of Disease Risk. Presented at the International Symposium on Modern Biostatistics and Machine Learning, University of Pretoria, Pretoria, South Africa, 26-28 July 2022

CONTENTS

1	INTRODUCTION	1
1.1	Background	1
1.1.1	Rationale of the study	3
1.1.2	Brief overview of Spatial models	3
1.1.3	Dissertation outline	10
2	COMPARISON OF BAYESIAN SPATIAL MODELS FOR HIV MAPPING IN SOUTH AFRICA	11
2.1	Introduction	11
2.2	Methods and data	13
2.2.1	Skew-t spatial random effects distribution	13
2.2.2	Methods for comparing competing models	14
2.2.3	Implementation	15
2.2.4	Data	16
2.3	Results	17
2.4	Discussion	19
2.5	Conclusions	20
3	SKEWED INTRINSIC CONDITIONAL AUTOREGRESSIVE SPATIAL MODELS AND THEIR APPLICATION FOR DISEASE MAPPING	22
3.1	Introduction	22
3.2	ICAR skew-normal and ICAR skew-Laplace distributions for modeling the structured spatial random effects and their applications for disease mapping	24
3.2.1	Elliptical distribution	24
3.2.2	Skew-Laplace distribution	27
3.2.3	Skew-Normal and Skew-Laplace structured spatial random effects distribution	27
3.2.4	Posterior distribution	29
3.3	Simulation	30
3.4	Mapping District HIV Prevalence in South Africa	35
3.5	Discussion	38
4	SKEW-NORMAL MULTIVARIATE INTRINSIC CONDITIONAL AUTOREGRESSIVE SPATIAL MODEL AND ITS APPLICATION FOR DISEASE MAPPING	41
4.1	Introduction	41
4.2	Multivariate Conditional Autoregressive (MCAR)	43

4.3	Multivariate ICAR-skew-normal random effects distribution for modelling the structured spatial random effects	44
4.3.1	Posterior distribution	46
4.4	Simulation	47
4.5	Mapping District HIV Prevalence and proportion on ART among HIV positive pregnant women who know their HIV status in South Africa .	51
4.6	Discussion	55
5	MULTIVARIATE BAYESIAN NONPARAMETRIC DISEASE MAPPING USING AREAL STICK-BREAKING PRIORS	58
5.1	Introduction	58
5.2	Dirichlet process mixture	60
5.2.1	Dirichlet distribution	60
5.2.2	Dirichlet process	61
5.3	Multivariate disease mapping using areally-referenced spatial stick-breaking prior	62
5.4	Posterior distribution	63
5.5	Truncated Dirichlet process	64
5.6	Data	65
5.7	Mapping District HIV Prevalence and proportion on ART among HIV positive individuals in South Africa	66
5.8	Results	67
5.9	Discussion	68
6	CONCLUSION AND FUTURE WORK	70
6.1	Conclusion and future work	70
6.2	Future work and limitations	73
	Bibliography	74
	Appendix	85

LIST OF FIGURES

Figure 2.1	Map of HIV prevalence by district in South Africa before and after adjusting the data for zero positive tests in some districts .	17
Figure 2.2	Estimated HIV prevalence by district in South Africa with Co- variates (first row) and without covariates (second row)	19
Figure 3.1	Bar graph of spatially structured random effects simulated from municipal map of South Africa using ICAR-t distribution with outliers	32
Figure 3.2	Bar graph of spatially structured random effects simulated from municipal map of South Africa using ICAR-normal distribution with outliers	32
Figure 3.3	Map of spatially structured random effects simulated from mu- nicipal map of South Africa using ICAR-t distribution with out- liers	32
Figure 3.4	Map of spatially structured random effects simulated from mu- nicipal map of South Africa using ICAR-normal distribution with outliers	33
Figure 3.5	HIV prevalence by district in South Africa using 2016 SADHS data	38
Figure 4.1	Bar graph of spatially structured random effects simulated from multivariate ICAR-t distribution with outliers	48
Figure 4.2	Bar graph of structured spatial random effects simulated from multivariate ICAR-normal distribution with outliers	48
Figure 4.3	Map of structured spatial random effects simulated from mul- tivariate ICAR-t distribution with outliers	49
Figure 4.4	Map of structured spatial random effects simulated from mul- tivariate ICAR-normal distribution with outliers	49
Figure 4.5	Map of proportion of pregnant women who know they HIV sta- tus and the proportion on ART among these women by district in 2017, South Africa.	55
Figure 5.1	Map of HIV prevance by distict (a) and proportion of ART cov- erage among HIV positive individuals in South Africa, 2017 . .	68

LIST OF TABLES

Table 2.1	Comparison of the fitted models using DIC and CPO	18
Table 3.1	Estimated values of parameters of models for structured spatial random effects simulated from ICAR-normal and ICAR-t distributions with outliers	34
Table 3.2	Comparison of the fitted models using DIC and CPO/ LS_{cv} . . .	37
Table 4.1	Estimated values of parameters of models for structured spatial random effects simulated from multivariate ICAR-normal and ICAR-t distributions with outliers	50
Table 4.2	Posterior mean and 95% HPD intervals for the parameters of interest, and DIC and CPO/ LS_{cv} values for the different models in this study	54
Table 5.1	Estimates of parameters with their 95% confidence interval for the Bayesian nonparametric model.	67

LIST OF ABBREVIATIONS

AIDS	Acquired Immunodeficiency Syndrome
ARMSE	Average Root Mean Squared Error
ANC	Antenatal Sentinel
ART	Antiretroviral therapy
CAR	Conditional Autoregressive
CPO	Conditional Predictive Ordinates
EA	Enumeration Area
DIC	Deviance information criteria
GIS]Geographical Information System
GWR	Geographically Weighted Regression
HIV	Human Immunodeficiency Virus
ICAR	Intrinsic Conditional Autoregressive
IG	Inverse Gaussian
IW	Inverse Wishart
MCAR	Multivariate Conditionally Autoregressive Distribution
MICAR	Intrinsic Multivariate Conditionally Autoregressive Distribution
MCMC	Markov Chain Monte Carlo
OLS	Ordinary Least Square
PSU	Primary Sampling Units
RMSE	Root Mean Squared Error
SADHS]South African Demographic and Health Survey
SAL	Small Area Layer
SAR	Simultaneous Autoregressive
STI	Sexually Transmitted Infection
UNAIDS	Joint United Nations Program on HIV/AIDS

INTRODUCTION

1.1 BACKGROUND

Spatial statistics is one of the areas of statistics that is focused on the application of statistical techniques to data that are collected over space and time (Ripley, 2005; Haines & Thiart, 2021). In spatial statistics the major focus is modelling of the spatial pattern/variation of the outcome variable of interest. Thus spatial analysis considers distance, adjacency, elevation, environmental, climatic and other location information to model spatial pattern (Sha, 2018; Manda et al., 2020). Depending on the nature of the available data different statistical approaches are employed to model the spatial pattern of the outcome (Blangiardo & Cameletti, 2015). For example, geostatistical models is used for modelling a continues point referenced data, point pattern data are modelled using poison process and aggregated areal data are modelled mostly using conditional autoregressive approach (Banerjee et al., 2003; Haines & Thiart, 2021) which is the focus of this dissertation.

Estimation of outcomes at lower administrative level from aggregated data are one of the most popular approaches in spatial modelling as aggregated data are widely available and easily accessible. In addition, the demand of estimates such as health or other indicators at local administrative level is increasing overtime as there is a growing interest of the use of such outcomes to make informed decisions, appropriately allocate resources and evaluate the impact of interventions especially in low resource countries (Manda et al., 2020; Giorgi et al., 2018). The most commonly used method that has been used for estimating local level outcomes is a linear mixed effect model. The random effects in a linear mixed effect model are used as surrogates for spatially correlated factors that are not observed as data and are known as spatial random component (Lawson et al., 2003; Besag et al., 1991; Knorr-Held & Best, 2001). Spatial modelling approaches are thus focused on modelling this random component using Intrinsic conditional autoregressive method (ICAR) citepbesag1991bayesian,banerjee2003hierarchical. Thus the ICAR approach offers a mechanism for borrowing information from random components of neighboring areas as these areas tend to have similar environmental and socio-demographic factors (Besag et al., 1991).

The ICAR spatial model specification that are implemented in several statistical software and geographical information systems packages are premised on the assumption that the random effects are normally distributed. And as a result of advancement in statistical and computational techniques a hierarchical Bayesian approach is used for estimating the values of interest (Knorr-Held & Best, 2001; Feltbower & Manda, 2012; Waller & Carlin, 2010). However, the ICAR spatial model specification that are implemented in several statistical software and geographical information systems packages are premised on the assumption that the random effects are normally distributed. However, it may not be always right to assume that the random effects follow normal distribution and/or some known parametric distribution as there is a possibility that it could follow skewed, multimodal distributions or the distribution may be unknown at all. Misspecification of the distribution of the random effects may result in estimates that are biased (Ghosh et al., 2007; Verbeke & Lesaffre, 1996) and could have implications on inferences about the parameters of interest. It also becomes difficult to determine the impact of the covariates on the scale and shape of the random effects distribution (Heckman & Singer, 1984; Waller & Carlin, 2010).

A study by Manda et al. (2020) about the assessment of the spatial analysis approaches used by authors for analyzing health survey data in sub-Saharan Africa indicated that most of the authors are focused on analyzing their data using the existing spatial models and available statistical packages; and are short of validating, critically assessing and interrogating the existing spatial models and developing appropriate spatial models though the data at hand may not fit appropriately to existing methods. Therefore, Manda et al. (2020) and Haines & Thiart (2021) suggested the need for the development of robust and advanced spatial modelling approaches. In addition, there is an ever-growing of national survey data such as Demographic and Health Survey, Malaria Indicator Survey, Multiple Indicator Cluster Surveys, Antenatal Surveillance HIV Surveys and Population HIV Impact Assessment Surveys which constitutes a large number of indicators. The existing spatial models may not be suitable for modelling some of the indicators determined from these data; thus increase in availability of spatially reference data demands for the availability more advanced and flexible spatial models.

In response to the issues raised in the above paragraphs a number of alternative approaches were suggested in modelling the spatial component; for example (Manda, 2014) and (Lunn et al., 2013) proposed a double exponential and a mixture of ICAR normal, and ICAR double exponential respectively to the spatial random component which better capture distributions of data with wider tails and narrow distributions at the center of the data than a normal distribution. Similarly, Nathoo & Ghosh (2013) proposed a skew-t distribution to capture skewness and/or heavy tailedness in data. Though these approaches are used as an alternative to normal distribution for modelling the spatial component; there are quite a number of features that may not be

captured by the above distributions such as data with skew-normal, skew-Laplace and multimodal distributions. Thus, flexible approaches that allows the spatial distribution to follow skew-elliptical and/or infinitely many distributions would provide more flexibility and correct dependence of spatial data in space. Therefore, the aim of this dissertation is to develop and validate a set of flexible parametric and nonparametric spatial models in the spatial smoothing of areal data.

1.1.1 *Rationale of the study*

The standard spatial ICAR model assumes that the random effects are normally distributed. It may not be always right to assume that the spatial effects follow some known parametric distribution as there is a possibility that it could follow skewed and multimodal distributions. This study develops and validates flexible spatial model for modelling data that drifts away from symmetry or whose distribution are unknown. Thus, epidemiologists and practicing statisticians can use this generalized approach for conducting spatial analysis and making appropriate decisions using the results estimated using this method. The findings from this study will help governmental and non-governmental originations, and the private sector to know the level of the epidemics at lower administrative level, and thus prioritize and plan appropriate public health programs tailored to each community and evaluate the combined impact of national and local public health programs targeting HIV and other diseases.

1.1.2 *Brief overview of Spatial models*

In order to study the geographical distributions of events such as disease incidence, mortality, and other outcomes geographically referenced data are presented in the form of maps (Leroux et al., 2000; Besag et al., 1991; Manda, 2014). However, the presentation of data in the form of maps was minimal until recently due to shortage of geo-referenced data and limited availability of geographically information system software (GIS). These days geographical information are collected in most of surveys and surveillance activities and GIS software are now widely available which increases the presentation of data in the form of maps (Saran et al., 2020; Graham et al., 2011). However, mapping of disease incidence, mortality rates and other outcomes of interest at lower administrative level using direct estimates is complicated by the fact that direct estimates are unstable since survey and surveillance are not powered to produce reliable estimates at lower administrative level.

There are three broad categories of spatial data though sometimes it is not easy on how to classify geo-referenced data into these categories; as such methods used for analyzing one class of data can be used for analyzing another class of data (Cressie, 2015). Geographically referenced data relates to its location and the information from its neighbors (Leroux et al., 2000; Lawson et al., 2000). The three main categories of

spatial data are point referenced (geostatistical) data, areal data of regular (lattice) or irregular shape and point pattern data. Point referenced data is defined as a random measure about the outcome of interest over a spatial region (D) and the spatial points at which the measurement was taken varies continuously (Banerjee et al., 2003; Blangiardo & Cameletti, 2015). The spatial points are defined by a vector of latitude and longitude. These type of data arise in problems related to climatology, environmental monitoring, and geology (Banerjee et al., 2003; Cressie, 2015). Point pattern data is generated if we are interested in the occurrence or nonoccurrence of the event of interest in a random location. For example, one may be interested in studying the locations of particular species of trees, animals etc. in a natural forest (Cressie, 2015). This type of data are used for studying if the event under study is occurring at random, or show some form of clustering, or happen in some form of pattern (Cressie, 2015; Banerjee et al., 2003). Areal data is an aggregate value of the variable of interest over a finite areal unit defined over a region (Blangiardo & Cameletti, 2015). The areal units can be districts, counties, subdistricts etc. These data are used for studying proportions and incidence in an area. In this section we widely review methods of analysis of areal data.

Limitations associated with direct estimates of indicators determined from national survey and surveillance data necessitated in the development of advanced statistical models which help to produce reliable and stable estimates from these data (Manda et al., 2015; Besag et al., 1991). These statistical models use covariate information and borrow relevant data from surrounding regions (as closer areas may share similar characteristics that could affect the outcome variable) in order to reduce the instability associated with scattered spatial events (Manda et al., 2015). These statistical models smooth out the white noise in the data and display the underlying patterns of the data and reduces the impact of administrative boundaries which are not demarcated in relation to the outcomes variables (Lawson et al., 2003; Manda et al., 2015).

Depending on the nature of spatial data statistical models use different approaches to borrow information (to model spatial association) from the surrounding regions. For point referenced data the distance between spatial locations are used for modelling the spatial association (Diggle et al., 1998; Banerjee et al., 2003). Mostly an exponential function that decays with distance is used for these type of data (Diggle et al., 1998; Cressie, 1993). Point pattern data are modelled if the event of interest is clustered more, or less than expected in a given region than would be expected under a completely random situation. As this is a count of events in a region point pattern data are modelled through a homogeneous Poisson process (Diggle et al., 1998; Banerjee et al., 2003). The aggregated areal data are modelled by borrowing information from adjust regions. In order to model the spatial association a neighborhood structure is introduced constructed based on shared border. According to Banerjee et al. (2003) the most commonly used modelling approaches that take into account

the neighborhood information are the simultaneously and intrinsic conditional autoregressive models (SAR and ICAR). The SAR model is used for modelling areal data using likelihood methods, whereas the ICAR model is convenient for modeling areal data using Bayesian approach and this method is the focus of this chapter.

1.1.2.A *Spatial models for disease mapping*

Areal spatial models are used for generating estimates at lower administrative level from survey data and hence used for mapping disease occurrence or other events at lower area level. The analysis of geographical variation of health-related events and mapping estimates of public health outcomes is used for studying the geographical distribution of diseases, formulating hypothesis about the aetiology of disease, identify areas that have excess risk than expected under normal circumstance (Best et al., 2005; Manda, Feltbower & Gilthorpe, 2012; Bernardinelli & Montomoli, 1992). Direct or maximum likelihood estimates of health outcomes such as prevalence or incidence may lead to wrong conclusions and are less important since direct estimates are likely to be affected by random noise especially if the event under study in a small area is rare, if the sample size is not sufficient, or the area has low population which could result in extreme and unstable estimates (Clayton & Kaldor, 1987; Langford et al., 1999). As a result different statistical approaches to disease mapping were suggested; and these approaches primarily smooth out the random noises associated with the data and try to determine the true spatial patterns of a disease under study (Manda, Feltbower & Gilthorpe, 2012; Lawson et al., 2000).

In an effort to reduce the instability associated with direct estimates Clayton and (Clayton & Kaldor, 1987) proposed an empirical Bayesian estimation method. In this approach the observed number of events given the relative risk are assumed to follow a Poisson distribution and the conditional distribution of the relative risk given the observed number of cases are assumed to follow a gamma distribution which resulted in the so-called Poisson-Gamma model. The empirical bayes estimates based on the Poisson-Gamma model are thus a weighted average between direct estimate and an estimate from the posterior information determined from prior distributions (Lawson et al., 2000; Clayton & Kaldor, 1987). Similarly, Tsutakawa et al. (1985) assumed a normal distribution for the logit of the relative risk. Thus, the estimates from the empirical bayes estimates are relatively stable and reliable which could provide more epidemiological sense. The major shortcoming with these approaches is that they fail to incorporate the spatial and covariate information in the estimation of the relative risk (Banerjee et al., 2003; Clayton & Kaldor, 1987; Tsutakawa, 1985). The other limitation with the above empirical Bayesian approach is that unlike the full Bayesian approach the uncertainty associated with the model hyper parameters are not taken into account and hence the variance of the estimated relative risk is low which in turn results narrow confidence interval and hence wrong conclusions (Manda, Feltbower

& Gilthorpe, 2012; Tsutakawa, 1985).

The occurrence of an event or disease in an area is a countable variable and these count data can have a Poisson or binomial distribution depending on if the disease is rare or not. Hence the relative risk or prevalence computed from these data are modelled using generalized linear models. For example, Wakefield (2007) and Gutreuter et al. (2019) used generalized linear model, which incorporates covariates, to map incidence rates of lip cancer in Scotland and district level HIV prevalence in South Africa. The major limitation with the generalized linear models is that the spatial correlation is not taken into consideration in the estimation of the outcome variable. As a result, a linear mixed model was suggested to overcome the limitation associated whose generalized linear model and thus the random effect in the linear model is used to denote the spatial effects. These spatial effects can be used as surrogates of unknown or unobserved variables in the model (Lawson et al., 2003; Bernardinelli et al., 1995). Usually, these models are presented in hierarchical form, in the first stage the distribution of the data are specified and the distribution of the random effect is presented in second stage. The simplest approach in modelling the random effects is assuming that the random effects are independent and exchangeable (Lawson et al., 2000; Gutreuter et al., 2019; Bernardinelli & Montomoli, 1992). Another alternative and improved approach for modelling the random effects is by assuming that the random effects have a multivariate distribution, specifically multivariate normal, which enables to model the spatial correlation that exist in the random effects (Lawson et al., 2003; Tsutakawa, 1985). Multivariate normal distribution is the most commonly used distributional assumption used for modelling random effects for disease mapping. The spatial correlations are introduced through the variance-covariance matrix of the multivariate normal distribution which are based on geographical proximity. Thus, for point referenced data the variance-covariance matrix is determined using distance-based function among observed points whereas for areal or lattice data it is determined based on a neighborhood approach or distances between centers of areas (Besag et al., 1991). For variance-covariance matrix determined based on distance-based functions one needs to make sure that the resulting matrix is positive definite (Ripley, 1981). And the most widely used function for determining the variance covariance matrix is the exponential decay function (Best et al., 2005). The use of this approach for diseases mapping for a region having several hundreds of areas is computationally expensive since Markov chain Monte Carlo algorithm needs inversion of the variance-covariance matrix at each iteration (Best et al., 2005).

The unobserved covariates represented by the random effects in a linear mixed model, some may show spatial pattern, and some may not thus the random effects can be split into one that captures spatial correlation and one that do not show spatial pattern and are mostly known as structured spatial random component and unstructured spatial random components respectively (Besag et al., 1991; Breslow,

1984; Bernardinelli & Montomoli, 1992). This type of linear mixed model is known convolution model. The unstructured spatial random component is used to capture the extra-Poisson variability or overdispersion in the marginal distribution of the observed data (Breslow, 1984; Waller & Carlin, 2010). Mostly models of this nature are analyzed using full Bayesian approach and Gaussian Intrinsic conditional autoregressive model (ICAR) is used to model the structured spatial random component (Besag et al., 1991). In the full Bayesian approach the variability associated with the parameters in the model are accounted for which results in reasonable estimates of standard errors which could avoid apparent significant results (Bernardinelli & Montomoli, 1992; Manda, Feltbower & Gilthorpe, 2012).

Gaussian Intrinsic conditional autoregressive model model is the most widely used disease mapping model following a seminal work by Besag et al. (1991). In ICAR model the conditional distribution of the spatial random effect given all the others is a normal distribution whose mean is a weighted mean of all the other random effects, and whose variance is a weighted value of the overall variability. A weight value of one is assigned if two regions are neighbors and a weight of value zero is assigned if they are not neighbors. Haining (2003) has presented a wide-range of weighting methods that are widely used. Some of the weighting methods that are presented are exponential function of distance, common border weight $W_{ij} = (l_{i,j}/l_i)^x$ where $l_{i,j}$ is the length of the common border between i and j , and l_i is the length of the border of i and $x \geq 0$; distance weight $w_{(i,j)} = (d_{i,j})^{-y}$ where $d_{i,j}$ denotes the distance between i and j and the parameter $y \geq 0$ and Combined border and distance weighting: $W_{(i,j)} = (l_{i,j}/l_i)^x (d_{i,j})^y$. Lu et al. (2007) used different approach for generating spatial weights where observed socio-demographic, topological and distance based information are used for calculating weight. Thus, unlike the methods discussed above their approach produces random weights since they model the weight in a logit link function with covariates included. Following the work by Besag & Kooperberg (1995) the spatial component presented by a ICAR formulation can be presented in a multivariate normal distribution.

In ICAR model one of the limitations is that the overall variability in the data is used to represent the variabilities for both spatially structured and unstructured random components (Leroux et al., 2000). As a result they suggested an alternative formulation to that of Besag et al. (1991). Thus unlike the convolution model which has two random components, they introduced a normally distributed random component in their model that contains both the structured spatial and heterogenous information. The variability of the random component depends on a precision matrix which also contains a parameter, whose value is between zero and one, which controls the spatial dependency. The parameters in the model including the one that control the spatial dependency is determined using penalized quasi likelihood; on the other hand MacNab & Dean (2000) used a parametric bootstrap approach for estimation which enables

to test a hypothesis about the presence and absence of spatial dependence. This approach performs better than the ICAR model if the spatial association is low (Waller & Carlin, 2010). Alternatively Green & Richardson (2002) used a Hidden Markov random field approach to model relative risks spatially and followed an approach similar to a cluster analysis where relative risks in an area is assigned to a cluster based on assignment variable. The assignment variable is modelled assuming it follows a correlated spatial process. In this approach the spatial dependence parameter is not assumed constant unlike most spatial models where the spatial component included in the covariance matrix is fixed, and hence this approach avoids the challenge of over smoothing which hides local variability of rates and performs better than the ICAR model where there is higher discontinuity in the relative risks.

All the above models are used for analyzing univariate data however there could be a situation where more than one variable could be observed and the interest could be in modelling these variables jointly. Several authors have suggested different joint modelling approaches when more than one disease is observed. This approach produces relatively precise estimates as it pools information from different correlated diseases, also this approach is used to identify common risk factors of diseases by modelling multiple outcomes together (Manda, Feltbower & Gilthorpe, 2012). The simplest approach in joint modelling is the one suggested by Bernadinelli et al. (1997). In their approach the prevalence/incidence of the other disease is included in the model of the disease of interest as a covariate after controlling or smoothing its randomness. Wang & Wall (2003) instead of including the other disease as a covariate they tried to consider all diseases as outcome variables. They introduced a spatially correlated latent factor in their model which is common for all the diseases. This approach does not consider the fact that different disease may have different spatial factors and so is the strength of correlation that exists among these spatial factors.

Kim et al. (2001) developed an approach which was different from those presented above. In this approach a spatial random effect was included for each model and a twofold CAR model was used to model the spatial correlation which enables to share data obtained from different outcome variables. The spatial effect for a particular area is divided into three: one that denotes correlation to its neighbors of the same outcome variable, the other one denotes correlation to its neighbors because of the other outcome variable and the third one is the correlation between the two outcome variables in the same area. However, this model is used for modelling two outcome variables, difficult to extend for more than two outcomes and it is very complex. Carlin & Banerjee (2003) and Gelfand & Vounatsou (2003) extended the univariate CAR model to a multivariate setting following the theoretical description of a multivariate normal Markov Random Field by Mardia (1988). The precision matrix in the joint formulation is a Kronecker product between the univariate form and a symmetric positive definite matrix of the same dimension as the outcomes of interest. And they

determined the multivariate distribution from the full conditional distributions. This model was further extended by Jin et al. (2005) by taking into account the spatial correlation that exist between a given disease in an area with those of the other diseases in the neighboring areas. And based on multivariate normal theory they determined the joint distribution for multivariate normal Markov Random Field from conditional and marginal distributions. Another approach to joint modelling is the shared component model which was developed by Knorr-Held & Best (2001) for modelling two disease, and this approach was later extended to model more than two diseases by Held et al. (2005). The key idea behind the formulation of the shared component model is that most diseases share common risk factors as such these diseases have similar spatial patterns. Therefore, the risk component of two diseases that have common risk factors can be divided into one that is shared by both diseases, and risk factor that are specific to each disease (Knorr-Held & Best, 2001). These shared components are used to represent the unknown spatially structured factors that affect either the risk of both or one of the two diseases (Manda, Feltbower & Gilthorpe, 2012). Langford et al. (1999) and Leyland et al. (2000) tried to model multiple disease jointly by presenting the spatial model in multi-level model. The concept of multiple membership classification approach was used to account for the spatial correlation. Thus the spatial component in a given area i is a weighted sum of random components of neighbors of area i drawn from a normal distribution with zero mean and variance. And parameter estimation in their model was conducted using iterative generalized least square approach.

The structured spatial random components are modelled using ICAR normal prior. One of the main reasons for using a normal prior is because of its technical convenience. However, the spatial random effects are not observed as data; hence the type of distribution of the random components are unknown. The random components may have a skewed distribution, multimodal distribution, Laplace distribution etc. A number of studies have tried to relax the ICAR normal assumption by using different approaches. For example, Lunn et al. (2013) used a ICAR Laplace distribution to model structured spatial random components. In another effort to model the structured spatial random components Manda (2014) used a mixture of intrinsic conditional autoregressive (ICAR) normal and ICAR double exponential prior for the structured spatial random effects. This approach is a relatively flexible compared to CAR-normal distribution for modelling random effects that have longer tails. Nathoo & Ghosh (2013) proposed a ICAR-skew-t distribution prior for modelling the spatial random effects. They presented the ICAR-skew-t distribution as a scaled mixture of CAR normal and standard normal distributions. They indicated that their approach is flexible for modelling the structured spatial random effects in the presence of outliers and discontinuities. A Bayesian nonparametric approach for modelling the spatially structured random components was another flexible technique used in spatial modelling of structured spatial random effects. For example, Li et al. (2015) used the spatial-stick breaking approach to analyze a univariate areally-referenced data by adopting the work of

Reich & Fuentes (2007) which were developed for modelling point referenced data. The spatial dependence was modelled by adding additional component to the mixing weights. This additional weight included in the stick-breaking process is modelled by using ICAR prior. Similarly, Hossain et al. (2013) used a stick-breaking approach for modelling spatially structured random effects for areally referenced data. In this approach the spatial dependence was introduced by defining a spatial model through the mixing weights of the stick-breaking prior. And covariate dependent kernel function is included in the mixing weights of the stick breaking prior in order to introduce the spatial dependence between areas.

1.1.3 *Dissertation outline*

In the next chapter we presented the different existing spatial models, showed their application using complex survey data and compared their performance in modelling district level HIV in South Africa. Then in Chapter 3 we extend the common univariate ICAR approach to spatial modeling to a univariate ICAR-skew elliptical modelling approach; followed by a simulation analysis and application of our approach to a real data. The multivariate extension of the univariate approach developed in chapter 3 was presented in chapter 4 together with a simulation analysis and its application to a complex survey data. A Bayesian nonparametric approach developed by Li et al. (2015) for modelling univariate areal data was extended to a multivariate setting in chapter 5 and fitted to a complex survey data to show its application. Then this dissertation concluded in chapter 6 by presenting the conclusion, future work and limitations.

COMPARISON OF BAYESIAN SPATIAL MODELS FOR HIV MAPPING IN SOUTH AFRICA

2.1 INTRODUCTION

Governments in sub-Saharan Africa (SSA), in collaboration with non-governmental organizations and private sectors, design national strategic plans and policies, allocate resources and implement programs in the fight against the HIV/AIDS epidemic (UN-AIDS, 2015; UPsEPfA, 2021). Such efforts are designed to reduce HIV-related infection, morbidity and mortality. As well as understanding the level of the HIV epidemic at the national level, most governments in the region have implemented a decentralized approach to governance and service provision. Thus the need for reliable local (district)-level HIV statistics to support decision making regarding the delivery of HIV care, treatment and prevention services (Manda et al., 2015; Hallett et al., 2016). Most of the countries in SSA rely on data obtained from national HIV surveys for monitoring the level of the HIV epidemic and subsequent responses. However, the national HIV surveys are mostly empowered to produce reliable HIV estimates at national and provincial level. Crude HIV estimates at small area level could be exaggeratedly estimated due to small numbers, resulting in unstable variances (Tanser et al., 2009; Niragire et al., 2015; Chimoyi & Musenge, 2014). Consequently, HIV prevention and treatment programs tailored to small areas could be based on unreliable evidence (Houlihan et al., 2010).

As a result, modelling approaches are used for generating local-level estimates from survey data that are originally meant to provide reliable estimates at national and provincial levels (Johnson, 2004) (Leyland et al., 2000). The most used approach has been using spatial smoothing models where spatial components are incorporated in the model as random effects. The spatial models produce reliable disease rates with improved accuracy for small areas with few sparse observations by incorporating information from local, spatially contiguous areas. The structured random effect in spatial models represents clustering of diseases over geo-graphical areas, unobserved environmental or frailty factors which are spatially correlated but are not included as covariates in a model (Lawson et al., 2003; Knorr-Held & Best, 2001; Besag et al., 1991; Carlin & Banerjee, 2003). Structured spatial random effects (which consider the local effects) are mostly modelled using the intrinsic conditional autoregressive normal

(ICAR-normal) model (Besag et al., 1991; Carlin & Banerjee, 2003). The ICAR-normal model offers greater flexibility for modelling the spatial correlation than the linear mixed effects model, with only a global random effect. However, a normal spatial distribution on the structured spatial effect could be restrictive, as there could be a possibility that the normality assumption could be misspecified (Ghosh et al., 2007). Misspecification of the distribution of the random effects may result in estimates of diseases rates that are biased (Verbeke & Lesaffre, 1996; Lunn et al., 2013). The usual approach is to transform the data to normality, for example by performing a logarithm of the rates. However, if there was an appropriate theoretical model, transformation could be avoided, as it is difficult to interpret results from transformed data. In addition, the transformation could result in the loss of information (Verbeke & Lesaffre, 1996).

A few approaches have been proposed to reduce the impact of a normal distribution assumption for spatial random components. For example, Lunn et al. (2013) and (Manda, 2014) proposed a double exponential and a mixture of ICAR-normal and ICAR-double exponential, respectively, to better capture possible wider tails for the spatial random effects. Kim & Mallick (2004) and (Azzalini & Capitanio, 1999) considered a skew-normal spatial model for point referenced data. However, the structured spatial skewed random fields suffer identifiability problems (since the skewness parameter may be unknown) (Genton & Zhang, 2012) and must be determined uniquely (Gelfand & Sahu, 1999). To solve this identifiability problems, (Zhang & El-Shaarawi, 2010) defined a skewed stationary Gaussian process for spatial random effect based on the work by Azzalini & Capitanio (1999). In addition, Allard & Naveau (2007) and Zareifard & Khaledi (2013) introduced a skew-normal spatial random field based on Dominguez-Molina et al. (2003) and Palacios & Steel (2006), respectively, for point referenced data. Other skewed spatial distributions are the skew-normal by Rantini et al. (2021) and (Fernández & Steel, 1998).

Our aim, in this study, is to model the district-level HIV prevalence in South Africa using spatial smoothing methods. There is ample evidence of substantial small area variation in the distribution of HIV prevalence in Sub-Saharan Africa (Dwyer-Lindgren et al., 2019; Kim et al., 2021). Similarly evidence has also been found in South Africa by Kim et al. (2021) and Gutreuter et al. (2019). The distribution of the district HIV prevalence could be skewed and non-normal. Thus, we estimated the spatial distribution of the HIV prevalence among the districts in South Africa using the ICAR-normal (Besag et al., 1991), ICAR skew-t distribution (Nathoo & Ghosh, 2013) and ICAR-Laplace (Lunn et al., 2013) using the 2016 South African Demographic and Health Survey data. The next section presents the description of the spatial models used and the HIV data. Section 3 contains the results obtained from fitting the models to the data. We discuss the results in Section 4 and conclude in Section 5.

2.2 METHODS AND DATA

2.2.1 Skew- t spatial random effects distribution

Let Y_i be the number of HIV positive individuals out of a sample of size n_i in district i ($i = 1, \dots, 52$). Both Y_i and n_i are adjusted to account for the survey design to become the effective number of HIV cases, Y_i^* , and the effective sample size, n_i^* (Kish, 1995; Chen et al., 2014; Vandendijck et al., 2016). A three-stage Bayesian hierarchical spatial smoothing model for a binary HIV outcome uses a binomial distribution at stage one as:

$$Y_i^* / p_i \sim \text{Binomial}(n_i^*, p_i) \quad (2.1)$$

where $i = 1, \dots, 52$, p_i is the proportion (prevalence) of HIV in district i and is modelled at the second stage by a logit link function using a set of district-level predictor variables, X_i , and both unstructured and spatially structured random effects, as introduced by Besag et al. (1991).

$$\text{logit}(p_i) = \beta_0 + X_i \beta + u_i + v_i \quad (2.2)$$

where β_0 is the intercept; β is a vector of regression coefficients for predictor variable in X_i ; u_i is the unstructured random component and it is assumed to follow a normal distribution, $v_i \sim N(0, \sigma_v^2)$; u_i is the structured spatial random component for district i .

The structured spatial random effects could be modelled using an intrinsic conditional autoregressive normal (ICAR-normal) prior Besag et al. (1991), Knorr-Held & Best (2001), and Carlin & Banerjee (2003) as,

$$u_i / u_{-i} \sim \text{ICARN}(\mu_u, \sigma_u^2) = N\left(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i}\right) \quad (2.3)$$

where m_i is the number of neighbours of district i . Lunn et al. (2013) suggested an alternative model based on a Laplace/double exponential distribution (ICAR-Laplace), which is given as $u_i / u_{-i} \sim \text{ICARL}(\mu_u, \sigma_u^2)$.

However, in situations where the distribution of HIV prevalence data could be non-normal and asymmetric, alternative spatial smoothing models that are robust and flexible could fit the data better. As a result, Nathoo & Ghosh (2013) suggested the skew- t (ICAR-skew- t) spatial smoothing model, defined as:

$$u_i/u_{-i} \sim ST_v \left(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i}, \delta_u \right) \quad (2.4)$$

For easy implementation in most Bayesian statistical software, [Sahu et al. \(2003\)](#) presented a suitable representation of skew-t distribution with k degrees of freedom. Suppose $y \sim \text{skew} - t(k)$, then it could be expressed as $y = \eta^{-\frac{1}{2}} (\Delta |X_0| + X)$, where $X_0 \sim N(0, 1)$, $X \sim N(\mu, \sigma^2)$, Δ is the skewness parameter and $\eta \sim \text{gamma}(\frac{k}{2}, \frac{k}{2})$. The hierarchical set-up of this stochastic representation can be given as $Y/w \sim N(\mu + \Delta w, \frac{\Sigma}{\eta})$, where $|X_0| = w \sim N(0, I_k) I(w > 0)$. Thus, the ICAR-skew-t for the structured spatial random effect can be expressed as:

$$u_i \sim N \left(\frac{\sum_{j \sim i} s_j}{m_i} + \delta_u w_i, \frac{\sigma_s^2}{\eta * m_i} \right) \quad (2.5)$$

where $w_i \sim N(0, I) I(w_i > 0)$, $s_i/s_{-i} \sim N(\frac{\sum_{j \sim i} s_j}{m_i}, \frac{\sigma_s^2}{m_i})$, σ_s^2 and δ_u is the variance of s_i and is the skewness parameter, respectively. The hierarchical representation of the ICAR-skew-t model is shown in the Appendix of this chapter.

2.2.2 Methods for comparing competing models

In this study, we used the deviance information criterion (DIC) and conditional predictive ordinates (CPO) for comparing models. The deviance information criterion was developed by [Spiegelhalter et al. \(2002\)](#) as a method used for comparing models in a Bayesian framework. It is a measure of a model's goodness of fit or adequacy adjusted for a measure of model complexity measured as effective number of parameters. Let θ and $y = y_1, \dots, y_1$ be the model parameter and data, then DIC is expressed as:

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\bar{\theta})$$

where $\bar{D} = E_{\theta/y}[D(\theta)] = E_{\theta/y}[-2\log p(y/)]$ and is the posterior mean deviance that measures goodness of fit or adequacy $p_D = \bar{D} - D(\bar{\theta}) = E_{\theta/y}[D(\theta)] - D(E_{\theta/y}[\theta]) = E_{\theta/y}[-2\log p(y/)] - [-2\log p(y/\bar{\theta}(y))]$ is a measure of the effective number of parameters and measures model complexity, larger values of p_D suggests higher complexity of the model. It is also defined as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest in other words it is considered as the expected excess of the true residuals over the estimated residuals in the data conditional on the parameter θ ([Ghosh et al., 2007](#)). Let $\theta^1, \dots, \theta^k$ be parameter estimates from converged Markov chain then \bar{D} is estimated as $\frac{1}{k} \sum_{i=1}^k D(\theta^i)$ and $D(\bar{\theta}) = D(\frac{1}{k} \sum_{i=1}^k \theta^i)$.

The CPO is a leave-one-out cross validation approach that measures the posterior probability of observing y_i when the model is fitted to all data excluding y_i and it measures the predictive ability of the fitted model. Let $Y = Y_1, Y_2, \dots, Y_n$ be the $n \times 1$ data vector, Y_{-i} be the data vector without y_i . Then the conditional predictive ordinate for observation y_i is given as:

$$CPO_i = f(y_i / \mathbf{y}_{-i}) = \int f(y_i / \boldsymbol{\theta}) P(\boldsymbol{\theta} / \mathbf{y}_{-i}) d\boldsymbol{\theta} = E_{\boldsymbol{\theta} / \mathbf{y}} \left[\frac{1}{f(y_i / \boldsymbol{\theta})} \right]$$

where $\boldsymbol{\theta}$ is the parameter vector, y_i is the i th observation and \mathbf{y}_{-i} is the observed data set except y_i . Thus, one can estimate the value of the inverse of CPO_i by averaging the inverse probability function evaluated at y_i for each θ^k produced from the posterior density. And thus the CPO_i values could be easily determined from the standard MCMC output which is give as:

$$CPO_i = \left[\frac{1}{K} \sum_{k=1}^K \frac{1}{f(y_i / \theta^k)} \right]^{-1},$$

which is the harmonic mean of the probability density function evaluated at y_i for each θ^k , where K is the number of iterations. For discrete data, comparison of CPO_i with the relative frequency determined from data without y_i (\mathbf{y}_{-i}) enables to assess the predictive capacity of the fitted model to the data. In order to compare two or more competing models the overall CPO values of each model are assessed which is given as, $CPO = \prod_i CPO_i$ and a model with higher CPO value suggests better predictive performance than the other models and hence this model is preferred over other models. Mostly, the CPO value is close to zero thus negative of the sum of the log of the CPO_i is used as indicated by [Cai et al. \(2013\)](#), and is given by $LS_{cv} = -\sum_{i=1}^n \log CPO_i$. Thus, a model with the lowest LS_{cv} value is the best model in terms of its predictive capacity.

2.2.3 Implementation

The model parameters were determined using a Bayesian estimation approach via Markov Chain Monte Carlo (MCMC) as implemented in OpenBUGS ([Spiegelhalter et al., 2003](#)). The prior distributions for the regression coefficients and the unstructured random component were the same for all the three models. The prior distribution for the intercept was $\beta_0 \sim \text{uniform on } (-\infty, \infty)$, and the prior for the regression coefficients was $\beta_q \sim N(0, 0.00001)$ where $q = 1, 2, 3, 4$; the variance parameters σ_u^2 and σ_v^2 were given as inverse gamma prior distributions with shape and scale parameters set at 20 and 2000, respectively. The skewness parameters for ICAR-skew-t were assigned $\delta_u \sim N(0, 0.01)$ prior. We conducted a sensitivity analysis to determine the impact of the hyper-parameters of the priors on the outcome variable; for this, we chose

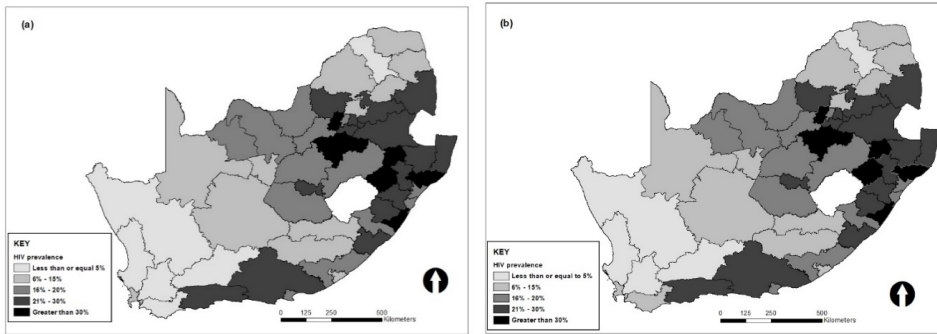
the most commonly used hyper-parameters, such as $IG(1000, 1000)$, $IG(10, 10)$, $IG(1, 10)$, and $IG(2, 2000)$. Since prior distributions with larger variances are considered in the model, the estimates from this analysis are expected to be relatively robust. Moran's I test was conducted on the model residuals to determine the presence of spatial correlation (Moran, 1950). We ran 100,000 iterations for each model to make inferences. We determined the number of initial iterations that needed to be discarded by assessing the history plots of each model and for each parameter. Similarly, we also investigated the autocorrelation plots of each model and each parameter to determine the selection intervals to avoid correlation problems in the generated chains.

2.2.4 Data

The data analyzed were obtained from the 2016 South African Demographic and Health Survey (SADHS 2016). The SADHS 2016 was conducted for evaluating the country's health programs by monitoring key milestones such as mortality, fertility, maternal and child health, nutrition, HIV, gender-based violence, etc. The data for measuring these indicators were collected by asking respondents relevant sociodemographic and behavioral characteristic questions and by collecting biological specimens. The SADHS 2016 survey employed a multistage stratified cluster sampling design to select households and/or respondents for the sample. All women between the age of 15 and 49 and men between the ages of 15 and 59 were included in the survey. Interview data were collected from a total of 8514 women and 3618 men and 6912 individuals were tested for HIV seropositivity. More information about SADHS 2016 can be obtained from the full study report (National Department of Health et al., 2019).

The observed district level HIV prevalence was computed by taking the survey design into account. The effective sample sizes in each district was determined by dividing the observed number of sample size at each district by the design effect (Kish, 1995); the effective number of HIV cases is thus the product of effective sample size and the weighted prevalence. The number of HIV tests conducted in the survey by district varied substantially, with a sample size of between 8 tests and 455 tests, with a median sample size of 111 tests. There were some districts with zero count of HIV positive individuals in the sample. For this, we assigned them the average of the simulated data from a normal distribution with mean value equal to the average of the log of prevalence in the neighboring districts and variance as the variance of the log of the prevalence p_i calculated from all the neighboring districts divided by the number of neighbors, shown in Figure 2.1b; the map in Figure 2.1a shows the raw data not adjusted for zero positive cases. A skewness test was conducted on the prevalence, with and without adjusting for zero HIV prevalence, but no significant skewness was found.

Figure 2.1: Map of HIV prevalence by district in South Africa before and after adjusting the data for zero positive tests in some districts



The covariates included in the models are the multidimensional poverty index constructed using the 2016 community survey data (Fransman & Yu, 2019), HIV prevalence among pregnant women obtained from the 2017 National Antenatal Sentinel Survey report (Woldesenbet et al., 2018), population density and male condom distribution coverage obtained from the 2017 district health barometer report (Massyn et al., 2017). Previous studies indicate that these factors are associated with HIV prevalence ecologically as well as individually (van Schalkwyk et al., 2021; Manda et al., 2015).

2.3 RESULTS

The skewness parameters for ICAR-skew-t were not significant, perhaps suggesting that the spatial component is lighter tailed (see Table 2.1). The model with the lowest LS_{cv} and DIC values was deemed to be the best model in its predictive performance and goodness of fit, respectively. Thus, as can be seen in Table 2.1, the model with the lowest LS_{cv} (170.5) is the ICAR-skew-t model, followed by the ICAR-normal model ($LS_{cv} = 172.4$). The ICAR-normal model and the ICAR-Laplace model have the lowest (291.3) and second lowest (315) DIC values, respectively. The difference in the DIC values between these models is more than five, suggesting that there is substantial difference between the two models in terms of goodness of fit to the data, according to Spiegelhalter et al. (2002); however, a study by De la Cruz & Branco (2009) indicated that DIC is not appropriate for such type of complex models. Thus, based on the LS_{cv} values, the ICAR-skew-t model was the best in terms of its predictive capacity as compared to the other two models used in this study.

As a sensitivity analysis, we ran the analysis using different sets of hyper-parameters for priors of the precision parameters. Thus, the mean difference in the values of the outcome variables at different choices of hyper-parameter values was observed at the third digit after the decimal point, which suggests the absence of a significant impact on the outcome variable. The Moran's I test statistic was significant (p-value = 0.000001), suggesting that residuals were spatially clustered. As shown in Table 2.1,

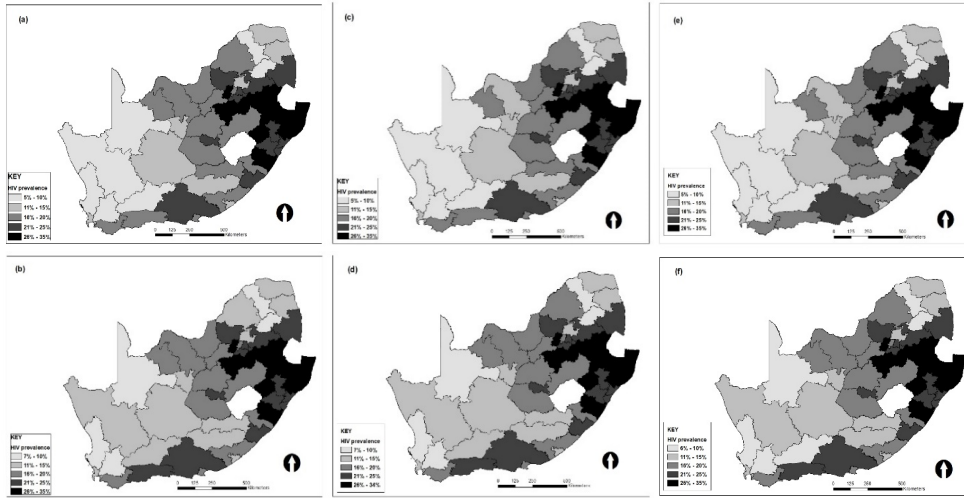
district-level ANC prevalence is the strong predictor of district-level HIV prevalence determined from the 2016 SADHS data, whereas the other covariates were not statistically significant.

Table 2.1: Comparison of the fitted models using DIC and CPO

Covariates	ICAR-normal	ICAR-Laplace	ICAR-Skew-t
Intercept	2.414 (-3.249,-1.52)	-2.526(-3.352,-1.677)	-2.606(-3.699,-1.478)
Population density	-0.0001(-0.0003, 0.0002)	-0.0001(-0.0003, 0.0002)	0.0001(-0.0003, 0.0002)
Male condom distribution	-0.0071(-0.0197, 0.0044)	-0.0059(-0.0192, 0.0054)	-0.0071(-0.0203, 0.0047)
Multidimensional poverty index	0.7312(-3.432, 4.84)	0.2791(-3.687, 4.34)	0.97(-3.129, 5.141)
ANC HIV prevalence	3.686 (1.497, 5.681)	3.964 (2.025, 5.804)	3.9 (1.767, 6.096)
σ_u^2	0.0058(0.0006, 0.7974)	0.0041(0.0005, 1.0150)	0.3453(0.0282, 0.7363)
σ_v^2	0.0055(0.0005, 0.2474)	0.01132(0.0011, 0.3004)	0.1584(0.0191, 0.4253)
δ_u			0.1265 (-0.585, 0.685)
DIC	281.9	280.7	339.4
LS_{cv}	168.5	173.2	174.1

Figure 2.2e, shows the prevalence of HIV by district in South Africa estimated using the ICAR-skew-t spatial model (best model). According to the estimates from this model, most of the districts with high levels of HIV prevalence are located in southeastern parts of the country, while low levels of HIV prevalence are in the southwestern parts. This pattern is the same for all the maps (Figure 2.2) produced using estimates from different models with or without covariates. Maps (a), (c) and (e) in Figure 2.2 are estimates of the ICAR-normal, ICAR-Laplace and skew-t models with covariates, respectively; the spatial pattern of HIV prevalence is the same for these models, except the estimate from the ICAR-normal model for one district in the north-western part. Maps (b), (d) and (f) are estimates of the ICAR-normal, ICAR-Laplace and skew-t models without covariates and the pattern of HIV prevalence by district is the same for the estimates determined using these models. One notable difference for the pattern of estimates with and without covariates for the models is that the level of HIV prevalence is lower for estimates with covariates than those without covariates in two districts in the western part.

Figure 2.2: Estimated HIV prevalence by district in South Africa with Covariates (first row) and without covariates (second row)



2.4 DISCUSSION

HIV is a leading cause of disease burden in sub-Saharan Africa. In the era of decentralized approach to governance and service provision, designing effective HIV intervention programs and monitoring strategies at local administrative levels requires reliable estimates of local variation in HIV burden. Our study compared three spatial smoothing models, namely, the intrinsically conditionally autoregressive normal, Laplace and skew-t (ICAR-normal, ICAR-Laplace and ICAR-skew-t) in the estimation of the HIV prevalence across 52 districts in South Africa. It analyzed HIV prevalence data from the 2016 South African Demographic and Health Survey. The models were fitted using the Markov Chain Monte Carlo method in OpenBUGS, a freely available Bayesian statistical package. We found that the ICAR-skew-t distribution was the best spatial smoothing model for the estimation of HIV prevalence in our study.

We found that the districts with high levels of HIV prevalence were in the southeastern parts of the country, while low levels of HIV prevalence corresponded to the southwestern parts. Our findings are similar to those by [Gutreuter et al. \(2019\)](#) and [\(Woldesenbet et al., 2018\)](#). The estimates of HIV prevalence by district in South Africa could help governmental and non-governmental organization, as well as the private sector, to know the level of the epidemics at lower administrative level, thus prioritizing and plan appropriate public health programs tailored to each community and evaluating the combined impact of national and local public health programs.

A major weakness of our study could be that there were no HIV data in some of the sparsely populated districts; hence, we simulated data from neighboring districts to estimate prevalence of HIV in such districts; thus, the estimates for these districts may not be reliable and should be interpreted with caution. In addition, a limited

number of predictors was included in the model; hence, some important predictors of district-level HIV prevalence might be missing.

2.5 CONCLUSIONS

In conclusion, alternative spatial distributions to ICAR-normal should be considered for modeling spatial disease outcomes. The spatial random effects could be skewed or non-normal and misspecification of the distribution of random effects could lead to estimates that are biased. This could lead to implications in the estimation of disease burden, adversely impacting policy derivations. In our study, we found that the intrinsic conditional autoregressive skew-t (ICAR-skew-t) model was the best in predicting district-level HIV prevalence compared to the ICAR-normal and ICAR-Laplace spatial models based on an analysis of the 2016 South African Demographic and Health Survey (2016 SADHS) data. District antennal clinic HIV prevalence was the most influential predictor of the district-level 2016 SADH HIV prevalence.

Appendix

Hierarchical representation of the disease mapping model presented in section 2.1 assuming the spatial random components follows skew-t distribution is given as follows:

Let $Y_i^* = Y_1^*, Y_2^*, \dots, Y_n^*$ be a one-dimensional random variable with binomial distribution

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{u}_i + v_i \\ v_i &\sim N(0, \sigma_v^2) \\ u_i | S_i, \sigma_u^2, \delta_u, w_i &\sim N\left(\frac{\sum_{j \sim i} u_j}{m_i} + \delta_u w_i, \frac{\sigma_s^2}{\eta * m_i}\right) \\ s_i | S_{-i} &\sim N\left(\frac{\sum_{j \sim i} s_j}{m_i}, \frac{2}{m_i}\right) \\ w_i &\sim N(0, I) I(w_i > 0), \\ \eta &\sim \text{gamma}\left(\frac{k}{2}, \frac{k}{2}\right) \end{aligned}$$

$\beta_i \sim N(\beta_0, \Lambda), i=0,1,2, \dots, k$ where k is the number of covariates

$$\begin{aligned} \sigma_v^2 &\sim IG(\Omega, v) \\ \delta_u &\sim N(0, \Gamma) \\ \sigma_s^2 &\sim IG(\Omega, u) \\ k &\sim \text{Exp}(k_0) I(k > 2) \end{aligned}$$

where p_i the weighted prevalence corresponding to Y_i^* $i = 1, 2, \dots, 52$, σ_u^2 and σ_v^2 are variance of the spatial and the heterogeneous random component and $I(w_i > 0)$ is an indicator function, IG is inverse gamma, Exp is exponential.

Based on the likelihood distribution and the above prior specifications the posterior distribution of all the parameters assuming conditional independence between the response variable and the hyper parameters is given as:

$$\begin{aligned}
 p(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \sigma_u^2, \sigma_v^2, \delta_u, w, k, \eta, s / y^*) &\propto L(y / \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \sigma_s^2, \sigma_v^2, \delta_u, w, s) P(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \sigma_s^2, \sigma_v^2, \delta_u, w, k, \eta) \\
 &= \prod_i p(y_i^* / \mu_i) \prod_j (p(\beta_j / \Lambda) p(\Lambda)) p(u / \sigma_s^2) p(\sigma_s^2) p(v / \sigma_v^2) p(\sigma_v^2) p(s / \sigma_s^2) p(w) p(\delta_u) p(k) p(\eta).
 \end{aligned}$$

SKEWED INTRINSIC CONDITIONAL AUTOREGRESSIVE SPATIAL MODELS AND THEIR APPLICATION FOR DISEASE MAPPING

3.1 INTRODUCTION

Local-level estimates of indicators are increasingly being sought from data that are originally meant to provide reliable estimates at national and regional levels (Johnson, 2004; Leyland et al., 2000). The most commonly used approach that has been used for estimating local level estimates from national survey data is a linear mixed model where the spatial component is incorporated as random effect. The random effects are assumed to collectively represent covariates that are not collected/observed as data in areas under study (and hence not considered in the modeling process as covariates) that have some form of spatial pattern (Lawson et al., 2003; Besag et al., 1991; Knorr-Held & Best, 2001). Thus areas that are closer to each other or neighboring areas are likely to have similar environmental and socioeconomic conditions and hence the estimates from these areas are expected to show some level of spatial consistency (Besag et al., 1991; Feltbower & Manda, 2012).

Most disease mapping methods focus on modeling the spatially structured random component. In disease mapping the most widely used approach for modeling the spatially structured random component (which considers the local effects) is the intrinsic conditional autoregressive normal (ICAR-normal) model suggested by Besag et al. (1991); Banerjee et al. (2003). The ICAR-normal model is a type of linear mixed effect model with structured spatial and unstructured/heterogeneous random components. The unstructured random component represents those unobserved factors with no spatial structure. The structured spatial random component is assumed to have an intrinsic conditional autoregressive normal distribution and that of the heterogeneous random component is assumed to follow a normal distribution because of its technical convenience. However, it may not be always right to assume that the random effects, both the spatial and heterogeneous component, follow some normal distribution as there is a possibility that it could follow skewed and multimodal distributions. Thus the normality assumption may not capture the distribution of the data

(Arellano-Valle et al., 2007).

Estimates of models may be biased if the distribution of random effects is wrongly specified, and it could have implications on inferences about the parameters of interest (Ghosh et al., 2007; Verbeke & Lesaffre, 1996). Also, it becomes difficult to determine the impact of the covariates on the scale and shape of the random effect distribution (Heckman & Singer, 1984; Laird, 1978; Walker & Mallick, 1997). Therefore, the most widely used approach is to transform data to distributions that are simpler for modelling such as normal by applying appropriate transformation methods. Data transformation has its own limitations such as loss of information, a model suggested for the original data set may no longer work to the transformed data set, individual data transformation may not guarantee joint normality and a data transformation method may be needed for each data set. Thus, transformation may be avoided if the distribution of the given data fits some form of theoretical distributions (Jara et al., 2008).

Consequently a double exponential, and a mixture of ICAR-normal and ICAR-Laplace/double exponential distributions were used for the structured spatial random components by Lunn et al. (2013) and Manda (2014) respectively in an effort to reduce the impact of normal assumption on the models estimates. However, there are also different types of data sets that do not follow normal, double exponential and a mixture of normal and double exponential distributions. As a result, Nathoo & Ghosh (2013) proposed a robust approach for modelling data with skewness and/or heavy tail by using a skew-elliptical distribution to the structured spatial random component. In particular they assumed that the marginal distribution of the random component follows a skew-t distribution and implemented this distribution in semi-parametric Bayesian approach. Still there are data sets which have got different structures than the normal, double exponential, mixture of these two and skew-t distributions. In this study we propose an alternative flexible approach to the ICAR-normal model for modelling the spatial random component, in a hierarchical Bayesian framework. More specifically we develop disease mapping models where the random effects follow ICAR-skew-normal and ICAR-skew-Laplace distributions.

The family of skew-normal distribution which provide an alternative robust approach for modelling asymmetric data which are analytically tractable, accommodate practical values of asymmetry have been introduced by Azzalini (1985, 1986); Azzalini & Valle (1996); Henze (1986); Azzalini & Capitanio (1999); Branco & Dey (2001) and Sahu et al. (2003). And skew-Laplace distribution was presented by (Arslan, 2010; Kotz et al., 2001; Kozubowski et al., 2013) and Okhli et al. (2017). The skew-normal and skew-Laplace distributions developed by Sahu et al. (2003) and Arslan (2010) are easier to implement in a Bayesian framework. The application of skew-normal and skew-Laplace distributions in regression analysis for modelling random effects was studied by Jara et al. (2008); Dagne (2013); Arellano-Valle et al. (2007); Kazemi et al.

(2013); Sahu et al. (2003); Lachos et al. (2009); Cancho et al. (2010); Yu & Moyeed (2001); Yavuz & Arslan (2018); Huang et al. (2016) and Galarza et al. (2017). In this study we used the skew-normal and skew-Laplace distribution developed by Sahu et al. (2003) and Arslan (2010) to develop ICAR-skew-normal and ICAR-skew-Laplace models.

The rest of this paper is organized as follows: the next section discusses about skew-normal and skew-Laplace distribution followed by a presentation on extending existing spatial models using skew-normal and skew-Laplace distribution. Then a simulation study to show the impact of wrongly specifying the distribution of random effects on spatial models on the estimates, and application of our proposed approach to real data set, using the 2016 South African Demographic and Health Survey data, are presented. The chapter concludes with a discussion, future work and limitation.

3.2 ICAR SKEW-NORMAL AND ICAR SKEW-LAPLACE DISTRIBUTIONS FOR MODELING THE STRUCTURED SPATIAL RANDOM EFFECTS AND THEIR APPLICATIONS FOR DISEASE MAPPING

3.2.1 Elliptical distribution

Let Y be a k -dimensional random vector with k -dimensional location parameter μ and a positive definite scale matrix $\Sigma_{k \times k}$, following Sahu et al. (2003)) the elliptical distribution of Y is given as:

$$f(y/\mu, \Sigma, g^k) = |\Sigma|^{-\frac{1}{2}} g^k[(y - \mu)^T \Sigma^{-1} (y - \mu)], y \in R^k \quad (3.1)$$

where $g^k(u) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ and defined by:

$$g^k(u) = \frac{\Gamma(k/2)}{\pi^{k/2}} \frac{g(u, k)}{\int_0^\infty r^{\frac{k}{2}-1} g(r, k) dr},$$

and it is called the density generator of the random variable Y , where $g(u, k) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that $\int_0^\infty r^{\frac{k}{2}-1} g(r, k) dr$ exists. Symbolically the above distribution is given as: $Y \sim El(\mu, \Sigma, g^k)$ and the corresponding cumulative density function is $F(y/\mu, \Sigma, g^k)$. In addition, the function $g(u, k)$ denotes the kernel of Y and the other terms in $g^k(u)$ constitutes the normalizing constant for the probability density function and may depend on other parameters, for example on the degrees of freedom if Y has a t-distribution. The two special cases of $g(u, k)$ that results in the multivariate normal and t-distributions are, $g(u, k) = \exp(-u/2)$ and $g(u, k, v) = (u/v)^{-(v+k)/2}$ respectively, where $v > 0$ and is the degrees of freedom. Substituting $g(u, k) = \exp(-u/2)$ in the density generator function and simplifying it results:

$g^k(u) = e^{-u/2}/(2\pi)^{k/2}$. Therefore, the probability density function, pdf of Y is:

$$f(y/\mu, \Sigma, g^k) = \frac{1}{(2\pi)^{\frac{k}{2}}} |\Sigma|^{-\frac{1}{2}} \exp[-1/2(y - \mu)^T \Sigma^{-1} (y - \mu)], y \in \mathbb{R}^k \quad (3.2)$$

which results in a k -variate normal density function, denoted as $Y \sim N(\mu, \Sigma)$.

3.2.1.A Skew-Elliptical distribution

Sahu et al. (2003) developed a new class of parametric skewed probability distributions by adding a shape parameter for skewness from the elliptically symmetric distributions. Thus following Sahu et al. (2003) a skew-elliptical distribution of a k -dimensional random vector Y is given as:

$$f(y/\mu, \Sigma, \Delta, g^k) = 2^k f(y/\mu, \Sigma + \Delta^2, g^k) P(V > 0), \quad (3.3)$$

where μ is a vector of location parameters, Σ is a covariance matrix, Δ is a diagonal matrix of skewness parameters with elements $\delta = (\delta_1, \delta_2, \dots, \delta_k)^T$, and

$$V \sim El(D(\Sigma + \Delta^2)^{-1} (y - \mu), I_k - \Delta(\Sigma + \Delta^2)^{-1} \Delta; g_{q(y_*)}^k),$$

$$g_a^k(u) = \frac{\Gamma(k/2)}{\pi^{k/2}} \frac{g(a+u, 2k)}{\int_0^\infty r^{\frac{k}{2}-1} g(a+r, 2k) dr}, a > 0 \text{ and}$$

$$q(y_*) = (y - \mu)^T (\Sigma + \Delta^2)^{-1} (y - \mu).$$

This distribution is symbolically denoted as:

$$Y \sim SE(\mu, \Sigma, \Delta, g^k).$$

If $k=1$ then $\Sigma = \sigma^2$, $\Delta = \delta$ and the distribution reduces to a univariate distribution. And the density of Y is:

$$f(y/\mu, \sigma^2, \delta, g^1) = \frac{2}{\sqrt{\sigma^2 + \delta^2}} g^{(1)} \left(\frac{(y - \mu)^2}{\sigma^2 + \delta^2} \right) F \left(\frac{\delta}{\sigma} \frac{y - \mu}{\sqrt{\sigma^2 + \delta^2}} / 0, 1; g_a^{(1)} \right),$$

where $g^1(u)$ and $g_a^{(1)}(u)$ are as defined above, and $a = (y - \mu)^2 / (\sigma^2 + \delta^2)$. The marginal probability density function of a subsets of Y is determined as $Y_i/Z > 0$, not Y_i/Z_i ; the marginal probability density function of m_1 components of Y is given as:

$$f(y^{m_1}/\mu^{m_1}, \Sigma_{11}, \Delta_{11}, g^{m_1}) = 2^{m_1} f(y^{m_1}/\mu^{m_1}, \Sigma_{11} + \Delta_{11}, g^{m_1}) P(V > 0).$$

3.2.1.B Skew-Normal distribution

A k -dimensional random variable Y is said to have a skew-normal distribution if its probability density function is defined as:

$$f(y/\mu, \Sigma, \Delta) 2^k |\Sigma + \Delta^2|^{-1/2} \phi_k \left\{ \left(\Sigma + \Delta^2 \right)^{-1/2} (y - \mu) \right\} P(V > 0), \quad (3.4)$$

where

$$V \sim N_k \left\{ \Delta \left(\Sigma + \Delta^2 \right)^{-1} (y - \mu), I_k - \Delta \left(\Sigma + \Delta^2 \right)^{-1} \Delta \right\},$$

and $\phi_k(\cdot)$ is the probability density function of a k -variate standard normal distribution. This distribution is denoted as, $SN(\mu, \Sigma, \Delta)$. An important characteristic of the above expression is that it gives independent marginal when $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2)$. And the density function is simplified as:

$$f(y/\mu, \Sigma, \Delta) = \prod_{i=1}^k \left\{ 2 \left(\sigma_i^2 + \delta_i^2 \right)^{-1/2} \phi \left(\frac{y_i - \mu_i}{\sqrt{\sigma_i^2 + \delta_i^2}} \right) \Phi \left(\frac{\delta_i}{\sigma_i} \frac{y_i - \mu_i}{\sqrt{\sigma_i^2 + \delta_i^2}} \right) \right\},$$

where ϕ and Φ are the probability and cumulative density function of a standard normal distribution. The mean and variance of a k -variate random variable Y with a skew normal distribution are given as:

$$E(Y) = \mu + \left(\frac{2}{\pi} \right)^{1/2} \Delta \text{ and } \text{cov}(Y) = \Sigma + \left(1 - \frac{2}{\pi} \right) \Delta^2.$$

A suitable stochastic representation of a k -dimensional random vector Y having a skew distribution as suggested by [Sahu et al. \(2003\)](#) and [Arellano-Valle et al. \(2007\)](#) is given as:

$$Y = X + \Delta |X_0| \quad (3.5)$$

where X and X_0 are two independent random vectors and $X \sim N_k(\mu, \Sigma)$ if Y follows $SN(\mu, \Sigma, \Delta)$ distribution. Let $w = |X_0|$, then w has a normal distribution, $w \sim N_k(0, I_k)$ truncated in the space $w > 0$. The hierarchical set-up of the above stochastic representation is:

$$Y/w \sim N_k(\mu + \Delta w, \Sigma),$$

$$w \sim N_k(0, I_k) I(w > 0) \text{ if } Y \text{ has a } SN(\mu, \Sigma, \Delta).$$

This hierarchical specification is important for modelling a skew distribution using a Bayesian approach.

Note: a skew normal distribution reduces to normal distribution if the value of the skewness parameter $\Delta=0$.

3.2.2 Skew-Laplace distribution

A random variable Y is said to have a k -dimensional skew Laplace distribution if its density function is given by:

$$f(y/\mu, \Sigma, \Delta) = \frac{|\Sigma|^{-1/2}}{2^p \pi^{\frac{p-1}{2}} \alpha \Gamma(\frac{p+1}{2})} e^{-\alpha \sqrt{(y-\mu)^T \Sigma^{-1} (y-\mu)} + (y-\mu)^T \Sigma^{-1} \Delta}, \quad (3.6)$$

where, $y \in R^p$, $\mu \in R^p$ is the location parameter, $\Delta \in R^p$ is the skewness parameter, Σ is the positive definite scatter parameter and $\alpha = \sqrt{1 + \Delta^T \Sigma^{-1} \Delta}$. Symbolically it is given as $Y \sim SL_k(\mu, \Sigma, \Delta)$.

Alternatively Y can be defined by using the normal variance-mean mixture approach which introduces randomness into the parameters of normal distribution using a mixing random variable, $w > 0$ [Arslan \(2010\)](#). The normal variance-mean mixture distribution introduces skewness to the scale mixture of normal distribution by mixing the normal distribution with different means and different variances. Based on the variance-mean mixture approach Y is defined as:

$$Y = \mu + w\Delta + \sqrt{\Sigma w} Z, \quad (3.7)$$

where $Z \sim N_k(0, I_k)$ and $w \sim \text{Gamma}(\frac{k+1}{2}, 2)$. The conditional distribution of Y given $w=w$ is denoted as:

$$f(y/w, \mu, \Sigma, \Delta) \sim N_k(\mu + w\Delta, w\Sigma).$$

If the skewness parameter, $\Delta=0$ then the probability density function of Y becomes the multivariate Laplace probability density function. The mean and variance of Y are given as $\mu + (k+1)\Delta$ and $(k+1)(\Sigma + 2\Delta\Delta^T)$ respectively.

3.2.3 Skew-Normal and Skew-Laplace structured spatial random effects distribution

The commonly used method for modeling area spatial data discussed by [Besag et al. \(1991\)](#), [Knorr-Held & Best \(2001\)](#) and [Banerjee et al. \(2003\)](#) have two random components: structured spatial and unstructured spatial random effects. One of the fundamental assumptions in these models is that the random components are assumed to follow a normal distribution. However, in situations where the distribution of data drifts away from normality and/or symmetry this assumption may lack robustness and flexibility. Therefore, in this section we relax this stringent assumption by assum-

ing that the spatially structured random effects follow a skew elliptical distribution specifically skew-normal and skew-Laplace distributions. Let $Y = Y_1, Y_2, \dots, Y_n$ be a random variable with binomial distribution then according to [Besag et al. \(1991\)](#), [Knorr-Held & Best \(2001\)](#) and [Banerjee et al. \(2003\)](#) the formulation of a spatial model is:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \mathbf{X}_i\boldsymbol{\beta}u_i + v_i, i = 1, 2, \dots, n; \quad (3.8)$$

where n is the number of regions, p_i is the proportion/prevalence in the i^{th} region, β_0 is the intercept in the regression model, $\boldsymbol{\beta}$ s are vector of regression coefficients, \mathbf{X} s are ecological covariate risk vector, the vector of spatially structured random component, u_i follows intrinsically conditional autoregressive normal distribution, $u_i \sim ICAR(\mu_u, \sigma_u^2)$ and the vector of unstructured spatial random components follow a normal distribution, $v_i \sim N(0, \sigma_v^2)$.

The assumption on u_i/u_{-i} can be relaxed by assuming that it follows ICAR-skew-normal distribution:

$$u_i/u_{-i} \sim SN\left(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i}, \delta_u\right);$$

or ICAR-skew-Laplace distribution:

$$u_i/u_{-i} \sim SL_k\left(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i}, \delta_u\right).$$

As discussed in sections 3.2.1.B and 3.2.2 ICAR-skew-normal and ICAR-skew-Laplace distributions can be represented by means of transformation and conditioning using suitable positive random vector. Thus $u_i/u_{-i} \sim SN\left(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i}, \delta_u\right)$ is alternatively expressed as:

$$u_i/u_{-i}, \sigma_u^2, \delta_u, w_i \sim N\left(\frac{\sum_{j \sim i} s_j}{m_i} + \delta_u w_i, \frac{\sigma_s^2}{m_i}\right), \quad (3.9)$$

where $w_i \sim N(0, I) I(w_i > 0)$; and for ICAR-skew-Laplace $u_i/u_{-i} \sim SL_k\left(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i}, \delta_u\right)$ can also be given as:

$$u_i/u_{-i}, \sigma_u^2, \delta_u, w_i \sim N\left(\frac{\sum_{j \sim i} s_j}{m_i} + w_i \delta_u, w_i \frac{\sigma_s^2}{m_i}\right) \quad (3.10)$$

where $w_i \sim \text{Gamma}(1, 2)$, $s_i/s_{-i} \sim N\left(\frac{\sum_{j \sim i} s_j}{m_i}, \frac{\sigma_s^2}{m_i}\right)$, σ_s^2 is the variance of s_i and δ_u is the skewness parameter.

A Bayesian estimation approach is used to determine the unknown parameters from the above disease mapping model. We follow the above hierarchical set-up of a stochastic representation of a skew random variable in order to implement the Markov Chain Monte Carlo (MCMC) parameter estimation procedure. Therefore, the hierarchical representation of the above disease mapping model assuming the random components follows a skew-normal distribution is given as follows:
let $Y = Y_1, Y_2, \dots, Y_n$ be a one-dimensional random variable with binomial distribution

$$\text{logit}(p_i) = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + u_i + v_i$$

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

$$u_i/u_{-i}, \sigma_u^2, \delta_u, w_i, \sim N\left(\frac{\sum_{j \sim i} s_j}{m_i} + \delta_u w_i, \frac{\sigma_s^2}{m_i}\right)$$

$$u_i/u_{-i}, \sigma_u^2, \delta_u, w_i, \sim N\left(\frac{\sum_{j \sim i} s_j}{m_i} + w_i \delta_u, w_i \frac{\sigma_s^2}{m_i}\right) \text{ if } u_i/u_{-i} \text{ has skew-Laplace distribution}$$

$$w_i \sim N(0, I)I(w_i > 0), \text{ or } w \sim \text{Gamma}(1, 2) \text{ if } u_i/u_{-i} \text{ has skew-Laplace distribution}$$

$$s_i/s_{-i} \sim N\left(\frac{\sum_{j \sim i} s_j}{m_i}, \frac{\sigma_s^2}{m_i}\right)$$

$$\beta_i \sim N(\beta_0, \Lambda), i=0,1,2, \dots, k \text{ where } k \text{ is the number of covariates}$$

$$\sigma_v^2 \sim IG(\Omega, v)$$

$$\delta_u \sim N(0, \Gamma)$$

$$\sigma_u^2 \sim IG(\Omega, u)$$

where $i = 1, 2, \dots, n$, σ_u^2 and σ_v^2 are variance of the spatial and the heterogeneous random component and $I(w_i > 0)$ is an indicator function, IG is inverse gamma.

3.2.4 Posterior distribution

Based on the likelihood distribution and the above priors specification the joint posterior distribution of all the parameters assuming conditional independence between the response variable and the hyper parameters is given as:

$$\begin{aligned} p\left(\mu, \boldsymbol{\beta}, \mathbf{u}, \{v, \sigma_u^2, \sigma_v^2, \delta_u, w/y\}\right) &\propto L(y/\boldsymbol{\beta}, \mathbf{u}, v, \sigma_u^2, \sigma_v^2, \delta_u, w)P(\boldsymbol{\beta}, \mathbf{u}, v, \sigma_u^2, \sigma_v^2, \delta_u, w) \\ &= \prod_i p(y_i/\mu_i) \prod_j (p(\beta_j/\Lambda)p(\Lambda))p(u/\sigma_u^2)p(\sigma_u^2)p(v/\sigma_v^2)p(\sigma_v^2)p(w)p(\delta_u). \end{aligned}$$

However, it is difficult to determine the joint posterior distribution and the marginal posterior distribution for the parameters of interest from the above expression. Thus estimating the parameters of interest will be quite complicated, as a result one needs to follow MCMC approaches in this case. The full conditional posterior distributions

are needed for the required parameters to implement the MCMC approaches and the Gibbs sampling algorithm is used in our case since the posterior conditional distributions are known and are in a closed form. The conditional posterior distributions are given as:

$$\beta_i / \mu_i, \sigma_v^2, \delta_u, Y \sim N(A_\beta^{-1}a_\beta, A_\beta^{-1})$$

$$\text{where } A_\beta^{-1} = \Lambda^{-1} + \sigma_v^{-2}X^T X \text{ and } a_\beta = \Lambda^{-1}\beta_0 + X^T(\mu - \mathbf{u})/\sigma_v^2$$

$$u_i/u_{-i}, \sigma_u^2, \delta_u, w_i, \sigma_v^2, \sigma_u^2, \mu_i, \beta = N\left(\frac{\gamma_v(\mu - X\beta) + \gamma_u(\sum_{j \sim i} \omega_{ij}(s_j + \delta_u w_i))}{\sigma_v^{-2} + \gamma_u m_i}, \frac{1}{\sigma_v^{-2} + \gamma_u m_i}\right)$$

$$\text{where } \gamma_u = \frac{1}{\sigma_u^2}, \gamma_v = \frac{1}{\sigma_v^2}, \omega_{ij} \text{ is the weight matrix of between area } i \text{ and } j$$

$$v_i/\sigma_u^2, \sigma_v^2, \mu_i, \beta = N\left(\frac{\gamma_u(\mu - X\beta)}{\sigma_v^{-2} + \gamma_u m_i}, \frac{1}{\sigma_v^{-2} + \gamma_u m_i}\right)$$

$$w_i/u_i, \delta_u^2, \delta_u = N(A_w^{-1}a_w, A_w^{-1}) I(w_i > 0)$$

$$\text{where } A_w = \delta_u^2 \gamma_u m_i + 1 \text{ and } a_w = \delta_u \gamma_u u_i \text{ or}$$

$$w_i/u_i, \delta_u^2, \delta_u \sim G(1 + \frac{(\mu_i - X_i \beta - v_i)^2}{2}, 2 + \frac{n}{2}) \text{ if } u_i/u_{-i} \text{ has skew-Laplace distribution}$$

$$\delta_u/u_i, \sigma_u^2, w_i \sim N(A_{\delta_u}^{-1}a_{\delta_u}, A_{\delta_u}^{-1}) \text{ where } A_{\delta_u} = \Gamma^{-1} + \sum_{i=1}^n \frac{w_i^2}{\sigma_u^2} \text{ and } a_{\delta_u} = \sum_{i=1}^n \frac{w_i u_i}{\sigma_u^2}$$

$$\sigma_v^2/\mu_i, \beta_i, u_i, Y \sim IG(\Omega + \frac{n}{2}, v + \frac{\sum_{i=1}^n (\mu_i - X_i \beta - u_i)^2}{2})$$

$$\sigma_u^2/\mu_i, \beta_i, u_i, Y \sim IG(\Omega + \frac{n}{2}, u + \frac{\sum_{1 \leq i \leq j \leq n} w_{ij}(u_i - u_j)^2}{2})$$

3.3 SIMULATION

In order to show the impact on the estimates of spatial models as a result of using symmetric distribution on structured spatial random effects that are skewed and to better understand the advantages and benefits of the models we proposed over the ICAR-normal and ICAR-Laplace (CAR.L1) we present a simulation study. In this section we simulated spatially structured random components that have ICAR-normal and ICAR-t distributions with outlying observation which results in data having ICAR-skew-normal and ICAR-skew-t distributions respectively. We fit all the four models; ICAR-normal, ICAR-Laplace, ICAR-skew-normal and ICAR-skew-t, to these simulated data sets and determine the best model that least violates the assumptions.

Without loss of generality, we assume that there are no covariates in the models. The spatially structured random components are simulated on the municipal area of South Africa and hence the local municipal neighborhood and weight matrix is used for simulating the data. The spatially structured random component with ICAR-normal distribution is generated from a conditionally autoregressive, CAR-normal distribution which is given as, $s \sim N(0, \Sigma)$, with $\Sigma = M(I - \rho W)^{-1}$, $M = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_{234}^2)$, W is the neighborhood (weight) matrix reflecting the local spatial effect determined from the municipal map of South Africa (Banerjee et al., 2003). The CAR distribution is well defined if Σ symmetric and positive definite which entails $\frac{w_{ij}}{\sigma_i^2} = \frac{w_{ji}}{\sigma_j^2}$ and ρ is a parameter included to enrich the spatial model and grossly indicates the global strength of the spatial effect, ρ is between $(1/\lambda_{234}, 1/\lambda_1)$ where $\lambda_1 < \lambda_2 < \dots < \lambda_{234}$ are the eigenvalues of W (Banerjee et al., 2003; Gelfand & Vounatsou, 2003). The smallest and largest eigen values of W are -3.042195 and 5.957226 respectively thus ρ is within $(-0.32871, 0.1678634)$; here we use 0.1678633 as a value of ρ to mimic ICAR prior. A value of 2 is used for σ_{is} , thus $M = \text{diag}(4)$. Therefore, we simulate a spatial random component s_i from a multivariate normal distribution with mean value zero and variance-covariance matrix, $\Sigma = 4(I - 0.1678633W)^{-1}$. The R package CAR.simWmat was used to generate the structured spatial random component (Sha, 2018). A t-distribution can be presented as a scale mixture of normal distribution (Mallows, 1974; Choy & Smith, 1997), thus a spatial random component with a t-distribution is determined by dividing a random variable s_i drawn from CAR-normal distribution by V , where $V \sim \Gamma(\frac{\nu}{2}, \frac{\nu}{2})$ and V is independent. For this exercise we used 5 as a value for ν .

The spatially structured random components generated using the above procedures are ordered into increasing order and the largest 20 observations were multiplied by 3 to introduce outliers and/or to skew the data (Nathoo & Ghosh, 2013). And for the unstructured spatial random component a data set with 234 (number of municipalities in South Africa) observations were simulated from a standard normal distribution, $v \sim N(0, 1)$. Assuming a binomial distribution for the count of disease in each municipality; we use the logit model to generate the odds and hence the prevalence from the simulated random components. The number of infected individuals is simulated using the prevalence generated above and the number of individuals in each municipality which is sampled randomly between 300 and 600.

Using the above procedures we generated a 100 data sets with 234 observations. As an example, some selected bar graphs (Figure 3.1) and (Figure 3.2) and maps (Figure 3.3) and (Figure 3.4) of the spatially structured random components having ICAR-skew-t and ICAR-skew-normal distributions on the local municipal map of South Africa are shown below. As can be seen the map of the ICAR-skew-t spatial random component (Figure 3.3) and map of the ICAR-skew-normal spatial random component (Figure 3.4) show some form of clustering and spatial correlation. And the bar graph of the ICAR-skew-t spatial random component (Figure 3.1) and bar graph of

the ICAR-skew-normal spatial random component (Figure 3.2) indicate that the structured spatial components have some outlying observations and/or are skewed. As can be seen the bar graphs the spatial random components are skewed to the right, which is what we have expected since the largest 20 simulated observations were multiplied by 3.

Figure 3.1: Bar graph of spatially structured random effects simulated from municipal map of South Africa using ICAR-t distribution with outliers

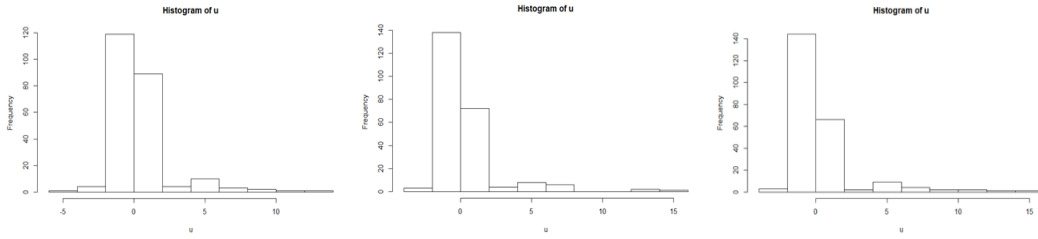


Figure 3.2: Bar graph of spatially structured random effects simulated from municipal map of South Africa using ICAR-normal distribution with outliers

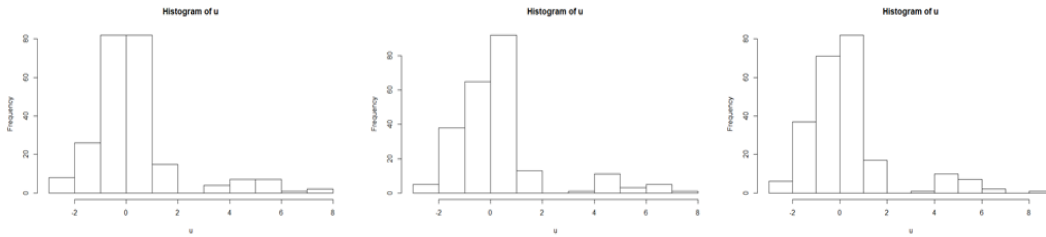
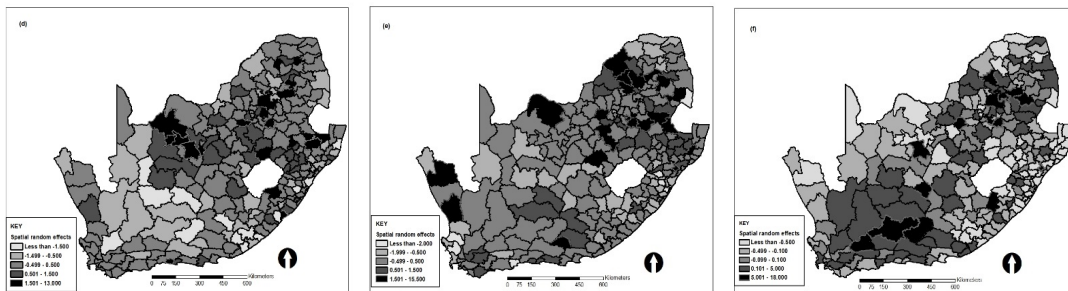
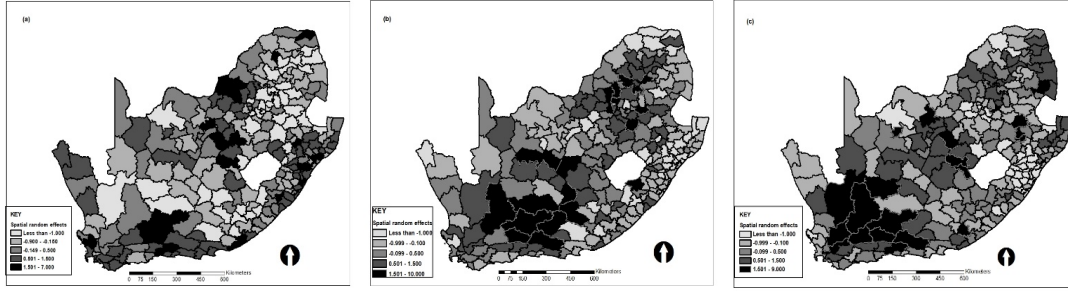


Figure 3.3: Map of spatially structured random effects simulated from municipal map of South Africa using ICAR-t distribution with outliers



The simulated data were analyzed in a hierarchical Bayesian framework. The spatially structured data simulated using ICAR-skew-t and ICAR-skew-normal distributions were analyzed using ICAR-normal model assuming that the structured random effects are normally distributed, ICAR-Laplace model assuming double exponential (Laplace) distribution for the spatially structured random components and using the

Figure 3.4: Map of spatially structured random effects simulated from municipal map of South Africa using ICAR-normal distribution with outliers



method we proposed assuming that the structured spatial components were drawn from ICAR-skew-normal and ICAR-skew-Laplace distributions. A standard normal distribution was assigned as a prior for the unstructured random component. The data analysis was conducted using OpenBUGS 3.2 statistical package.

The root mean squared error (RMSE) between the estimated prevalence and the true prevalence values (simulated following the above procedures) were computed to compare the impact of wrongly specifying the distribution of random effects on spatial models on the accuracy of the estimates. Thus, for 100 data sets the average of the root mean square error is given as:

$$ARMSE = \frac{1}{100} \left(\sum_{i=1}^{100} \sqrt{\left(\frac{1}{234} \sum_{j=1}^{234} (\hat{p}_{ij} - p_{ij})^2 \right)} \right) \quad (3.11)$$

where $i = 1, 2, \dots, 234$ and $j = 1, 2, \dots, 100$. A model with small values of ARMSE provides more accurate results. The average root mean square error of the estimates determined using models that assume the structured spatial random components are drawn from ICAR-normal, ICAR-Laplace, ICAR-skew-normal and ICAR-skew-Laplace distributions while the structured spatial random component were drawn from skew-t distribution is 0.00153, 0.00182, 0.00154 and 0.00157 respectively as shown in Table 3.1 below. Similarly, the mean square error of estimates generated using models that assume the structured spatial random components are drawn from ICAR-normal, ICAR-Laplace, ICAR-skew-normal and ICAR-skew-Laplace distributions while the simulated structured spatial random component were drawn from ICAR-skew-normal is 0.00144, 0.00146, 0.00147 and 0.00150 respectively. For both data sets the difference in root mean square error among the model estimates were observed at the fourth decimal point, and hence the difference among these errors is very small and hence can be ignored; thus on average all the models produce quite identical results. The skewness parameters are positive for both data sets suggesting that both data sets are skewed to the right and hence presense of outliers Table 3.1.

Table 3.1: Estimated values of parameters of models for structured spatial random effects simulated from ICAR-normal and ICAR-t distributions with outliers

Parameter	ICAR-Normal	ICAR-skew-Normal	ICAR-Laplace	ICAR-skew-Laplace
Estimates of parameters for structured spatial random effects simulated from ICAR-normal distribution with outliers				
Average RMSE	0.00144	0.00146	0.00147	0.00150
Percent of times LS_{cv} is lower compared to other models	12%	31%	39%	18%
σ_v^2	0.524	1.597	0.671	6.594
σ_u^2	8.338	9.500	2.186	7.674
δ_u		0.352		0.632
Estimates of parameters for structured spatial random effects simulated from ICAR-t distribution with outliers				
Average RMSE	0.00153	0.00154	0.00182	0.00157
Percent of times LS_{cv} is lower compared to other models	10%	20%	26%	44%
σ_v^2	3.026	0.485	0.562	2.113
σ_u^2	53.569	61.090	49.553	55.386
δ_u		0.258		0.508

The other approach used for comparing a model that assume a normal distribution and a skewed distribution for the spatial random components is done using conditional predictive ordinates (CPO). In this study we have determined the LS_{cv} values using the data simulated based on procedures presented above for each competing models. And the LS_{cv} values of models that assumed ICAR-normal, ICAR-Laplace, ICAR-skew-normal and ICAR-skew-Laplace distributions for structured spatial random components that were simulated from ICAR-normal distribution with outliers were lower than the rest in 12%, 39%, 31% and 18% of the time respectively. Thus the models that assumed ICAR-Laplace and ICAR-skew-normal distributions for the structured spatial random components that are simulated from ICAR-normal distribution with outliers observation are the best and second best model in terms of its predictive performance. And the LS_{cv} values of models that assumed ICAR-normal, ICAR-Laplace, ICAR-skew-normal and ICAR-skew-Laplace distributions for the structured spatial random components that were simulated from ICAR-t distribution with outlying observation were lower than the rest of the models in 10%, 26%, 20% and 44% of the time respectively. This indicates that models that assumed ICAR-skew-Laplace and ICAR-Laplace distributions for the structured spatial random components are the best and second-best models in capturing the underlying features of the ICAR-skew-t data.

A sensitivity analysis was conducted to determine the impact of the choice of parameters on the gamma distribution, that was used in the scale mixture on normal distribution to determine a t-distribution, on the accuracy and predictive capacity of the competing models. We used 50, 20, 10, 5 and 2 as a value for v . According to the sensitivity analysis the ARMSE of the ICAR-normal, ICAR-Laplace, ICAR-skew-normal

and ICAR-skew-Laplace were 0.00155, 0.00157, 0.00156 and 0.00160 respectively. Similarly, the LScv values were less than 40%, 30%, 20% and 10% of the time compared to the rest of the models for ICAR-skew-Laplace, ICAR-Laplace, ICAR-skew-normal and ICAR-normal models respectively. This suggests that the choice parameters on the gamma distribution may not have impact on the choice of models in terms of accuracy and predictive capacity.

3.4 MAPPING DISTRICT HIV PREVALENCE IN SOUTH AFRICA

The models presented in the above section are illustrated by applying district level HIV prevalence determined from complex survey data in South Africa and the best model that fits the data is used as the final model, and used for generating estimates of HIV prevalence at district level. Data obtained from the 2016 South African Demographic and Health Survey (SADHS2016) are used for this purpose. The SADHS 2016 was conducted for evaluating the country's health programmes by monitoring key milestones such as mortality, fertility, maternal and child health, nutrition, HIV, gender-based violence etc. The data for measuring these indicators are collected by asking respondents relevant sociodemographic and behavioral characteristic questions and by collecting biological specimens.

The SADHS 2016 survey employed multistage stratified cluster sampling design to select households and/or respondents for the sample. Stratification was done by dividing each province into urban, rural and farm areas. The master sampling frame prepared by Statistics South Africa for the Census 2011 was used as primary sampling units (PSUs) and each master sampling frame could be an enumeration area (EA), a group of small EAs or part of a large EA. In the first stage 750 primary sampling units were selected and the number of PSUs selected from each stratum was determined based on probability proportional to PSU size. In the second stage equal number (20) of dwelling units were selected from each selected primary sampling units. In order to obtain a nationally representative sample, 15 000 dwelling units were selected from the master sampling frame.

All households in the selected dwelling units are included in the survey. All women between the ages of 15 and 49 were asked to collect relevant sociodemographic and behavioral characteristic information from all sampled households. In every second dwelling unit relevant sociodemographic and behavioral characteristic information were also collected among all men between the ages of 15-59 in all households, in addition biological specimens and relevant health information were collected among all adults above the age of 14 years residing in the households from these dwelling units. Overall interview data were collected from a total of 8514 women and 3618 men, and HIV test was conducted among 6912 individuals. More information about SADHS 2016 can be obtained from the full study report ([National Department of Health et al.,](#)

2019).

Direct estimates of HIV prevalence by district was computed from this survey data by taking the survey characteristics into account; and then the effective sample sizes in each district were determined which are needed in modelling using OpenBUGS. And thus the effective sample sizes and corresponding number of HIV positive cases for each district was determined based on the method explained in Chapter 2 (Kish, 1995). In this survey there were districts with zero HIV positive cases due to small sample size; and we followed a similar procedure as in Chapter 2 to resolve the issue of zero HIV positive cases.

The covariates included in the models are the multidimensional poverty index constructed using the 2016 community survey data (Fransman & Yu, 2019), HIV prevalence among pregnant women obtained from the 2017 National Antenatal Sentinel Survey report (Woldesenbet et al., 2018), population density and male condom distribution coverage obtained from the 2017 district health barometer report (Massyn et al., 2017).

We fitted the ICAR-normal model, models that assume the spatial random component have ICAR-skew-normal, ICAR-Laplace/double exponential (car.l1) and ICAR-skew-Laplace to the data. And these models are defined respectively as follows:

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u_i + v_i; \quad (3.12)$$

$u_i/u_{-i} \sim N(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i})$ for ICAR-normal model; for ICAR-skew-normal model the prior distribution of the structured spatial component is replaced by:

$$u_i/u_{-i} \sim SN(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i}, \delta_u);$$

for ICAR-Laplace-model it is replaced by:

$$u_i/u_{-i} \sim L(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i});$$

and for skew-Laplace model it is assigned with:

$$u_i/u_{-i} \sim SL_k\left(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\sigma_u^2}{m_i}, \delta_u\right),$$

and $v_i \sim N(0, \sigma_v^2)$ for all the models; where p_i is the prevalence of HIV at district i and β_i s are the regression coefficients, u_i is the structured spatial random component and v_i is the unstructured spatial random component. The model parameters were determined using a Bayesian estimation approach. Prior distributions were assigned to the

model parameters and random components. The prior distributions for the regression coefficients and unstructured spatial random component were the same for all the four models. The prior distribution for the intercept is $\beta_0 \sim \text{uniform on } (-\infty, \infty)$, the prior for the rest of the regression coefficients is $\beta_i \sim N(0, 0.00001)$ where $i = 1, 2, 3, 4$; for σ_u^2 and σ_v^2 the prior was inverse gamma distribution with shape parameter set to be 20 and scale parameter equal to 2000 and prior for the skewness parameters is $\delta_u \sim (0, 0.01)$. Since prior distribution with larger variances are considered in the model, the estimates from this analysis is expected to be relatively robust. The analysis was conducted using OpenBUGS 3.2 statistical package. We run 100,000 Markov Chain Monte Carlo (MCMC) iterations for each model to make inferences. We determined the number of initial iterations that need to be discarded by assessing the history plots of each model and for each parameter. Similarly, we also investigated the autocorrelation plots of each model and each parameter to determine the selection intervals to avoid correlation problems in the generated chains.

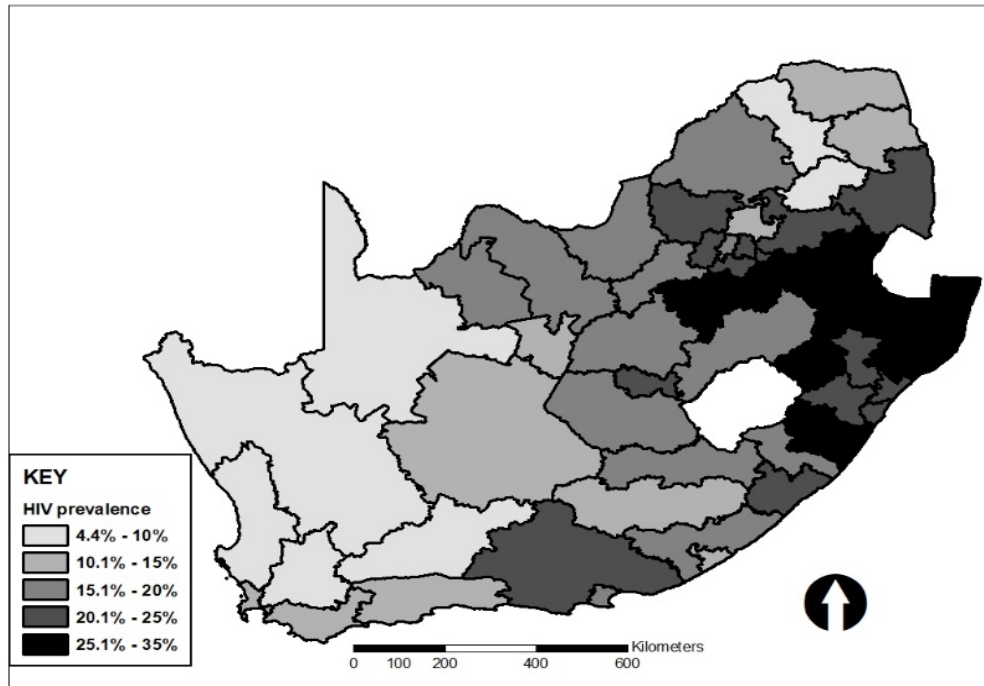
The LS_{cv} and DIC values were used to compare the models. The skewness parameters for both ICAR-skew-normal and ICAR-skew-Laplace models are not significant; perhaps suggesting that the spatial component is lighter tailed (see Table 3.2). The DIC values of the ICAR-normal and the ICAR-skew-Laplace models were the lowest (288.8) and second lowest (295) respectively. The difference in the DIC values between these two models is greater than 5 suggesting the ICAR-normal model is the best model in terms of fitting the data (Spiegelhalter et al., 2002). The ICAR-skew-normal model has the lowest LS_{cv} value (170) followed by the LS_{cv} value of ICAR-normal model suggesting that the ICAR-skew-normal model is the best model in terms of its predictive capacity. However, De la Cruz & Branco (2009) indicated that DIC is not appropriate for comparing such type of complex models. Thus, based on the LS_{cv} values the ICAR-skew-normal model is the best model as compared to the other competing models included in this study.

Table 3.2: Comparison of the fitted models using DIC and CPO/ LS_{cv}

Covariates	ICAR-Normal	ICAR-skew-Normal	ICAR-Laplace	ICAR-skew-t
Intercept	-2.473 (-3.285, -1.658)	-2.58 (-3.72, -1.453)	-2.531 (-3.319, -1.703)	-2.514 (-3.541, -1.548)
Population density	-0.0001 (-0.0003, 0.0002)	0.0001 (-0.0003, 0.0002)	-0.0001 (-0.0003, 0.0002)	-0.0001 (-0.0003, 0.0002)
Male condom distribution	-0.0070 (-0.0180, 0.0038)	-0.0067 (-0.0178, 0.0037)	-0.0064 (-0.0175, 0.0038)	-0.0069 (-0.0177, 0.0030)
Multidimensional poverty index	0.9253 (-2.964, 4.912)	0.7788 (-2.106, 4.81)	0.544 (-3.118, 4.339)	0.8523 (-2.19, 4.558)
ANC HIV prevalence	3.768 (1.79, 5.822)	3.848 (1.884, 5.583)	3.954 (2.024, 5.731)	3.897 (1.861, 5.909)
σ_v^2	0.0069 (0.0007, 0.2232)	0.0028 (0.0004, 0.1676)	0.0010 (0.0009, 0.2433)	0.0042 (0.0005, 0.1900)
σ_u^2	0.0057 (0.0005, 0.6748)	0.0066 (0.0007, 0.5841)	0.0062 (0.0006, 0.9443)	0.0062 (0.0006, 1.4747)
δ_u		0.0987 (-0.6748, 7472)		0.05 (-0.6, 0.62)
DIC	288.8	314	315	295
LS_{cv}	172	170	172.2	176.8

Figure 3.5 shows the prevalence of HIV by district in South Africa estimated using the ICAR-skew-normal spatial model. According to the estimates from this model most of the districts with the highest level of HIV prevalence are located in KwaZulu Natal, Gauteng and Mpumalanga; whereas the districts with the lowest level of HIV prevalence are located in the provinces of Northern Cape and Western Cape.

Figure 3.5: HIV prevalence by district in South Africa using 2016 SADHS data



3.5 DISCUSSION

Spatial disease mapping models have spatial random components that are often assumed to follow normal distributions for analytical tractability (Banerjee et al., 2003; Besag et al., 1991; Knorr-Held & Best, 2001). Besag et al. (1991), however, suggested that alternative distributional assumptions for the random effects could be considered. This has been the case in several research papers on spatial disease mapping models and application (Manda, 2014; Lunn et al., 2013). Mixtures of ICAR-normal and ICAR-double exponential and ICAR-Laplace distributions have been used. Kottas et al. (2008); Li et al. (2015); Hossain et al. (2013); Gelfand et al. (2005) and Gelfand et al. (2007) used a Bayesian nonparametric approach for modelling the structured spatial random component.

In this study we present alternative approaches to those presented above by assuming the structured spatial random components are assumed to have ICAR-skew-

normal and ICAR-skew-Laplace distributions. These approaches are more general and very flexible as compared to the standard approaches as the skew-normal and skew-Laplace distribution tends to normal and Laplace/double exponential distributions as the skewness parameter approaches zero; and are important for modelling data that drifts away from symmetry (Arellano-Valle et al., 2007; Sahu et al., 2003).

The proposed approaches are fitted using Markov Chain Monte Carlo methods. The skew-normal distribution presented by Sahu et al. (2003) and the skew-Laplace distribution suggested by Arslan (2010) were used in our approach as these approaches are computationally simple and easy to conduct the analysis in a Bayesian framework using OpenBUGS statistical package (Spiegelhalter et al., 2003). Conditioning and transformation of the Skew-normal and the skew-Laplace distribution made it possible to do the analysis using OpenBUGS as skew distributions are not standard form of distributions in OpenBUGS (Arellano-Valle et al., 2007). And analyzing using OpenBUGS (Freely available statistical package) makes our approach powerful and accessible to practicing statisticians and epidemiologists.

We conducted a simulation study to determine the impact of wrongly specifying the distribution of random components on the precision of the estimates and the predictive capacity of models by comparing the mean square errors of the estimates and CPO values respectively among competing models. According to the simulation analysis we found that the ARMSE values among competing models were very small for both data sets simulated using ICAR-skew-t and ICAR-skew-normal distributions suggesting that the models included in this study produced almost identical results on average. Based on the values of LS_{cv} we found that the ICAR-skew-Laplace and ICAR-Laplace models are the best models as they showed a better predictive capacity in 44% and 39% of the time as compared to models considered in this study for data sets generated using ICAR-t and ICAR-normal distributions with outlier observations respectively. Therefore, ICAR-skew-Laplace and ICAR-Laplace distributions assumption for the structured spatial random effect are robust approach for modelling when the spatial random component has ICAR-skew-t and ICAR-skew-normal distributions respectively.

The ICAR-skew-normal and ICAR-skew-Laplace distributions were fitted to the 2016 SADHS data to illustrate the applications of these models to real data set. Comparison of these models with the standard ICAR-normal and ICAR double exponential/Laplace models using DIC and CPO values indicates that the model that assumed ICAR-skew-normal distribution is the best model in terms of its predictive capacity. And a similar study that used Fernandez–Steel skew normal (FSSN) CAR model also suggested better predictive performance as compared to its CAR-normal counterpart (Rantini et al., 2021). This model is used for generating the final estimates of HIV prevalence by district in South Africa. The outputs from this model help governmen-

tal and non-governmental originations, and the private sector to know the level of the epidemics at lower administrative level, and thus prioritize and plan appropriate public health programs tailored to each community and evaluate the combined impact of national and local public health programs.

These models are used for modeling the structured spatial random component that are drawn from ICAR-skew-normal and ICAR-skew-Laplace distributions and may not be used for modelling data that are drawn from distributions that have low and high peaked distributions relative to normal distributions and heavy tailed distributions. Our approaches may not also be used for modeling data with multimodal distributions. In this study we assumed that the unstructured spatial components are drawn from a normal distribution and hence our approaches can be extended by assuming the heterogenous random component have skew-normal and skew-Laplace distributions; and a more flexible approach can also be developed for modeling random components drawn from multimodal distributions.

SKEW-NORMAL MULTIVARIATE INTRINSIC CONDITIONAL AUTOREGRESSIVE SPATIAL MODEL AND ITS APPLICATION FOR DISEASE MAPPING

4.1 INTRODUCTION

Most of the time more than one outcome/disease can be observed in a single geographical area. These diseases may have common environmental and frailty factors and hence may show some form spatial pattern (Lawson, 2008). For example, a number of studies have indicated that HIV/AIDS is fueling the TB epidemic, and viral and bacterial STIs are increasing the likelihood of HIV acquisition and transmission during sexual intercourse (Adeiza et al., 2014; Hagan et al., 2010; Middelkoop et al., 2015; Simon et al., 2006; Anderson & Maher, 2001) as such these diseases may have similar spatial patterns. The simplest and most common approach to model these diseases is to fit a univariate spatial model for each disease and include the other diseases as covariates in the model. However, the univariate spatial model doesn't take into account the association that exists among diseases. An estimation process that ignores correlations that exist among diseases may result estimates that are biased, distort regression coefficients and could lead to wrong parameter inferences (Congdon, 2007).

Consequently, in order to overcome the above problems diseases that are associated are modelled jointly (Leyland et al., 2000). Joint spatial modelling unlike univariate modeling takes into consideration the correlation that exists among diseases in the estimation process by borrowing information from other correlated diseases (Assunção & de Castro, 2004; Liu & Zhu, 2017). The joint spatial modeling of public health and epidemiological data enables to understand disease aetiology and the ability to explore shared and divergent trends in disease risk as well as increased in precision of the estimates in each collection of disease risks as it utilizes the information obtained from all other correlated diseases (Assunção & de Castro, 2004). If the interest lies in estimating the prevalence of rare disease or prevalence at lower administrative level a joint model helps to produce relatively precise and stable estimates by incorporating information from a relatively common and related diseases (Knorr-Held & Best, 2001; Manda, Feltbower & Gilthorpe, 2012). A number of joint disease mapping approaches have been proposed for example see Carlin & Banerjee (2003); Dabney & Wakefield

(2005); Knorr-Held & Best (2001); Langford et al. (1999). The shared component and multivariate spatial methods are the two most common approaches used in joint spatial modelling.

The key idea behind the formulation of the shared component model is that the risk component of diseases that have common risk factors can be divided into one that is shared by all diseases, and a risk factor that are specific to each disease (Knorr-Held & Best, 2001). These shared components are used to represent the unknown spatially structured factors that affect the risk of all of the diseases (Manda, Lombard & Mosala, 2012). This approach allows one to observe joint and specific patterns of different disease risks over geographical areas (Dabney & Wakefield, 2005; Manda, Feltbower & Gilthorpe, 2012). The multivariate spatial modelling approach accounts the correlation among disease in the modelling process by introducing a covariance matrix to the structured spatial components. The multivariate spatial model considers that the structured spatial components have a multivariate conditionally autoregressive distribution (MCAR). The MCAR model is based on the assumption of normality and this model may lack robustness and flexibility when the data drifts away from normality; as a result in the previous chapter we proposed a more flexible framework for modelling the spatial component for a single disease.

In this chapter we extend our approach presented in the previous chapter to a more generalized form for modelling diseases in a multivariate setting. Therefore, the structured spatial components in multivariate spatial model are assumed to have multivariate skew distribution specifically multivariate ICAR-skew-normal distribution with a corresponding covariance matrix to take into account the correlation among diseases. We conducted a simulation analysis to show the impact of wrongly specifying the distribution of random effects on the predictive capacity and accuracy of estimates of multivariate spatial models. In addition in order to illustrate the application of our approach to real data, we analyzed the proportion of HIV positive pregnant women who know their HIV positive status and the proportion on ART among these women by using the 2017 National Antenatal Sentinel HIV Survey data to show the application of the models we proposed.

This chapter is organized as follows: in section 2 we reviewed the methods of multivariate conditional autoregressive model. In section 3 we discussed about extending the standard multivariate spatial model to a spatial model with multivariate skew-normal distribution for the structured spatial components. A simulation study with its procedures and results were presented in section 4. An illustration of our approach to real data set were presented in section 5. In section 6 this chapter concluded with discussion, limitation and future work.

4.2 MULTIVARIATE CONDITIONAL AUTOREGRESSIVE (MCAR)

Mardia (1988) extended the theoretical explanations of a univariate CAR models presented by Besag (1974) to a multivariate setting. The theoretical development of MCAR by Mardia was further clarified and simplified by Gelfand & Vounatsou (2003) and Carlin & Banerjee (2003). Consider a vector $\boldsymbol{\phi}^T = (\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T, \dots, \boldsymbol{\phi}_n^T)$, each $\boldsymbol{\phi}_i$ is a $p \times 1$ vector and hence $\boldsymbol{\phi}$ is an $np \times 1$ vector. Assuming the $\boldsymbol{\phi}_i$ s follow a normal distribution, the density of $\boldsymbol{\phi}$ with mean zero and a $np \times np$ precision parameter \mathbf{B} is given as:

$$p(\boldsymbol{\phi}) = (2\pi)^{-np/2} |\mathbf{B}|^{1/2} \exp\left(\frac{1}{2} \boldsymbol{\phi}^T \mathbf{B} \boldsymbol{\phi}\right). \quad (4.1)$$

One needs to consider a conditional distribution from the above multivariate distribution as CAR distribution is a conditional distribution. Based on the theory of normal distribution the full conditional distribution of $\boldsymbol{\phi}_i$ is defined as:

$$p(\boldsymbol{\phi}_i / \boldsymbol{\phi}_{-i}) \propto \exp\left(-\frac{1}{2} \left(\boldsymbol{\phi}_i - B_{ii}^{-1} \sum_{j \neq i} (-B_{ij}) \boldsymbol{\phi}_j\right)^T B_{ii}^{-1} \left(-\frac{1}{2} \left(\boldsymbol{\phi}_i - B_{ii} \sum_{j \neq i} (-B_{ij}) \boldsymbol{\phi}_j\right) \boldsymbol{\phi}_j\right)\right). \quad (4.2)$$

This explanation is the same as $p(\boldsymbol{\phi}_i / \boldsymbol{\phi}_{-i}) \sim N(\sum_{j \neq i} \frac{B_{ij}}{B_{ii}} \boldsymbol{\phi}_j, B_{ii}^{-1})$. Visualizing \mathbf{B} as a nxn block matrix with $p \times p$ block we can consider B_{ij} as $p \times p$ matrix. The simplest way to understand the density of the full conditional distribution $p(\boldsymbol{\phi}_i / \boldsymbol{\phi}_{-i})$ is, the quadratic form $\boldsymbol{\phi}^T \mathbf{B} \boldsymbol{\phi}$ can be presented as $\sum_{i=1}^n \boldsymbol{\phi}_i^T B_{ii} \boldsymbol{\phi}_i + \sum_{i=1}^n \sum_{j \neq i} \boldsymbol{\phi}_i^T B_{ij} \boldsymbol{\phi}_j$, and considering only the terms that contains $\boldsymbol{\phi}_i$ and since \mathbf{B} is symmetric, $\boldsymbol{\phi}_i^T B_{ij} \boldsymbol{\phi}_j$ is equal to $\boldsymbol{\phi}_j^T B_{ij}^T \boldsymbol{\phi}_i$, therefore the density of $p(\boldsymbol{\phi}_i / \boldsymbol{\phi}_{-i}) \propto \exp(-\frac{1}{2} (\boldsymbol{\phi}_i^T B_{ii} \boldsymbol{\phi}_i + 2 \sum_{j \neq i} \boldsymbol{\phi}_i^T B_{ij} \boldsymbol{\phi}_j))$; by completing the square for the quadratic form one can get the above expression. And the full conditional distribution results in a unique joint distribution as shown by Besag (1974) using Brook's Lemma. Let $C_{ij} = \frac{-B_{ij}}{B_{ii}}$, $B_{ii}^{-1} = \Sigma_i$ where each C_{ij} is a $p \times p$ matrix, as each Σ_i s are positive definite and denotes the variance-covariance matrix of the conditional distribution, $C_{ii} = 0_{p \times p}$ and $\boldsymbol{\Sigma}$ is a block diagonal matrix with Σ_i s as blocks. Then the precision matrix, $\mathbf{B} = \boldsymbol{\Sigma}^{-1}(\mathbf{I} - \mathbf{C})$ where \mathbf{I} is the identity matrix and \mathbf{C} is a partitioned matrix with blocks C_{ij} . Thus the unique joint distribution determined from the conditional distribution, $p(\boldsymbol{\phi}_i / \boldsymbol{\phi}_{-i}, j \neq i, \Sigma_i) = N_p(\sum_{j \neq i} C_{ij} \boldsymbol{\phi}_j, \Sigma_i)$ is $N(0, (\mathbf{I} - \mathbf{C})^{-1} \boldsymbol{\Sigma})$. The requirement from the $(\mathbf{I} - \mathbf{C})^{-1} \boldsymbol{\Sigma}$ matrix is that it needs to be symmetric. The symmetric condition is satisfied if $C_{ij} \Sigma_j = \Sigma_i C_{ji}^T$. As the matrix \mathbf{C} is modelled directly in the modelling process, $\boldsymbol{\Sigma}$ needs to be specified appropriately to ensure symmetry in $(\mathbf{I} - \mathbf{C})^{-1} \boldsymbol{\Sigma}$.

According to [Besag \(1974\)](#) the most widely used CAR model for the analysis of areal data is based on neighborhood approach. The matrix C is therefore denoted as $c_{ii} = 0$ and $c_{ij} = 1/m_i$ if area j is neighbor to i and zero if not neighbors, where m_i is the number of neighbors of area i . The values of a neighborhood matrix, W is defined as $w_{ij} = 1$ if area i and j are neighbors and 0 otherwise, as area i can't be a neighbor to itself the value of $w_{ii} = 0$, then $C = W_s$ where $W_s = \text{diag}(1/m_i) W$. Based on the above expression the precision matrix B can be defined as $B = \lambda(\text{diag}(m_i) - W)$ thus $B_{ii} = \lambda m_i$ and $B_{ij} = -\lambda$. A unique joint distribution is determined from the conditional distribution with parameters C and Σ only if the expression $(I - C)^{-1}\Sigma$ is positive definite/nonsingular. Most of the time this positive definiteness assumption is not satisfied as singularity arises due to the fact that $\text{diag}(m_i) - W = 0$ or $W_s 1 = 1$ (the row sum of W_s all add up to 1). For practical purpose the ϕ_i are sampled from the full conditional distribution with a linear constraint imposed on it and hence the singularity problem is immaterial.

[Cressie \(1993\)](#) added a parameter α to the expected value of the conditional distribution, $E(\phi_i/\phi_{-i}) = \alpha \sum C_{ij}\phi_j$ to rectify the singularity problem. Therefore, the covariance matrix is given as $(I - \alpha C)^{-1}\Sigma$ and it is positive definite if α lies in the interval $(\lambda_{i \min}^{-1}, \lambda_{i \max}^{-1})$ which are the smallest and largest eigenvalues of C . And this distribution is denoted as $MCAR = (\alpha, \Sigma)$. Alternatively, according to [Carlin & Banerjee \(2003\)](#) the $B = \Sigma^{-1}(I - \alpha C)$ matrix is diagonally dominant and symmetric if $|\alpha| < 1$; according to [Harville \(1998\)](#) symmetric and diagonally dominant matrices are positive definite. The value of α needs to be close to 1 for the prior to show spatial clustering; which brings back the issue of impropriety. Some suggested a beta prior distribution, $\text{beta}(18, 2)$ for α though this approach is criticized by others.

4.3 MULTIVARIATE ICAR-SKEW-NORMAL RANDOM EFFECTS DISTRIBUTION FOR MODELLING THE STRUCTURED SPATIAL RANDOM EFFECTS

In this section we relax the multivariate ICAR-normal assumptions used for the structured spatial random effects in multivariate intrinsic conditionally autoregressive (MICAR) model by assuming that these random effects follow ICAR-skew elliptical distributions, specifically ICAR-skew-normal. Let $Y_i = Y_{i1}, Y_{i2}, \dots, Y_{ik}$ be a k dimensional random variable with binomial distribution, then according to [Besag et al. \(1991\)](#); [Knorr-Held & Best \(2001\)](#) and [Carlin & Banerjee \(2003\)](#) the standard formulation of a spatial model is:

$$\text{logit}(p_{ij}) = \beta_0 + X_i\beta + u_i + v_i, \quad (4.3)$$

where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, k$; n and k are the number of regions and diseases respectively, p_{ij} is the proportion of disease j in area i , β_0 s are vector of intercepts

in the regression model, β s are vector of regression coefficients, X s are ecological covariate risk vector, the vector of structured spatial random component u_i follow a multivariate intrinsically conditional autoregressive distribution, $u_i \sim MICAR(\mu_i, \Sigma_u)$ and the vector of heterogeneous random components follow a multivariate normal distribution, $v_i \sim MVN(0, \Sigma_v)$.

As discussed in the previous chapter the assumption on u_i/u_{-i} can be relaxed assuming that it follows a k-variate skew-normal distribution, $u_i/u_{-i} \sim SN_k\left(\frac{\Sigma_{j \sim i} u_i}{m_i}, \frac{\Sigma_u}{m_i}, \Delta_u\right)$. After transformation and conditioning using a suitable positive random vector as discussed in the previous chapter; $u_i/u_{-i} \sim SN_k\left(\frac{\Sigma_{j \sim i} u_i}{m_i}, \frac{\Sigma_u}{m_i}, \Delta_u\right)$ can be represented as:

$$u_i/u_{-i}, \Sigma_u, \Delta_u, w_{iu} \sim N_k\left(\frac{\Sigma_{j \sim i} s_j}{m_i} + \Delta_u w_{iu}, \frac{\Sigma_s}{m_i}\right), \quad (4.4)$$

where $w_{iu} \sim N_k(0, I_k) I(w_{iu} > 0)$. As discussed in the above section Σ_u is the covariance matrix and Δ_u is a diagonal matrix of skewness parameters with elements $\Delta_u = (\delta_1, \delta_2, \dots, \delta_k)^T$ and $s_i/s_{-i} \sim N\left(\frac{\Sigma_{j \sim i} s_j}{m_i}, \frac{\Sigma_s}{m_i}\right)$.

A Bayesian estimation approach is used to estimate the unknown parameters in the above disease mapping models. We follow the above hierarchical set-up of a stochastic representation of a multivariate skew-normal random variable in order to implement the Markov Chain Monte Carlo (MCMC) parameter estimation procedure. Therefore, the hierarchical representation of the above disease mapping models assuming the random components follow the above skew elliptical distribution is given as follows: let $Y_i = Y_{i1}, Y_{i2}, \dots, Y_{ik}$ be a k dimensional random variable with binomial distribution

$$\mu_i = \beta_0 + X_i \beta + u_i + v_i$$

$$v_i \sim MVN(0, \Sigma_v)$$

$w_{iu} \sim N_k(0, I_k) I(w_{iu} > 0)$ if u_i/u_{-i} has a skew-normal distribution

$u_i/u_{-i}, \Sigma_u, \Delta_u, w_i \sim N_k\left(\frac{\Sigma_{j \sim i} s_i}{m_i} + \Delta_u w_{iu}, \frac{\Sigma_s}{m_i}\right)$ if u_i/u_{-i} has a skew-normal distribution

$$s_i/s_{-i} \sim N\left(\frac{\Sigma_{j \sim i} s_j}{m_i}, \frac{\Sigma_s}{m_i}\right)$$

$$\beta_i \sim N(\beta_0, \Lambda),$$

$$\Sigma_v \sim IW(\Omega, h)$$

$$\Delta_u \sim N(0, \Gamma)$$

$$\Sigma_u \sim IW(D, h)$$

where $i = 1, 2, \dots, n$, Σ_u and Σ_v are covariances of the spatial and the heterogeneous random components and $I(w_{iu} > 0)$ is an indicator function, IW is inverse Wishart.

4.3.1 Posterior distribution

In Bayesian estimation posterior distributions are required to estimate and make inferences about the parameters of interest. The prior distribution is combined with the data generating process (likelihood function) to determine the posterior distribution, and it is the distribution of the parameters after observing the data (Lesaffre & Lawson, 2012). Assuming conditional independence the response variable and the hyper parameters the joint posterior distribution is defined as:

$$p(\mu_i, \beta_i, u_i, v_i, \Sigma_u, \Sigma_v, \delta_u, w_{iu} / y_i) \propto L(y_i / \mu_i, \beta_i, u_i, v_i, \Sigma_u, \Sigma_v, \Delta_u, w_{iu}) P(\beta_i, u_i, v_i, \Sigma_u, \Sigma_v, \Delta_u, w_{iu})$$

$$= \prod_i p(y_i / \mu_i) \prod_j (p(\beta_j / \Lambda) p(\Lambda)) p(u_i / \Sigma_u) p(\Sigma_u) p(v_i / \Sigma_v) p(\Sigma_v) p(w_{iu}) p(\Delta_u).$$

Estimation of parameters is done using the MCMC approach and the full conditional distributions are required for this purpose. Thus, the conditional distributions of the parameters are presented below:

$$\beta_i / \mu_i, \Sigma_u, \Sigma_v, y_i \sim N(A_\beta^{-1} a_\beta, A_\beta^{-1})$$

$$\text{where } A_\beta^{-1} = \Lambda^{-1} + \Sigma_v X^T X \text{ and } a_\beta = \Lambda^{-1} \beta_0 + X^T (\mu - \mu_i) / \Sigma_v$$

$$u_i / \mu_i, \Sigma_u, \Delta_u, w_{iu}, \Sigma_v, \mu_i, \beta_i = N \left(\frac{\gamma_v (\mu_i - X \beta_i) + \gamma_u (\sum_{j \sim i} \omega_{ij} (s_j + \Delta_u w_{iu}))}{\Sigma_v^{-1} + \gamma_u m_i}, \frac{1}{\Sigma_v^{-1} + \gamma_u m_i} \right)$$

$$\text{where } \gamma_u = \Sigma_u^{-1}$$

$$v_i / \Sigma_u, \Sigma_v, \mu_i, \beta_i = N \left(\frac{\gamma_u (\mu_i - X \beta_i)}{\Sigma_v^{-1} + \gamma_u m_i}, \frac{1}{\Sigma_v^{-1} + \gamma_u m_i} \right)$$

$$w_{iu} / u_i, \Sigma_u, \Delta_u = N(A_w^{-1} a_w, A_w^{-1}) I(w_i > 0)$$

$$\text{where } A_w = \Sigma_u \gamma_u m_i + 1 \text{ and } a_w = \Delta_u \gamma_u u_i$$

$$\Delta_u / u_i, \Sigma_u, w_{iu} \sim N(A_{\Delta_u}^{-1} a_{\Delta_u}, A_{\Delta_u}^{-1}) \text{ where } A_{\Delta_u} = \Gamma^{-1} + \sum_{i=1}^n \frac{w_{iu}^2}{\Sigma_u} \text{ and } a_{\Delta_u} = \sum_{i=1}^n \frac{w_{iu} u_i}{\Sigma_u}$$

$$\Sigma_v / \mu_i, \beta_i, u_i, \mu_i = IW \left(\Omega + \sum_{i=1}^n (\mu_i - u_i - X \beta_i) (\mu_i - u_i - X \beta_i)^T, h + n \right)$$

$\Sigma_u/\mu_i, \beta_i, u_i, \mu_i = IW(D + u'(D_w - W)u, k + n)$, where D_w is a diagonal matrix whose elements are the sum of neighbors of a region and W is the spatial proximity matrix.

And the Gibbs sampling algorithm is used in our case since the posterior conditional distributions are known and are in a closed form.

4.4 SIMULATION

We conducted a simulation study to show the performance of our proposed approach for modeling multivariate spatially structured random components that have outliers/wider tails, and its performance is compared against the common modeling approaches such as MICAR. In this section we considered two scenarios where in the first case data are simulated by assuming that the structured spatial random components are generated from a bivariate ICAR-normal distribution with outliers and in the second case assuming that the spatially structured random components are drawn from a bivariate ICAR-t distribution with outliers with spatial pattern. In both scenarios without loss of generality we assume that there are no covariates in the models.

We used the formulation by [Gelfand & Vounatsou \(2003\)](#) and [Banerjee et al. \(2003\)](#) about MCAR to simulate the spatially structured random effects. Therefore, let $S^T = s_1, s_2, \dots, s_n$ where each s_i is $p \times 1$ vector; then S^T is said to have MCAR distribution if its density function is given as $S \sim N(0, (D_w - \rho W)^{-1} \otimes \Sigma)$; where D_w is a diagonal matrix whose elements are the sum of neighbors of region i (w_{i+}), W is the neighborhood/spatial proximity matrix and whose value is 1 if i and j are neighbors and 0 otherwise, Σ is a variance covariance matrix of S with $p \times p$ dimension, ρ is a value included to overcome singularity problem in Σ^{-1} since $(D_w - W)1 = 0$ and it measures spatial association; $\rho \in (\lambda_{max}^{-1}, \lambda_{min}^{-1})$ where λ_{min} and λ_{max} are the minimum and maximum eigenvalues of $D_w^{-1}W$, $\lambda_{max} = 1$ and $\lambda_{min} < 0$. The weight matrix is determined from the municipal map of South Africa, $\rho = 0.99$ to mimic the MICAR prior, and $\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$. The R package MVS.CARleroux was used with slight change to generate S . The bivariate spatially structured random components with t-distribution were determined from S using the approach presented in the previous chapter section 3. The bivariate spatially structured components with normal and t-distributions were ordered into increasing order and the largest 20 observation were multiplied by 3 to generate spatially structured bivariate normal and t-distributed random components with outliers.

Since a convolution model is used, the bivariate unstructured random components are drawn from a bivariate normal distribution. Assuming a binomial distribution for the count of each disease in each municipality of South Africa; we use the logit model to generate the odds and hence the prevalence from the simulated random

components. The number of infected individuals by each disease is simulated using the prevalence generated above and the number of individuals in each municipality which is sampled randomly between 400 and 700.

Using the above procedures we generated a bivariate 100 data sets with 234 observations for each variable. To have a visual inspection of the data the figures below show bar graphs and maps for some spatially structured random components simulated using ICAR-Skew-normal and ICAR-skew-t distributions (each row of graphs and maps represents a set of joint observations). As can be seen in the bar graphs the spatially structured random components are skewed to the right, and these random components exhibit some form of spatial clustering and joint associations, see Figure 4.1 and Figure 4.2.

Figure 4.1: Bar graph of spatially structured random effects simulated from multivariate ICAR-t distribution with outliers

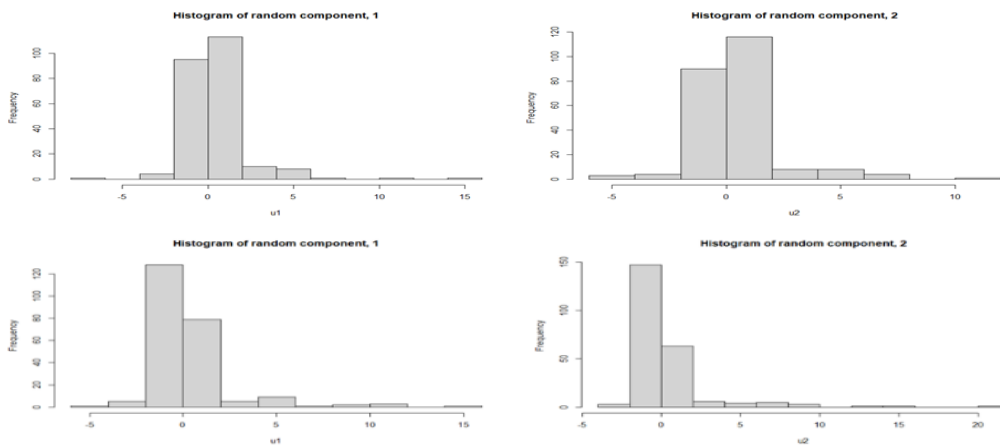


Figure 4.2: Bar graph of structured spatial random effects simulated from multivariate ICAR-normal distribution with outliers

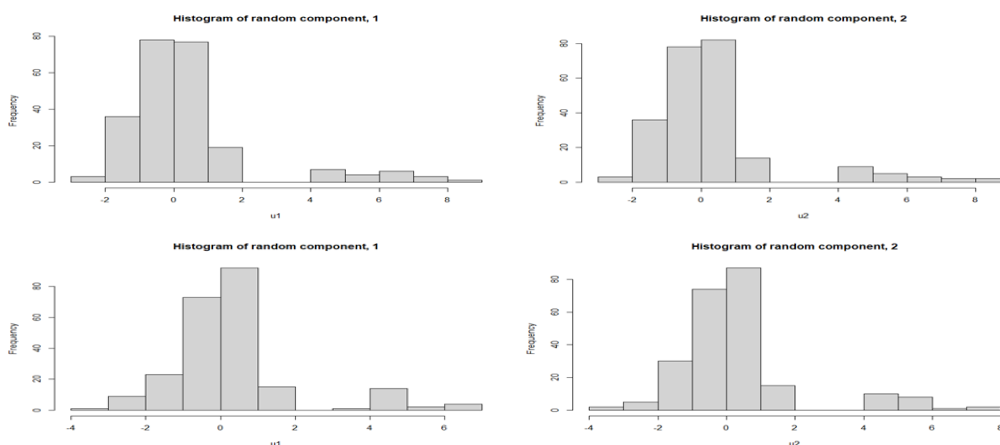


Figure 4.3: Map of structured spatial random effects simulated from multivariate ICAR-t distribution with outliers

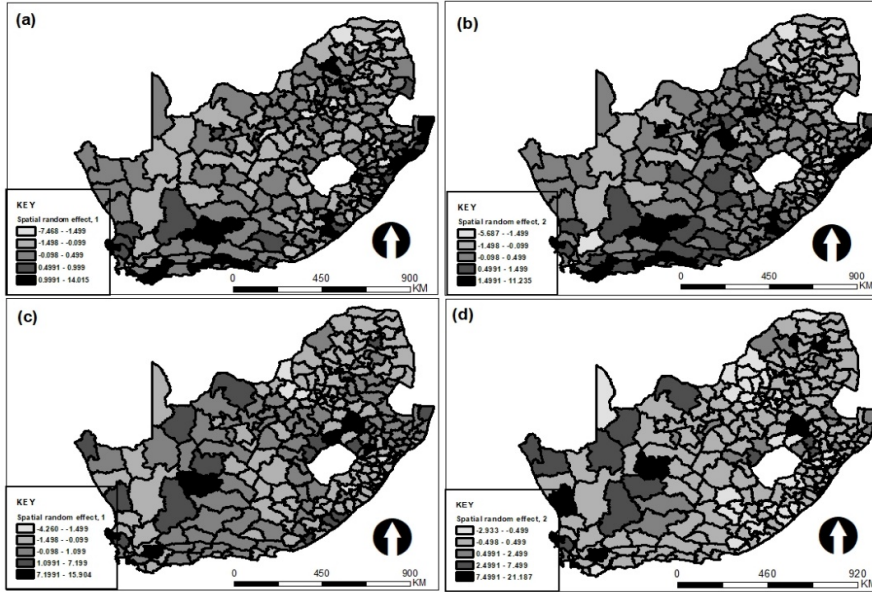
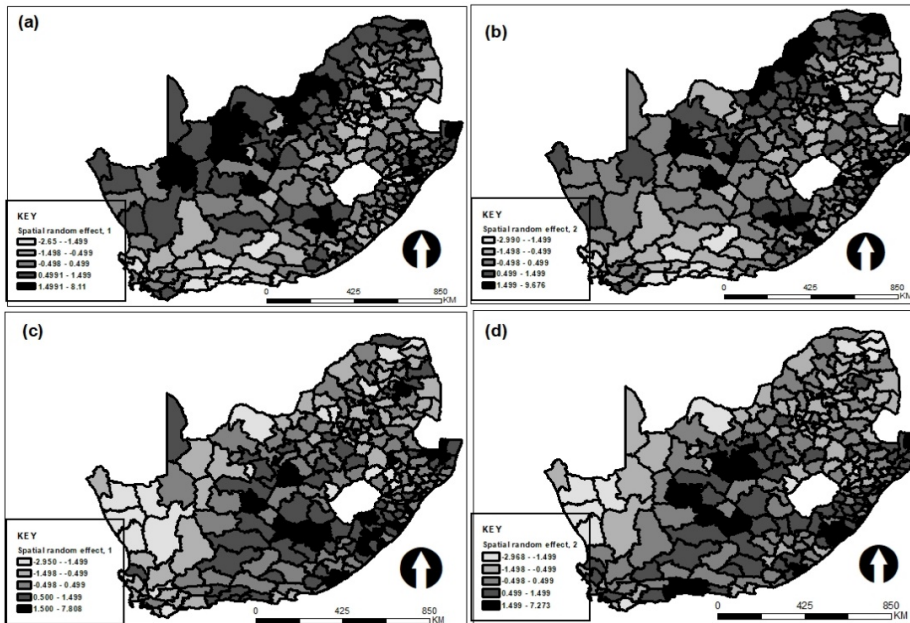


Figure 4.4: Map of structured spatial random effects simulated from multivariate ICAR-normal distribution with outliers



We analyzed the simulated data using a Bayesian approach. The simulated spatially structured random components generated using the above procedures were analyzed using areal spatial models. The MICAR-normal model that assumes the structured spatial random components are drawn from multivariate normal distribution and our proposed approach that assumes the random components are drawn from multivari-

ate skew-normal (MICAR-skew-normal) distribution were fitted to the data. And the nonspatial random components were assumed to be drawn from a bivariate normal distribution. Non-informative priors are used for all parameters used in the model. The statistical package used for data analysis is OpenBUGS 3.2.

Table 4.1: Estimated values of parameters of models for structured spatial random effects simulated from multivariate ICAR-normal and ICAR-t distributions with outliers

Parameter	MICAR-Normal model	MICAR-skew-Normal model
Estimates of parameters for structured spatial random effects simulated from MICAR-normal distribution with outliers		
Average RMSE	0.00176	0.00177
Percent of times LS_{cv} is lower compared to other models	42%	58%
Σ_{u11}	8.694	9.242
Σ_{u12}	-9.254	-9.876
Σ_{u21}	-9.254	-9.876
Σ_{u22}	12.046	12.785
Σ_{v11}	12.432	101.982
Σ_{v12}	-11.972	-4.119
Σ_{v21}	-11.972	-4.119
Σ_{v22}	12.379	85.631
δ_{u1}		0.361
δ_{u2}		0.359
Estimates of parameters for structured spatial random effects simulated from MICAR-t distribution with outliers		
Average RMSE	0.00190	0.00191
Percent of times LS_{cv} is lower compared to other models	31%	69%
Σ_{u11}	84.862	86.357
Σ_{u12}	-48.389	-44.654
Σ_{u21}	-48.389	-44.654
Σ_{u22}	99.132	102.628
Σ_{v11}	1.924	20.738
Σ_{v12}	-1.585	0.837
Σ_{v21}	-1.585	0.837
Σ_{v22}	1.856	24.486
δ_{u1}		0.317
δ_{u2}		0.348

As discussed in section 3 from the previous chapter, the root mean square error (RMSE) and negative of log of conditional predictive ordinate (LCsv) methods were used for comparing the methods we proposed fitting for multivariate skewed spatially structured random components with existing approaches. As shown in Table 4.1, the value of ARMSE for the MICAR-normal model and the MICAR-skew-normal model fitted to the simulated data drawn from the MICAR-t distribution with outliers were 0.00190 and 0.00191 respectively. And the ARMSE value calculated from the MICAR-normal model and the MICAR-skew-normal model fitted to the simulated data drawn from the MICAR-normal distribution with outliers were 0.00176 and 0.00177 respectively. As we can see in Table 4.1 the ARMSE values are the same until the 4th digit

after the decimal point for both data sets and these differences are small enough and hence can be ignored. Thus this suggests that overall there is no difference between the two models in terms accuracy.

The LScv values of MICAR-skew-normal model fitted to the simulated structured spatial data drawn from the MICAR-normal distribution with outliers is lower than the LScv values of the MICAR-normal model fitted to the same data set 58% of the time, see Table 4.1. Similarly, LScv values were lower 69% of the time when structured spatial random components simulated from MICAR-skew-t distribution with outliers were fitted to the MICAR-skew-normal model than the MICAR-normal model. Thus, the MICAR-skew-normal model is better in terms of its predictive capacity as compared the MICAR-normal model.

4.5 MAPPING DISTRICT HIV PREVALENCE AND PROPORTION ON ART AMONG HIV POSITIVE PREGNANT WOMEN WHO KNOW THEIR HIV STATUS IN SOUTH AFRICA

In order to illustrate the application of the spatial models discussed in the above section to a real data we modelled direct survey estimates of proportion of pregnant women who know their HIV positive status and proportion of pregnant women who know their HIV positive status and are on ART in South Africa. The best model that fits these data are used as the final model and used for generating knowledge of HIV positive status among pregnant women and proportion of pregnant women on ART at district level. Direct estimates of proportion of pregnant women who know their HIV positive status and proportion pregnant women who know their status and are on ART at district level is computed from the 2017 National Antenatal Sentinel HIV survey data. This survey was conducted among pregnant women attending public antenatal clinics for measuring the distribution of HIV infection, to monitor trends of HIV prevalence, to provide scientific evidence to monitor development goals, to evaluate HIV prevention and treatment programs targeting pregnant women, for strategic response and planning, and data source for modelling HIV epidemic etc. The data collection procedure for measuring these indicators are through a brief interview, medical chart review and blood specimen collection.

The 2017 National Antenatal Sentinel HIV Survey is a cross-sectional survey of pregnant women attending antenatal services from public health facilities in South Africa. The survey selected a representative sample of 32,716 pregnant women from 1595 sentinel clinics distributed in all the 52 districts of South Africa. The study design of this survey is stratified cluster sampling design, taking districts as strata and sentinel clinics as clusters. The sample size in the survey was determined to estimate HIV prevalence by district at a level of precision of 3-5%, 10% error rate and a de-

sign effect of 1.5%. In order to ensure representativeness of the sample the facilities in each district were stratified by urban, rural and size (small, medium and large). The selected facilities were proportionally allocated to each stratum based on number of facilities.

Public health facilities that offer pregnancy testing and antenatal care services at least for 20 pregnant women attending the service for the first time per month, routinely collects blood among ANC clients and have facilities to store sera at four degree Celsius and able to transport collected blood specimens to the nearest laboratory within 24 hours are included in the study. Inclusion is also based on if staff in the facility have the capacity to conduct the study and willingness of the facility to be included in the study. All pregnant women between the ages of 15-49 accessing the ANC service in these facilities are invited to participate in the study regardless of whether this visit to the facility is first time or follow-up. The required data for the study are collected from consenting pregnant women attending the service consecutively and data collection continues until the sample size assigned to each clinic is achieved or until the end of the study period. Please refer the study report for more information about the 2017 ANC survey (Woldesenbet et al., 2018).

The survey data are weighted by taking the survey characteristics and response rates at each stratum into account in order to reduce bias in the direct district level estimates computed from the survey data. The effective sample sizes for each outcome variables are determined from the weighted direct estimates as they are required for modelling proportions in a Bayesian framework using OpenBUGS.

The multidimensional poverty index determined from the 2016 community survey data (Fransman & Yu, 2019), HIV prevalence among pregnant women obtained from the 2017 National Antenatal Sentinel Survey report (Woldesenbet et al., 2018), population density and male condom distribution coverage obtained from the 2017 district health barometer report (Massyn et al., 2017) are the factors included in the spatial model.

The spatial models fitted to the data are MICAR-normal and MICAR-skew-normal that assume a multivariate normal and multivariate skew-normal distributions for the structured spatial random components respectively. These models are defined respectively as:

$$\text{logit}(p_{ij}) = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + \beta_{14}X_4 + u_{1i} + v_{1i},$$

where $u_{1i}/u_{-1i} \sim N_k\left(\frac{\sum_{j \sim i} u_{1j}}{m_i}, \frac{\sum u}{m_i}\right)$ and $v_i \sim N_k(0, \Sigma_v)$;

$$\text{logit}(p_{ij}) = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 + \beta_{24}X_4 + u_{2i} + v_{2i},$$

where $u_{2i}/u_{-2i} \sim SN_k\left(\frac{\sum_{j \sim i} u_{2j}}{m_i}, \frac{\Sigma_u}{m_i}, \Delta_u\right)$, $v_i \sim N_k(0, \Sigma_v)$,

and $i = 1, \dots, 52$, $j = 1, 2$; i and j denotes number of districts and outcome variables respectively, p_{ij} is the proportion of pregnant women who know their HIV positive status and proportion of pregnant women who know their HIV status and are on ART at district i and β_{kq} s are vector of regression coefficients and k is the number of alternative models to be fitted to the data, β_{k0} s are vector of intercepts, u_{ki} is the vector of structured spatial random component and v_{ki} is the vector of unstructured spatial random component for knowledge of HIV status and proportion of pregnant women on ART. The model parameters were determined using a Bayesian estimation approach. Prior distributions were assigned to the model parameters and random components. The prior distributions for the regression coefficients and non-spatial random component were the same for all the models. The prior distribution for the intercept is $\beta_{k0} \sim dflat()$, the prior for the rest of the regression coefficients is $\beta_{kq} \sim N(0, 100000)$ where $q = 1, 2, 3, 4$; and for the precision parameters of the structured spatial and unstructured spatial random components the priors are $\Sigma_u \sim IW(D, v)$ and $\Sigma_v \sim IW(\Omega, v)$ respectively; where $D = \Omega = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$ is the scale matrix and $v = 2$ is the degrees of freedom. The prior for the skewness parameter is $\delta_{uk} \sim N(0, 100)$. The analysis was done using OpenBUGS 3.2 statistical package. We run 100,000 Markov Chain Monte Carlo (MCMC) iterations for each model to make inferences. We determined the number of initial iterations that need to be discarded by assessing the history plots of each model, and for each parameter and outcome variables. Similarly, we also investigated the autocorrelation plots of each output to determine the selection intervals to avoid correlation problems in the generated chains. The above models were compared using the LCsv and DIC values.

Based on the results of the analysis the value of the skewness parameters for the MICAR-skew-normal model are close to zero and insignificant which may suggest that the spatial components are lighter tailed. The model that fits the data best is the one with the lowest LCsv and DIC values. As shown in Table 4.2 the values of the LCsv for the multivariate ICAR-normal model is 286.9 and for the MICAR-skew-normal the value is 282.2 suggesting that the MICAR-skew-normal performs relatively better in terms of its predictive capacity as compared to the multivariate ICAR-normal. In addition, the DIC values for the MICAR-skew-normal model and MICAR-normal model are 487.1 and 508.8 respectively. The difference in the DIC values between the MICAR-normal model and the MICAR-skew-normal model is greater than 10 which is a strong evidence that the earlier model fits the data better than the earlier. According to Lunn et al. (2013) and Stone & Zhu (2015) DIC may not be used for comparing models if the posterior density reflects extreme skewness or multimodality. In our

study an assessment of the posterior density plots of the outputs of the analysis do not indicate presence of extreme skewness and multimodality. In general, we suggest that the model that assumes MICAR-skew-normal distribution for the structured spatial component is an adequate model to capture the underlying features of the data. The posterior estimates of the fixed effect parameters across all the models are closer as shown in Table 4.2 with corresponding confidence intervals.

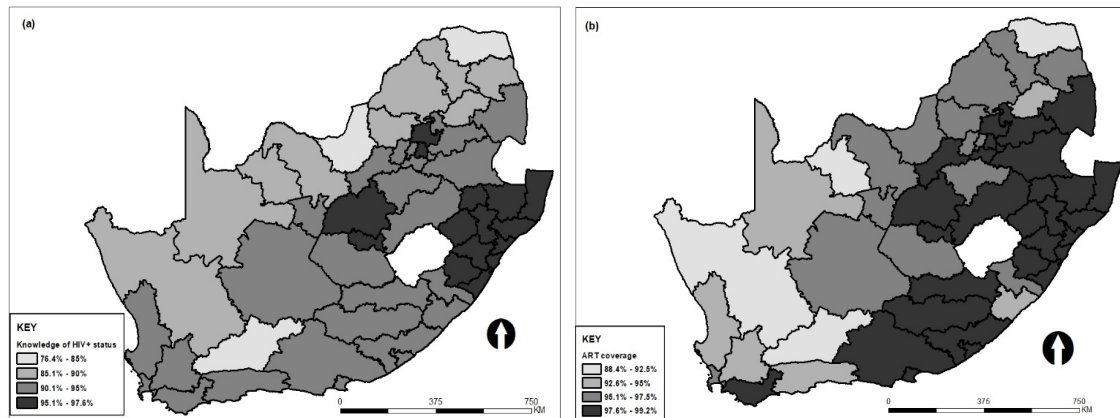
Table 4.2: Posterior mean and 95% HPD intervals for the parameters of interest, and DIC and CPO/ LS_{cv} values for the different models in this study

Covariates	MICAR-normal		MICAR-skew-normal	
	Proportion of pregnant women who know their HIV status	ART coverage among pregnant women	Proportion of pregnant women who know their HIV status	ART coverage among pregnant women
Intercept	2.037(1.06, 2.96)	2.345(1.285, 3.461)	1.851(0.6982, 2.8780)	2.35(1.077, 3.716)
Population density	-0.00005(-0.0003, 0.0002)	-0.0002(-0.0005, 0.0002)	-0.00003(-0.0003, 0.0002)	-0.0002(-0.0005, 0.0002)
Male condom distribution	-0.0062(-0.0174, 0.0051)	-0.0058(-0.0194, 0.0081)	-0.0052(-0.0163, 0.0061)	-0.0045(-0.0188, 0.0108)
Multidimensional poverty index	-0.4082(-4.802, 4.199)	-3.417(-8.550, 1.834)	-0.6929(-5.466, 3.920)	-2.734(-8.806, 3.91)
ANC HIV prevalence	2.821(0.351, 5.319)	5.746(3.344, 8.223)	3.108(0.6273, 5.4380)	6.035(3.269, 8.798)
Σ_{u11}	56.76 (2.311, 261.3)		69.39 (2.513, 363.2)	
Σ_{u12}	-77.26 (-357.5, 103.6)		-65.04 (-355.1, 146.4)	
Σ_{u21}	-77.26 (-357.5, 103.6)		-65.04 (-355.1, 146.4)	
Σ_{u22}	244.8 (16.96, 793)		252.3 (16.48, 774.2)	
Σ_{v11}	225 (10.15, 769.6)		200.3 (9.396, 765.8)	
Σ_{v12}	-8.09 (-262.3, 246.8)		-3.864 (-271.2, 255.8)	
Σ_{v21}	-8.09 (-262.3, 246.8)		-3.864 (-271.2, 255.8)	
Σ_{v22}	57.94 (3.855, 364.2)		166.1 (6.369, 638.7)	
Skewness parameter for prop. Knowledge of HIV status, δ_{u1}			0.09313 (-0.5774, 0.8614)	
Skewness parameter for prop. ART coverage, δ_{u2}			-0.2094 (-1.091, 1.032)	
DIC	286.9		282.2	
LS_{cv}	508.8		487.1	

Figure 4.5 below indicates maps of proportion of HIV positive pregnant women who know their HIV positive status (a) and proportion on ART among those who

know their HIV positive status (b) by district in 2017 South Africa, generated based on the best model that fits the underlying features of the data. According to the estimates using this model knowledge of HIV positive status among pregnant women is above 95% in all districts in KwaZulu Natal, and in a few districts in Gauteng and Free state provinces. The districts with the highest proportion of pregnant women who know they HIV positive status are Ugu (97.6%), Umzinyathi (96.7%) and iLembe (96.5%) and those districts with the lowest proportion are Vhembe (76.4%), Central Karoo (81.9%) and Ngaka Modiri Molema (83.5%). A higher percentage of districts in South Africa (71.2%) have proportion of pregnant women who know their HIV positive status less than the target set by UNAIDS to achieve the HIV Epidemic control by 2030. The proportion of HIV positive pregnant women who know their HIV positive status and are on ART is above 90% in all the districts except Central Karoo (81.9%). Almost 80% of districts in South Africa have proportion of pregnant women who know their HIV positive status above the target (95%) set by UNAIDS to achieve the HIV Epidemic control by 2030.

Figure 4.5: Map of proportion of pregnant women who know they HIV status and the proportion on ART among these women by district in 2017, South Africa.



4.6 DISCUSSION

In the previous chapter we presented a univariate spatial modelling approach assuming that the structured spatial components are drawn from a symmetric and skewed distribution. If more than one outcome is observed across each district univariate approach does not take into account the association that exist between outcomes. In this chapter we presented joint modelling of outcomes as this approach considers the association that exists between outcomes in the modelling process (Assunção & de Castro, 2004; Liu & Zhu, 2017). And the estimates based on this approach are relatively precise and stable in particular if the outcome under study is rare and if there is no sufficient sample size to generate reliable estimates (Feltbower & Manda, 2012;

Congdon, 2007).

We reviewed the common methods of joint spatial modelling of Knorr-Held & Best (2001) and Carlin & Banerjee (2003). The standard joint spatial methods assume the random components are normally distributed. This assumption may not be always right since in real world problem we could come across with data which have got multimodal and skewed distributions (Ghosh et al., 2007; Verbeke & Lesaffre, 1996). Thus, we suggested alternative generalized parametric approaches for modelling the spatial random components. Our approach is specifically used for modeling multivariate skewed-normal spatially structured random components. This approach is a generalization to the standard (symmetric) assumption as multivariate skew-normal distribution reduces to multivariate normal when the skewness parameter tends to zero.

In order to formulate our approaches to a spatial model we adopted a class of multivariate skew-elliptical distribution developed by Sahu et al. (2003) which is based on conditioning and transformation. In addition to the parameters that measures location and scatteredness, like symmetric distribution, the skew-normal distribution has additional parameter that controls skewness of the distribution. The skew-normal distribution used in this study is simple and convenient to formulate in a Bayesian framework and conduct the analysis using OpenBUGS statistical package.

Using a simulation analysis we have investigated the impact of wrongly specifying the distribution of multivariate structured spatial random effects in spatial models; and compared the performance of the MICAR-skew-normal model against MICAR-normal model in terms of accuracy and predictive capacity by evaluating their RMSEs and CPOs. Assessment of the CPO values suggest that the MICAR-skew-normal model that assumes multivariate skew-normal distribution for the structured spatial random components that are drawn from multivariate ICAR-skew-normal distribution or multivariate ICAR-skew-t distribution has a better predictive capacity as compared to MICAR-normal model that assumes a normal distribution for these random components.

The ARMSE values of the MICAR-skew-normal model that assumes multivariate skew-normal distribution for the structured spatial random components simulated from multivariate ICAR-skew-t and multivariate ICAR-skew-normal distributions were determined and compared with that of MICAR-normal model that assumes a multivariate normal distribution for these random effects. The difference in the ARMSE values between these models for both data sets simulated from multivariate ICAR-skew-normal and multivariate ICAR-skew-t distributions were very small and can be ignored. Thus, there is no difference between the MICAR-normal model and the MICAR-skew-normal model in terms of accuracy. And perhaps suggesting

that using normal distribution assumption for spatially structured random components that are skewed may not have negative impact on the accuracy of estimates as compared to the other model considered in this study. Similarly a simulation study conducted by [Kim & Mallick \(2004\)](#) for point referenced data indicated that a model that assumed skew-normal distribution to random effects that are drawn from skewed distribution has a better predictive performance as compared to a model that assumed normal for the random effects.

We have also illustrated the use and feasibility of our approaches for modelling real data. As an example, we model ART use and knowledge of HIV status among pregnant women by district using the 2017 ANC survey data assuming the spatially structured components have a multivariate skew-normal distribution using Bayesian smoothing techniques. We have investigated and compared the performance of MICAR-skew-normal and MICAR-normal spatial models using DIC and CPO values. According to the values of these indicators the spatial model (MICAR-skew-normal) that assumes the structured spatial components are drawn from multivariate skew-normal distribution appears to perform better than the MICAR-normal model; suggesting that the MICAR-skew-normal model has a better predictive capacity and fits the district ART coverage and knowledge of HIV status data better as compared to the MICAR-normal model. The posterior estimates of the fixed effect parameters between the two spatial models are relatively closer; which is also the case in other linear mixed models ([Ghidey et al., 2004](#); [Jara et al., 2008](#); [Zhang & Davidian, 2001](#)). The estimates of the MICAR-skew-normal are used for generating the maps shown in section 5, and hence relevant stakeholders can use these results to determine the level of the epidemic among pregnant women, monitor and evaluate the impact of different HIV interventions at district level in South Africa and design appropriate HIV intervention programme tailored to each community.

In this chapter we have considered extending multivariate ICAR-normal distribution assumptions in spatial model to multivariate ICAR-skew-normal distributions. These approaches may not be used for modelling spatially structured components that are drawn from data that have high and low peaked distributions as well as heavy tailed distribution. Moreover, our approach may not be used for modelling random components drawn from multimodal and mixture distributions. Therefore, our approach can be extended for modelling structured spatial random components that have mixture distributions and high or low peaked distributions; and it can also be extended for modelling the unstructured spatial random component. In addition the simulation study was conducted by generation data using positive skewness parameter; and thus our approach can be tested and verified by simulating data that have negative skewness or opposite skewness.

MULTIVARIATE BAYESIAN NONPARAMETRIC DISEASE MAPPING USING AREAL STICK-BREAKING PRIORS

5.1 INTRODUCTION

Data are becoming increasingly available as rates, summary of counts aggregated over different geographical areas such as census tracts, post or zip codes, districts, or counties etc. in order to protect patient confidentiality (Li et al., 2015). These data sets show some form of spatial pattern, and modeling of these data needs to take this spatial pattern into account. The most common form of modelling for such data is to introduce the structured spatial component as random effect, and mostly the estimation is done in a hierarchical Bayesian framework (Gelfand et al., 2005; Manda, Feltbower & Gilthorpe, 2012). In Bayesian hierarchical spatial modeling the spatially structured random effects are assumed to follow exchangeable prior distributions which are taken from a specific family of parametric distributions mostly normal, Gamma, Beta etc. (Jara et al., 2009; Walker & Mallick, 1997).

However, it is difficult to determine the accuracy of this assumption since the random effects are not measurable and hence most of the time the distribution of the random effects are unknown (Jara et al., 2009; Müller & Quintana, 2004). The reason for assuming a particular parametric distribution to the structured spatial random effects is because of its technical convenience (Jara et al., 2009; Müller et al., 2015). Therefore, it may not be always right to assume that the spatially structured random effects follow some known parametric distribution as there is a possibility that it could follow multimodal distributions perhaps unknown number of modes.

Consequently, mixture distributions were suggested for the structured spatial random effects in an effort to reduce the impact of parametric/distributional assumption on the estimates of parameters in a spatial model. For example, in an effort to relax the above restrictive assumption; (Manda, 2014) assumed a mixture of conditionally autoregressive (CAR) normal and CAR double exponential for the spatially structured random effects. Similarly, Langford et al. (1999); Moraga & Lawson (2012) assumed that the structured spatial random effects are drawn from weighted normal distributions. However, the limitation with the above approaches is that it is difficult to

determine the number and type of mixture distributions used for modelling the structured spatial random component (Manda, 2014).

Thus a more flexible approach that reduce the impact of misspecification of the distribution of structured spatial random effects on the estimates is Bayesian nonparametric modelling of spatial random effects (Gelfand et al., 2005; Heckman & Singer, 1984; Manda, 2011). In Bayesian nonparametric modelling the random effects are assumed to follow some unknown distributions (Müller & Quintana, 2004). In other words it is assumed that the random effects can best be described by an infinite dimensional parametric family; and hence the prior becomes a probability model on the infinite dimensional space which could capture different possible distributions (DeYoreo & Kottas, 2015). For example if θ_i is the random effect then its distribution is given as $\theta_i/G \sim^{iid} G$, with a Bayesian nonparametric prior for the unknown G ; and unlike the parametric case the distribution of the random effects becomes an unknown distribution. The Bayesian nonparametric prior for G is: $G/\eta \sim \pi(.|\eta)$, where η is the hyper-parameter matrix. Mostly the Dirichlet process is used as a prior for the family of distributions of G which involves a baseline distribution, H and a concentration parameter, α and it is denoted as $G \sim DP(\alpha, H)$ (Ferguson, 1973; Kottas et al., 2008; Gelfand et al., 2005).

Duan et al. (2007); Griffin & Steel (2006) and Gelfand et al. (2005) applied Bayesian nonparametric model with Dirichlet process prior to model spatially referenced data. In Gelfand et al. (2005) spatial dependency was introduced through the zero mean Gaussian base distribution from which the mixing components are drawn. Duan et al. (2007) generalizes that of Gelfand et al. (2005) assuming that the spatially referenced data and hence the spatially structured random effects are coming from a variety of surfaces (and hence different weights) instead that they are coming from a randomly selected single surface (common weight). Whereas Griffin & Steel (2006) introduced spatial dependence through mixing weights using order based stick-breaking prior. They determined the ordering assuming that distributions of similar covariate values have similar ordering. Hossain et al. (2013) used a Bayesian nonparametric method for modelling areal data by using stick-breaking prior, which is one of the representations of the Dirichlet process and it is discussed in the sections below. They introduced the spatial dependence by defining a spatial model through the mixing weights of the stick-breaking prior. A covariate dependent kernel function is included in the mixing weights of the stick breaking prior in order to introduce the spatial dependence between areas. Li et al. (2015) also used stick-breaking priors for modelling a univariate areally-referenced data. They introduced the spatial dependence by including a conditionally autoregressive prior on the weight function of the stick-breaking process. In this study we extend this model to a multivariate setting.

The work in this chapter is structured as follows: section 2 briefly defines Dirichlet distribution and explains Dirichlet process including Dirichlet process mixture distribution. Multivariate disease mapping using areally referencing stick-breaking approach is presented in section 3. A posterior distribution about the model parameters and a discussion about truncated Dirichlet process were discussed in section 4 and 5. Application of the method we proposed to a real data followed by discussion of results was presented in section 7 and 8 respectively; and finally the chapter concluded with a discussion in section 9.

5.2 DIRICHLET PROCESS MIXTURE

5.2.1 Dirichlet distribution

Dirichlet distribution is a distribution over $n-1$ dimensional simplex and is considered as a multivariate generalization of a beta distribution (Kotz et al., 2004). Let $p = p_1, p_2, \dots, p_k$ be a k dimensional random variables such that $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$, and let $(\alpha_1, \alpha_2, \dots, \alpha_k)$ be a k dimensional parameter vector, $\alpha_i \geq 0 \forall k$ and $\sum_{i=1}^k \alpha_i > 0$, then the Dirichlet distribution of p is given as,

$p = (p_1, p_2, \dots, p_k) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_k) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)} \prod_{i=1}^k p_i^{\alpha_i-1}$ and is denoted as:

$$p \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_k).$$

Mixture distributions are used for modeling clustered data with multimodal distributions, and Dirichlet distribution prior is used for modelling the weights in a mixture model. Therefore, we consider the following scheme which provides a basis to define a suitable prior; let's consider a two-component symmetric Dirichlet distribution and hence the scaling parameter α is divided equally between the two components:

$$p^2 = (p_1^{(2)}, p_2^{(2)}) \sim \text{Dirichlet}(\frac{\alpha}{2}, \frac{\alpha}{2}).$$

According to the expansion rule the components are divided as:

$$\theta_1^{(2)}, \theta_2^{(2)} \sim \text{Beta}(\frac{\alpha}{2} \times \frac{1}{2}, \frac{\alpha}{2} \times \frac{1}{2})$$

$$p^4 = (\theta_1^{(2)} p_1^{(2)}, (1 - \theta_1^{(2)}) p_1^{(2)}, \theta_2^{(2)} p_2^{(2)}, (1 - \theta_2^{(2)}) p_2^{(2)}) \sim \text{Dirichlet}(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}).$$

Thus for k -components we get, $p^k \sim \text{Dirichlet}(\frac{\alpha}{k}, \dots, \frac{\alpha}{k})$; similarly a prior over an infinite space of distribution is determined as $k \rightarrow \infty$.

5.2.2 Dirichlet process

Dirichlet process is defined using two parameters, H which denotes the central/base-line distribution over some space that defines the expectation of the process and α which is defined as a concentration/precision parameter. The Dirichlet process is denoted as $G \sim DP(\alpha, H)$ and thus α measures the variability of G around the base measure H . Larger values of α implies that G is closer to the base distribution H . Based on the above scheme, let $p \sim \lim_{K \rightarrow \infty} \text{Dirichlet}(\frac{\alpha}{K}, \dots, \frac{\alpha}{K})$; for each point in this Dirichlet distribution, randomly select values, θ_k from the base distribution, i.e. $\theta_k \sim H$ for $k = 1, \dots, \infty$; thus alternatively G is defined as, $G = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}$, which is an infinite discrete distribution over the base distribution, H which is continuous; where θ'_k s are atoms drawn from the base distribution, δ_{θ_k} is the Dirac measure (point mass) associated with θ_k .

The most important representation of the Dirichlet Process is the stick-breaking construction defined by (Sethuraman, 1994). In this construction the weights are determined using a Beta distribution with parameters 1 and α . Thus from the discrete representation of the Dirichlet Process, $G = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}$, the value of $p_1 = V_1$ and for any $k > 1$, $p_k = V_k \prod_{l=1}^{k-1} (1 - V_l)$ with $V_l \sim \text{iid Beta}(1, \alpha)$. The weights are generated as, break a stick of unit length randomly according to a Beta distribution, $\text{Beta}(1, \alpha)$ and assign length of the break part V_1 to p_1 , again breaking the remaining part of the stick, $1 - V_1$ randomly according to a beta distribution $\text{Beta}(1, \alpha)$ and assign $p_2 = V_2(1 - V_1)$ etc. and this process shows that the value of p_k approaches zero at a higher rate. As a result, most of the time the infinite sum is replaced by a sum of the first m terms after which the value of $V_l = 1$ and hence the value p_k is zero. The number of terms, m is determined as number that results some empty components when MCMC is running or by examining the size of the last weight under the prior.

Since the base distribution is continuous one may assume the Dirichlet process as a continuous process, however, the Dirichlet Process generates discrete distributions that are mixtures of point masses whose values are drawn randomly from the base distribution and whose weights are determined using the stick-breaking process with probability one (Duan et al., 2007; Fuentes & Reich, 2013; Gelfand et al., 2005; Kottas et al., 2008). Dirichlet process prior could result in estimates that are inconsistent if it is used for modelling continuous distribution in some hierarchical models (Diaconis & Freedman, 1986). This unattractive property makes the Dirichlet Process difficult to use as a prior for continuous distributions. This limitation can be resolved by convolving the Dirichlet Process with some continuous kernel, or more generally, by using a Dirichlet Process to define a mixture distribution. This is known as a Dirichlet process (DP) mixture (Jara et al., 2009; Müller et al., 2015; Phadia, 2013). And it is defined as:

$$Y_i \sim p(.|\theta_i), \theta_i \sim G, G \sim DP(\alpha H), \text{ where } p(.|\theta_i) \text{ is some parametric distribution.}$$

Based on the stick-breaking representation it is also given as: $Y_i \sim p(.|\theta_i)$, $\theta_i \sim \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$, where $\theta_k^* \sim H$ and π_k is determined using a stick-breaking construction. Thus as an example; let $Y_i \sim N(\mu_i, \sigma_i^2)$, $(\mu_i, \sigma_i^2) \sim \sum_{k=1}^{\infty} \pi_k \delta_{(\mu_k^*, \sigma_k^{2*})}$, where the base distribution H is: $\mu_k^* / \sigma_k^{2*} \sim N(m_k, s_k^{2*})$ and $\sigma_k^{2*} \sim IG(c_k, d_k)$. It can also be rewritten as: $Y_i \sim \sum_{k=1}^{\infty} \pi_k N(\mu_k^* \sigma_k^{2*})$ since each (μ_i, σ_i^2) must be equal $(\mu_k^* \sigma_k^{2*})$ for some k in clustering.

5.3 MULTIVARIATE DISEASE MAPPING USING AREALLY-REFERENCED SPATIAL STICK-BREAKING PRIOR

Reich & Fuentes (2007) developed a Bayesian nonparametric model using a Dirichlet process mixture prior which was based on a stick-breaking construction for analyzing geostatistical data. As discussed in the above section they assigned unknown spatially smoothed prior distributions for modelling the structured spatial random effects. Using a kernel function they extended the stick-breaking construction for modelling spatial data by adding a spatial component through the mixing weights of the stick-breaking construction. Thus, let $u(s)$ be the spatially structured random effect assigned a prior distribution, $u(s) \sim F(s)$. Based on stick-breaking representation the prior for $u(s)$ is given as, $F(s) \sim \sum_{k=1}^m p_k(s) \delta_{\theta_k}$, where $p_1(s) = V_1(s)$, and $p_k(s) = V_k(s) \prod_{l=1}^{k-1} (1 - V_l(s))$ for $k > 1$, and $V_k(s) = w_k(s) V(k)$. The spatially dependent kernel function $w_k(s)$ lies between $[0, 1]$ and is distributed with mean, and variance which depends on the spatial location of the observations.

Li et al. (2015) adopted the above spatial stick-breaking approach to analyze a univariate areally-referenced data whose spatial patterns are determined based on neighborhood approaches of regions. Similarly, they modelled the spatial dependence of areal data by adding additional component in the mixing weights of the stick-breaking process which has conditional autoregressive, CAR priors. Therefore, in this section we extend the work of Li et al. (2015) to a multivariate ICAR setting. Let $\mathbf{Y}_i = Y_{i1}, Y_{i2}, \dots, Y_{ih}$ be h dimensional random variable with binomial distribution (For example number of cases of diseases); then the areally-referenced stick-breaking model of the occurrence of diseases of interest in region i truncated at m terms is given as:

$$\text{logit}(\mathbf{p}_i) = \boldsymbol{\beta}_0 + \boldsymbol{\beta} X_i^T + \mathbf{u}_i + \mathbf{v}_i ,$$

where $i = 1, 2, \dots, n$ and n is the total number of regions, $\mathbf{u}_i = u_{i1}, u_{i2}, \dots, u_{ih}$ and are the spatially structured components, \mathbf{p}_i is the prevalence of disease j , where $j = 1, \dots, h$ (number of diseases) in region i , $\mathbf{v}_i = v_{i1}, v_{i2}, \dots, v_{ih}$ and are the spatially unstructured components having the following prior distribution,

$$\mathbf{v}_i \sim MVN(0, \boldsymbol{\Sigma}_v),$$

$u_i \sim G^{(i)}$, where $G^{(i)} = \sum_{k=1}^m p_{ik} \delta_{\theta_k}$ and is the stick-breaking prior for the spatial random effect.

$Z_{i1}, \dots, Z_{im} \sim \text{categorical}(p_{i1}, \dots, p_{im})$, where Z_{ik} s are clustering or classification variables

$$\theta_k \sim \text{MVN}(0, \Sigma_\theta),$$

$p_{i1} = V_1 w_{i1}$, $p_{ik} = V_k w_{ik} \prod_{l=1}^{k-1} (1 - V_l w_{il})$ for $k > 1$, thus the mixing weights depend not only on the V_k s but also on the w_{ik} which are the "location" weight parameters and $w_{ik} = w_{ik1}, w_{ik2}, \dots, w_{ikh}$. Since the ICAR distribution has the support over the entire real line, the location weight parameter is transformed into a logit scale, $\text{logit}(w_{ik}) = \mathbf{Z} \mathbf{C}_{ik}$ and the $\mathbf{Z} \mathbf{C}_{ik}$ s are assumed to have as multivariate ICAR-normal prior distribution, and $V_k \sim \text{iid Beta}(1, \alpha)$. For the sake of completing the Bayesian model, the priors for the other parameters are $\beta \sim N(\mu_0, \Lambda)$, $\Sigma_\theta \sim \text{IW}(\Omega_\theta, v)$.

5.4 POSTERIOR DISTRIBUTION

Posterior inferences about the parameter of interest are based on MCMC approach and hence the full conditional posterior distributions are needed to implement this approach. The blocked Gibbs sampler developed by [Ishwaran & James \(2001\)](#) which directly samples from the posterior of the random measure is used for sampling from the posterior. Thus, let Z_1^*, \dots, Z_N^* be the set of N unique values of \mathbf{Z} ; the blocked Gibbs sampling is used to simulate the parameters of interest from the following conditional posterior distributions: for $Z_j \in \mathbf{Z} - \{Z_1^*, \dots, Z_N^*\}$ simulate $\theta_j \sim \text{MVN}(0, \Sigma_\theta)$ and draw $\theta_{Z_j^*}$ from the density

$$f(\theta_{Z_j^*} / Y, \mathbf{Z}, \Sigma_\theta) \sim N(\theta_{Z_j^*}, \mathbf{0}, \Sigma_\theta) \prod_{(i: Z_i = Z_j^*)} p(\mathbf{y}_i / \theta_{Z_j^*})$$

$$\sim N(\mu_\theta^*, \Sigma_\theta^*) \text{ where } \mu_\theta = \left(\sum_{\{i: Z_i = Z_j^*\}} \frac{1}{\Sigma_e^2} + \frac{1}{\Sigma_\theta} \right)^{-1} \sum_{\{i: Z_i = Z_j^*\}} (\mathbf{y}_i - (\beta_0 + \beta \mathbf{X}_i^T))$$

and

$$\Sigma_\theta^* = \left(\sum_{\{i: Z_i = Z_j^*\}} \frac{1}{\Sigma_e^2} + \frac{1}{\Sigma_\theta} \right)^{-1}.$$

The conditional for Z_i is given as:

$$(Z_i / \theta, \mathbf{y}) \sim \sum_{k=1}^m p_{k,i} \delta_{\theta_k} \text{ where } i = 1, 2, \dots, n$$

where $(p_{1,i}, p_{2,i}, \dots, p_{m,i}) \propto (p_{1i} p(\mathbf{y}_i / \theta_1, \beta_0, \beta), p_{2i} p(\mathbf{y}_i / \theta_2, \beta_0, \beta), \dots, p_{mi} p(\mathbf{y}_i / \theta_m, \beta_0, \beta))$

$$p_{1i} = V_1^* w_{1i}, p_{ki} = V_k^* w_{1i} \prod_{l=1}^{k-1} (1 - V_l^* w_{il}), k = 2, 3, \dots, m-1, i = 1, 2, \dots, n,$$

conditional for $\logit(w_{ik}) = ZC_{ik}$ is:

$$ZC_{ik} \sim \sum_{l=1}^m p_{ik} \prod_{(s < l)} p_{il} \exp\left(-\frac{n_i \tau_{ZC}}{2} (ZC_{is} - \overline{ZC_{is}})^2\right)$$

Conditional for τ_{ZC} is: $\tau_{ZC} \sim \text{Gamma}(a_k + \frac{nm}{2}, a_k + \frac{1}{2} \sum_l^m \sum_{i \sim j} (ZC_{il} - ZC_{il})^2)$
 $V_k^* \sim \text{Beta}(1 + r_k, \alpha + \sum_{l=k+1}^m r_l)$, for $k = 1, 2, \dots, m-1$ and r_k records the number of Z_i values which equals k .

The conditional posterior distribution for the remaining coefficients were the same as those presented in the previous chapters.

5.5 TRUNCATED DIRICHLET PROCESS

Having infinite number of clusters is computationally expensive and is a challenge to generate practical Markov Chain Monte Carlo algorithms (Ishwaran & Zarepour, 2000; Reich & Fuentes, 2007). As a result, the infinite Dirichlet process mixture is approximated by a finite Dirichlet process mixture that truncates at m . Thus $\sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} = \sum_{k=1}^m \pi_k \delta_{\theta_k}$ for Dirichlet process, and $\sum_{k=1}^{\infty} \pi_k p(\cdot/\theta_i) = \sum_{k=1}^m \pi_k p(\cdot/\theta_i)$ for Dirichlet process mixture. In order to have a proper probability distribution the sum of the weights, $p_1 + p_2 + \dots + p_m = 1$ in other words $p_m = 1 - \sum_{k=1}^{m-1} p_k$; thus to ensure this property the value of $V_m = 1$. A Dirichlet process is truncated at m if the difference between the expected value of p_m and p_{∞} is small enough to be ignored. Thus, Ishwaran & James (2001) and Ishwaran & James (2002) presented a method that relates the sample size with the number of clusters and investigated the absolute difference between the expected value of p_m and p_{∞} by taking different values of sample size (n), and α . And their method is defined as; $|E(p_m - p_{\infty})| \leq 4 \left(1 - E \left[\left(\sum_{k=1}^{m-1} p_k \right)^n \right] \right)$ this is simplified as,

$$|E(p_m - p_{\infty})| \sim 4n \exp(-(m-1)/\alpha). \quad (5.1)$$

Similarly, Ohlssen et al. (2007) suggested an approach that approximates the number of cluster in a Dirichlet process; in this approach the basic idea is to set m which results in a small $E(p_m)$; thus $E(p_m) = E(1 - \sum_{k=1}^{m-1} p_k) \approx \varepsilon$. And, $E \left(1 - \sum_{k=1}^{m-1} p_k \right) = \left(\frac{\alpha}{\alpha+1} \right)^{m-1} \approx \varepsilon$, after rearranging and simplifying we get;

$$m \approx 1 + \frac{\log \varepsilon}{\log[\alpha/(1+\alpha)]}. \quad (5.2)$$

Fixing ε to a certain constant value the relationship between α and m is almost linear; since $\log[\alpha/(1+\alpha)]$ is approximated as $-1/\alpha$ for moderate value of α , thus

$m \approx 1 - \alpha \log \varepsilon$. For example after setting $\varepsilon = 0.01$, a conservative approximation to the number of clusters needed to approximate Dirichlet process is, $m = 5\alpha + 2$ according to [Ohlssen et al. \(2007\)](#). A value of α close to zero suggests that the data has only one cluster which is modelled by using a common mean model whereas a large value, $\alpha \rightarrow \infty$ indicates each data set (random effect if used it for modelling random component) may represent a separate cluster. The challenge is in determining the value of α and identify if it is large or small for modelling a given data set ([Ohlssen et al., 2007](#)). Perhaps a flexible approach may be to specify a prior for α which allows it to adapt to the data. Thus [Ohlssen et al. \(2007\)](#) suggested a uniform prior for $\alpha \sim \text{Unifrom}(0.3, 10)$. The upper bound 10 results in a reasonably large value of m and a lower bound of 0.3 is chosen to overcome computational issues due to small value for p_k . A study by [Liu \(1996\)](#) showed a relationship between α and number of occupied clusters k , thus the conditional expectation of k is, $E(k/\alpha, M) \approx \alpha \ln(\frac{\alpha+M}{\alpha})$ and its standard deviation is, $SD(k/\alpha, M) \approx \sqrt{\alpha(\ln(\frac{\alpha+M}{\alpha}) - 1)}$; where M is the number of random components. For example, if $\alpha = 5$ and $M=100$, $E(k/\alpha, M) \approx 15$ and $SD(k/\alpha, M) \approx 3.2$. In this study we use this expression to determine the number of occupied clusters.

5.6 DATA

In order to show the application of the multivariate Bayesian nonparametric approach for spatial modelling presented in section 5.3 we use data obtained from the 2017 South African National HIV Prevalence, Incidence, Behavior and Communication Survey. The 2017 South African National HIV Prevalence, Incidence, Behavior and Communication Survey was conducted primarily to measure HIV prevalence, incidence, proportion of males who are circumcised, proportion of HIV positive individuals who are on ART, to measure if the country is on track to achieve the 90%-90%-90% targets set by UNAIDS, risk factors of HIV infection etc. The data for measuring these indicators were collected by asking respondents related questions and collecting blood specimens from respondents. The survey interviewed a total of 36, 609 individuals of which 22368 agreed to provide dried blood spots for HIV testing. The survey was powered (80% power) to measure a five percent change in HIV prevalence overtime by sex, age group, race, locality type and province with a precision level of less than $\pm 5\%$ and a design effect of 2. Moreover 16 districts were oversampled to estimate HIV prevalence in these districts. The survey employed a multistage stratified cluster sampling design to select households and/or respondents. A sample of 1000 primary stage units/census enumeration areas, EAs were selected from a list of 84,907 small area layers (SALs) and these selected enumeration areas were stratified by province and locality type in order to ensure representativeness of the sample population. The SALs were disproportionately assigned to strata that were determined based on province and geography type to get sufficient sample for the white, colored and Indian community.

From the selected SALs a systematic random sample of 15 visiting points/households were selected and person of all ages living in the selected households were invited to participate in the survey. More information about the 2017 South African National HIV Prevalence, Incidence, Behaviour and Communication Survey can be obtained from the full study report (Simbayi et al., 2019).

For this study we computed direct estimates of district level observed HIV prevalence and proportion of HIV positive individuals who are on ART using data from this survey by taking the survey design and response rates into account.

5.7 MAPPING DISTRICT HIV PREVALENCE AND PROPORTION ON ART AMONG HIV POSITIVE INDIVIDUALS IN SOUTH AFRICA

The application of the model discussed in the above section is demonstrated by applying it to direct estimates of district level HIV prevalence and ART coverage among HIV positive individuals in South Africa. The covariates included in the model are the multidimensional poverty index constructed using the 2016 community survey data (Fransman & Yu, 2019), HIV prevalence among pregnant women obtained from the 2017 National Antenatal Sentinel Survey report (Woldesenbet et al., 2018), population density and male condom distribution coverage obtained from the 2017 district health barometer report (Massyn et al., 2017).

The model fitted to the district level HIV prevalence and ART coverage data is given as:

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u_i + v_i$$

where p_{ij} is the proportion of HIV positive individuals and proportion of HIV positive individuals who are on ART at district i , $j = 1, 2$ and denotes the outcome variables, β_q is vector of regression coefficients; where $q = 1, 2, 3, 4$, β_0 is vector of intercepts, u_i is vector of structured spatial random component and v_i is vector of unstructured spatial random component. Estimation of model parameters is done using Bayesian approach, via Markov Chain Monte Carlo procedure. Thus prior distribution for the random components and the model parameters need to be defined. The prior distribution for the intercept is $\beta_0 \sim \text{uniform on } (-\infty, \infty)$ the prior for the rest of the regression coefficients is $\beta_q \sim N(0, 0.00001)$ where $q = 1, 2, 3, 4$; prior for the precision matrices $\Sigma_\theta^{-1} \sim \text{Wishart}(v, \Omega_\theta)$ and $\Sigma_v^{-1} \sim \text{Wishart}(v, \Omega_v)$ with $v = 2$ degrees of freedom and $\Omega_\theta = \Omega_v = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$, prior for $\alpha \sim \text{uniform}(0.3, 10)$. And priors for the remaining parameters is the same as that presented in section 3. The analysis was done using OpenBUGS 3.2 statistical package. We run 50,000 MCMC iterations to estimate the required values and hence to make inferences. We discarded 7000 initial iterations as a burn in samples by assessing the the history plots of each

model and for each parameter. And we also investigated the autocorrelation plots of each parameter to determine the selection intervals to avoid correlation problems in the generated chains. The correlation plots of the precision parameters are very high suggesting wider selected interval; thus we selected every 30th values which resulted in lower autocorrelation.

5.8 RESULTS

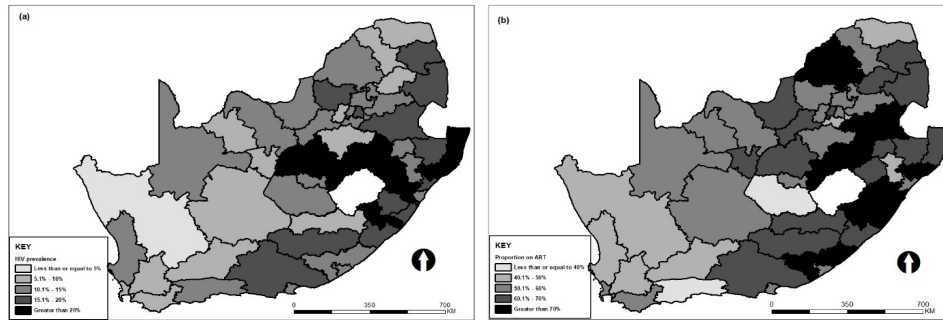
Table 5.1 shows estimates of parameters of the Bayesian nonparametric model. As can be seen in this table the ANC HIV prevalence computed from the 2017 National Antenatal Sentinel HIV survey data is a significant predictor of district level HIV prevalence and ART coverage in South Africa whereas the other covariates included in the model are not significant predictors of the outcome variables. And the covariance of the random components are significantly different from zero suggesting the presence of significant association between the corresponding random components included in each model.

Table 5.1: Estimates of parameters with their 95% confidence interval for the Bayesian nonparametric model.

Covariates	HIV-prevalence	ART-coverage
Intercept	-2.838 (-3.433, -2.273)	-0.3652 (-1.314, 0.5121)
Population density	-0.00006(-0.00025, 0.00017)	-0.00001(-0.00026, 0.00028)
Male condom distribution	-0.00369 (-0.01200, 0.00483)	-0.003105(-0.01519, 0.00992)
Multidimensional poverty index	1.775 (-1.01, 4.62)	3.126(-1.256, 7.568)
ANC HIV prevalence	3.211 (2.046, 4.459)	2.424 (0.4742, 4.32)
Σ_{v11}	149.1 (9.007, 603.9)	
Σ_{v12}	3.984 (-293, 302.3)	
Σ_{v21}	3.984 (-293, 302.3)	
Σ_{v22}	138.1 (5.345, 622.7)	
Σ_{θ_k11}	178.2 (7.919, 637)	
Σ_{θ_k12}	-22.33 (-322.3, 278.6)	
Σ_{θ_k21}	-22.33 (-322.3, 278.6)	
Σ_{θ_k22}	115.8 (2.691, 521)	
α	6 (1, 10)	
k	14 (8, 20)	

As shown in map (a) and (b), Figure 5.1 the HIV prevalence and ART coverage among HIV positive individuals is higher among districts located in the Eastern Cape, Free State, Mpumalanga, KwaZulu Natal province followed by districts in Gauteng, Limpopo, North West and Western Cape provinces. Thus, this suggests that HIV prevalence and ART coverage are strongly associated justifying modelling these two variables jointly.

Figure 5.1: Map of HIV prevalence by district (a) and proportion of ART coverage among HIV positive individuals in South Africa, 2017



The estimated value of α is 6 and thus the infinite Dirichlet process mixture model is approximated by a finite Dirichlet process mixture model that truncates at $m=32$. And the number of occupied clusters determined based on the expression given by Liu (1996) is $k=14$. The sum of the mixing weight estimates, $(p_i; s)$ is approximately close to 1, which is one of the requirements of the truncated Dirichlet process mixture model. All these indicate that an assumption of a normal distribution to the structured spatial random component may not be appropriate, suggesting a mixture of distributions (a nonparametric approach) as a prior for the structured spatial component. In addition, the LS_{cv} value of the multivariate nonparametric ICAR (384.5) model is less than that of the multivariate ICAR-normal model (388.1) suggesting the nonparametric ICAR model is better in terms of its predictive capacity as compared to the multivariate ICAR-normal.

5.9 DISCUSSION

In this study we have presented a flexible nonparametric approach for modelling multivariate spatial structured random effects by extending the univariate approach presented by Li et al. (2015). Thus, in this approach prior uncertainty for the structured random effects is specified at the level of distribution function (Manda, 2011). The nonparametric approach discussed in this chapter used a mixture Dirichlet process prior with a stick-breaking process. The dependency for the multivariate spatial structured random effect was introduced through the mixing weights by using areal spatial stick breaking process. We have illustrated this model by jointly modelling HIV prevalence and ART coverage using data determined from the 2017 South African Na-

tional HIV Prevalence, Incidence, Behavior and Communication Survey. The models were fitted using the Markov Chain Monte Carlo method and computation was carried out using OpenBUGS, a freely available Bayesian statistical package.

According to estimates of the outcome variables using this model, district level HIV prevalence and ART coverage are higher among districts located in the Southeastern parts of the country, whereas districts located in the Southwestern part of South Africa has low level of ART coverage and HIV prevalence. And the findings of this study are similar to our previous work [Ayalew et al. \(2021\)](#), those of [Gutreuter et al. \(2019\)](#) and [Woldesenbet et al., 2018](#)). This suggests that HIV prevalence and ART coverage have got similar spatial pattern and hence modelling them using joint spatial model is a plausible approach.

We used the pragmatic approach proposed by [Ohlssen et al. \(2007\)](#) to determine the number of truncation for the mixture Dirichlet process prior as the infinite Dirichlet process prior is computationally infeasible. Thus, according to the estimates based on this method the Dirichlet process is truncated at $m = 32$. And the sum of the first $m - 1$ mass points is close to 1 which is a requirement to ensure proper mixture Dirichlet distribution ([Ohlssen et al., 2007](#)). Accordingly the number of occupied clusters was 14 as estimated by the method proposed by ([Liu, 1996](#)).

Significant predictors of the outcome variables are the same for both the nonparametric model and the parametric model (ICAR-normal and ICAR-skew-normal) and estimates of the regression coefficients and their 95% confidence intervals are also similar in both nonparametric and parametric scenarios. Thus, one may argue the rationale behind the use of more complex model over the relatively easier parametric model, however when we do not have prior information to believe that parametric MICAR-normal and MICAR-skew-normal distributions are adequate to describe the distribution of the structured spatial random effects the argument may be irrelevant ([Manda, 2011](#); [Hjort et al., 2010](#)). Thus, if the spatial random effects had arbitrary distributions the MICAR-skew-normal and MICAR-normal model might not be adequate to model the structured spatial random effects.

One of the limitations of this study is that it may not be appropriate to model infinite mixture Dirichlet process distribution as the method we proposed is a truncated/finite mixture Dirichlet process distribution. The other limitation of this study is that the number of predictor variables included in the model were few and hence some relevant variables important for predicting the outcome variables may be missing. The model presented in this study can be extended by modelling the spatially unstructured spatial random component or in the spatio-temporal component which we are busy at the moment.

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION AND FUTURE WORK

Intrinsic conditional autoregressive (ICAR) model is used for estimating outcomes at lower administrative level. ICAR model is based on the assumption that the structured spatial random effects are normally distributed. However, the type of distribution for structured spatial random effects may be unknown since random effects are not observed as data and thus the reasonableness of the normality assumption may be in question. The normality assumption is used because of its computational convenience. Therefore, [Besag et al. \(1991\)](#) (developers of the ICAR model) suggested that other forms of distributions can be used in ICAR model for modelling the structured spatial random effects. Estimates may be biased and inferences may be misleading if erroneous distributions of random effects are used in modeling process. Mostly data transformation is used so that the transformed data resembles that of normality; due to limitation associated with modeling transformed data, transformation should be avoided if there is a theoretical distribution. Therefore, the aim of this study is to develop and validate flexible spatial model specifically for modelling multimodal and skewed spatially structured random effects. Thus, in this dissertation we reviewed and investigated alternative flexible and nonparametric spatial models, conducted simulation analysis to investigate the impact of wrongly specifying the distribution of random effect and compared the performance of our suggested models against the existing ones and applied the model we proposed to district level HIV burden and ART coverage obtained from national survey data in South Africa.

This dissertation begins by presenting a literature review of spatial models starting from the basic ones to the most complex and commonly used spatial models. In addition, it also presents a brief overview of developments for extending the existing spatial models so that these approaches can be used as an alternative to normal assumption in ICAR spatial modeling. In chapter 2, we briefly presented the existing modelling approaches to spatial data that assume ICAR-normal, ICAR-Laplace and ICAR-skew-t for the structured spatial random effects. And we tried to show the application of these models using the 2016 Demographic and health survey data conducted in South Africa and estimated the district level HIV prevalence in the country.

In order to identify the best model that captures the underlying features of the data we used CPO and DIC. Estimation of model parameters were conducted in a Bayesian framework and OpenBUGS statistical package was used for conducting the analysis. As DIC was less appropriate for evaluating complex models such as skewed spatial models (De la Cruz & Branco, 2009) we used CPO values as the main criteria for evaluating the performance of these models for modelling the district level HIV prevalence obtained from the 2016 SADHS survey data. Thus, according to the CPO values the skew-t spatial model is the best model in predicting the district level HIV prevalence data as compared to the ICAR-normal and ICAR-Laplace model.

In chapter 3 we developed and validated alternative flexible approaches that relax the symmetric distributional assumption, ICAR-normal and ICAR-Laplace distributions, for modelling structured spatial random effects. In this chapter we used ICAR-skew-normal and ICAR-skew-Laplace distributions for modelling the spatially structured random effects in spatial models. These distributions tend to ICAR-normal and ICAR-Laplace when the skewness parameter approaches to zero. Our approach is suitable for handling data that are skewed and/or contains outlying observations. We simulated two spatially structured random effects from ICAR-t and ICAR-normal distributions with outlier observations on the municipal map of South Africa to show the impact of wrongly specifying the distribution of the spatially structured random effects in spatial models, and hence the performance of our model against the existing ones. We fitted the ICAR-normal, ICAR-skew-normal, ICAR-Laplace and ICAR-skew-Laplace models to each of the simulated data sets and the analysis was done using a Bayesian approach. The mean square error (MSE) and CPO were used to compare the performance of these competing models. The MSE seems to suggest that there is no difference among competing models in terms of precision for both data sets. Whereas the CPO values indicated that the ICAR-Laplace and ICAR-skew-Laplace model are the best models for predicting the outcome variables for structured spatial random effects that are drawn from ICAR-normal and ICAR-t distributions with outlier observations, respectively. Also, we presented the application of our approach to a real data set by modelling the HIV burden by district in South Africa. Analysis of district level HIV burden data obtained from the 2016 SADHS data using our approach and existing methods indicated that the ICAR-skew-normal model seems to better capture the underlying features of the data as compared to the other models.

Two or more outcome variables could be observed in a given area and these variables may show some form of spatial consistency over the entire region; and hence modeling these outcomes jointly improves the precision of the outcome of interest and helps to explore shared and specific trends in disease risks. Thus the univariate ICAR-skew-normal approach that we developed in chapter 3 was extended to model two or more variables in chapter 4. As in chapter 3 we simulated spatially structured random effects from multivariate ICAR-normal and multivariate ICAR-t distributions

with outlier observations on the municipal map of South Africa. We analyzed these simulated data in a Bayesian framework to show that our approach (Multivariate ICAR-skew-normal) has performed better than the multivariate ICAR-normal if the data has skewed and/or outlying observations. Comparison of these models using the CPO values suggested that the multivariate ICAR-skew-normal model is the best model in predicting the values of the outcome variable for both data sets simulated from multivariate ICAR-normal and multivariate ICAR-t distributions with outlier observations. There was no difference between the two models in terms of precision for both data sets as per MSE values. To show the application of the multivariate ICAR-skew-normal approach to a real data set we modelled knowledge of HIV status among pregnant women and proportion of pregnant women who know their HIV status and are on ART by district in South Africa. The MICAR-skew-normal model performs relatively better than the MICAR-normal in terms of predicting the values of the outcome variable according to the CPO values.

The approaches we developed and presented in chapter 3 and 4 are suitable for modelling data with unimodal distributions. In addition, the commonly used modelling approaches and those we developed assumed that the random effects follow some form of parametric distributions. However, there could be instances where data are multimodal or even the distribution of data may be unknown thus in such scenarios our approaches may not be appropriate for modeling such data. One could use multimodal distributions in such situation, but it may be difficult to determine the number of mixtures especially for modeling random components since random components are not observed as data. Such situation motivates the need to use Bayesian nonparametric approach in spatial model for modelling the structured spatial random effects. Therefore, in chapter 5 we extended the univariate Bayesian nonparametric approach for modelling areal spatial data developed by [Li et al. \(2015\)](#) to a multivariate setting. Areal stick-breaking representation was used to develop our approach in a Bayesian nonparametric framework. To show its application to a real data set we fitted this model to a district level HIV prevalence and proportion of HIV positive individual who are on ART determined from the 2017 South African National HIV Prevalence, Incidence, Behaviour and Communication Survey. According to the estimates produced using this model HIV prevalence and ART coverage are higher in districts located in the eastern and central parts of the country.

Studies recommended that sampling weights in complex survey need to be taken into account in spatial modelling as it adjusts over/under representation of sampling units ([Vandendijck et al., 2016](#); [Chen et al., 2014](#); [Watjou et al., 2017](#)). Failure to account sampling weights in spatial modelling could result in biased estimates ([Chen et al., 2014](#); [Mercer et al., 2014](#); [Cassy et al., 2022](#)). Spatial modelling approaches that considers sampling weights produce stable estimates with narrow confidence interval ([Cassy et al., 2022](#); [Mercer et al., 2014](#)). In this dissertation we used complex survey

data to show the application of our approach to a real data set. Thus we took into account the survey design and sampling weights which takes into account non response at each stage of sampling units in determining the direct estimates of the outcomes (HIV prevalence and ART coverage) at district level which is in turn used for calculating weighted sample size and weighted cases in each district which are needed for modelling prevalence/proportion using Bayesian approaches in OpenBugs statistical package.

6.2 FUTURE WORK AND LIMITATIONS

Application of our approaches were demonstrated by modeling district level data of South Africa and hence may not show variations at lower granular level such as at sub-district level. Thus further study can be conducted using our approaches to model spatial data at finer geographical level such as sub-district level in the case of South Africa. We used health survey data (specifically data about HIV) to show the application of our approach to a real data set as these data are readily available to us, and hence didn't show the application of our approach using non-health data. Therefore, a study can be conducted to show the application of our approach outside of health data. Our approaches are focused on modeling data with outlying/skewed observations and multimodal distributions and hence are not by no means exhaustive as the approach we developed may not be suitable for modelling leptokurtic and platykurtic distributions. Thus, future research can be conducted in modelling spatially structured random effects using these distributional approaches both in the univariate and multivariate settings. We used our approach for modelling the structured spatial random components in spatial model and assumed that the unstructured spatial random components are normally distributed. Thus, future work can be done by using our approaches for modelling the unstructured spatial random components. Moreover, our approaches both parametric and nonparametric are developed and tested for modelling the structured spatial random component; hence our methods can be extended for modelling the structured temporal component in spatiotemporal modelling approaches (currently we are busy modelling the temporal using Bayesian nonparametric approach). As a limitation, the Bayesian nonparametric approach we proposed is appropriate for modelling truncated Dirichlet process and it may not be suitable for modeling infinite Dirichlet process. And the explanatory variables included in our model are few and thus some variables important for explaining the variability of the model may be missing; so, this may have an impact on the accuracy of the estimates generated by models.

BIBLIOGRAPHY

- Adeiza, M. A., Abba, A. A. & Okpapi, J. U. (2014), 'Hiv-associated tuberculosis: A sub-saharan african perspective', *Sub-Saharan African Journal of Medicine* **1**, 1.
- Allard, D. & Naveau, P. (2007), 'A new spatial skew-normal random field model', *Communications in Statistics—Theory and Methods* **36**(9), 1821–1834.
- Anderson, S. & Maher, D. (2001), An analysis of interaction between tb and hiv / aids programmes in sub-saharan africa.
- Arellano-Valle, R., Bolfarine, H. & Lachos, V. (2007), 'Bayesian inference for skew-normal linear mixed models', *Journal of Applied Statistics* **34**(6), 663–682.
- Arslan, O. (2010), 'An alternative multivariate skew laplace distribution: properties and estimation', *Statistical Papers* **51**(4), 865–887.
- Assunção, R. M. & de Castro, M. S. M. (2004), 'Multiple cancer sites incidence rates estimation using a multivariate bayesian model.', *International journal of epidemiology* **33** 3, 508–16.
- Ayalew, K. A., Manda, S. & Cai, B. (2021), 'A comparison of bayesian spatial models for hiv mapping in south africa', *International Journal of Environmental Research and Public Health* **18**(21), 11215.
- Azzalini, A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian journal of statistics* pp. 171–178.
- Azzalini, A. (1986), 'Further results on a class of distributions which includes the normal ones', *Statistica* **46**(2), 199–208.
- Azzalini, A. & Capitanio, A. (1999), 'Statistical applications of the multivariate skew normal distribution', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(3), 579–602.
- Azzalini, A. & Valle, A. D. (1996), 'The multivariate skew-normal distribution', *Biometrika* **83**(4), 715–726.
- Banerjee, S., Carlin, B., Gelfand, A., Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A. & West, M. (2003), 'Hierarchical multivariate car models for spatio-temporally correlated survival data (with discussion)', *Bayesian Statistics* **7**, 45–63.

- Bernardinelli, L., Pascutto, C., Best, N. & Gilks, W. (1997), 'Disease mapping with errors in covariates', *Statistics in medicine* **16**(7), 741–752.
- Bernardinelli, L., Clayton, D. & Montomoli, C. (1995), 'Bayesian estimates of disease maps: how important are priors?', *Statistics in medicine* **14**(21-22), 2411–2431.
- Bernardinelli, L. & Montomoli, C. (1992), 'Empirical bayes versus fully bayesian analysis of geographical variation in disease risk', *Statistics in medicine* **11**(8), 983–1007.
- Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Journal of the royal statistical society series b-methodological* **36**, 192–225.
- Besag, J. & Kooperberg, C. (1995), 'On conditional and intrinsic autoregressions', *Biometrika* **82**(4), 733–746.
- Besag, J., York, J. & Mollié, A. (1991), 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the institute of statistical mathematics* **43**(1), 1–20.
- Best, N., Richardson, S. & Thomson, A. (2005), 'A comparison of bayesian spatial models for disease mapping', *Statistical methods in medical research* **14**(1), 35–59.
- Blangiardo, M. & Cameletti, M. (2015), *Spatial and spatio-temporal Bayesian models with R-INLA*, John Wiley & Sons.
- Branco, M. D. & Dey, D. K. (2001), 'A general class of multivariate skew-elliptical distributions', *Journal of Multivariate Analysis* **79**(1), 99–113.
- Breslow, N. E. (1984), 'Extra-poisson variation in log-linear models', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **33**(1), 38–44.
- Cai, B., Lawson, A. B., Hossain, M. M., Choi, J., Kirby, R. S. & Liu, J. (2013), 'Bayesian semiparametric model with spatially-temporally varying coefficients selection', *Statistics in medicine* **32**(21), 3670–3685.
- Cancho, V. G., Lachos, V. H. & Ortega, E. M. (2010), 'A nonlinear regression model with skew-normal errors', *Statistical papers* **51**(3), 547–558.
- Carlin, B. P. & Banerjee, S. (2003), Models for spatio-temporally correlated survival data, in 'Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting', Oxford University Press, p. 45.
- Cassy, S., Manda, S. & Marques, F. (2022), 'Martins, mdro accounting for sampling weights in the analysis of spatial distributions of disease using health survey data, with an application to mapping child health in malawi and mozambique', *Int. J. Environ. Res. Public Health* **19**, 6319.

Bibliography

- Chen, C., Wakefield, J. & Lumely, T. (2014), 'The use of sampling weights in bayesian hierarchical models for small area estimation', *Spatial and spatio-temporal epidemiology* **11**, 33–43.
- Chimoyi, L. A. & Musenge, E. (2014), 'Spatial analysis of factors associated with hiv infection among young people in uganda, 2011', *BMC public health* **14**(1), 1–11.
- Choy, S. & Smith, A. (1997), 'Hierarchical models with scale mixtures of normal distributions', *Test* **6**(1), 205–221.
- Clayton, D. & Kaldor, J. (1987), 'Empirical bayes estimates of age-standardized relative risks for use in disease mapping', *Biometrics* pp. 671–681.
- Congdon, P. (2007), *Bayesian statistical modelling*, John Wiley & Sons.
- Cressie, N. (1993), 'Statistics for spatial data, revised edition wiley', *New York, NY.[Google Scholar]* .
- Cressie, N. (2015), *Statistics for spatial data*, John Wiley & Sons.
- Dabney, A. R. & Wakefield, J. (2005), 'Issues in the mapping of two diseases', *Statistical Methods in Medical Research* **14**, 112 – 83.
- Dagne, G. A. (2013), 'Bayesian inference for skew-normal mixture models with left-censoring', *Journal of biopharmaceutical statistics* **23**(5), 1023–1041.
- De la Cruz, R. & Branco, M. D. (2009), 'Bayesian analysis for nonlinear regression model under skewed errors, with application in growth curves', *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **51**(4), 588–609.
- DeYoreo, M. & Kottas, A. (2015), 'A fully nonparametric modeling approach to binary regression', *Bayesian Analysis* **10**(4), 821–847.
- Diaconis, P. & Freedman, D. (1986), 'On the consistency of bayes estimates', *The Annals of Statistics* pp. 1–26.
- Diggle, P. J., Tawn, J. A. & Moyeed, R. A. (1998), 'Model-based geostatistics', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **47**(3), 299–350.
- Dominguez-Molina, J., Gonzalez-Farias, G. & Gupta, A. (2003), 'The multivariate closed skew normal distribution', *Technical Report 03-12* .
- Duan, J. A., Guindani, M. & Gelfand, A. E. (2007), 'Generalized spatial dirichlet process models', *Biometrika* **94**(4), 809–825.
- Dwyer-Lindgren, L., Cork, M. A., Sligar, A., Steuben, K. M., Wilson, K. F., Provost, N. R., Mayala, B. K., VanderHeide, J. D., Collison, M. L., Hall, J. B. et al. (2019), 'Mapping hiv prevalence in sub-saharan africa between 2000 and 2017', *Nature* **570**(7760), 189–193.

- Feltbower, R. G. & Manda, S. O. (2012), Bayesian bivariate disease mapping, in 'Modern methods for Epidemiology', Springer, pp. 141–155.
- Ferguson, T. S. (1973), 'A bayesian analysis of some nonparametric problems', *The annals of statistics* pp. 209–230.
- Fernández, C. & Steel, M. F. (1998), 'On bayesian modeling of fat tails and skewness', *Journal of the american statistical association* **93**(441), 359–371.
- Fransman, T. & Yu, D. (2019), 'Multidimensional poverty in south africa in 2001–16', *Development Southern Africa* **36**(1), 50–79.
- Fuentes, M. & Reich, B. (2013), 'Multivariate spatial nonparametric modelling via kernel processes mixing', *Statistica Sinica* **23**(1).
- Galarza, C. E., Lachos, V. H. & Bandyopadhyay, D. (2017), 'Quantile regression in linear mixed models: a stochastic approximation em approach', *Statistics and its Interface* **10**(3), 471.
- Gelfand, A. E., Guindani, M. & Petrone, S. (2007), 'Bayesian nonparametric modelling for spatial data using dirichlet processes'.
- Gelfand, A. E., Kottas, A. & MacEachern, S. N. (2005), 'Bayesian nonparametric spatial modeling with dirichlet process mixing', *Journal of the American Statistical Association* **100**(471), 1021–1035.
- Gelfand, A. E. & Sahu, S. K. (1999), 'Identifiability, improper priors, and gibbs sampling for generalized linear models', *Journal of the American Statistical Association* **94**(445), 247–253.
- Gelfand, A. E. & Vounatsou, P. (2003), 'Proper multivariate conditional autoregressive models for spatial data analysis.', *Biostatistics* **4** **1**, 11–25.
- Genton, M. G. & Zhang, H. (2012), 'Identifiability problems in some non-gaussian spatial random fields', *Chilean Journal of Statistics* **3**(2), 171–179.
- Ghidey, W., Lesaffre, E. & Eilers, P. (2004), 'Smooth random effects distribution in a linear mixed model', *Biometrics* **60**(4), 945–953.
- Ghosh, P., Branco, M. D. & Chakraborty, H. (2007), 'Bivariate random effect model using skew-normal distribution with application to hiv-rna', *Statistics in medicine* **26**(6), 1255–1267.
- Giorgi, E., Diggle, P. J., Snow, R. W. & Noor, A. M. (2018), 'Geostatistical methods for disease mapping and visualisation using data from spatio-temporally referenced prevalence surveys', *International Statistical Review* **86**(3), 571–597.

Bibliography

- Graham, S. R., Carlton, C., Gaede, D. & Jamison, B. (2011), 'Student column: The benefits of using geographic information systems as a community assessment tool', *Public Health Reports* **126**(2), 298–303.
- Green, P. J. & Richardson, S. (2002), 'Hidden markov models and disease mapping', *Journal of the American statistical association* **97**(460), 1055–1070.
- Griffin, J. E. & Steel, M. J. (2006), 'Order-based dependent dirichlet processes', *Journal of the American statistical Association* **101**(473), 179–194.
- Gutreuter, S., Igumbor, E., Wabiri, N., Desai, M. & Durand, L. (2019), 'Improving estimates of district hiv prevalence and burden in south africa using small area estimation techniques', *PLoS One* **14**(2), e0212445.
- Hagan, H., Jenness, S. M., Wendel, T., Murrill, C., Neaigus, A. & Gelpí-Acosta, C. (2010), 'Herpes simplex virus type 2 associated with hiv infection among new york heterosexuals living in high-risk areas', *International Journal of STD & AIDS* **21**, 580 – 583.
- Haines, L. M. & Thiart, C. (2021), 'The impact of spatial statistics in africa', *Spatial Statistics* p. 100580.
URL: <https://www.sciencedirect.com/science/article/pii/S22111675321000774>
- Haining, R. (2003), *The nature of spatial data*, Cambridge University Press, p. 43–88.
- Hallett, T. B., Anderson, S.-J., Asante, C. A., Bartlett, N., Bendaud, V., Bhatt, S., Burgert, C. R., Cuadros, D. F., Dzangare, J., Fecht, D. et al. (2016), 'Evaluation of geospatial methods to generate subnational hiv prevalence estimates for local level planning', *Aids* **30**(9).
- Harville, D. A. (1998), *Matrix algebra from a statistician's perspective*.
- Heckman, J. J. & Singer, B. (1984), 'Econometric duration analysis', *Journal of econometrics* **24**(1-2), 63–132.
- Held, L., Natário, I., Fenton, S. E., Rue, H. & Becker, N. (2005), 'Towards joint disease mapping', *Statistical Methods in Medical Research* **14**(1), 61–82. PMID: 15691000.
- Henze, N. (1986), 'A probabilistic representation of the 'skew-normal' distribution', *Scandinavian journal of statistics* pp. 271–275.
- Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G. (2010), *Bayesian nonparametrics*, Vol. 28, Cambridge University Press.
- Hossain, M., Lawson, A. B., Cai, B., Choi, J., Liu, J., Kirby, R. S. et al. (2013), 'Space-time stick-breaking processes for small area disease cluster estimation', *Environmental and ecological statistics* **20**(1), 91–107.

- Houlihan, C. F., Mutevedzi, P. C., Lessells, R. J., Cooke, G. S., Tanser, F. C. & Newell, M.-L. (2010), 'The tuberculosis challenge in a rural south african hiv programme', *BMC infectious diseases* **10**(1), 1–9.
- Huang, Y., Chen, J. & Lu, X. (2016), 'Bayesian approach to nonlinear mixed-effects quantile regression models for longitudinal data with non-normality and left-censoring', *Journal of Advanced Statistics* **1**(3), 109.
- Ishwaran, H. & James, L. F. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**(453), 161–173.
- Ishwaran, H. & James, L. F. (2002), 'Approximate dirichlet process computing in finite normal mixtures: smoothing and prior information', *Journal of Computational and Graphical statistics* **11**(3), 508–532.
- Ishwaran, H. & Zarepour, M. (2000), 'Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models', *Biometrika* **87**(2), 371–390.
- Jara, A., Hanson, T. E. & Lesaffre, E. (2009), 'Robustifying generalized linear mixed models using a new class of mixtures of multivariate polya trees', *Journal of Computational and Graphical Statistics* **18**(4), 838–860.
- Jara, A., Quintana, F. & San Martín, E. (2008), 'Linear mixed models with skew-elliptical distributions: A bayesian approach', *Computational statistics & data analysis* **52**(11), 5033–5045.
- Jin, X., Carlin, B. P. & Banerjee, S. (2005), 'Generalized hierarchical multivariate car models for areal data', *Biometrics* **61**(4), 950–961.
- Johnson, G. D. (2004), 'Small area mapping of prostate cancer incidence in new york state (usa) using fully bayesian hierarchical modelling', *International Journal of Health Geographics* **3**(1), 1–12.
- Kazemi, I., Mahdihyeh, Z., Mansourian, M. & Park, J. J. (2013), 'Bayesian analysis of multivariate mixed models for a prospective cohort study using skew-elliptical distributions', *Biometrical Journal* **55**(4), 495–508.
- Kim, H.-M. & Mallick, B. K. (2004), 'A bayesian prediction using the skew gaussian distribution', *Journal of Statistical Planning and Inference* **120**(1-2), 85–101.
- Kim, H., Sun, D. & Tsutakawa, R. K. (2001), '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.
- Kim, H., Tanser, F., Tomita, A., Vandormael, A. & Cuadros, D. F. (2021), 'Beyond hiv prevalence: identifying people living with hiv within underserved areas in south africa', *BMJ global health* **6**(4), e004089.

Bibliography

- Kish, L. (1995), 'Methods for design effects', *Journal of official Statistics* **11**(1), 55.
- Knorr-Held, L. & Best, N. G. (2001), '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.
- Kottas, A., Duan, J. A. & Gelfand, A. E. (2008), 'Modeling disease incidence data with spatial and spatio temporal dirichlet process mixtures', *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **50**(1), 29–42.
- Kotz, S., Balakrishnan, N. & Johnson, N. L. (2004), *Continuous multivariate distributions, Volume 1: Models and applications*, Vol. 1, John Wiley & Sons.
- Kotz, S., Kozubowski, T. & Podgórski, K. (2001), *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*, number 183, Springer Science & Business Media.
- Kozubowski, T. J., Podgórski, K. & Rychlik, I. (2013), 'Multivariate generalized laplace distribution and related random fields', *Journal of Multivariate Analysis* **113**, 59–72.
- Lachos, V. H., Dey, D. K. & Cancho, V. G. (2009), 'Robust linear mixed models with skew-normal independent distributions from a bayesian perspective', *Journal of Statistical Planning and Inference* **139**(12), 4098–4110.
- Laird, N. (1978), 'Nonparametric maximum likelihood estimation of a mixing distribution', *Journal of the American Statistical Association* **73**(364), 805–811.
- Langford, I. H., Leyland, A. H., Rasbash, J. & Goldstein, H. (1999), 'Multilevel modelling of the geographical distributions of diseases', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **48**(2), 253–268.
- Lawson, A. B. (2008), *Bayesian disease mapping: Hierarchical modeling in spatial epidemiology*.
- Lawson, A. B., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P. & Divino, F. (2000), 'Disease mapping models: an empirical evaluation. disease mapping collaborative group', *Statistics in medicine* **19**(17), 2217–41.
- Lawson, A. B., Browne, W. J. & Rodeiro, C. L. V. (2003), *Disease mapping with WinBUGS and MLwiN*, Vol. 11, John Wiley & Sons.
- Leroux, B. G., Lei, X. & Breslow, N. (2000), Estimation of disease rates in small areas: a new mixed model for spatial dependence, in 'Statistical models in epidemiology, the environment, and clinical trials', Springer, pp. 179–191.
- Lesaffre, E. & Lawson, A. B. (2012), *Bayesian biostatistics*, John Wiley & Sons.

- Leyland, A. H., Langford, I. H., Rasbash, J. & Goldstein, H. (2000), 'Multivariate spatial models for event data', *Statistics in Medicine* **19**(17-18), 2469–2478.
- Li, P., Banerjee, S., Hanson, T. A. & McBean, A. M. (2015), 'Bayesian models for detecting difference boundaries in areal data', *Statistica Sinica* **25**(1), 385.
- Liu, H. & Zhu, X. (2017), 'Joint modeling of multiple crimes: A bayesian spatial approach', *ISPRS Int. J. Geo Inf.* **6**, 16.
- Liu, J. S. (1996), 'Nonparametric hierarchical bayes via sequential imputations', *The Annals of Statistics* **24**(3), 911–930.
- Lu, H., Reilly, C. S., Banerjee, S. & Carlin, B. P. (2007), 'Bayesian areal wombling via adjacency modeling', *Environmental and ecological statistics* **14**(4), 433–452.
- Lunn, D., Jackson, C., Best, N., Thomas, A. & Spiegelhalter, D. (2013), 'The bugs book', *A Practical Introduction to Bayesian Analysis*, Chapman Hall, London .
- MacNab, Y. & Dean, C. (2000), 'Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models', *Statistics in medicine* **19**(17-18), 2421–2435.
- Mallows, C. (1974), 'Scale mixtures of normal distributions. journal of the royal statistical society', *Series B* **36**, 99–102.
- Manda, S., Feltbower, R. & Gilthorpe, M. (2012), 'Review and empirical comparison of joint mapping of multiple diseases', *Southern African Journal of Epidemiology and Infection* **27**(4), 169–182.
- Manda, S., Haushona, N. & Bergquist, R. (2020), 'A scoping review of spatial analysis approaches using health survey data in sub-saharan africa', *International journal of environmental research and public health* **17**(9), 3070.
- Manda, S., Masenyetse, L., Cai, B. & Meyer, R. (2015), 'Mapping hiv prevalence using population and antenatal sentinel-based hiv surveys: a multi-stage approach', *Population health metrics* **13**(1), 1–15.
- Manda, S. O. (2011), 'A nonparametric frailty model for clustered survival data', *Communications in Statistics—Theory and Methods* **40**(5), 863–875.
- Manda, S. O. (2014), Macro determinants of geographical variation in childhood survival in south africa using flexible spatial mixture models, in 'Advanced techniques for modelling maternal and child health in Africa', Springer, pp. 147–168.
- Manda, S. O. M., Lombard, C. J. & Mosala, T. I. (2012), 'Divergent spatial patterns in the prevalence of the human immunodeficiency virus (hiv) and syphilis in south african pregnant women.', *Geospatial health* **6** 2, 221–31.

Bibliography

- Mardia, K. V. (1988), 'Multi-dimensional multivariate gaussian markov random fields with application to image processing', *Journal of Multivariate Analysis* **24**, 265–284.
- Massyn, N., Padarath, A. & Peer, N. (2017), *District health barometer 2016/17*, Health Systems Trust.
- Mercer, L., Wakefield, J., Chen, C. & Lumley, T. (2014), 'A comparison of spatial smoothing methods for small area estimation with sampling weights', *Spatial Statistics* **8**, 69–85.
- Middelkoop, K., Mathema, B., Myer, L., Shashkina, E., Whitelaw, A. C., Kaplan, G., Kreiswirth, B. N., Wood, R. & Bekker, L.-G. (2015), 'Transmission of tuberculosis in a south african community with a high prevalence of hiv infection.', *The Journal of infectious diseases* **211** **1**, 53–61.
- Moraga, P. & Lawson, A. B. (2012), 'Gaussian component mixtures and car models in bayesian disease mapping', *Computational Statistics & Data Analysis* **56**(6), 1417–1433.
- Moran, P. A. (1950), 'Notes on continuous stochastic phenomena', *Biometrika* **37**(1/2), 17–23.
- Müller, P. & Quintana, F. A. (2004), 'Nonparametric bayesian data analysis', *Statistical science* **19**(1), 95–110.
- Müller, P., Quintana, F. A., Jara, A. & Hanson, T. (2015), *Bayesian nonparametric data analysis*, Springer.
- Nathoo, F. S. & Ghosh, P. (2013), 'Skew-elliptical spatial random effect modeling for areal data with application to mapping health utilization rates', *Statistics in medicine* **32**(2), 290–306.
- National Department of Health, N., Statistics South Africa, (Stats SA) and South African Medical Research Council, S. & ICF (2019), 'South africa demographic and health survey 2016', *Pretoria, South Africa and Rockville, Maryland, USA*.
- Niragire, F., Achia, T. N., Lyambabaje, A. & Ntaganira, J. (2015), 'Bayesian mapping of hiv infection among women of reproductive age in rwanda', *PloS one* **10**(3), e0119944.
- Ohlssen, D. I., Sharples, L. D. & Spiegelhalter, D. J. (2007), 'Flexible random-effects models using bayesian semi-parametric models: applications to institutional comparisons', *Statistics in medicine* **26**(9), 2088–2112.
- Okhli, K., Mozafari, M. & Naderi, M. (2017), 'Skew laplace finite mixture modelling', *Journal of The Iranian Statistical Society* **16**(2), 97–110.
- Palacios, M. B. & Steel, M. F. J. (2006), 'Non-gaussian bayesian geostatistical modeling', *Journal of the American Statistical Association* **101**(474), 604–618.

- Phadia, E. G. (2013), 'Prior processes and their applications', *Nonparametric Bayesian estimation* **6**.
- Rantini, D., Iriawan, N. et al. (2021), 'Fernandez–steel skew normal conditional autoregressive (fssn car) model in stan for spatial data', *Symmetry* **13**(4), 545.
- Reich, B. J. & Fuentes, M. (2007), 'A multivariate semiparametric bayesian spatial modeling framework for hurricane surface wind fields', *The Annals of Applied Statistics* **1**, 249–264.
- Ripley, B. (1981), 'Spatial statistics john wiley & sons', New York, New York .
- Ripley, B. D. (2005), *Spatial statistics*, John Wiley & Sons.
- Sahu, S. K., Dey, D. K. & Branco, M. D. (2003), 'A new class of multivariate skew distributions with applications to bayesian regression models', *Canadian Journal of Statistics* **31**(2), 129–150.
- Saran, S., Singh, P., Kumar, V. & Chauhan, P. (2020), 'Review of geospatial technology for infectious disease surveillance: use case on covid-19', *Journal of the Indian Society of Remote Sensing* **48**(8), 1121–1138.
- Sethuraman, J. (1994), 'A constructive definition of dirichlet priors', *Statistica sinica* pp. 639–650.
- Sha, Z. (2018), 'mclcar: an r package for maximum monte carlo likelihood estimation of conditional auto-regression models'.
- Simbayi, L., Zuma, K., Zungu, N., Moyo, S., Marinda, E., Jooste, S., Mabaso, M., Ramlagan, S., North, A., Van Zyl, J. et al. (2019), 'South african national hiv prevalence, incidence, behaviour and communication survey, 2017: towards achieving the un-aids 90-90-90 targets'.
- Simon, V., Ho, D. D. & Karim, Q. A. (2006), 'Hiv/aids epidemiology, pathogenesis, prevention, and treatment', *The Lancet* **368**, 489–504.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the royal statistical society: Series b (statistical methodology)* **64**(4), 583–639.
- Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003), 'Winbugs user manual, version 1.4 mrc biostatistics unit', Cambridge, UK .
- Spiegelhalter, D., Best, N., Carlin, B. & Van der Linde, A. (2002), 'Bayesian measures of model complexity and fit (with discussion)', *Journal of the Royal Statistical, Series B* **64**, 583–616.

Bibliography

- Stone, C. A. & Zhu, X. (2015), *Bayesian analysis of item response theory models using SAS*, Sas Institute.
- Tanser, F., Barnighausen, T., Cooke, G. S. & Newell, M.-L. (2009), 'Localized spatial clustering of hiv infections in a widely disseminated rural south african epidemic', *International journal of epidemiology* **38**(4), 1008–1016.
- Tsutakawa, R. K. (1985), 'Estimation of cancer mortality rates: a bayesian analysis of small frequencies', *Biometrics* pp. 69–79.
- Tsutakawa, R. K., Shoop, G. L. & Marienfeld, C. J. (1985), 'Empirical bayes estimation of cancer mortality rates', *Statistics in medicine* **4**(2), 201–212.
- UNAIDS (2015), 'Unaid 2016–2021 strategy: On the fast-track to end aids'.
- UPsEPfA, R. (2021), 'Pepfar 2021 country and regional operational plan (cop/rop) guidance for all pefpar countries', *US State Department* .
- van Schalkwyk, C., Dorrington, R. E., Seatlhodi, T., Velasquez, C., Feizzadeh, A. & Johnson, L. F. (2021), 'Modelling of hiv prevention and treatment progress in five south african metropolitan districts', *Scientific reports* **11**(1), 1–10.
- Vandendijck, Y., Faes, C., Kirby, R., Lawson, A. & Hens, N. (2016), 'Model-based inference for small area estimation with sampling weights', *Spatial statistics* **18**, 455–473.
- Verbeke, G. & Lesaffre, E. (1996), 'A linear mixed-effects model with heterogeneity in the random-effects population', *Journal of the American Statistical Association* **91**(433), 217–221.
- Wakefield, J. (2007), 'Disease mapping and spatial regression with count data', *Biostatistics* **8**(2), 158–183.
- Walker, S. G. & Mallick, B. K. (1997), 'Hierarchical generalized linear models and frailty models with bayesian nonparametric mixing', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(4), 845–860.
- Waller, L. A. & Carlin, B. P. (2010), 'Disease mapping', *Chapman & Hall/CRC handbooks of modern statistical methods* **2010**, 217.
- Wang, F. & Wall, M. M. (2003), 'Generalized common spatial factor model', *Biostatistics* **4**(4), 569–582.
- Watjou, K., Faes, C., Lawson, A., Kirby, R., Aregay, M., Carroll, R. & Vandendijck, Y. (2017), 'Spatial small area smoothing models for handling survey data with non-response', *Statistics in medicine* **36**(23), 3708–3745.

- Woldesenbet, S., Kufa, T., Lombard, C., Manda, S., Ayalew, K., Cheyip, M. & Puren, A. (2018), 'The 2017 national antenatal sentinel hiv survey, south africa, national department of health [cited: 02/03/19]', *Avaiable from: https://www.nicd.ac.za/wp-content/uploads/2019/07/Antenatal_survey-report_24July19.pdf*.
- Yavuz, F. G. & Arslan, O. (2018), 'Linear mixed model with laplace distribution (llmm)', *Statistical Papers* **59**(1), 271–289.
- Yu, K. & Moyeed, R. A. (2001), 'Bayesian quantile regression', *Statistics & Probability Letters* **54**(4), 437–447.
- Zareifard, H. & Khaledi, M. J. (2013), 'Non-gaussian modeling of spatial data using scale mixing of a unified skew gaussian process', *Journal of Multivariate Analysis* **114**, 16–28.
- Zhang, D. & Davidian, M. (2001), 'Linear mixed models with flexible distributions of random effects for longitudinal data', *Biometrics* **57**(3), 795–802.
- Zhang, H. & El-Shaarawi, A. (2010), 'On spatial skew-gaussian processes and applications', *Environmetrics: The official journal of the International Environmetrics Society* **21**(1), 33–47.

APPENDIX: R CODES FOR SIMULATING ICAR-NORMAL WITH OUTLIERS AND WINBUGS CODE FOR ANALYZING THE SIMULATED DATA FOR CHAPTER 3

R Codes for simulating ICAR-skew-normal structured spatial effect

```
# Packages required
library(maps)
library(maptools)
library(spdep)
library(rgdal)
sf_use_s2(FALSE)
library(mclcar)
library(dplyr)
library("writexl")
shape <- readOGR(dsn = "C:/Users/ylo8/Desktop/Data for GIS/SA Demarcation
Board Shapefiles/Local Munics1", layer = "LocalMunicipalities2011")
W.nb <- poly2nb(shape)
W<- nb2mat(W.nb, style="B")
e<-eigen(W)
max_e<-max(e$value)
rho<-1/max_e
s1<- CAR.simWmat(rho = 0.1678633, prec = 1/2, W = W)
s<-(s1-mean(s1))/sqrt(var(s1)) #standardised to avoid errors & warnings in generating
data from rbinom
y<-1:234
m<-data.frame(y,s)
k<-m[with(m,order(s)),] # sort s from largest to smallest

for (i in 214:234){
  k$s[i] <- -k$s[i] * 3}
#thenmergetherandomeffectsusingthelowercodes.

df1= m %>% inner _join(k,by="y")
u<-df1$s.y

nu<-rnorm(234) # the nonspatial structured effect
nu1<-(nu-mean(nu))/sqrt(var(nu)) #standardised to avoid errors warnings in gener-
ating data from rbinom
```

```

p1<-exp(u+nu1)
p2<-1+exp(u+nu1)
p<-p1/p2
n<-0
for (i in 1:234){
n[i] <- sample(300 : 600,1,replace = FALSE)
}
x <- -0
for(i in 1 : 234){
x[i] <- -rbinom(1,n[i],p[i])
}
ux <- -data.frame(u,nu1,x,n,p)
write_xlsx(ux,"C : /Users/ylo8/Desktop/filename.xlsx")

```

WinBUGS Codes for ICAR-skew-normal model

```

model {
for(i in 1 : 234){
y[i] ~ dbin(p[i],n[i])
logit(p[i]) <- -u[i] + v[i]
v[i] ~ dnorm(0,tau.v)
wu[i] ~ dnorm(0,1)I(0,)
u[i] <- -m[i] + deltau * wu[i]
LL[i] <- -logfact(n[i]) - (logfact(y[i]) + logfact(n[i] - y[i])) + y[i] * log(p[i]) + (n[i] -
y[i]) * log(1 - p[i])
ppo[i] <- -exp(LL[i])
icpo[i] <- -(1/ppo[i])
cpo[i] <- -1/icpo[i]
}
m[1 : 234] ~ car.normal(adj[],weights[],num[],tau.u)
for(kin1 : sumNumNeigh)weights[k] <- -1
tau.v ~ dgamma(0.05,0.0005)
tau.u ~ dgamma(0.05,0.0005)
sigma.v <- -sqrt(1/tau.v)
sigma.u <- -sqrt(1/tau.u)
deltau ~ dnorm(0,0.01)
}
#data
#INITIALS
list(tau.u = 1,tau.v = 1,deltau = 0,...)

```